

TerrSet 2020

Geospatial Monitoring and Modeling System

MANUAL

J Ronald Eastman

CLARK LABS



Source Code

© 1987-2020

J. Ronald Eastman

Production

© 1987-2020

Clark University

www.clarklabs.org clarklabs@clarku.edu

CHAPTER ONE

INTRODUCTION

Welcome to the TerrSet Geospatial Monitoring and Modeling System. TerrSet incorporates the IDRISI GIS and Image Processing tools and offers a constellation of vertical applications focused on monitoring and modeling the earth system for sustainable development. The full constellation includes:

- The Land Change Modeler (LCM) for analyzing land cover change, empirically modeling its relationship to explanatory variables and projecting future changes. LCM also includes special tools for the assessment of REDD (Reducing Emissions from Deforestation and forest Degradation) climate change mitigation strategies.
- The Habitat and Biodiversity Modeler (HBM) for habitat assessment, landscape pattern analysis and biodiversity modeling. HBM also contains special tools for species distribution modeling.
- GeOSIRIS – a unique tool for national level REDD (Reducing Emissions from Deforestation and forest Degradation) planning, developed in close cooperation with Conservation International. With GeOSIRIS one can model the impact of various economic strategies on deforestation and carbon emissions reductions.
- The Ecosystem Services Modeler (ESM) for assessing the value of various ecosystem services such as water purification, crop pollination, wind and wave energy, and so on. ESM is based closely on the InVEST toolset developed by the Natural Capital Project.
- The Earth Trends Modeler (ETM) – a tool for the analysis of time series of earth observation imagery. With ETM one can discover trends and recurrent patterns in fundamental earth system variables such as sea surface temperature, atmospheric temperature, precipitation, vegetation productivity and the like. ETM is an exceptional tool for the assessment of climate change in the recent past (e.g., the past 30 years).
- The Climate Change Adaptation Modeler (CCAM, pronounced “see cam”) – a tool for modeling future climate and assessing its impacts on sea level rise, crop suitability and species distributions.
- The IDRISI GIS Analysis tools – a wide range of fundamental analytical tools for GIS analysis, primarily oriented to raster data. Special features of the IDRISI tool set include a suite of multicriteria and multiobjective decision procedures and a broad range of tools for statistical, change and surface analysis. Special graphical modeling environments are also provided for dynamic modeling and decision support. IDRISI also provides a scripting environment and an extremely flexible application programming interface (API) that allows the ability to control IDRISI using languages such as C++, Delphi and Python. Indeed, all TerrSet components make very extensive use of this API.
- The IDRISI Image Processing System – an extensive set of procedures for the restoration, enhancement, transformation and classification of remotely sensed images. IDRISI has the broadest set of classification procedures in the industry including

both hard and soft classification procedures based on machine learning (such as neural networks) and statistical characterization.

The relationship between TerrSet and IDRISI is thus quite intimate. IDRISI provides two of the major components of the system (GIS Analysis and Image Processing) as well as the foundation for all components. All components use the IDRISI API and the IDRISI data file structures. While IDRISI was once a standalone product, it is now incorporated within the TerrSet system.

Exploring TerrSet

For those generally familiar with raster geospatial software such as a Geographic Information System (GIS) with raster capability or an Image Processing System (IPS) for analysis of remotely sensed imagery, the **Quickstart** discussion in the next chapter may be adequate to get started. However, the **TerrSet System Overview** chapter provides a more complete description of the TerrSet system including the organization of data, map and image display, file structures and georeferencing conventions. Perhaps the easiest introduction to TerrSet is through the **Tutorial** which can be accessed through the File|Help menu.

Following this general orientation, specific chapters are provided regarding the operation of each member in the TerrSet system. While they can be read in any order, we generally recommend becoming familiar with the IDRISI GIS Analysis tools early since they provide fundamental tools that are useful to other components.

License Agreement

The software described in this document is furnished under a license and may be used or copied only in accordance with the terms of this license.

The TerrSet software described in this document is protected by the United States Copyright Law and International Treaty provisions. This software remains the property of Clark Labs, Clark University. However, Clark Labs grants the purchaser non-exclusive license to use this software subject to the terms outlined in this statement.

The purchaser of a single-user license of TerrSet is licensed to install and use the software on no more than one single-user computer system. The purchaser is also permitted to make a backup copy of the TerrSet distribution media for the sole purpose of protecting the purchaser's investment from loss. The purchaser may not rent or lease the software but may transfer the license to another user upon written agreement from Clark Labs. The user may not reverse-engineer, decompile, or disassemble the TerrSet software or any of its associated software programs contained on the distribution media.

The PDF manuals that accompany this software are also protected by United States Copyright Law and International Treaty provisions. All rights are reserved. No part of the manuals may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopy, microfilm, recording, or otherwise without written consent of Clark Labs. Direct translation of any part of the manuals is also prohibited without the written permission of the copyright holder.

Warranty

This software is sold "as is," and the information in this manual is subject to change without notice. Furthermore, the Clark Labs assumes no responsibility for any errors that may appear in this document, or in the software it describes.

CLARK LABS DISCLAIMS ALL OTHER WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT TO DEFECTS IN THE DIGITAL MEDIA AND/OR DOCUMENTATION, AND THE PROGRAM LICENSE GRANTED HEREIN IN PARTICULAR, AND WITHOUT LIMITING OPERATION OF THE PROGRAM LICENSE WITH RESPECT TO ANY PARTICULAR APPLICATION, USE, OR PURPOSE. IN NO EVENT

SHALL CLARK LABS BE LIABLE FOR ANY LOSS OF PROFIT OR ANY OTHER COMMERCIAL DAMAGE, INCLUDING BUT NOT LIMITED TO SPECIAL, INCIDENTAL, CONSEQUENTIAL OR OTHER DAMAGES.

Trademarks

TerrSet and IDRISI are trademarks and registered trademarks respectively of Clark University. Windows and Access are trademarks of Microsoft Corporation. ArcGIS and its related components are trademarks of Esri, Inc. MapInfo is a registered trademark of MapInfo Corporation. Adobe and the Adobe logo are trademarks of Adobe Systems Incorporated. GfK Macon's digital maps are protected by copyright (Copyright GfK Macon GmbH). All other product names mentioned in this volume may be trademarks or registered trademarks of their respective companies and are hereby acknowledged.

About Clark Labs

Clark Labs is dedicated to the research and development of geospatial technologies for effective and responsible decision making for environmental management, sustainable resource development and equitable resource allocation.

Clark Labs is best known for development of pioneering geospatial software products. IDRISI was the first GIS software system specifically targeted at a microcomputer platform. Since its introduction in 1987, it has grown to become one of the most widely used raster systems of its type. More recently, Clark Labs has introduced major new products such as the Land Change Modeler and the Earth Trends Modeler. TerrSet continues this trend and consolidates them into a single integrated system.

Based within the world-renowned Graduate School of Geography at Clark University, Clark Labs is known for pioneering advancements in areas such as decision support, uncertainty management, classifier development, change and time series analysis, land change modeling and dynamic modeling. Partnering with such organizations as the Gordon and Betty Moore Foundation, Esri Inc., Google.org, USDA, the United Nations and Conservation International, Clark Labs leverages its academic base to develop innovative and customized research tools, provide software solutions to organizations in need and apply geospatial expertise to a range of real-world problems.

Contacting Clark Labs

Clark Labs. Clark University. 950 Main Street. Worcester, MA. 01610-1477. USA

Customer Support: +1.508.793.7526

Fax: +1.508.793.8842

Email: clarklabs@clarku.edu

Website: www.clarklabs.org

Facebook: Clark Labs

Our office hours are from 09:00-17:00 hours US Eastern Time (-5 hours GMT Winter/-4 hours GMT Summer) Monday through Friday.

Clark Labs Return Policy

Returns are only accepted due to installation difficulties and must receive prior authorization from Customer Support.

Clark Labs Technical Support

Clark Labs is dedicated to providing registered users with quality technical support. We maintain free software updates (within the same version) that any registered user can download from our website at www.clarklabs.org. Check the version number of your installed software from the Help menu About box and then visit our web site's download area to compare with the update's release number. If your installed version has a lower release number, download and install the update. After updating, check to see if your problem has been resolved.

If you are experiencing a technical problem with a specific module or you are receiving unexpected results, the problem resolution may be found within the Help System. Check the context-sensitive Help for the module you are running. Information concerning the module's limitations may be found in the respective Notes section.

If the update and the Help System information have not resolved your problem, contact our Technical Assistance Staff via email at CLARKLABS.ORG. You must provide the following in your initial contact with us:

- Your Customer ID number, name, phone and e-mail address (if available).
- The name and version number of the Clark Labs product you are using.
- A description of your hardware and operating system.
- A detailed description of the problem you are experiencing.

This should include a list and description of the data sets involved, a description of the operation you are trying to accomplish, and a step by step description of what you have tried so far (with the specific values you entered into the module's dialog box). You should also include the exact text of any error messages you receive.

CHAPTER TWO

QUICKSTART

Projects and Session Files

Projects are an organizing concept throughout TerrSet. Projects contain folders which in turn contain files. Projects are similar to the concept of a geodatabase in some other systems – a collection of geographic data for use in a particular project.

TerrSet recognizes two kinds of folders – a *working folder* where all outputs are placed and *resource folders* from which files can only be read. Projects can contain only one working folder but an unlimited number of resource folders. Important data that will be used again and again without being changed should be placed in resource folders.

All TerrSet components need to use projects. They are created and modified with TerrSet Explorer – the collapsible tool adjacent to the left edge of the screen. To start a new project, click on the Projects tab within Explorer and right-click to access a menu of choices. The lower panel of Explorer allows you to modify the folders associated with a project.

Some of the vertical applications in TerrSet (such as the Land Change Modeler and the Earth Trends Modeler) tend to be used repeatedly with the same data and parameter settings. In these cases, a *session* file is used to record these settings. Session files record information about layers, their role and associated parameters. However, session files do not record the locations of layers. That is the role of the project. Using this logic allows session files to be moved to different locations. Only a few TerrSet components require session files.

Layers and Map Compositions

The files in TerrSet project folders define two forms of geospatial data – layers and map compositions. Layers represent elementary geographic themes such as elevation, roads, cities and so on. Map compositions contain one or more layers along with ancillary information such as titles, legends, a scale bar, etc. TerrSet recognizes six basic layer types – raster layers (images), point vector layers, line vector layers, polygon vector layers, text vector layers and photo vector layers. Each layer is actually stored as a pair of files – a data file and an associated metadata file that contains vital information about the layer such as its georeferencing system and bounding coordinates.

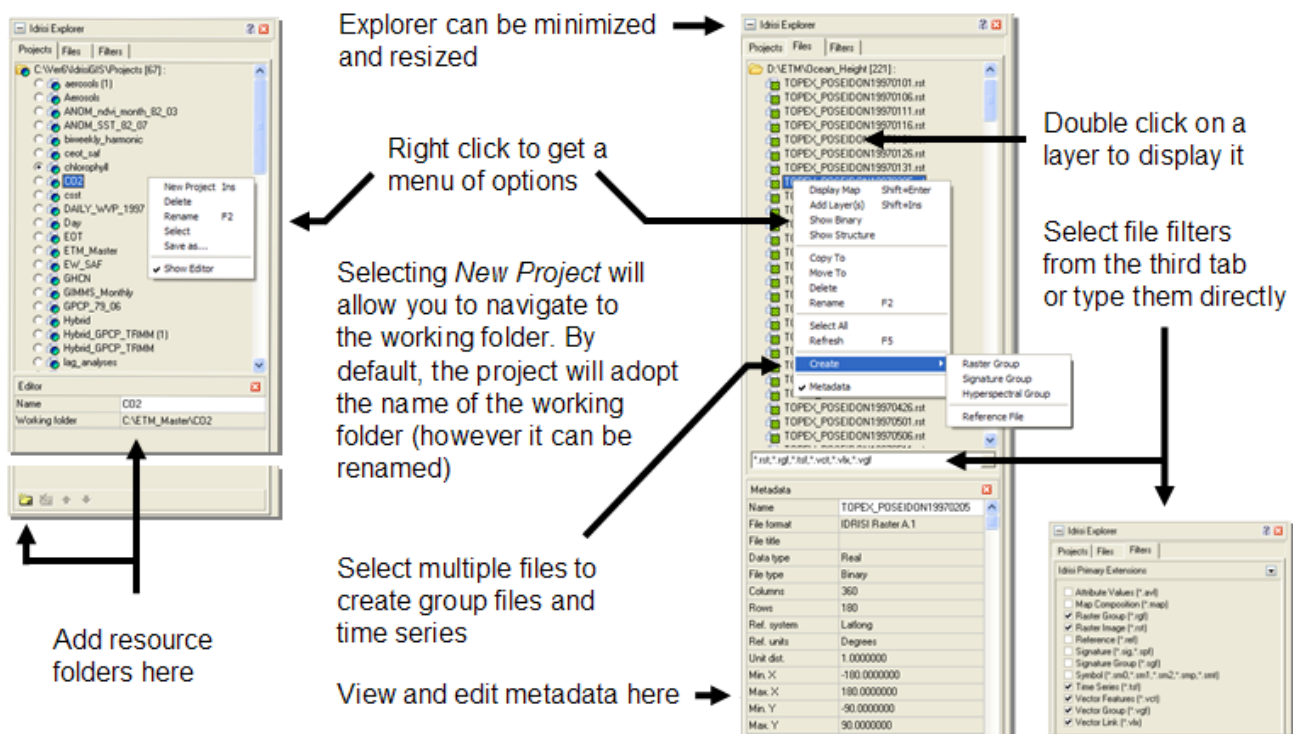
Map compositions are begun as soon as a single map layer is displayed. Additional layers can then be added to the composition along with ancillary information, as will be described below.

Importing Data

TerrSet uses open file formats that are specially optimized for display and analysis. External formats need to be converted using the various import routines in the File|Import menu. These are very quick and support most common formats. Similarly, TerrSet provides Export utilities for moving data to other systems.

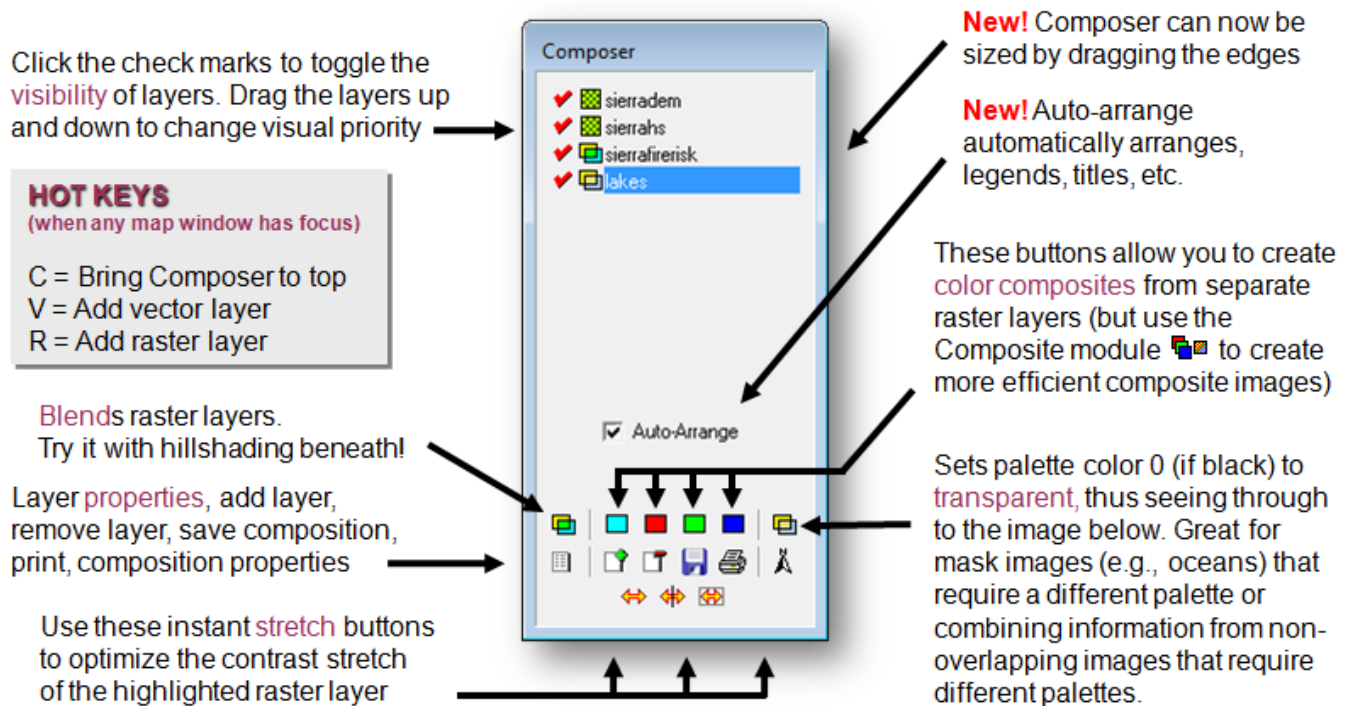
Explorer

TerrSet Explorer allows you to manage both projects and the layers and compositions they contain. The Filters tab allows you to control what files are viewed while the Files tab shows the filtered results for all project folders. In Explorer, layers are shown as objects with their associated metadata in the lower panel. Explorer can also be used to display map layers (double click or right-click) as well as many other options (right-click) such as adding the layer to the map window that currently has focus, deleting the layer, renaming it, etc. The graphic below summarizes these options:



Displaying Layers and Compositions

Layers can be displayed either by using DISPLAY Launcher (accessed by clicking its toolbar icon or from the File|Display menu) or by launching them in Explorer (double click or right-click). To assist in this process, a special tool known as Composer is launched whenever a new map window is opened. Composer allows you to modify the symbolization of map layers as well to save the map composition for later display. The figure below summarizes these options:



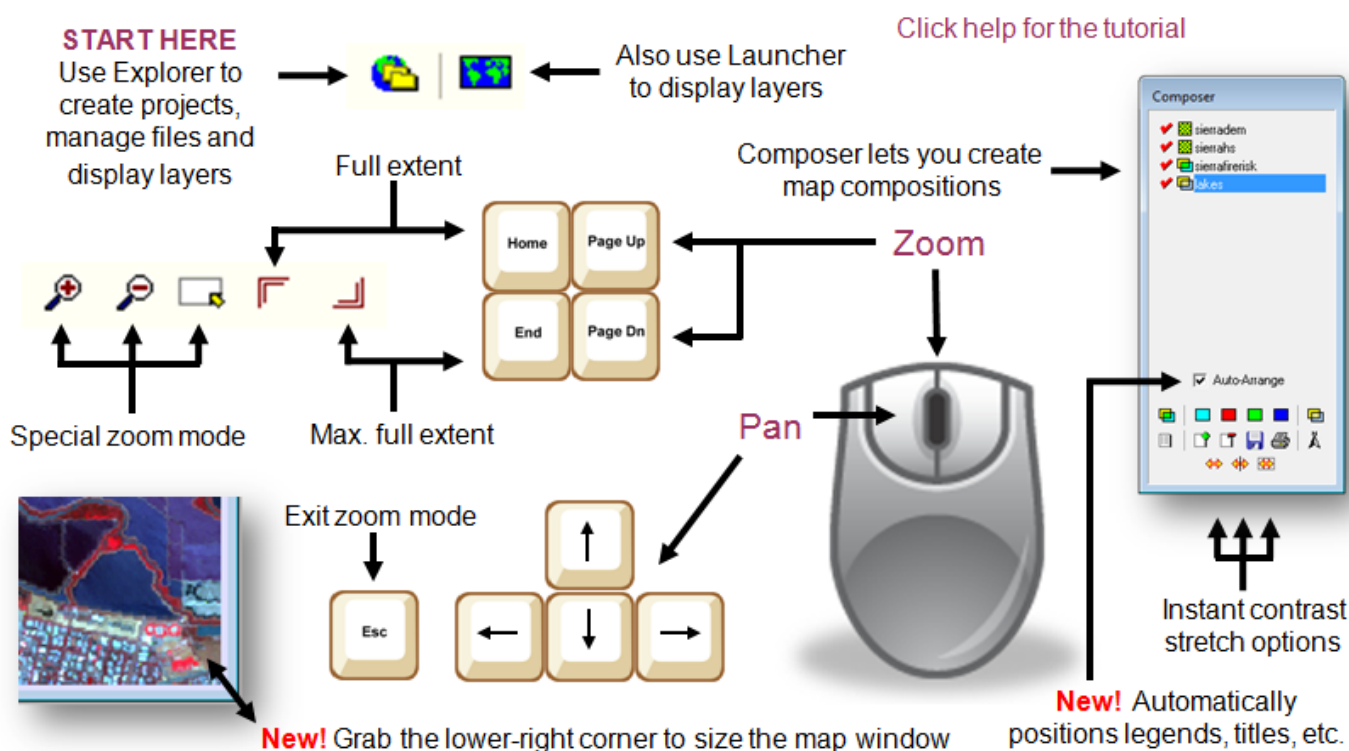
Composer and DISPLAY Launcher both provide options to change the way a map layer is symbolized. While a large library of symbol files is provided, users can create their own with SYMBOL WORKSHOP. Five types of symbol files are used by TerrSet – palette files for raster images, and point, line, polygon and text layer symbol files.

Note that when multiple layers are displayed in a composition, some interesting options are provided. To create color composites dynamically from three bands in the composition, use the color composite buttons to designate the red, green and blue bands of the composites. Image layers can also be blended with layers below them (a great option for adding hill shading) or be displayed with transparency. Transparency makes palette color 0 transparent – a tremendously useful procedure whenever you wish to remove the background from the visual display.

Finally, make special note of the stretch buttons provided at the bottom of the Composer dialog. The left-most optimizes the display of raster layers by using a linear stretch with saturation (if you are unfamiliar with the concept, it is described in the next chapter). The middle stretch button does the same, but stretches the image symmetrically around 0. This is designed for images with both positive and negative values. The right-most stretch button does the same as the left-most button, but calculates the stretch based only on the pixels currently in view. This can be especially valuable with newer remotely sensed images with very high radiometric resolution.

Display Navigation and Inquiry

TerrSet provides a wide variety of options for display navigation to provide compatibility with other systems. The most flexible option is to use a mouse with a wheel. Unlike some systems that require you to change modes between pan mode (where the display is moved horizontally) and zoom mode (where the scale is changed), TerrSet merges these into a single mode. Hold the mouse down and move, pans the display. Scroll the wheel zooms. To inquire about the value at a location in the layer highlighted in Composer, simply click with the mouse. To inquire about the value for all layers in a composition, click the IDENTIFY button on the menu. The figure below highlights the major tools for display navigation and inquiry.



Pyramids, Collections and Time Series

The volume of data associated with raster images is substantial. This can impact the speed of display quite substantially. To handle this, many systems, including TerrSet use *pyramids*. Pyramids refer to the storage of an image at several resolutions to facilitate display. If you are experiencing a slow display response, find the image involved in Explorer and right-click to select the option to create pyramids. It doesn't take long and the result can greatly facilitate image display.

Finally, note that many of the special features of TerrSet operate on collections of map layers. The most fundamental of these is the *Raster Group File* (RGF). Group files allow you to refer to multi-band images using a single identifier, greatly simplifying the entry of file names. These are indispensable when using the IDRISI Image Processing tools. When members of the same group are displayed in separate map windows, the toolbar offers the option of having them synchronously zoom and pan. Again, Explorer provides the easiest (but not the only) way of creating group files.

A special form of group file is a time series file. A time series file is a group of raster images with additional metadata that documents the time dimension. The Earth Trends Modeler is a special tool designed specifically for the analysis of image time series.

Making Complex Tasks Easier – Graphical and Programmatic Modeling Tools

The IDRISI GIS Analysis and Image Processing tools provided in TerrSet provide a very extensive set of basic analysis tools. Used in combination, they can form the basis for an unlimited set of possible analyses. However, complex sequences can be both tedious and difficult to execute without errors. Fortunately, TerrSet provides a number of important tools for automating complex analyses.

For many, the most important automation tool is MACRO MODELER. MACRO MODELER is a graphical programming environment where analytical sequences can be linked together in a visually intuitive environment. Of special note is that MACRO MODELER also provides feedback loops and the ability to iterate analyses over groups of layers (great for when you have a tedious operation to perform on many layers). Also note that a similar graphical modeling tool is provided for the special case of multi-criteria/multi-objective decision making. This facility is known as the SPATIAL DECISION MODELER (SDM).

TerrSet is a COM server. This is a fundamental Microsoft Windows protocol whereby client applications can request that TerrSet execute various actions in a specified sequence. These actions include virtually all of the features of the IDRISI GIS Analysis and Image Processing tools as well as the ability to access TerrSet's projects and display facilities. Tools that can be used to access the COM server include scripting languages such as Python or high-level programming languages such as Delphi and C++. Collectively these services are known as the TerrSet *Application Programming Interface* (API). For reasons of backward compatibility and in recognition of the fact that the API is almost exclusively providing access to the IDRISI GIS and Image Processing tool sets, the API is referenced by COM-compliant languages using the server name IDRISI32.

The TerrSet Constellation

TerrSet provides an unprecedented set of tools for monitoring and modeling the earth system. They can be accessed from the main menu buttons. The IDRISI GIS Analysis and Image Processing facilities access a traditional menu of basic tools, and as such constitute *horizontal* applications – basic tool sets without a single focus. These are the core of TerrSet. The remaining components of the TerrSet constellation are *vertical* applications – applications focused on a specific task sequence. For example, the Land Change Modeler is focused on the specific sequence of operations associated with modeling and predicting land cover change. At present, the full TerrSet constellation includes:

- IDRISI GIS Analysis
- IDRISI Image Processing
- The Land Change Modeler (LCM)
- The Earth Trends Modeler (ETM)
- The Habitat and Biodiversity Modeler (HBM)
- The Ecosystem Services Modeler (ESM)
- The Climate Change Adaptation Modeler (CCAM)
- GeOSIRIS

Subsequent chapters will introduce each of these TerrSet components.

CHAPTER THREE

TERRSET SYSTEM OVERVIEW

TerrSet consists of a main interface program (with a menu, toolbar and display system) and eight basic components. Two of these form the backbone of the system – the IDRISI GIS Analysis tools and the IDRISI Image Processing tools. IDRISI was the first GIS and Image Processing System specifically developed for a microcomputer platform¹. Over the decades that followed, IDRISI has matured into a research grade system for geographic analysis with ground breaking developments in areas such as multi-criteria/multi-objective decision making, machine learning and time series analysis. The IDRISI collection of over 300 program *modules* provide the basis for a virtually unlimited range of geospatial analytical operations that can be choreographed with the assistance of a variety of graphical and programmatic modeling tools. While these two components are general purpose tool sets, TerrSet also provides a collection of six vertical applications that focus on monitoring a modeling the earth system: the Land Change Modeler (LCM), the Habitat and Biodiversity Modeler (HBM), the Ecosystem Services Modeler (ESM), the Earth Trends Modeler (ETM), the Climate Change Adaptation Modeler (CCAM) and GeoSIRIS – a special tool for national level REDD (Reducing Emissions from Deforestation and Forest Degradation) planning.

In TerrSet, geographic data are described in the form of *map layers*—elementary map components that describe a single theme. Examples of map layers might include a roads layer, an elevation layer, a soil type layer, a remotely sensed solar reflectance layer and so on. All analyses act upon map layers. For display, a series of map layers may be brought together into a *map composition*.

Because geographic data may be of different types, TerrSet incorporates two basic forms of map layers: *raster image layers* and *vector layers*.² The former consist of dense grids of numeric data that describe the condition of the land (or ocean) at each cell location (called a *pixel*), while the latter describe features in space by means of a set of vertices (described by numeric coordinates) that delineate their position, course or boundary. For all of its operations, TerrSet uses the well-known IDRISI raster and vector data structures. Thus a wide range of import and export procedures are provided (under the File menu) for converting data from one format to other formats.

While TerrSet uses both raster and vector data layers, its focus is decidedly raster. The simple structure of raster images with its implicit topology provides an exceptional basis for geographic analysis. In addition, raster is the native structure of remotely sensed data. Thus TerrSet is strongly oriented to monitoring and modeling the earth system with extensive capabilities for working with remotely sensed data. In contrast to systems that focus on data management and the mapping of geospatial features, TerrSet focuses on space/time as a continuum and the analysis of variations in condition.

¹ Eastman, J.R., Warren, S., (1987) "IDRISI : A Collective Geographic Analysis System Project", Proceedings, *AUTOCARTO 8*, 421-430.

² For an in-depth discussion of raster and vector data structures, see the chapter **Introduction to GIS** in this volume.

Interface Elements

The TerrSet Application Window

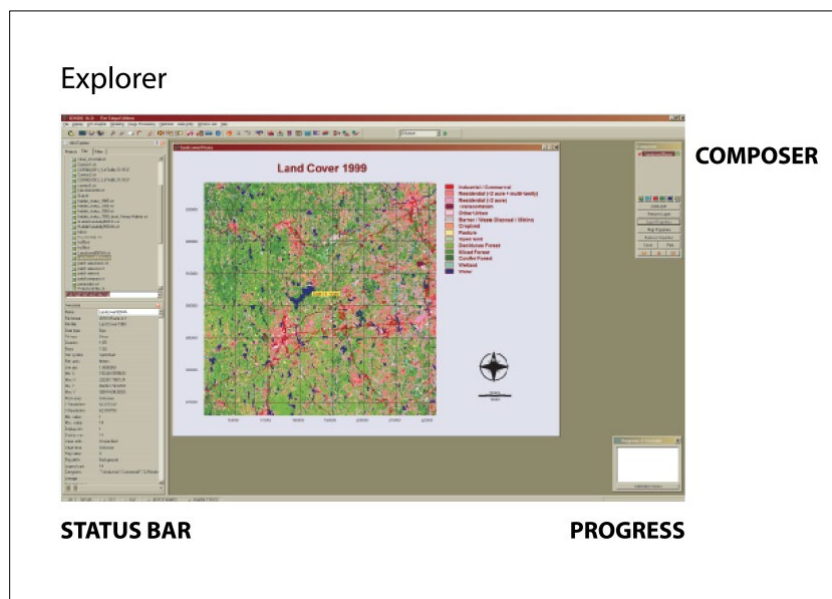
When TerrSet is opened, the application window will completely occupy the screen (in Windows terminology, it is automatically maximized). Though not required, it is recommended that it be kept maximized because many of the dialog boxes and images you will display will require a substantial amount of display space. The TerrSet application window includes the menu, the toolbar, TerrSet Explorer, and the status bar. In addition, when vertical applications such as the Land Change Modeler are activated, they will occupy additional vertical panels similar to Explorer.

The Menu System

The menu system is at the top of the application window. You can access it using the mouse or similar pointer device. If you select a menu option that includes a right-pointing arrow, a submenu will appear. Clicking on a menu option without a right-pointing arrow will cause the selected action (such as opening the dialog box for a module) to appear.

The Toolbar

Just below the menu is a set of buttons that are collectively known as the toolbar. Each button represents either a program module or an interactive operation that can be selected by clicking on that button with the mouse. Some of these buttons toggle between an active and an inactive state. When active, the button appears to remain depressed after being clicked. In these cases, the button can be released by clicking it again. You will also notice that some buttons may not always be available for use. This is indicated when the button icon appears grey. Hold the cursor over an icon to cause the name of the function or module represented by that icon to momentarily appear. The set of icons represents interactive display functions as well as some of the most commonly-used modules.



The Status Bar

At the bottom of the screen is the status bar. The status bar provides a variety of information about program operation.

When maps and map layers are displayed on the screen and the mouse is moved over one of these windows, the status bar will indicate the position of the cursor within that map in both column and row image coordinates and X and Y map reference system coordinates. In addition, the status bar indicates the scale of the screen representation as a *Representative Fraction* (RF).³

The status bar will also indicate the progress of the most recently launched analytical operation with a graphic progress bar as well as a *percent done* measure. Since TerrSet has been designed to permit multitasking of operations, it is possible that more than one operation may be working simultaneously. To see a listing of all active processes and their status, simply double click on the progress bar panel at the bottom right of the screen. Modules may also be terminated from here.

TerrSet Explorer

TerrSet Explorer is a general purpose utility to manage and explore TerrSet files and projects. Use TerrSet Explorer to set your project environment, manage your group files, review metadata⁴, display files, and organize your data. Tools are provided to copy, delete, rename, and move layers. TerrSet Explorer can also be used to view the internal structure of map layers and to drag and drop files into TerrSet dialog boxes. TerrSet Explorer is permanently docked to the left of the TerrSet desktop. It cannot be moved but it can be horizontally resized and minimized against the left margin. Note also that the relative space given to listing files versus displaying their metadata can also be resized.

Modules and Applications

TerrSet provides access to both elementary operations, called *modules* and more complex analytical sequences known as *applications*. For example, OVERLAY and RECLASS are modules that perform vary basic operations while the Land Change Modeler (LCM) is an application. Modules and vertical applications may be accessed in three ways:

1. by selecting the module in the menu structure and activating it with a click of the mouse,
2. by selecting its program icon (if one exists) on the toolbar just below the menu, or
3. by typing in or selecting the module/application name from the alphabetical list provided by the Shortcut utility on the right-side of the tool buttons.

Data Organization

Modules and applications act upon data—map layers and tabular data stored in data files. These files are stored in folders (also known as directories) and sub-folders (sub-directories) on hard drives either in the computer or available over a network. The location of any specific data file is thus designated by a name consisting of the filename plus the drive, folder and sub-folder locations. Further, filenames have an extension (such as “.rst”) that indicate what type of file they are. Collectively, these are known as the *path*

³ A *Representative Fraction* expresses the scale of a map as a fractional proportion. Thus, an RF of 1/10,000 indicates that the image on the screen is one ten-thousandth of the actual size of the corresponding features on the ground. TerrSet automatically determines your screen resolution and dimensions, which, in combination with the minimum and maximum X and Y coordinates of the image or vector layer, are used to determine the RF for any display window.

⁴ *Metadata* refers to the properties associated with a data set (effectively, data about data).

(since they specify the route to a particular data file). For example, "c:\massachusetts\middlesex\census_tracts.vct" might designate a vector layer of census tracts contained in the Middlesex County sub-folder of the Massachusetts folder on hard disk drive C.

TerrSet can work with files from any folder (including network folders). However, it can often be a nuisance to continually specify folder and sub-folder names, particularly when a specific project typically accesses data in a limited set of folders. To simplify matters, TerrSet allows you to specify these folders as belonging to a specific *Project*. When the paths to these folders are designated in a Project, TerrSet finds data very easily and with minimal user intervention.

The Project Working Folder

Within any Project, the most important folder is the *Working Folder*. Users may choose to store all of the data for a project in the Working Folder (particularly for smaller projects). However, for larger projects, or projects requiring libraries of carefully organized and protected data sets, additional *Resource Folders* can also be added to the Project.

While the user is always able to specify any location for input or output data, designating the most commonly-used folders in a Project facilitates the use of TerrSet. For instance, if input filenames are given without a path designation, TerrSet automatically looks in the Working Folder first, then in each Resource Folder in turn to locate the file. Similarly, if no alternative path is specified, output data are automatically written to the Working Folder.

Project Resource Folders

Resource Folders contain data that can be accessed quickly through the TerrSet *Pick List* (see below). A Project can contain any number of Resource Folders. To the user, the Working Folder and all Resource Folders function as a single Project. For example, one might ask TerrSet to display a raster map layer named "landuse". If the full path is not supplied as part of the filename, TerrSet first looks for the file in the Working Folder, and then each Resource Folder in turn, ultimately displaying the first instance it finds. This is very convenient as it allows the user to enter filenames without paths. However, if duplicate filenames exist in separate folders that are part of the same project, the user must exercise caution and remember the order in which TerrSet accesses the project folders.

Creating or Modifying Projects

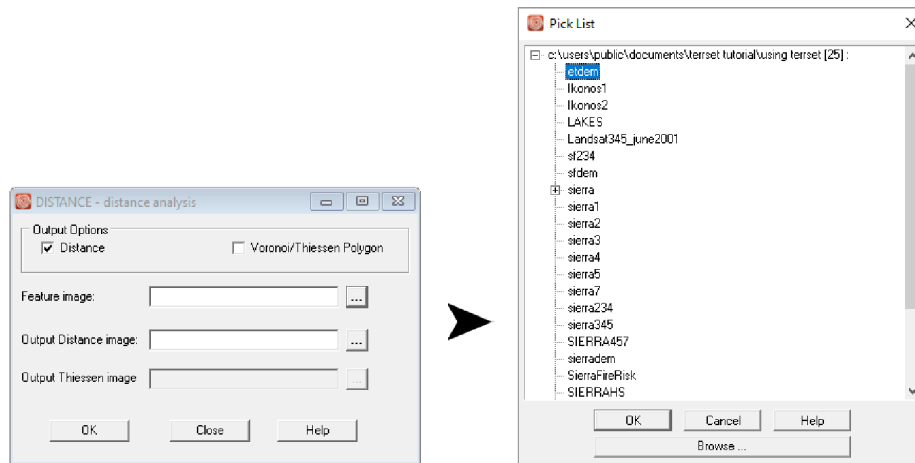
Establishing a Project and its associated Working and Resource Folders can be set in TerrSet Explorer. Explorer is always present and can be minimized or maximized using the [+] or [-] button at its top. Within Explorer, use the Projects tab to select or create projects and their associated working and resource folders. Creating a new project is done by right-clicking within the Projects tab and selecting the New Project option. You will then be presented with a dialog from which you can either navigate to the folder you wish to use as the Working Folder, or create a new folder to be used as the Working Folder. Once your Working Folder is set, the metadata window in the lower part of Explorer will allow you to add any number of Resource Folders. The structure of projects can be changed at any time with this tool.

Working with TerrSet Dialog Boxes

When a module is activated, a dialog box appears with information on data or option choices that are required for the module to run. Virtually all input boxes must be filled with information before the module can be run. Only the title and measurement units input boxes under the Output Documentation button can be left blank (although it is recommended that these boxes also be filled). In some cases, input boxes already contain values, or will fill in with values as other information is entered. These are default values that may be freely edited. In those cases where a set of choices should be made, the most common settings are normally pre-selected. These

should be examined and changed if necessary. The system will display an error message in cases where a needed element of data has been left out.

Pick Lists



With the vertical applications such as the Land Change Modeler, the analytical sequence is broken down into major stages organized as tabs. Within each tab, basic tasks are organized into collapsible panels, each of which is effectively identical in character to a module dialog box.

Whenever a dialog box requires the name of an input data file (such as a raster image), you have two choices for input of that value. The first is simply to type the name into the input box. Alternatively, you can activate a *Pick List* of choices by either double-clicking into the input box, or clicking the *Pick List* button to the right of the input box (the small button with the ellipses characters "..."). The Pick List is organized as a tree directory of all files of the required type that exist in the Working and Resource Folders. The Working Folder is always at the top of the list, followed by each Resource Folder in the order in which they are listed in the Project. The files displayed in the Pick List depend upon the particular input box from which the Pick List was launched. All TerrSet file types are designated by unique filename extensions. If an input box requires a raster image, for example, the Pick List launched from that input box will display only those files that have an .rst extension.

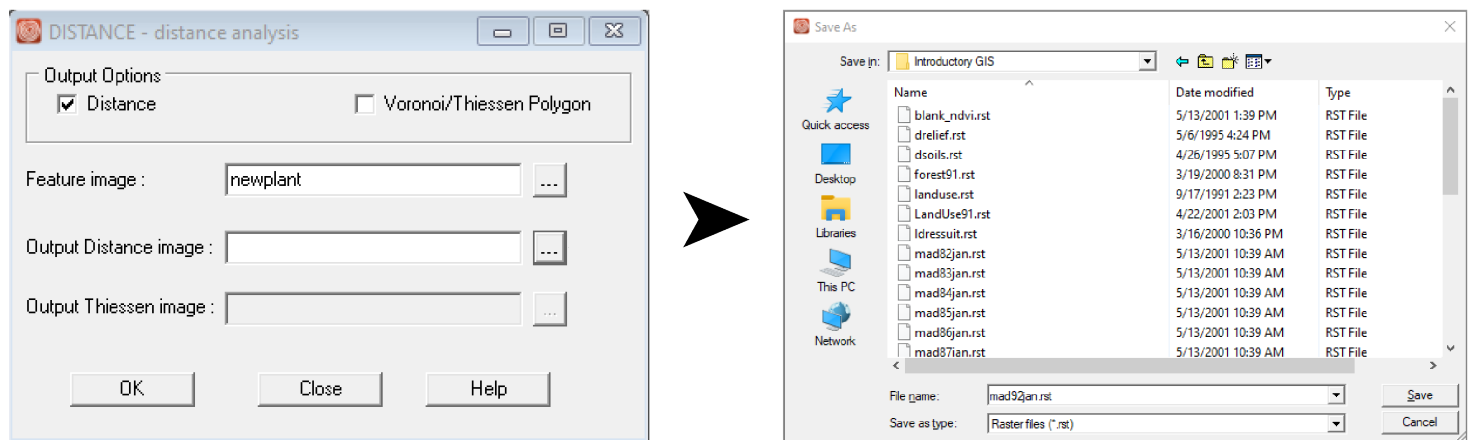
You may select a file from the Pick List by first highlighting it with a click of the left mouse button, and then selecting it by clicking the OK button or pressing the Enter key. Alternatively, double-clicking the highlighted entry will also cause it to be selected. The Pick List will then disappear, and the name of that file will appear in the input box. Note that you can also choose a file that is not in the current project environment by clicking on the Browse button at the bottom of the Pick List.

Output File Names

You will normally want to give output filenames that are meaningful to you. It is common to generate many output files in a single analysis and descriptive filenames are helpful for keeping track of these files.⁵ File names may be long and can contain spaces and most keyboard characters (all except / \ : * ? " < > and |). It is not necessary for the user to specify the three-letter filename

⁵ The documentation file for output raster and vector data files includes in the lineage field the command line from which the layer was created. This may be viewed with the Metadata utility in Explorer.

extension as TerrSet takes care of this. By default, TerrSet will internally add whatever path information is required to store the output in the Working Folder. This can be bypassed by entering a full path in the output filename box. However, we do not recommend this. You will find that your experience with TerrSet will be much smoother if you establish a proper Project and allow it to write files automatically to the Working Folder.



Clicking on the Pick List button to the right of the output filename box will bring up a standard Windows *Save As* dialog box with an automatically-generated output filename as a suggestion. The name and location can be freely changed. However, as stated previously, we do not recommend that you change the path. To change the name, simply click into the filename box and type the desired name.

As stated above, users typically give descriptive filenames to help avoid confusion as the database grows. However, TerrSet can also generate automatic names for output data files. Automatic output names are intended for the quick naming of temporary files.

Automatic output filenames begin with a user-defined three-letter prefix (set to TMP by default),⁶ followed by a three digit number. By default, filenames begin with TMP000 and progress up to TMP999. As these files are normally temporary, overwrite protection (see below) is disabled for files with names that begin with the designated prefix followed by three digits. Note that after exiting TerrSet, the cycle will begin again with TMP000 the next time TerrSet is used. Since the numbering sequence starts from 000 with each new session, the data in such files is likely to be lost in a subsequent session unless it is intentionally saved under a new name.

To use the automatic output filename feature, move to the output filename box concerned, and either double-click in the box or click the Pick List button. If the double-click is used, an automatic filename will be generated and placed into the output filename box. Since no path is specified with this method, output will automatically go to the Working Folder.

In the case where the pick button is clicked beside an output filename box, a standard Windows *Save As* dialog box will appear with the temporary name (e.g., TMP001) already selected.

Overwrite Protection

By default, TerrSet checks if the given output file already exists. If it does, it will ask whether you wish to overwrite it. If not, you will be returned to the dialog box where you can enter a different name.

⁶ The automatically-generated output file prefix may be changed in the User Preferences option of the File menu.

There are three exceptions to this logic. The first is when an automatic output name is generated (see the section immediately above). The second is when the user removes overwrite protection in the User Preferences dialog under the File menu. And the third is when modules are run in macro mode.

Getting Help

While using TerrSet, help is always close at hand. The TerrSet System contains an extensive Help System that can be accessed in a variety of ways. The most general means of accessing help is to select it from the main TerrSet menu or the special Help button on the toolbar. However, each program module and application panel also contains a help button. In this case, help is context sensitive, accessing the section in Help associated with that particular module. In either case, it is worth becoming familiar with the many options the Help System provides. To do so, select the Using Help option from the main Help menu.

Data Files and Structures

Map layers are the fundamental units of display and analysis in GIS. A map layer is an elementary theme, a single phenomenon that can be mapped across space. Examples would include a landuse layer, a roads layer, an elevation layer, a soils layer, and so on. Map layers, then, are somewhat different from traditional maps. Traditional maps usually consist of several map layers. For example, a topographic map sheet would typically consist of a contour layer, a forest cover layer, an administrative boundaries layer, a roads layer and a settlements layer. In TerrSet, this same concept is applied. A *map* is a graphic representation of space that is composed of one or more map *layers* using a tool called *Composer*.

This breakdown of the map into its elementary constituents offers important advantages. Clearly it allows for the simple production of highly customized maps; we simply pull together the layers we wish to see together. However, the more important reason for storing data this way is that layers are the basic variables used by the GIS in analytical modeling. Thus, for example, we might create a map of soil erosion as a mathematical function of soil type, slope, landcover, and precipitation layers. Indeed, the breakdown of geographic data into layers allows for the production of an extraordinary range of mathematical and logical models.

Map Layer Types

As detailed in the **Introduction to GIS** section, there are two basic types of layers in GIS: raster image layers, and vector layers.⁷ Some systems deal exclusively with one or the other type. TerrSet incorporates both since they each have special advantages. However, while they have equal stature in terms of display and database query, TerrSet offers a far broader range of analytical operations for raster layers. This is partly because of the historic development of TerrSet based on the foundation of the IDRISI engine and partly because the range of analytical operations possible in GIS is far greater with raster layers (as a consequence of their strong implicit topology).

All TerrSet raster layers have the same basic structure. Sub-types exist only in terms of the numeric data type that is used to record cell values. As one can easily imagine, raster images are high in data volume. Thus, small changes in the underlying data type can make huge differences in storage requirements. TerrSet supports raster images stored with byte (1 byte per pixel), integer (2 bytes per pixel), and real (4 bytes per pixel) numbers as well as a special format for the storage of full-color images (RGB24).

Vector layers describe the location and character of distinct geographic features. Sub-types include point, line, polygon, text and photo layers. Point features are phenomena that can be described by a single point, such as meteorological station data or (at small scales) town locations. As the name suggests, line vector layers describe linear features such as rivers and roads. Polygons refer to

⁷ In TerrSet, the terms raster layer, raster image layer and image are all used interchangeably.

areas. Because the boundaries of the areas are made up of straight line segments between points, the figure formed is technically a polygon. Text layers are used to describe the location and orientation of text labels. Finally, photo layers indicate the positions and orientation of geographically encoded photos. Photo symbols can be clicked on to display the image. This layer type is intended to facilitate ground-truthing operations for remotely sensed images.

Map Layer Names

All analytical functions in TerrSet act on map layers. Thus, the input and output filenames that are given in TerrSet operations are typically those of map layers. In your normal use of TerrSet, map layers are specified with simple names that do not need to express their data path (the TerrSet Project defines the paths to your Working and Resource folders). It is also not necessary to specify their filename extensions (since TerrSet looks for and creates specific extensions automatically).

Consistent with the 32-bit variants of the Windows operating system, layer names may contain any letter of the alphabet as well as numerals, spaces and most symbols (excluding \ / : * ? " < > and |). For example, layers might have names such as "soils" or "landuse 1996." In principle, filenames can be up to 255 characters in length. However, in practical use, it is recommended that they be only as long as necessary so the entire filename will be visible in the Pick Lists and dialog input boxes.

In normal use, no distinction is made between image and vector layer names. However, their type is always known by the system—by the context in which they were created and by the file extension that the system uses internally to store each layer. Thus a raster image layer named "soils" would be stored as "soils.rst" internally, while a vector layer named "soils" would be stored as "soils.vct". As a result, it is possible to have raster and vector layers of the same name. However, most users of TerrSet allow this to occur only in cases where the two files store vector and image manifestations of the same underlying data.

When filenames are input into TerrSet dialog boxes, the system always knows the appropriate layer type in use. Indeed, only those files of the appropriate type for that operation will display in the Pick List.

Map Layer Data Files

While we refer to a map layer by a simple name, it is actually stored by TerrSet as a pair of data files. One contains the spatial data, while the second contains information about those data (i.e., documentation or *metadata*). Thus, while raster images are stored in data files with an ".rst" extension, they each have an accompanying documentation file with an ".rdc" extension (i.e., raster documentation). Similarly, while vector data files have a ".vct" extension, each has a companion documentation file with a ".vdc" extension (i.e., vector documentation). In use, however, you would refer to these layers by their simple layer names (i.e., "soils" and "districts" respectively). The contents of data documentation files are described in detail below in the section on file structures.

Vector Layer Collections

Sometimes it makes sense to group map layers together into collections. A good example is when you have a set of vector features associated with a data table. For instance, a vector file might be created of census tracts within a city. Each of these tracts possesses multiple attributes such as median income, population density, median educational attainment, and so on. Each of these attributes can be linked to a vector file defining the geography of the census tracts and displayed as a map layer. Thus the combination of the vector file and the entire database table constitutes a collection of map layers.

A vector collection is produced by linking a geographic or feature definition vector file (such as the digital representation of census tracks mentioned above) with a database table.⁸ This is called a linked-table collection and is established with a vector link file (.vlx) and is created in Database Workshop. All the elements of the collection must reside in the same folder.

Vector link files (.vlx) associate a feature definition vector file with a database table. As a consequence, each of the fields (columns) of data in the data table becomes a layer, with the entire set of fields producing a linked-table collection. As you will see, TerrSet has special features for the display and analysis of relationships between layers in such a collection. All that is required is a link file that can establish four important properties of the collection:

1. the feature definition (i.e., spatial frame) vector file;
2. an associated attribute database;
3. a specific table within that database; and
4. the database field that contains the identifiers which can associate database records with vector features (the *link* field).

In Database Workshop, the Establish Display Link utility is used to construct these link files. Only vector point, line or polygon layers can be specified as a feature definition vector file. The database files and associated tables must be Microsoft Access Database Engine files with an ".accdb" extension.

Raster Group File

A group of independent raster layers can be associated into a raster group file (.rgf). As the name suggests, a group file is simply a listing of a set of independent layers that one wishes to associate together. For example, a group file might consist of the seven bands of imagery that make up a Landsat satellite image. Many of the image processing procedures in TerrSet allow you to specify a group file as input rather than each individual band to be used. This saves a lot of time when the same group of images is used in several procedures. Additionally, when more than one member of a group is displayed on the screen, you have the option (by selecting the Group Link button on the toolbar) to treat them identically during zoom and pan operations. Thus when one member is zoomed or panned, all other members of that group that are on the screen will zoom or pan simultaneously.

Raster Group Files can be created in two ways. One is to use a facility under the File menu called Collection Editor. The second is to use TerrSet Explorer where a group of raster layers can be highlighted, and a right click can be used to select the option to create a raster group file. It will create the file with a default name which can subsequently be renamed.

Other Forms of Group Files

The need to refer to groups of files is fairly common in TerrSet. While Raster Group Files (RGF) are the most commonly encountered, other forms exist. *Vector Group Files* (VGF) are vector counterparts to RGF's in that they contain a set of layers with completely different geographies (as opposed to vector collections that contain multiple layers regarding the same set of geographic features). However, they are rarely needed. Signature files used in image classification may also be collected together in *Signature Group Files* (SGF) and *Hyperspectral Signature Group files* (HSG). These group files are also created with TerrSet Explorer. Note that some early versions of this software used an additional form of group file known as a Time Series. This use has been deprecated. Image time series are now specified as a combination of an RGF with a special form of documentation file known as a *Time Series File* (TSF).

⁸ Feature definition vector files and image files describe the geography of features but not their attributes. Rather, the features have identifier values that are used to associate attributes stored in database tables or attribute values files.

Collection and Group Member Naming Conventions

To specify a vector layer that belongs to a collection, one uses the "dot" convention—i.e., the reference is a combination of the collection name and the layer name, separated by a dot. For example, suppose one had a link file named `CensusTractData99` associating a feature definition vector file with a database table of census statistics. Each of the layers in the collection would be referenced by a combination of the collection name and the appropriate field name in the table. Thus, the `Median_Income` layer would be referred to as:

`CensusTractData99.Median_Income`

Similarly, one might create a group file of the seven multispectral bands in a Landsat image. Perhaps the bands have names such as `Band1`, `Band2`, etc. If the group file that was created to associate them was named `Oahu`, then a reference to the second band would be as follows:

`Oahu.Band2`

Pick Lists for Group File and Collection Display

The Pick List facility that is used for all input boxes is *collection-aware* and *group-aware*. For example, when a raster image is required, the Pick List will display raster group files as well as raster images. Group files are immediately evident in a Pick List by the "+" sign at their point of connection to the directory tree. If you click on the group file, it will expand to show each of the members. These members are shown with their simple names. However, if you choose one, the TerrSet System will enter its full reference (using the dot convention) into the input box. In some cases, a group filename may itself be selected as input for a dialog box.

Attribute Files

An attribute file contains information about the characteristics of features, but not the geographic definition of those features. TerrSet supports two forms of attribute files, database tables and ASCII values files.

Data Tables

The TerrSet System incorporates the Microsoft Access Database Engine within its Database Workshop facility. Thus, this form of attribute file contains tables and has an ".accdb" extension, and can also be used and modified with Microsoft Access or any other Access-compatible system. The linked layer collection can only be created with this type of attribute file. Microsoft Access Database files may contain multiple tables, each one of which can be linked to form a layer collection. See the chapter **Database Workshop** for more information about using database tables with TerrSet.⁹

Values Files

The second form of attribute file that TerrSet uses is an ASCII values file. This has a very simple file structure that contains the values for a single attribute. It is stored in ASCII text format and consists of two columns of data separated by one or more spaces. The first column contains an identifier that can be used to associate the value with a feature (either raster or vector), while the second column contains the attribute value. Attribute values files have an ".avl" extension. Some modules create these as output files while others

⁹ The .accdb format supersedes the earlier Microsoft .mdb format. If you have data files in this earlier format, they will need to be converted to the newer format. Please refer to the help system for information on how this should be done.

require (or can use) them as input for analyses. Attribute values files can be created with the EDIT module – simply type the two columns directly or paste the two columns from Excel. They may also be imported or exported to or from a database table using Database Workshop. In addition, TerrSet provides a procedure (ASSIGN) to assign those values to a set of features included in a raster image or vector file, and a companion procedure (EXTRACT) to create a new values file as a statistical summary of the values in another image for the features of a feature definition image.

Map Layer File Structures

Raster Layers (.rst)

Raster images are the most fundamental and important data type in TerrSet. They are automatically assigned an ".rst" file extension by the system. Raster layers are both simple in structure and regular in their organization, allowing an extraordinary range of analytical operations. The data structures that TerrSet uses for storing raster images are optimized for simplicity and efficiency.

Raster images define a rectangular region of space by means of a fine matrix of numeric data values that describe the condition or character of the landscape in each cell of a fine grid. These numeric grid cell values are used not only for analysis, but also for display. By assigning specific colors (in a *palette*) to designated numeric ranges, a very fine matrix-like color image is formed (so fine that typically the individual cells cannot be seen without zooming in to a very large scale).

In this grid-cell structure used by TerrSet, rows and columns are numbered starting from zero. Thus an image of 1000 columns and 500 rows has columns numbered from 0-999 and rows numbered from 0-499. Unlike the normal Cartesian coordinate system, the 0,0 cell is in the top left-hand corner. Cells are numbered left-to-right as you might expect, but rows are numbered top-to-bottom. This is common with most raster systems because of the directionality of many output devices, particularly printers, that print from top to bottom.

While the logical structure of an image file is a grid, the actual structure, as it is stored, is a single stream of numbers (which might be imagined as a single column). For instance, an image consisting of 3 rows by 5 columns is stored as a single stream (column) of 15 numbers. It is the image's documentation file that allows TerrSet modules to reconstruct the grid from this list.

The raster documentation file, containing the number of rows and columns, allows the image to be correctly recreated for display and analysis.

As mentioned earlier, the major sub-types of raster images are differentiated on the basis of the data types that are used to represent cell values. TerrSet recognizes four raster data types: integer, byte, real, and RGB24.

1. Integers are numbers that have no fractional part and lie within the range of -32768 to +32767. Integer files are sometimes called 16-bit signed integer files since they require 16 bits (i.e., 2 bytes) of memory per value (and therefore per pixel). Integer values can be used to represent quantitative values or can be used as codes for categorical data types. For example, a soils map may record three soil types in a particular region. Since TerrSet images are stored in numeric format, these types can be given

An image that displays like a grid:

10	15	9	10
1	14	10	11
14	13	11	10

actually has an image
file that consists of a
column of values:

10
15
9
10
1
14
10
11
14
13
11
10

integer codes 1, 2 and 3. The documentation file records a legend for these, on the relationship between the integer codes and the actual soil types.

2. Byte values are positive integer numbers ranging from 0 to 255. The byte data type is thus simply a subrange of the integer type. It is used in cases where the number range is more limited. Since this more limited range only requires 8 bits of memory per value for storage (i.e., 1 byte per pixel), only half as much hard disk or memory space for each image is needed compared to integer.
3. Real numbers have a fractional part such as 3.14. Real numbers are used whenever a continuous (rather than discrete) data variable is being stored with great precision, or whenever the data range exceeds that of the integer data type. The real data type of TerrSet raster images is known as single precision real numbers. Thus it can store values within a range of $\pm 1 \times 10^{38}$ with a precision of 7 significant figures. Single precision values provide a nice balance between range, precision and data volume. Each number (and thus each pixel) requires 4 bytes of storage. Note that real numbers have no discrete representation (e.g., 4.0 is the same as 3.9999999999999999 at some level of precision). Thus, while it is possible to store whole numbers in a real data type, they cannot then be used as feature identifiers (e.g., for ASSIGN or EXTRACT) or in modules that require byte or integer values (e.g., GROUP).
4. RGB24 data files use 24 bits (3 bytes) for each pixel to encode full-color images. The internal structure is band-interleaved-by-pixel and can encode over 16 million separate colors. RGB24 images are constructed using the COMPOSITE module (in the File|Display menu). The TerrSet display system allows for interactive and independent contrast manipulation of the three input bands of a displayed RGB24 image. Note, however, that when three or more raster layers are stacked in a map composition, Composer allows you to designate three images that should form the red, green and blue primaries of a color composite, without the need to create an RGB24.

The documentation file associated with each image records the file data type. When a new image is created, it maintains the prevailing data type of the input image, or it produces a data type that is logical based upon standard mixed arithmetic rules. Thus, dividing one integer data image by a second integer data image yields a real data image. TerrSet accepts integer, byte and real data types for almost all operations that would logically allow this. Some modules, though, do not make sense with real data. For example, the GROUP operation that extracts contiguous groups can only do so for categorical data coded with integer numbers. If the data type in an image is incorrect for a particular operation, an error message will be displayed.

The CONVERT module can be used at any time to convert between the integer, byte, and real data types. In the case of conversion from real numbers to either the integer or byte formats, CONVERT offers the option of converting by rounding or by truncation.

The data type described above indicates the type of numbers stored in an image. Another parameter, the file type, indicates how these numbers are stored. As with the data type, the file type of an image is recorded in its documentation file. Image files may be stored in ASCII or binary formats, although only the binary form is recommended (for reasons of efficiency). Binary files are those which are stored in the native binary encoding format of the operating system in use. In the case of TerrSet, this will be one of the Windows operating system variants. Thus, for example, the binary coding of a real number is that adopted by Windows and the Intel hardware platform.

Binary files are efficient in both hard disk space utilization and processing time. However, the format is not universal and not always very accessible. As a consequence, TerrSet also provides limited support for data recorded in ASCII format. A file in ASCII format is also referred to as a *text* file, and can be viewed directly with any text editor (such as the Edit module in TerrSet). The ASCII file type is primarily (although rarely) used to transfer files to and from other programs, since the coding system is a recognized standard (ASCII = American Standard Code for Information Interchange). ASCII files are not an efficient means of storing data and they cannot be displayed. You must convert your files to binary format before continuing with your analyses. The CONVERT module converts files from ASCII to binary or binary to ASCII. Although binary files cannot be examined directly, TerrSet provides strong data examination facilities through TerrSet Explorer. Binary files may be viewed as a matrix of numbers using the Show Structure utility of TerrSet Explorer. There is rarely a need then to convert images to ASCII form and it is generally no longer recommended.

Raster Pyramids

The display of very large raster images can sometimes be slow as a result of the sheer volume of data required. In cases where the display seems sluggish, it is suggested that you create an image pyramid. An image that has been converted to contain a pyramid contains multiple versions of the file at different resolutions. This facilitates rapid display at varying resolutions.

```
file format : IDRISI Rater A.1
file title : Major Soils Groups
data type : byte
file type : binary
columns : 512
rows : 480
ref. systems : US83TM18
ref.units : m
unit dist. : 1
min. X : 503000
max. X : 518360
min. Y : 4650000
max. Y : 4664400
pos'n error : unknown
resolution : 30
min. value : 0
max. value : 3
display min. : 0
display max. : 3
value units : classes
value error : 0.15
flag value : 0
flag def'n : background
legend cats : 3
code 1 : Podzol soils
code 2 : Brown podzolic soils
code 3 : Gray-brown podzolic soils
lineage : Soil polygons derived from 1:5000 scale color
air photography and ground truth, with the
final compilation being adjusted to the map
base by hand.
comment : Value error determined by statistical accuracy
assessment based on a stratified random
sample of 37 points
```

Raster Documentation Files (.rdc)

Each of the primary file types used by TerrSet (Raster, Vector and Attribute) is associated with a companion documentation file. Raster documentation files are automatically assigned an ".rdc" file extension by TerrSet.

Documentation files are always stored in text format and are created automatically by any import routine or analytical module with an image output. They can be viewed and modified using the Metadata utility in TerrSet Explorer. After specifying the file you wish to view from the file list, the contents of its documentation file will then be displayed. The documentation file consists of a series of lines containing vital information about the corresponding image file. The first 14 characters describe the contents of the line, while the remaining characters contain the actual data.

The graphic is an example of the documentation file for a soils image (soils.rst). This file contains information on major soils groups and is in byte format stored as a binary file. The image contains 512 columns and 480 rows, for a total of 245,760 values. The image is georeferenced with a reference system named US83TM18¹⁰ with each coordinate unit representing 1 meter. The minimum and maximum X and Y coordinate values indicate the reference system coordinates of the left, right, top and bottom edges.

The position error indicates how close a feature's actual position is to its mapped position in the image. It is marked in the example as unknown. If known, this field should record the RMS (Root Mean Square) error of locational references derived from the bounding rectangle coordinates. This field is for documentation purposes only and is not currently used analytically by any module.

The resolution refers to the inherent resolution of the image. In most cases, it should correspond with the result of dividing the range of reference coordinates in X (or Y) by the number of columns (or rows) in the image. However, there are some rare instances where it might differ from this result. A good example is the case of Landsat TM Band 6 (Thermal) imagery. The resolution of these data is actually 120 meters and would be recorded as such in this field. However, the data is distributed in an apparent 30-meter format to make them physically match the dimensions of the other bands in the image. This is done by duplicating each 120-meter pixel 4 times in X and then each row 4 times in Y. The resolution field is a way of correctly indicating the underlying resolution of these data.

In most images, the resolution in X and Y will be equal (i.e., pixels will be square). However, when this is not the case, the coarser resolution is recorded in the documentation file.¹¹ For example, with Landsat MSS data, the pixel resolution is 80 meters in X and 60

¹⁰ The reference system indicated in the documentation file is the name of a reference system parameter file (.ref). The parameters of this file are detailed in the chapter **Georeferencing** in this volume.

¹¹ Earlier versions of IDRISI calculated and reported the resolution field as that of the X dimension only.

meters in Y. In this case, the documentation file would show a resolution value of 80. Rectangular pixels will display correctly, however, preserving the true resolution in both X and Y. In addition, all analytical modules using measures of distance or area calculate resolution in X and Y independently. They, therefore, do not rely on the resolution value recorded in the documentation file.

The minimum and maximum value fields record the actual range of data values that occurs in image cells, while the display min and display max values designate what are sometimes called the *saturation points* of the data range. Values less than or equal to display min are assigned the lowest color in the palette sequence, while values greater than or equal to display max are assigned the highest color in the palette sequence. All values in between are assigned palette colors according to their position in the range (i.e., with a linear stretch). It is very common for the display min and max to match the min and max values. However, in cases where one wishes to alter the brightness and contrast of the image, altering the saturation points can produce a significant enhancement.¹² Saturation points may also be manipulated interactively in the display system (see the **Display System** section below).

The value units are *classes* in this example to indicate that the numbers are simply qualitative codes for soil classes, not quantitative values. The value units field is informational only, therefore any description can be used. It is suggested that the term *classes* be used for all qualitative data sets. However, when standard linear units are appropriate, use the same abbreviations that are used for reference units (m, ft, mi, km, deg, rad).

The value error field is very important and should be filled out whenever possible. It records the error in the data values that appear in image cells. For qualitative data, this should be recorded as a proportional error. In the example, the error is recorded as 0.15, indicating cell accuracy of 85% (i.e., what is mapped is expected to be found on the ground 85% of the time). For quantitative data, the value here should be an RMS error figure. For example, for an elevation image, an RMS error of 3 would indicate 68% of all values will be within ± 3 meters of the mapped value, that approximately 95% will be within ± 6 meters, and so on. This field is analytical for some modules (e.g., PCLASS) and should therefore be specified whenever the information is known.

The flag value and flag definition fields can be used to indicate any special meaning that certain cell values might carry. The flag value indicates the numeric value that indicates a flag while the flag definition describes the nature of the flagged areas. The most common data flags are those used to indicate background cells and missing data cells. These flag definition field entries are specifically recognized and used analytically by some modules (e.g., SURFACE). Other terms would be considered only informational at this stage. Note that some modules also output data flags. For example, when SURFACE is used to create an aspect image, aspects derived from a slope of zero are flagged with a -1 (to indicate that the aspect cannot be evaluated). In the output documentation file, the flag value field reads -1 while the flag definition field reads "aspect not evaluated, slope=0".

The legend cats field records the number of legend category captions that are recorded in the file. Following this, each legend caption is described, including the symbol code (a value from 0-255) and the text caption.

Finally, the image documentation file structure allows for any number of occurrences in four optional fields: comment, lineage, consistency and completeness. These fields are for information only and are not read by the TerrSet modules (although some modules will write them). Note that the lineage, consistency and completeness fields are intended to meet the recommendations of the U.S. National Committee for Digital Cartographic Data Standards (NCDCCDS). Along with the positional (pos'n) error and value error fields, they provide a means of adhering to the current standards for reporting digital cartographic data quality.¹³ Multiple occurrences of any of these field types can occur (*but only at the end of the file*) as long as they are correctly indicated in the 14 character descriptive field to the left. The lineage field can be used to record the history of an image. The command line used to generate a new file is recorded in a lineage field of the documentation file. The user may add any number of additional lineage lines. The consistency field is used to report the logical consistency of the file; it has particular application for vector files where issues of topological errors would be reported. The completeness field refers to the degree to which the file comprehensively describes the subject matter indicated. It might record, for example, the minimum mapping unit by which smaller features were eliminated. Finally, the comment field can be used for any informational purpose desired. Note that the Metadata utility in TerrSet Explorer and

¹² Note that one would normally only change the saturation points with quantitative data. In the example illustrated here, the numeric values represent qualitative classes. In these cases, the saturation points should match the actual minimum and maximum values.

¹³ For further information about the NCDCCDS standard, refer to the January 1988 issue of *The American Cartographer*, Volume 15, No. 1.

CONVERT module both read and maintain these optional fields. The Metadata utility in Explorer also allows one to enter, delete or update fields of this nature (as well as any documentation file fields).

Vector Layers (.vct)

The TerrSet System supports five vector file types: point, line, polygon, text and photo. All are automatically stored with a ".vct" file extension. They are stored in a feature-encoded structure in which each feature is described in its entirety before the next is described. The File Structures section of the Help System provides specific details about the structures of each vector file type. In addition, vector files may be viewed as ASCII representations using the Show Structure option in TerrSet Explorer. However, the following descriptions provide all the information that most users will require.

Common Features of All Vector Files

All vector files describe one or more distinct features. Unlike raster images that describe the totality of space within a rectangular region, vector files may describe only a small number of features within a similarly defined rectangular region.

Each feature is described by means of a single numeric attribute value and one or more X,Y coordinate pairs that describe the location, course or boundary of that feature. These points will be joined by straight line segments when drawn. Thus, curved features require a great number of closely spaced points to give the appearance of smooth curves.

The numeric attribute values can represent either identifiers (to link to values in a data table) or actual numeric data values, and can be stored either as integer or real numbers (although identifiers must be integers).¹⁴

A special feature of vector files is that their data types (integer or real) are less specific in their meaning than raster files. The integer type has a very broad range and is compatible with both the byte and integer type of raster layers as well as the long integer type commonly used for identifiers in database tables. Roughly, the vector integer type covers the range of whole numbers from -2,000,000,000 to +2,000,000,000. Similarly, the vector real data type is compatible with the single precision real numbers of raster layers but is in fact stored as a double precision real number with a minimum of 15 significant figures. TerrSet's conversion procedures from vector to raster handle these translations automatically, so there is little reason to be concerned about this detail. Simply recognize that integer values represent whole numbers while real numbers include fractional parts.

Point Layer Files

Point files are used to represent features for which only the location (as a single point location designation) is of importance. Examples include meteorological station data, fire hydrants, and towns and cities (when their areal extent is not of concern). Each point feature is described with an attribute value that can be integer or real, and an X,Y coordinate pair.

file format : IDRISI Vector A.1
file title : Landuse / Landcover
id type : integer
file type : binary
object type : polygon
ref. systems : utm16spm
ref.units : m
unit dist. : 1
min. X : 296000
max. X : 316000
min. Y : 764000
max. Y : 775000
pos'n error : unknown
resolution : unknown
min. value : 1
max. value : 9
display min. : 1
display max. : 7
value units : classes
value error : 0.15
flag value : 9
flag def'n : Unknown: Obscured by Clouds
legend cats : 7
code 1 : Residential
code 2 : Industrial
code 3 : Commercial
code 4 : Other Urban or Built Up Land
code 5 : Open Water
code 6 : Barren
code 7 : Transitional

¹⁴ All features in a single vector layer should be encoded with numeric values of the same type. Thus, if integer identifiers are used, all features must have integer identifiers. Similarly, if real number attributes are encoded, all features in that layer must have real number attributes.

Line Layer Files

Line files describe linear features such as rivers or roads. Each line feature in a layer is described by an attribute value that can be integer or real, a count of the number of points that make up the line, and a series of X,Y coordinate pairs (one for each point).

Polygon Layer Files

Polygon files describe areal features such as forest stands or census tracts. Each polygon feature in a polygon layer is described by an attribute value that can be integer or real, a count of the number of *parts* in the polygon, and for each part, a list of the points (by means of an internal index) that make up that part. This is then followed by an indexed list of the X,Y coordinate pairs of all points in the polygon.

Text Layer Files

Text vector files represent text captions that can be displayed as a layer on a map. They store the caption text, their position and orientation, and a symbol code that can be used to link them to a text symbol file. Text vector files can be created with the on-screen digitizing feature in TerrSet. Text symbol files are created with Symbol Workshop.

Photo Layer Files

Photo layers are a special form of vector text layers that record, in addition to any text information, a specially encoded file name of a photo in jpeg format, along with an optional azimuth. Photo layers can be created manually using the TerrSet on-screen digitizing system (see the Help Entry for Photo Layers for instructions). However, in cases where photos have been geocoded using GPS coordinates, a simpler procedure is available through an import module named EXIFIDRIS. To use this, put all the photos you wish to display in a single layer into a separate folder. Then run EXIFIDRIS and specify the folder. It will then create a photo layer of all geocoded jpeg files in the folder.

Vector Documentation Files (.vdc)

As with image files, all vector files are paired with documentation files. Vector documentation files have a ".vdc" extension. Any TerrSet module that creates or imports a vector file will automatically create a documentation file. The Metadata utility in TerrSet Explorer can be used to update documentation files as required.

A sample vector documentation file appears in the graphic above. As can be seen, the structure of a vector documentation file is virtually identical to that for a raster layer. The only differences are:

1. the row and column information is absent;
2. *data type* has been replaced by *id type*, although the intention is identical. Choices here are integer or real. The reason for the slightly different wording is that coordinates are always stored as double precision real numbers. The id's, though, can vary in their data type according to whether they encode identifiers or numeric attribute values.
3. the file type is always binary. (ASCII format for vector files is supported through the Vector Export Format.)
4. the minimum and maximum X and Y values recorded in the documentation file do not necessarily refer to the minimum and maximum coordinates associated with any feature but to the boundary or limits of the study area. Thus they correspond to the BND (boundary) coordinates in vector systems such as ArcGIS.

Attribute Files (.accdb and .avl)

In addition to the attributes stored directly in raster or vector layers, TerrSet permits the use of freestanding attribute files. Two types are recognized, data tables and values files. Only the former can be linked to a vector file for display or to produce a vector collection. The latter may be used to assign new values to a raster image and will be produced when summary values for features are extracted from a raster image. Fields from data tables may be exported as values files and values files may be imported into data tables using Database Workshop.

Data Tables (.accdb)

TerrSet data tables are Microsoft Access-compatible relational database files. Thus, TerrSet attribute tables have an ".accdb" extension¹⁵. Internally, each can carry multiple tables and can also be used and modified by TerrSet's Database Workshop and with Microsoft Access or any other Access-compatible system.

Values Files (.avl)

Values files contain the values for a single attribute. They are stored in ASCII-compatible text format with two columns of data separated by one or more spaces. The first column contains an identifier that can be used to associate the value with a feature (either raster or vector), while the second column contains the attribute value. Attribute values files have an ".avl" extension. The following is an illustration of a simple values file listing the populations (*1000) of the 10 provinces of Canada: where 1 = Newfoundland, 2 = Nova Scotia, 3 = Prince Edward Island, and so forth.

1	560
2	850
3	129
4	690
5	6360
6	8889
7	1002
8	956
9	2288
10	2780

The feature definition image to be used with this values file would have the value 1 for all pixels in Newfoundland, the value 2 for Nova Scotia, and so on. AVL files can be created very easily using the EDIT module. You can even cut and paste the columns from a spreadsheet such as EXCEL into EDIT. When you finish your edits and save the file, be sure to specify AVL as the file type.

Attribute Documentation Files (.adc)

As with images and vector files, attribute files (of either type) also carry documentation files, but this time with an ".adc" extension. ADC files are automatically created by modules such as EDIT or Database Workshop while the Metadata panel in TerrSet Explorer is the utility which can be used to modify the documentation file associated with an attribute file.

As with the other documentation files in TerrSet, those for attribute files are stored in ASCII format with the first 14 characters used purely for descriptive purposes.

¹⁵ The .accdb format supercedes the earlier Microsoft .mdb format. If you have data files in this earlier format, they will need to be converted to the newer format. Please refer to the help system for information on how this should be done.

A sample attribute documentation file appears in the graphic. This example shows the documentation file for an text attribute values file (indicated by the code ASCII – American Standard Code for Information Interchange). This type of file always has two fields. Database tables will normally have many more fields. For each field, the parameters shown are repeated.

In this version of TerrSet, the following file types are supported:

- ascii*: Simple 2-column ASCII-compatible text files¹⁶ where the left field contains integer feature identifiers and the right field contains data about those features. These values files have an ".avl" file extension.
- access*: A database file in Microsoft Access format having an ".accdb" extension. This format is the current resident database format supported by TerrSet, and is supported by Database Workshop. It is also the format used in the creation and display of vector collections.

In the case of the simple ASCII form (".avl" file), format information is unimportant, and thus reads 0 in this example. Spaces or tabs can be used to separate fields. With fixed length ASCII and database files, format information is essential. The format line simply indicates the number of character positions occupied by the field.

- byte and integer data*: A single number to indicate the number of character positions to be used.
- character string data*: A single number to indicate the maximum number of characters.

real number data: Two numbers separated by a colon to indicate the total number of columns and the number of those columns to be used for recording decimal values. For example, 5:2 indicates that five columns are to be used, two of which are for the decimal places. Note that the decimal itself occupies one of these columns. Thus the number "25.34" occupies this field completely.

For most other entries, the interpretation is the same as for raster image layers. Valid data types include *byte*, *integer*, *real* and *string* (for character data). However, for data tables (.accdb), many of the data types recognized by the Microsoft Access Database Engine (ACE) are supported. These include:

- real (-3.402823E38 to +3.402823E38 real numbers)
- byte (0-255, whole numbers)
- integer (-32768 to 32767, whole numbers)
- longint (long integer, -2,147,483,648 to 2,147,483,647, whole numbers)
- text (character strings)
- Boolean (true or false)

The TerrSet System will document these types automatically, and knows how to convert them when they are used in the context of display (in a vector collection) or when creating an ASCII values file.

file format :	IDRISI Value A.1
file title :	Roads
file type :	ASCII
records :	12
field :	2
field 0 :	IDR_ID
data type :	integer
format :	0
min. value :	105
max. value :	982
display min. :	105
display max. :	982
value units :	ids
value error :	unknown
flag value :	none
flag def'n :	none
legend cats :	0
field 1 :	ROAD_TYPE
data type :	integer
format :	0
min. value :	1
max. value :	3
display min. :	1
display max. :	3
value units :	clouds
value error :	unknown
flag value :	none
flag def'n :	none
legend cats :	3
code 1 :	Major Road
code 2 :	Secondary Road
code 3 :	Minor Road

Other File Types

¹⁶ The ANSI and UTF-8 coding standards supported by Windows Notepad are both ASCII compatible (as long as only 8-bit characters are used in the case of UTF-8).

While the majority of data files you will work with are those describing layers and their attributes, many others exist within the TerrSet system and some of them are described below. Other more specialized file types are described in the context of specific modules and in the File Structures section of the Help System.

Map Composition Files (.map)

Map composition files store the graphic instructions necessary to create a map composition using data from a set of map layers and associated symbol and palette files. They are described further in the **Display System** section below.

Symbol and Palette Files (.sm0, .sm1, .sm2, .smt, .smp)

In order to display map layers, it is necessary to set up an association between vector features or image values and particular graphic renditions. This is done through the use of *Symbol Files* and *Palette Files*. An extensive set of symbol and palette files is included with TerrSet, and will meet most user needs. However, for final output, it is often desirable to create custom symbol and palette files. This is done with the Symbol Workshop utility under the Display Menu. See the **Display System** section below for more information about Symbol Workshop.

Symbol files indicate the manner in which vector features should be symbolized (such as the type, thickness and color of lines or the font, style, size and color of text). Each symbol file lists the characteristics of up to 256 symbols that are identified by index numbers from 0 to 255. In all, four kinds of symbol files are used, one each for the vector feature types: point, line, polygon and text (since photo layers are a form of text layer, they also use text symbol files). They are stored with file extensions of ".sm0," ".sm1," ".sm2" and ".smt" respectively. As usual, TerrSet always knows the appropriate symbol file type to use. As a result, you never need to specify a symbol file with its extension.

For raster images, graphic renditions are specified by the use of palette files. Like symbol files, palette files also define up to 256 renditions identified by index numbers from 0 to 255. However, in this case, only the color mixture (defined by the relative amounts of red, green and blue primaries) is specified. Palette files are stored with an ".smp" extension.

Reference System Parameter Files (.ref)

Reference system parameter files record information about specific geographic referencing systems. They include data on the projection, ellipsoid, datum, and numbering conventions in use with a reference system. TerrSet includes over 500 such files. The user can modify these and also create new reference system parameter files with the Metadata utility in TerrSet Explorer. See the **Georeferencing** section below for more information about these files.

Display System

The previous section described map layers as each representing a single elementary theme. By combining map layers and giving them graphic renditions, we create a *map*. Map compositions may contain as few as one layer and as many as 32 layers. Map compositions are created as a natural consequence of working with the TerrSet display system. As you work, TerrSet keeps track of all changes and additions you make to the composition. You can then save the composition at any time. The composition is stored in a file with a ".map" extension and is simply called a map file.

The display system in TerrSet consists of several separate but interdependent program components, each of which is described in this chapter.

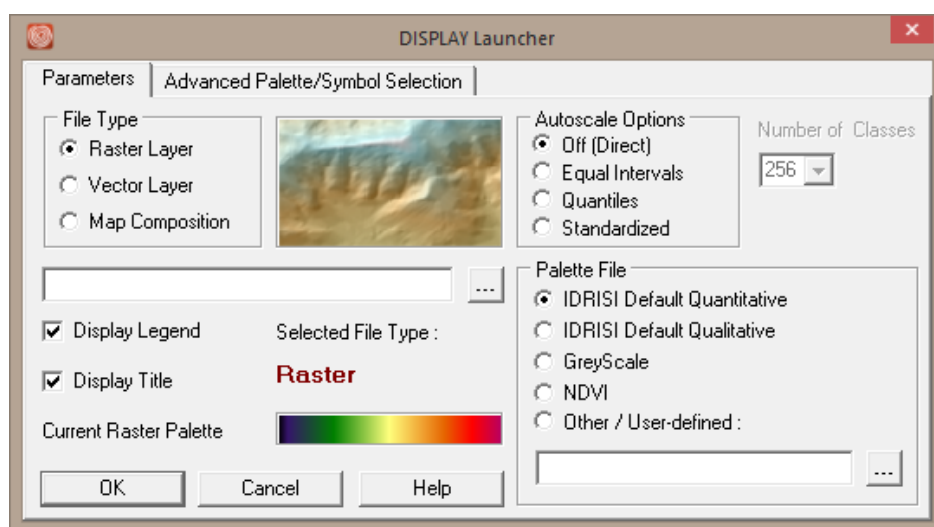
DISPLAY Launcher

DISPLAY Launcher is used to open a new display window. It begins the map composition process, and is always the first operation required to create a new map display. DISPLAY Launcher can be accessed from its toolbar icon (shown above) or by choosing it from the Display menu. Doing so opens a dialog box with options to display a raster layer, a vector layer, or an existing map composition.

When you select a raster or a vector layer, TerrSet uses a set of decision rules based on the values in the layer to suggest an appropriate palette or symbol file. You may change this selection. You can also specify if the layer should be displayed with a direct relationship between numeric values and symbol codes, or should be autoscaled (see below). In the case of a map composition, you will only be required to specify its name, since all display parameters are stored in the map file.

Palette and Symbol Files

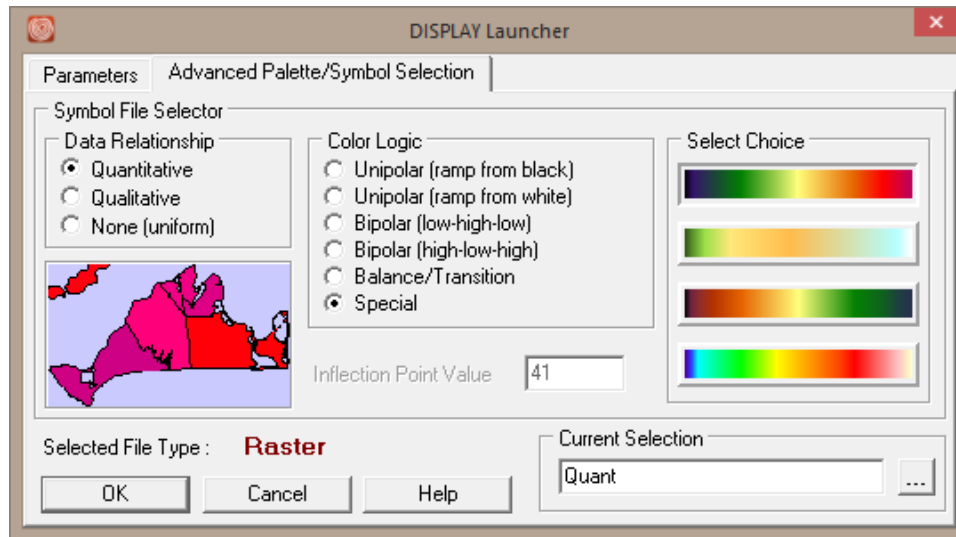
As described above, palette files define the way the information stored in image pixel values will be displayed on the screen. Similarly, symbol files define the way vector features with particular ID's or linked attribute values will appear. Each stores the details of up to 256 graphic renditions referenced by index numbers 0-255. Palette files store the Red, Green and Blue (RGB) color mixture for each index while symbol files store a variety of parameters related to specific vector object types (e.g., point size, line width, color, etc.). The section below on Symbol Workshop lists the parameters that may be defined for each type of symbol file.



The TerrSet System comes with a set of standard palettes and symbol files. These are installed in the symbols folder of the TerrSet program directory (e.g., C:\Program Files (x86)\TerrSet\Symbols). You may wish to extend that set, either by modifying existing palettes and symbol files, or by creating entirely new ones. Symbol Workshop can be used for this purpose. User-created palette and symbol files may be stored anywhere, including the TerrSet symbols directory. Typically, users store commonly-used symbol files in the Symbols directory. Files that are specific to particular data sets are usually stored with those data.

Advanced Palette/Symbol Selection

DISPLAY Launcher also offers an Advanced Palette/Symbol Selection tab that provides simple access to over 1300 palette and symbol files.



You should first indicate whether your data express quantitative or qualitative variations. Quantitative data express differences of degree (such as elevation) while qualitative data express differences of kind (such as landcover categories). Alternatively, you can indicate that you wish to symbolize all features with a uniform symbol. You will also need to decide on the color logic. For each color logic, you are provided four choices which can be selected by pressing the appropriate button. For Qualitative and Uniform symbol schemes, the choices of the color logic will be straightforward. However, the choices for quantitative data warrant additional explanation:

- Unipolar schemes are used to symbolize data that progress either from low to high or high to low.
- Bipolar schemes have two high or two low ends, with a distinct inflection somewhere in between. For example, population change data may range from +20% to -20%. In this case we have two high ends – high positive change and high negative change, with a clear inflection point at 0%. The colors in bipolar schemes are designed to be perceived by the visual system as falling into two qualitatively different groups while maintaining a quantitative relationship in sequence.
- Balance schemes are less common. These are used in cases where two variables perfectly co-vary – i.e., declines in one are perfectly balanced by increases in the other. For example, imagine a country in which two languages are used. Districts with 40% of one language group would thus have 60% of the other. The colors in balance schemes are designed to be seen as varying mixtures of two distinctive components.

Autoscaling

Autoscaling concerns the relationship between raster cell values and palette indices (and, similarly, vector ID's or linked attribute values and symbol indices). By default, a direct relationship between them is assumed (i.e., a cell with a numeric value of 12 should be displayed with the color defined by palette index 12). However, not all images contain integer values that fall nicely within the allowable range of values for palette indices (0-255). As a result, it is often necessary to scale the actual range of values into this more limited range. For example, we might have an image layer of temperature anomalies ranging from (-7.2) degrees to (+4.6) degrees. These values cannot be directly displayed because there is no palette index of (-7.2). Autoscaling offers a solution to this problem.

Four options for autoscaling are provided: Off (Direct), Equal Intervals, Quantiles and Standard Scores.

- Off (Direct). When autoscaling is off, it is assumed that there is a direct relationship between the numeric values of pixels or features in the layer and the numeric codes in palettes and symbol files. Thus if a pixel contains the number 5 in a layer, it is

symbolized with the palette color 5. This option is only permitted in cases where all pixels or features have integer values between 0-255.

- **Equal Intervals.** This is the default for all cases where direct scaling is not possible - i.e., for cases where some data values are less than zero or greater than 255, and all cases where the layer contains real numbers. In this case, the range of numeric values is automatically scaled (hence the name autoscaling) into a series of equal-width data classes that are subsequently assigned to symbols. Thus, for example, a layer containing elevations from 0 to 5000 meters might be automatically scaled into 10 classes, 500 meters in width. Therefore, the values from 0-499 would be assigned to symbol 0, 500-999 to symbol 1, 1000-1499 to symbol code 2, etc. DISPLAY Launcher will make an initial determination of the number of classes. However, you have the ability to specify any number from 2-256. Remember, all numeric series start from 0 in TerrSet. Thus, if you were to choose 256 classes, they would be understood to be symbolized by symbols codes 0 through 255. Note that this process of dividing the numeric data range into a series of symbolization classes is also referred to as classification by cartographers.
- **Quantiles.** A quantiles classification scheme places an equal number of pixels or features into each class, first by rank ordering the pixels or features and then assigning them to classes in rank order. Some quantiles schemes are known by special names. Thus, for example, a quantiles scheme with four classes is known as a quartiles scheme. Quantiles are a good choice whenever the data are strongly skewed or disproportionately loaded on a small number of equal interval classes.
- **Standardized.** A standardized classification scheme divides the data into symbol classes based on standard deviation units. With an even number of classes, the mean will represent the central boundary between classes, while with an odd number of classes, the mean will be at the center of the middle class. All classes are one standard deviation in width and the end classes always include all cases below its defining boundary. Thus, for example, the default case of 6 classes will be defined as:

$\leq -2sd$

$-2sd$ to $-1sd$

$-1sd$ to *mean*

mean to $+1sd$

$+1sd$ to $+2sd$

$\geq +2sd$

whereas a selection of 5 classes would yield:

$\leq -1.5sd$

$-1.5sd$ to $-0.5sd$

$-0.5sd$ to $+0.5sd$

$+0.5sd$ to $+1.5sd$

$\geq +1.5sd$

Standardized classes are appropriate whenever the data are approximately normally distributed and one wishes to differentiate unusual from common cases.

There are several further issues that should be noted about autoscaling. First, Equal Intervals autoscaling is automatically invoked whenever real numbers or integer values outside the 0-255 range must be displayed. Second, while the maximum range of palette and symbol indices is 0-255, some palette or symbol files may use a more limited range of values for the purpose of autoscaling (e.g., 1-100). The autoscaling range of a palette or symbol file can be inspected and changed with Symbol Workshop. Third, while Equal

Intervals autoscaling works well for many images, those with extremely skewed distributions of data values (i.e., with a very small number of extremely high or extremely low values) may be rendered with poor contrast. In these cases, you can alter the display min and display max values (known as the saturation points) using the Layer Properties option of Composer (see below) or by using one of the Instant Stretch options detailed below. Alternatively, Quantiles autoscaling can be used. Finally, note that autoscaling does not change the values stored in a data file; it only alters the display. To create a new raster image with contrast-stretched values, use the module STRETCH.

Instant Stretch Buttons in Composer

Composer has three very useful buttons at the bottom that provide the ability to easily modify the saturation points (display min and max). The left-most will change the display min and max such that the most extreme 1% of values are assigned the minimum and maximum symbol or palette entries. It generally works well to provide an image with good contrast. The middle button does the same except that it makes the stretch symmetrical about 0. The right-most works the same as the left-most except that it only considers the pixels visible on the screen when determining the stretch points.

Automatic Display

The TerrSet System includes an automatic display feature that can be enabled or disabled from User Preferences, under the File menu. With automatic display, the results of analytical operations are displayed immediately after the operation has finished. The system will determine whether to use the Default Qualitative or Quantitative palette (both of which are user-defined in the Preferences dialog) for display and if autoscaling should be invoked. The algorithm that is used to make these determinations is not fool-proof, however. Palette and autoscaling choices may be quickly corrected in the Layer Properties dialog of Composer, if necessary. The automatic display feature is intended as a quick-look facility, not as a substitute for DISPLAY Launcher.

Launching Maps and Layers from TerrSet Explorer

TerrSet Explorer can display layers (both individually or as part of collections), both vector and raster, and map compositions. Right-click on any layer to display or add it to the active map composition. You can also display layers within a group file or vector collection and they will be displayed with the *dot-logic* that signifies their membership in a collection. Alternatively, you can double-click on any layer to display it. The same logic applies to map compositions except that you cannot add a map composition to an existing one.

Explorer also allows you to highlight more than one file. If you then right-click the selected group and select the *display* option, the layers will be opened in separate map windows. However, if select the *add* option, they will be combined into a single map composition. Map compositions displayed from TerrSet Explorer are displayed exactly as they have been saved, since all palette and symbol file information is included in the map file. However, layers do not record information about palettes or symbol files and will thus be displayed in the default format specified in User Preferences (accessible from the File menu).

Map Windows, Layer Frames, Ancillary Information and Auto-Arrange

A map layer is displayed in a *Layer Frame*. This and all other components of the map composition (e.g., north arrow, legend) are placed in the *Map Window*. A Map Window includes a single Layer Frame and ancillary information such as a title, subtitle, caption, legend, north arrow, scale bar, and graphic insets. By default, the map composition system (which can be controlled by Composer) automatically arranges the composition to accommodate the various elements of the composition. It is generally recommended that you leave *auto-arrange* on, and only turn it off when you want a non-standard layout e.g., if you have resized or moved the legend or other components).

The other components that may be placed in a Map Window are described below in the section on Map Properties. The Map Window may be thought of as the graphic "page" upon which a map composition is arranged. Its color is set in the Map Properties dialog and it can be maximized or resized interactively by placing the cursor on the window border until it changes to a double arrow, then dragging it to the desired position. A layer frame can be moved by simply holding down the left mouse button and dragging the map to its new position.¹⁷

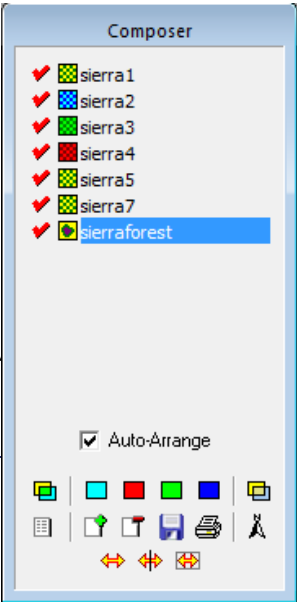
Note that there are several toolbar buttons that are useful in layer frame manipulations. These are shown in the table below. The first, Full Extent Normal, returns the map window and layer frame to their initially-displayed state. This action can also be triggered by pressing the Home key. The second icon, Full Extent Maximized, causes the map window and layer frame to expand as much as possible while still being fully visible (the button can also be activated by pressing the End key). Note that in doing so, the system reserves space for Composer. If you wish to ignore Composer, hold the shift key while pressing the End key. The third icon, Fit Map Window to Layer Frame, causes the map window to fit around layer frame only, while removing all components (e.g., title and legend).

Toolbar Icon	Keyboard	Action
	HOME	Full Extent Normal
	END	Full Extent Maximized (leave space for Composer)
	SHIFT+END	Full Extent Maximized (ignore Composer)
		Fit Map Window to Layer Frame

Composer

As soon as the first TerrSet map window has been opened, the Composer dialog box appears on the screen. Composer can be considered a cartographic assistant that allows one to:

- a) add or remove layers from the composition;



¹⁷ To drag a component, place the cursor over the component (not on the sizing buttons) and hold the left mouse button down until the component is at the desired position. "Drop" the component in place by releasing the mouse button.

- b) change the order in which layers are drawn (called the priority of a layer);
- c) set a layer to be part of a color composite;
- d) set a layer to have a transparent background or blend with the layer below it;
- e) temporarily toggle a layer to be invisible (i.e., hide it) without removing it;
- f) examine and alter the properties of a layer, including the symbol or palette file in use, display saturation points, and autoscaling;
- g) add, delete and modify a variety of map components in the map window;
- h) perform an instant histogram stretch on the active map;
- i) save the current composition (as displayed); and
- j) print the composition.

If you have several map windows open, you will notice that Composer displays the information for the map window that has *focus* (or the last one to have received focus if a non-map window, such as a dialog box or other window, currently has the focus). Focus refers to the ability of a window to receive input messages from the mouse and keyboard. Windows designates the window that has focus by displaying its banner with a specific color.¹⁸ To bring a window into focus, click on any part of it. To change the characteristics of a map composition, first give it focus. With many map windows open, Composer may be difficult to find. Hitting the “c” key on the keyboard will always bring Composer to the top of the TerrSet desktop.

Add Layer

To add a layer, click the Add Layer¹⁹ button on Composer and choose the desired file.²⁰ You will then be presented with a similar dialog to indicate the palette or symbol file to be used and whether the new layer should be autoscaled. All layers in a map composition must share a common reference system. If this is not the case, a warning message will appear indicating this and warning that the layer may not display properly. The bounds of files do not need to match to be displayed together in a map composition.

When a new layer is added, it is automatically given the highest priority (i.e., it is drawn on top of the other layers). The first layer displayed will thus, by default, have the lowest priority (priority 0) and will appear to be on the bottom of the composition. The layer with priority 0 is often referred to as the *base layer* in this documentation.

When multiple layers are present, their priority can be changed by dragging the name of the layer concerned, and dropping it in the desired position. Whenever a map window is redrawn, the layers are drawn in their priority order, starting with 0 and progressing to the highest value.

¹⁸ This is only one of many settings in the Windows system display that is defined when you choose a color scheme or define the display characteristics of individual Windows components.

¹⁹ Alternatively, pressing Ctrl-R when a map window has focus will bring up the Add Layer dialog set for adding a raster layer, and pressing Ctrl-V will bring up the Add Layer dialog set for adding a vector layer. Another option is to select a file in TerrSet Explorer, right-click on it and choose Add Layer, or Shift-Insert.

²⁰ Only one layer frame is present in any map window, so all layers are added to the same layer frame. Note that layers may be added from several paths for display. However, if the composition is to be saved as a map composition file, all the layers must exist in the Working Folder and/or Resource Folders of the project from which it is displayed again.

Note that if a raster layer is added it will obscure all the layers beneath it unless it is designated to have a transparent background or to blend with the layers below it (see the next section for further details).

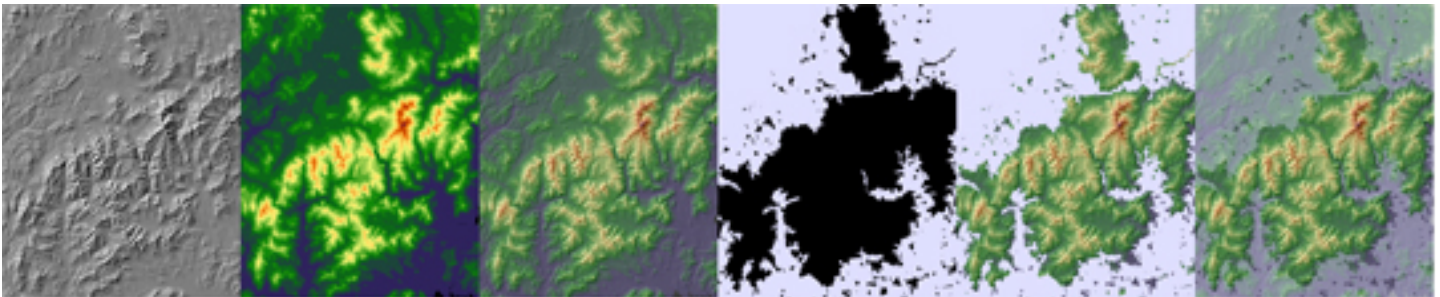
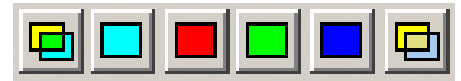
Auto-Arrange

The auto-arrange feature automatically arranges map elements such as titles, legends, scale bar, and inserts. By default, the auto-arrange feature is on for any map window, but it can be turned off, allowing the manual positioning of elements for a map composition. This is especially useful for printing.

When auto-arrange is on, whenever the map window is resized, all the map elements for that map window will be repositioned relative position to their original layer frame location. When auto-arrange is off, all map elements will hold their absolute position when the map window is resized.

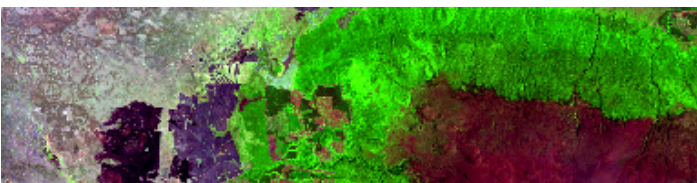
Layer Interaction Buttons

Immediately above the Add Layer button is a set of buttons that control layer interaction effects. The left-most is the Blend button. If the currently highlighted layer is raster, clicking Blend will cause the layer to blend with the visible layer group below it by 50%. To remove the blend effect, highlight the blended layer and click this button again. The button to the far right is the Transparency button. If the highlighted layer is raster, clicking it will cause the background color (whatever is color 0 in the palette) to become transparent. Note that layers can be both blended and transparent as illustrated below.



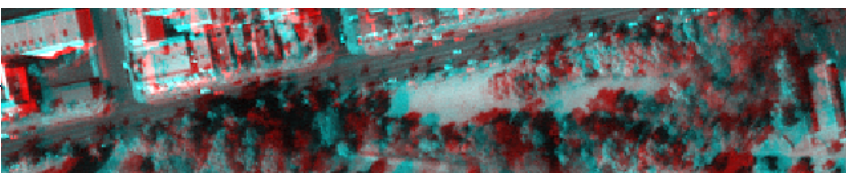
This sequence shows a hillshading layer and a digital elevation model being blended (third frame). In the fourth frame a mask layer has been added with black as color 0 and grey as color 1. In the fifth frame, the mask is made transparent allowing the elevation model to show through the background color. Finally, in the last frame, the mask is also set to blend allowing the topography to be partially visible through the grey mask.

The remaining buttons are for creating various color composites. The Red, Green and Blue buttons allow you to designate layers as the red, green and blue primary color components of a full color image.



Note that the layers need to be adjacent to each other for this to work, but that the specific order is unimportant. The Cyan button (second from the left) is used to create three dimensional anaglyphs.

The creation of an anaglyph requires stereo images -- two views of the same landscape taken from different



locations, and a set of anaglyphic 3-D glasses. When these are superimposed using Add Layer, assign the red primary to the left view and the cyan primary to the right view (this order may need to be reversed depending upon your glasses). Again, the layers should be next to each other in their priority in Composer for this to work.






Remove Layer

To remove a layer, select its name in the list of layers shown on Composer, then click the Remove Layer button. If you wish to only temporarily hide the layer, do not remove it, but rather, uncheck it to change its visibility.

Layer Names, Types, and Visibility

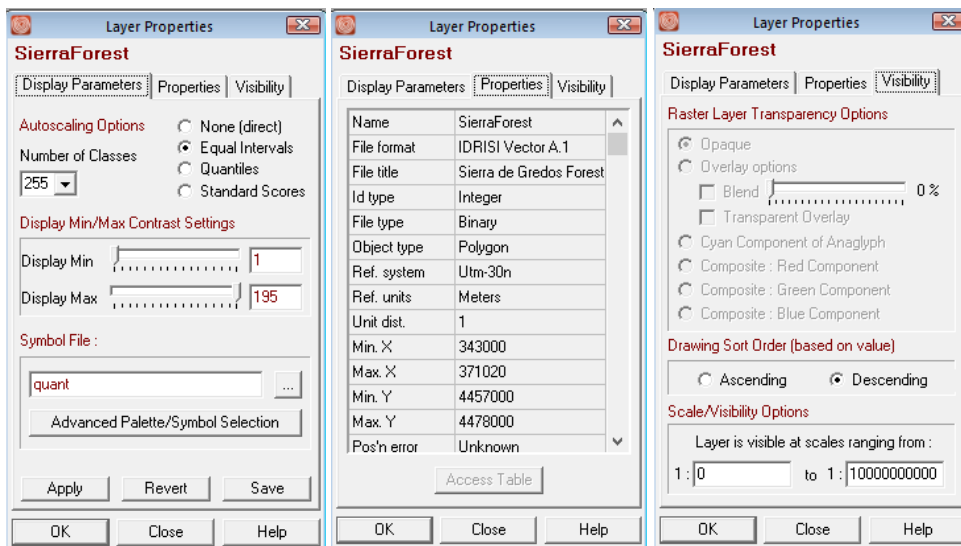
Composer displays one row for each layer in the composition. You can toggle a layer visible or invisible by clicking on the button to the left of its name. Note that toggling a layer off does not remove it from the composition. Rather, it simply makes it temporarily hidden. This can greatly facilitate visual analysis in some instances.

To the right of the layer name is a graphic symbol which indicates the layer's type: raster, vector point, vector line, vector polygon or vector text. These are illustrated in the table below. This information is used, for example, to indicate which file is which when a raster layer and a vector layer are in the same composition.

Layer Type	Icon on Composer
Raster	
Vector Point	
Vector Line	
Vector Polygon	
Vector Text	

Layer Properties

The properties of any layer, including key documentation information and display parameters, are shown in the Layer Properties dialog. To view this, highlight the desired layer with a single mouse click on the layer name in Composer, and then click the Layer Properties button. This will display the Layer Properties dialog associated with that layer.



The Layer Properties dialog summarizes several important elements of the layer's documentation file organized into three tab views. The first tab allows you to change the way the image is visually scaled for display, including the number of classes and the classification scheme. In addition, it allows you to change the contrast of the image if it has been autoscaled using Equal Intervals. This tab also allows you to change the palette or symbol file and gives access to the Advanced Palette/Symbol Selection page (see the section on DISPLAY Launcher).

The second tab highlights the Metadata for a selected file. If the layer is raster, a histogram of the layer may be launched using the Histogram button. If the layer is a vector feature definition file linked to an attribute table field, the full data table may be displayed using the Access Table button.

The third tab accesses the visibility properties, including the ability to set specific blends other than the default 50% accessible from the Blend button on Composer. The Drawing Sort Order affects vector layers only and controls the order in which vector features are drawn. This is used to establish the proper visual priority when vector features overlap. This dialog also allows you to establish at what scale the layer is visible. As you zoom in and out, layers will toggle off if the scale falls outside the visible range.

When Equal Intervals autoscaling is in effect, all values less than or equal to the display min recorded in the documentation file will be assigned the lowest symbol or palette color in the sequence. Similarly, all values greater than or equal to the display max will be assigned the highest symbol or color in the sequence.²¹ It is for this reason that these values are known as saturation points. Adjusting them will alter the brightness and contrast of the image. If an image is not autoscaled and you wish to adjust its contrast quickly, you can click on the Instant Stretch icon on Composer.

Two options exist for changing these values within the Layer Properties dialog. They can be edited within the appropriate input boxes, or they can be adjusted by means of the sliders. The sliders can be moved by dragging them with the mouse, or by clicking on the particular slider and then pressing either the left or right arrow key. The arrow keys move the slider by very small increments by default, but will yield larger movements if you hold down the shift key at the same time.²²

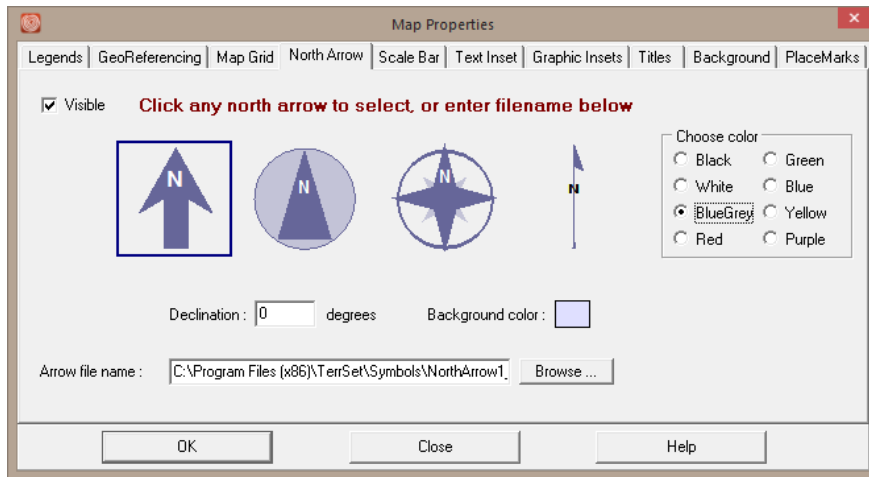
The best way to become familiar with the effects of changing the saturation points is to experiment. In general, moving the saturation points toward the center from their corresponding minimum and maximum data values will increase the contrast. In effect, you sacrifice detail at the tails of the distribution in order to increase the visibility of detail in the rest of the distribution. If the image

²¹ Note that the *lowest* and *highest* symbols or colors in the sequence are those set as the autoscale min and max in the symbol file or palette. These may be adjusted in Symbol Workshop.

²² If you always wish to display a layer with particular display min/max values that are not the actual min/max values of the layer, use the Metadata utility in IDRISI Explorer to set the display min and display max fields of the layer documentation file to the desired values. Whenever the file is displayed, these values will be used to set the range of values covered by the palette or symbol file. The actual data values remain unchanged.

seems too dark overall, try moving the display max down. Similarly, if the image seems too bright, move the display min up. Depending on the distribution of the original data values (which can be assessed with HISTO), interactively adjusting the saturation points and observing the effects can be very helpful in visual image analysis. This is particularly the case with 24-bit color composite images.

Map Properties



While the Layer Properties dialog displays information about a single map layer, the Map Properties dialog describes the entire map composition. It is used to set the visibility and characteristics of map components such as layer legends, north arrow and so forth.

The Map Properties dialog can be brought forward by clicking the Map Properties button on Composer or by clicking the right mouse button in a map window. The right-click method is context sensitive. For example, a right-click over a legend launches the Map Properties dialog with the legend options visible.

Map Properties is a tabbed dialog. Simply click on the appropriate tab to access the options associated with a particular map component. Ten tabs are provided: Legends, Georeferencing, Map Grid, North Arrow, Scale Bar, Text Inset, Graphic Insets, Titles, Background and Placemarks. The options for each are described below.

You can enlarge the Map Window at any time to increase the size of your graphic page and make space for new map components. In all cases, the components must be set to Visible to be displayed. All components are movable and many may be interactively resized. To do so, double-click on the component, make the necessary changes, then click any other element to release move/resize mode.

Legends

Up to five layers may have displayed legends in a map composition. Any raster layer or vector point, line or polygon layer can have a legend. The legend text entries are stored in the layer documentation file and may be entered or altered with the Metadata utility in TerrSet Explorer. Up to 256 entries are allowed. By default, the legend will display up to 20 categories (this can be changed in User Preferences under the File menu). However, whenever there are more than 20 legend entries in the documentation file, a vertical scrollbar will be attached to the legend, and you may scroll through the remaining entries. You can also lengthen the legend by double-clicking on it then extending the bottom edge.

For a quantitative data layer with byte or integer data type and a range of values of 20 or less, a legend will be formed and labeled automatically (i.e., no legend entries are needed in the documentation file). If the data are real or the data range is greater than 20, a special continuous legend will be displayed with automatically generated representative category labels.

Georeferencing

The Georeferencing tab of Map Properties shows the reference system and units of the current composition, as well as its bounding rectangle. It also shows the actual bounding rectangle of all features within the composition. The former can be changed using the input boxes provided. In addition, a button is provided that will allow you to set the composition bounding rectangle to the current feature bounds.

The Georeferencing tab is also used to set a very important property regarding vector text layers. Text sizes are set in *points*—a traditional printing measurement equal to 1/72 inch. However, in the dynamic environment of a GIS, where one is constantly zooming in and out (and thus changing scale), it may be more useful to relate point size to ground units. The Override of Text Points to Ref. System conversion option on the Georeferencing Tab does this. When it is enabled, text captions associated with text layers (but not map components, such as titles) will change size as you zoom in and out, appearing as if the text is attached to the landscape. The scaling relationship between point size and map reference units is also user-defined. When this option is disabled, the text captions retain their original size as you zoom in and out of the map layers.

Map Grid

A map grid can be placed on the layer frame. The intervals may be specified as well as the starting values in X and Y (all in reference units). You can also choose the type of grid, color and width of the grid lines and the label font. The grid will be automatically labeled.

North Arrow

A north arrow can be added to a composition using several default styles, or one can create their own and import them as bitmaps or enhanced metafiles.

Scale Bar

Adding a scale bar can also be achieved with the Map Properties dialog. You set the length (in map reference system units), the number of divisions, color properties and text label of the scale bar. The scale bar automatically adjusts in width as you zoom.²³

Text Inset

The text inset can be used to display blocks of text. This is done by specifying the name of an ASCII text file. The text inset frame is sizeable and incorporates automatic word-wrap. Font style and background color may be set.

Graphic Insets

²³ This can sometimes prove to be a nuisance since the scale bar may need to expand to be greater than the width of the map window. In these cases, simply set its visibility to "off" using Map Properties. Otherwise, use Map Properties to reset the size of the scale bar to a smaller length.

Graphic insets can be used to hold inset maps, pictures, logos, or any other graphic image. Any Windows Bitmap (".bmp"), JPEG (.jpg) or Metafile (".wmf" or ".emf") may be used as a graphic inset. All graphic insets are resizable. However, if the Stretchable option is disabled, resizing is controlled to preserve the original aspect ratio of the inset.

Titles

A title, subtitle, and caption may be added to a map composition. These are not associated with any map layer, but rather belong to the map composition as a whole. However, the title of the base layer, read from its documentation file, will appear as the title text by default. This may be edited, and font properties may be set.

Background

The Background tab allows you to change the color of the backgrounds for the Layer Frame and the Map Window. The backgrounds of other components, such as legends and the scale bar, are set in their respective tabs. However, the Background tab also allows you to set the backgrounds of all components to match that of the Map Window. Note that setting the layer frame background may not be noticeable if a raster layer is present in the composition (since it may cover up the entire frame). Also note that there is a default background color that is set with the User Preferences dialog under the File menu.

Placemarks

Finally, the Placemarks tab keeps track of placemarks associated with the map composition. A placemark is a particular window of the map layers. The Placemarks tab allows you to define new placemarks, delete existing ones, and go to any specific placemark. Placemarks are described further below in the section on Interactive Display Features.

Identify

The Identify button on the toolbar invokes the Identify cursor mode, which brings up a table of properties associated with features identified with the mouse.

Save Composition

At any point in your use of Composer, it is possible to save your map composition in several formats. The Save Composition button brings up a dialog presenting the format choices. The first is to save the map composition as a MAP file. A MAP file contains a complete record of the composition in a file with a ".map" extension.²⁴ This can then be redisplayed at any time by starting DISPLAY Launcher and indicating that you wish to display a map composition file. With this form of composition storage, it is possible to restore the composition to its exact previous state, and then continue with the composition process.

The next four options allow you to save the composition to either a PNG, BMP, WMF or EMF graphics file format. These file types are used in desktop publishing, spreadsheet and word processing software programs. The PNG and BMP options are raster bitmaps, and you are provided the option of creating them either at screen resolution, or at a multiple of screen resolution for higher resolution. BMP is an older format and is not compressed. PNG is generally preferred since it is compressed with lossless compression. The EMF format is the Windows "Enhanced Metafile" structure. The WMF format is the "Windows Metafile" structure used through Windows

²⁴ More information about the MAP file structure may be found in the chapter **Map Layers, Collections and Data Structures**. Note especially that the MAP file contains the instructions to construct the map composition but does not include the map data layers themselves. These separate files are used to recreate the composition.

3.1 and is generally now obsolete. Both store the Windows instructions that can be used to reconstruct both the vector and raster elements of the map window. These can then be imported into a desktop publishing or graphics program. Whenever you have a choice between WMF and EMF, choose the EMF format.

The next option copies the map window to the Windows clipboard rather than to a file. This facilitates copying compositions immediately into other software programs.

Finally, the last option allows you to save the currently windowed region of the active layer as a new TerrSet layer. If the layer is raster, only the portion of the raster image that is currently displayed will be written to the new raster file. This is thus an interactive version of the WINDOW module. However, if the active layer is vector, the entire file will be copied, but the bounding coordinates will be altered to match the current window. The new file will display the window region, but all the original data outside that window still exists in the file.

Print Composition

To print a map composition, first display it. Ensure that the map window has focus, then click the Print Composition button on Composer. In the Print Composition dialog, you will be able to select the desired printer, access its setup options, choose to fit the map window as large as possible on the page or print to a user-specified scale. Page margins may be set and the line widths may be scaled. A preview of the page as it will print is shown. All Windows-compatible printing devices are supported.

Stretch

At the bottom of Composer are several image stretch buttons. These allow you to stretch the display of the active image in a map window for optimal display. The inherent values are not altered, just the display minimum and maximum values stored in the files metadata. All three options perform a saturation stretch, the first option applies it to the entire image, the second option applies the stretch symmetrically around zero, assuming you have values above and below zero. The last option is used to stretch within a zoomed in display. It will stretch using the values only within the zoomed window, but it will apply it to the entire image. When you zoom back to the default display, you can click the first icon to optimize for the entire image.

Symbol Workshop

The map layers in a composition are rendered by means of symbol and palette files. While a set of symbol files is included with TerrSet, commonly you will want to develop specific symbol files to optimize the impact of the information presented on your final maps.

Symbol and palette files are created and modified using Symbol Workshop,²⁵ available under the Display menu and through its toolbar icon. In all, five types of files can be created: point symbol files, line symbol files, polygon symbol files, text symbol files and palette files. Symbol files record the graphic renditions for up to 256 symbols, indexed 0-255. For example, a text symbol file might indicate that features assigned symbol index 5 are to have their names rendered using bold, italic, 10-point Times New Roman text in a red color.

From the Symbol Workshop File menu, you can choose to open an existing file or create a new one. If you choose the latter, you will need to indicate the symbol type: point, line, polygon, text, or palette. The 256 symbols are then arrayed in a 16 x 16 grid. To change a symbol, simply click within its symbol grid cell. A symbol-specific dialog will then appear, allowing you to alter any of the following settings:

²⁵ IDRISI for Windows had a separate utility named Palette Workshop for the manipulation of palettes. This was incorporated into Symbol Workshop.

Symbol File Type	Attribute
Point Symbols	Symbol Type
	Fill Style
	Size
	Fill Color
	Outline Color
Line Symbols	Line Style
	Line Width
	Color
Polygon Symbols	Fill Style
	Color
Text Symbols	Font
	Size
	Color
	Style (normal, bold, italic, underline)
Palette	Color

A very important feature of Symbol Workshop is the ability to copy or blend attributes. For example, imagine creating a point symbol file of graduated circles. Click on symbol 0 and set its properties to yield a very small yellow circle. Then move to symbol 255 and set it to be a large red circle. Now set the blend end points to be 0 and 255 and click the blend button. You will now have a sequence of circles smoothly changing in both size and color.

As a companion to the blend option, Symbol Workshop also provides a copy function. With both the blend and copy functions, it is possible to set options that will cause them only to copy or blend specific attributes (such as color or size). To get a sense of how the symbols will look on a particular background, you may also set a background color for the Symbol Workshop display.

Finally, the autoscaling range of a symbol file may be specified or altered in Symbol Workshop. All symbol files contain definitions for 256 symbols. However, by altering the autoscale range, it is possible to create sequences of more limited range. For example, to set

up a palette with 8 colors, set the autoscale min and max to be 0 and 7 respectively. Then define these 8 colors. In use, it will then appear that the palette has only 8 colors if the layer with which it is used is autoscaled.

Media Viewer

Media Viewer is a facility for creating and displaying video images composed of a series of TerrSet images. Media Viewer is located under the File|Display menu. When activated, it will present the basic control dialog with its own separate menu. Click on the File menu to create a new video or to open an existing video. If you choose to create a new video, you will be presented with a new dialog that will require the name of either a Raster Group File (".rgf") or a Time Series File (".tsf") that defines the images to be used and their order in the video. You will also be asked to specify a palette to be used and the time delay to be used between each successive image. The result of the operation will be an ".avi" multi-media video file that can be displayed at any time using the Media Viewer controls. Note that the viewer can be sized by dragging its borders. The image can be sized to fit within the viewer by selecting the *Fit to Window* option under the Properties menu. Also note that you can place a video into a continuous loop.

Interactive Display Features

One of the remarkable features of GIS is that map displays are not static. Rather, they provide a highly interactive medium for the exploration of map data. The TerrSet System includes a number of features that form the basis for this interaction.

Pan and Zoom

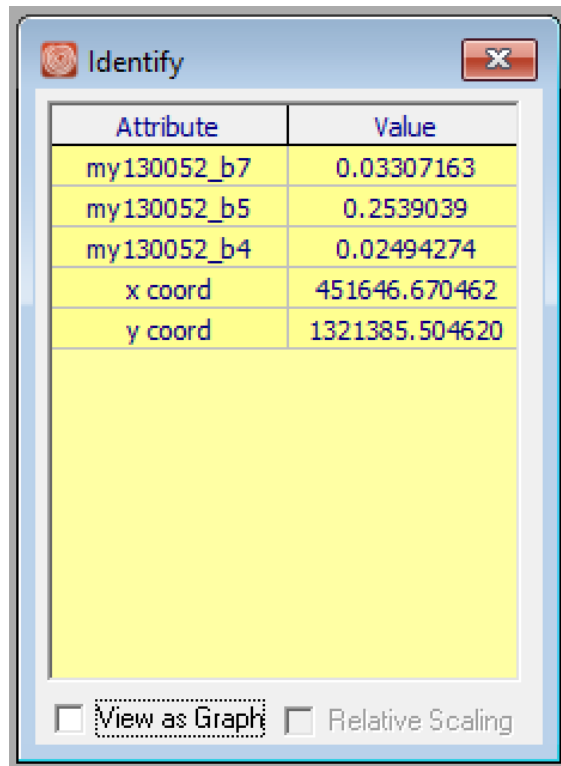
When the mouse is over the layer frame in a map window the cursor normally appears as a cross-hair. In this mode, the layer frame can be panned (moved horizontally) by holding the left button down while moving the mouse. Zooming (changing the map scale) in this mode can be achieved by moving the mouse wheel or by using the Page Up and Page Down keys.

The toolbar provides several other options for panning and zooming. Perhaps the most important is the Zoom Window mode. This allows you to define a rectangular region that should be zoomed. When you click this button, the cursor in the map window that currently has focus will change to indicate that Zoom Window mode is active. While it is active, move the mouse to the position of any of the corners of the region to be zoomed into, and then hold the right mouse button down and pull open the rectangle. When the rectangle has been defined to your satisfaction, release the left button. The layer frame will then be zoomed and you will exit from Zoom Window mode.

The toolbar also provides traditional Zoom In and Zoom Out buttons. These are provided for compatibility with other software systems.

Move and Resize

As discussed earlier, all map components (layer frame, title, legend, etc.) can be moved, and most can also be resized. However, to do so, the AutoArrange option on Composer must be checked off. Moving is achieved by placing the cursor over the component in question, double-clicking to enter move/resize mode, and then holding down the left mouse button to "drag" that element to its new location. Clicking on any other component, or the map window banner, will exit move/resize mode. While in move/resize mode, resizing is achieved by dragging one of the resizing buttons on the margins of that component.



Identify Mode



As you move the cursor over a map window, the status bar at the bottom of the screen will indicate the X and Y coordinates of the cursor in the geographic reference system (for raster layers, the column/row position is also shown). You can find out *what* is at that location by simply clicking with the left mouse button. However, the value displayed will only be that of the layer highlighted in Composer. If you wish to see the values for every layer in your composition, turn on Identify mode. To enable Identify mode, click on its button on the tool bar. The button will be highlighted. When this mode is active, you can query any location and it will show the value for all active layers, as well as the coordinates. Identify can remain active, but you may want to turn it off if you wish move and resize map components. To do so, simply click the Identify button again.

Pan and Zoom

Pan and zoom allow one to navigate around the display at varying scales (magnifications). The simplest method of pan and zoom is to use the mouse. If your computer has the facility, a center mouse wheel can be used to zoom and out. To pan, drag the image with the left mouse button. The keyboard arrow keys, and page up and page down keys may also be used. Functions are summarized in the table below. The easiest way to understand the actions of these keys is to imagine that you are in an airplane. Zooming in lowers your altitude, thus increasing your scale (i.e., you see less area, but with greater detail). Similarly, zooming out increases your altitude, thus reducing your scale (i.e., you will see more, but with less detail). The representative fraction (RF) in the status bar changes as you zoom in and out. A similar logic is associated with the arrow keys. Pressing the right arrow key (Pan Right) moves your imaginary airplane to the right, causing the scene to appear to move to the left

In addition to these continuous zoom and pan options, the tool bar also offers several additional options. The zoom in and zoom out icons will zoom in or out with each left mouse click, and recenter the image depending on the location of the mouse when it is clicked. A zoom window option allows you to draw a rectangle around the area you wish to zoom in to. To do so, click the Zoom Window

button (shown below) and move your cursor to any corner of the area to be zoomed into. Then hold down the left button and drag out the rectangle to the corner diagonally opposite the one you started with. When you release the left mouse button, the zoom will take place. The Home key will revert the map window and layer frame back to their original size.

Toolbar Icon	Keyboard	Action
		Zoom Window
	HOME	Restore Original Window

All zoom and pan operations take place within the context of the layer frame. See the earlier section of this chapter on Map Windows and Layer Frames for information about enlarging both of these.

Group Link

When a raster layer is part of a group, it is possible to have all displayed members of the group zoom and pan simultaneously and identically. To activate this, click the Group Link button on the tool bar. Notice that the button depresses and remains so until clicked again.

The Group Link works only when layers are launched with their complete collection reference. For layers belonging to raster group files, these layers exist in isolation as well as being referenced as part of a group. Thus, if you wish to use the group link feature, be sure to enter the full collection reference (e.g., "spotxs.band3" rather than simply "band3") or display them from within an .rgf in TerrSet Explorer.

Placemarks

Placemarks are the spatial equivalent of bookmarks. Any particular view of a composition can be saved as a placemark. To do so, either launch the Map Properties dialog and click the Placemarks tab or click the Placemarks icon on the tool bar. Then choose the appropriate option to name, rename, delete, or go to any placemark. Note that placemarks are saved with map compositions. They will not be saved if the composition is not saved.

Measure Length

Distances can be measured in map compositions. With an image displayed, click on the Measure Length icon. To begin measuring the distance, move the mouse to the desired start location and click the left mouse button. A balloon above the cursor will appear which will indicate the length as you move the mouse. If length is non-linear, click the left mouse button to indicate inflection points as you measure. End measuring by right clicking the mouse.

Measure Zone

Circular zones can be measured in map compositions. With an image displayed, click on the Measure Zone icon. To begin measuring circular zones, move the mouse to the desired start location and click the left mouse button. A balloon above the cursor will appear which will indicate the length from the center of the zone. A circular zone will also be drawn out as you move the mouse. End measuring by right clicking the mouse.

GPS Support

TerrSet also provides real-time GPS support, intended for use with a laptop computer. A Global Positioning System receiver receives continuous position updates in latitude/longitude²⁶ from active GPS navigation satellites. When the TerrSet GPS link is active, this position is shown graphically with a blinking cursor on all map windows that cover your current location and have a common reference system. TerrSet automatically projects the incoming positions to the reference system specified (so long as the specified projection is supported by PROJECT).²⁷ In addition, TerrSet will automatically save your route and allow you to save *waypoints*—positionally tagged notes.

Most receiver units available today support communication with a computer over an RS232C communications channel (e.g., the COM1 or COM2 port on your computer), and provide support for the NMEA (National Marine Electronics Association) communications protocol. Using a GPS with TerrSet involves the following steps:

1. Set the GPS to output NMEA data (this is usually an option in its setup menu and can remain as a permanent setting, avoiding this step in future use).
2. Connect the special communication cable that is designed for your GPS to one of the serial communication ports on your computer. For a laptop, this is typically a 9-pin port known as COM1. By default, TerrSet expects communication over COM1, but this can be changed (see below).
3. Display an image or map layer that includes your current location. Be sure that it has focus and then click the GPS button on the tool bar. You will then see the blinking position cursor appear simultaneously in all map windows that use the same reference system (as long as your actual position is within the currently displayed area).
 - a) When you have finished a GPS session, click the GPS button again to turn it off. At that point, it will give you the option of saving your route as a vector line layer and your waypoints (if any) as a vector point layer.

Saving Routes and Waypoints

While the GPS is connected and communicating with TerrSet, it automatically saves the route that was started when the GPS button was first clicked. It also keeps a record of any waypoints that are entered along the way. When you click the GPS button again to terminate GPS communication, you have the option of saving these as layers. The route will be saved as a vector line layer and the waypoint file will be stored as a text layer.

To save waypoints, simply press the "w" key at those positions for which you wish to record information. A dialog will then appear, allowing you to enter a text description. TerrSet will keep track of both the location and the text descriptor. Note that as a text layer, your waypoint information is easily displayed after the file has been saved. However, you will probably want to keep these waypoint descriptors very brief. Position readings continue to be recorded while waypoint information is being entered.

²⁶ Actually, the native reference system is a three-dimensional datum known as WGS84. However, the NMEA interface for communication of fixes converts these to latitude/longitude and elevation readings.

²⁷ See the chapter **Georeferencing** in this volume for a discussion of projections and reference systems.

Finally, note that TerrSet is not intended as a spatial database development tool. Thus, the route and waypoint saving features are more suited to simple ground truthing operations rather than database development.

How It Works

When you click the GPS button, TerrSet checks to see if there is a map layer with focus and with a valid reference system. This can be any system with a Reference System Parameter file that has a projection supported by the PROJECT module (this does not include the system labeled "plane"). This then becomes the output reference system, and all positions will be automatically converted into that form.


Next, TerrSet launches (if it is not already active) a special GPS server program named IDRNMEA.EXE from the folder called GPS located in the TerrSet program folder. Then it establishes communication with the GPS using communication parameters stored in a special file named "IDRNMEA.CFG," also found in the GPS folder of the TerrSet program folder. The default settings stored in this file (1,4800,n,8,1) are probably correct for your unit (since they assume the use of serial port 1 and comply with the NMEA standard) and can, in most cases, be used without modification. However, if you have trouble communicating with the GPS, this file can be edited using an ASCII text editor such as the TerrSet Edit module or Windows Notepad. Details can be found in the Help System.

Once communication has been established, TerrSet converts the native latitude/longitude format of the NMEA GPS fix to the designated TerrSet reference system, polls all map windows to find which ones are using this system, and then moves a blinking cursor to that location on each.

Interactive Screen Digitizing

Another very important interactive capability that TerrSet provides is the ability to digitize on screen. This facility is largely intended as a means of undertaking digitizing tasks such as the delineation of training sites for the classification of remotely sensed imagery or creating text vector layers in a very rapid fashion. Additional features allow for the editing of raster layers.

The Digitize button on the toolbar is shaped like a cross within a circle and can be found among a group of three related digitizing buttons.

Toolbar Icon	Action
	Digitize
	Delete Feature
	Save Digitized Data

To digitize on screen, make sure that the appropriate map window has focus, and then click on the Digitize button. If the active layer (the one highlighted in Composer) is a raster layer, you will be presented with a dialog box in which you can define the new vector layer to be created. If the active layer is a vector layer, you will first be asked whether you wish to create a new vector layer for the digitized features or append the new features to those existing in the active layer. If you choose to add to an existing layer, you will

notice that the file information is already filled out and you only need enter the ID or value. If you are digitizing a new file, you will be asked to specify the following:

Data Type

The data type refers to the attribute stored for each feature. This can be either integer or real. If you are digitizing identifiers to be associated with data in a table, this must be integer.²⁸ In all other cases, use integer for coding qualitative attributes and real for recording quantitative data (where a fractional number may be required), or whenever the value is expected to exceed the integer range.²⁹

Layer Type

Specify the nature of the features you wish to create. Note that if you choose to append to an existing layer, both the layer type and data type are predetermined to match those of the existing layer and you will not be able to change those settings. Layer type options include points, lines, polygons or text.³⁰

Automatic Index

This option is for those cases where you are storing identifiers for features (rather than some attribute directly). When checked, this option will automatically increment the identifier from one feature to the next. This offers speed when digitizing features whose ID's increment in this simple manner.

ID or Value / Index of First Feature

This allows you to specify the ID or attribute of the feature to be digitized. When the Automatic Index feature is specified, this will be used as the starting value for the numeric sequence.

Once you click OK to this startup dialog, a new layer will be added to your composition (unless you chose to add to an existing layer), and the digitizing cursor will appear in the Layer Frame. Note that if you were in Cursor Inquiry mode or Feature Properties mode, these will be temporarily disabled until you exit digitizing mode.

Mouse Button Functions When Digitizing

Once the Digitize dialog has been completed and you click on the OK button, you may begin digitizing. To digitize, use the left mouse button to identify points that define the position, course or boundary of a feature. You will notice it being formed as you digitize. To finish the feature (or the current sequence of points in the case of point features), click the right mouse button.

Deleting Features

²⁸ With vector data, no distinction is made between different storage formats of integer. Vector integer data can have values between $\pm 2,147,483,647$.

²⁹ Although vector layers can carry integer attribute values within a $\pm 2,147,483,647$ range, integer raster layers are restricted to a range of $\pm 32,767$. Thus, if you intend to rasterize your digitized data at a later stage, you may wish to specify real if the range is outside the $\pm 32,767$ range.

³⁰ When polygons are chosen, an option to digitize a flood polygon is presented. This is useful in image classification for delineating training sites. See the **Classification of Remotely Sensed Imagery** chapter for more information.

The Delete Feature button to the immediate right of the Digitize button on the toolbar allows you to select and delete vector features from the active vector layer. First click on the Delete Feature icon, then select the feature to be deleted (the cursor will become a hand). When the feature is selected, it will turn red. To delete it from the file, press the Delete key on the keyboard.

Digitizing Additional Features Within a Single File

Once you have finished a feature by clicking the right mouse button, you can continue to add further features to the same vector layer data file. Make sure the map window still has focus, and that the layer to which you wish to append another feature is highlighted, and then click the Digitize button on the toolbar. You will be asked whether you wish to add the new feature into the currently active layer or create a new layer. You will then set the ID for the next feature. If you are digitizing multiple features with the same ID, you can type "D" (without Ctrl), to skip the dialog box and to start to digitize another feature with the same ID.

Saving the Vector Layer

As you digitize features, they will each in turn be added to the vector layer designated. However, these data are not committed to disk until you click on the Save Digitized Data button. This can be done repeatedly throughout a session to save your work incrementally. If a map window is closed before a digitized layer is saved, a message will ask whether to save the layer or changes made to the layer.

Digitizing Multiple Layers

You may have multiple vector layers open for digitizing at the same time. Any actions you take will apply only to the active layer, the name of which is highlighted in Composer.

Using a Vector Layer to Update a Raster Image

You can update a raster image using the digitize facility. It will update a raster image identified in Composer with layer type point, line or polygon by assigning a new ID for the feature digitized. After launching the Digitize facility, select the option to use vector features to update a raster image and select the raster image in the list to update. The list will contain all raster images in Composer.

With the option to update all pixels within digitized features, all pixels falling within the digitized feature will change to the new class ID. The one class with digitized feature option will change all pixels within the digitized feature to the new class ID that have a specified original ID. The use mask file option will change all pixels within the digitized feature to the new class ID that also are non-zero values for that pixel in a secondary mask image. The line buffer option will only work with the line layer type and will add new features with a new class ID out from the digitized line feature by as many pixels widths specified.

A Special Note About Digitizing Point Features

With both line and polygon layers, you digitize a single feature at a time, right click, then click the Digitize button again to digitize the next feature. However, with point vector layers, you can digitize multiple features before ending a sequence with the right mouse button. To do this, select the Automatic Indexing option.

A Special Note About Digitizing Polygon Features

Polygons are defined by lines that join up to themselves. When digitizing polygonal features, clicking the right button not only terminates the definition of the feature, but it also adds a final point which is identical to the first, thereby closing the feature perfectly. As a result, it is not necessary to try to close the feature by hand—it will be done automatically.

A Special Note on Photo Layers

Photo Layer links photos to map layers which is very useful for linking photos from the field during a ground-truthing exercise. Photo Layers are IDRISI format text vector files that use a special syntax. Photos can be imported automatically or manually and can be in .jpg, .tif, .bmp or .png formats. The automatic option, using the module **Photo Layer**, will use the georeferencing information contained in the photo to create a point of the exact location. For the manual option, photo layers are created as text layers during the on-screen digitizing process, either through digitizing a new text layer or when laying down waypoints during GPS interaction. In both cases, entering the correct syntax for the text caption will create a photo layer. See the Help section on **Photo Layer** for details.

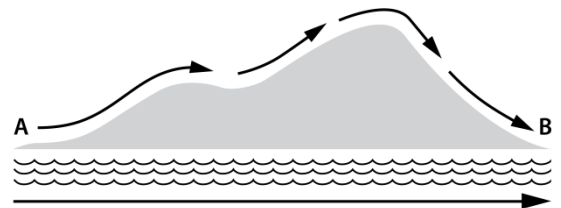
A Final Note

Perhaps the best way to learn more about the TerrSet display system is to work through the first few tutorial exercises which can be accessed from the Help menu. These tutorials give step-by-step instructions and guide you through the basics of map layer display and map composition in TerrSet.

Georeferencing

Georeferencing refers to the way map locations are related to earth surface locations. Georeferencing requires several ingredients:

- a logic for referring to earth surface locations—a concern of the field of Geodesy;
- a specific implementation of that logic, known as a Geodetic Datum—a concern of the field of Surveying;
- a logic for referring locations to their graphic positions—a concern of the field of Cartography; and
- an implementation of that logic, known as a data structure—a concern of GIS and Desktop Mapping software.

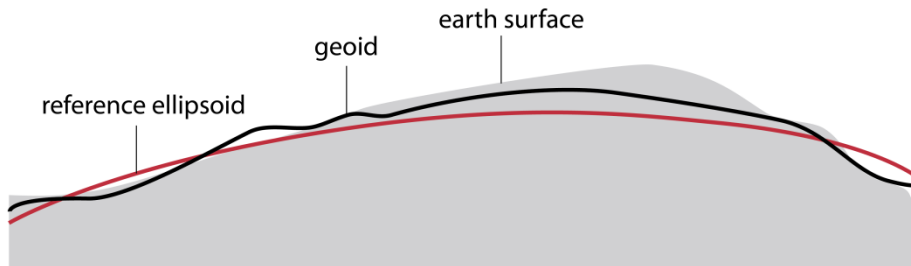


Geodesy

Geodesy is that field of study which is concerned with the measurement of the size and shape of the earth and positions upon it. The most fundamental problem that geodesists face is the fact that the earth's surface is irregular in shape. For example, imagine two locations (A and B) at either end of a thin straight beach bordered by a steep cliff (figure above). Clearly the distance one would determine between the two locations would depend upon the route chosen to undertake the measurement. Measuring the distance by the cliff route would clearly be longer than that determined along the coast. Thus irregularities (hills and valleys) along the measuring surface can cause ambiguities in distance (and thereby, location). To alleviate this situation, it has long been a common practice to reduce all measurements to a more regular measuring surface—a *reference surface*.

Geoid

The oldest reference surface used for mapping is known as the *geoid*. The geoid can be thought of as mean sea level, or where mean sea level would be if the oceans could flow under the continents. More technically, the geoid is an *equipotential surface of gravity* defining all points in which the force of gravity is equivalent to that experienced at the ocean's surface. Since the earth spins on its axis and causes gravity to be counteracted by centrifugal force progressively towards the equator, one would expect the shape of the geoid to be an oblate spheroid—a sphere-like object with a slightly fatter middle and flattened poles. In other words, the geoid would have the nature of an ellipse of revolution—an *ellipsoid*.

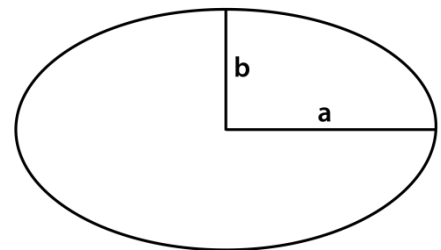


As a reference surface, the geoid has several advantages—it has a simple physical interpretation (and an observable position along the coast), and it defines the horizontal for most traditional measuring instruments. Thus, for example, leveling a theodolite or sextant is, by definition, a process of referring the instrument to the geoid.

Reference Ellipsoids

Unfortunately, as it turns out, the geoid is itself somewhat irregular. Because of broad differences in earth materials (such as heavier ocean basin materials and lighter continental materials, irregular distributions such as mountains, and isostatic imbalances), the geoid contains undulations that also introduce ambiguities of distance and location. As a result, it has become the practice of modern geodetic surveys to use abstract reference surfaces that are close approximations to the shape of the geoid, but which provide perfectly smooth reference ellipsoids (figure above). By choosing one that is as close an approximation as possible, the difference between the level of a surveying instrument (defined by the irregular geoid) and the horizontal of the reference ellipsoid is minimized. Moreover, by reducing all measurements to this idealized shape, ambiguities of distance (and position) are removed.

There are many different ellipsoids in geodetic use (see **Appendix 1: Ellipsoid Parameters**). They can be defined either by the length of the major (a) and minor (b) semi-axes³¹ (see figure), or by the length of the semi-major axis along with the degree of flattening [$f = (a-b) / a$]. The reason for having so many different ellipsoids is that different ones give better fits to the shape of the geoid at different locations. The ellipsoid chosen for use is that which best fits the geoid for the particular location of interest.



³¹ The major and minor semi-axes are also commonly referred to as the semi-major and semi-minor axes.

Geodetic Datums

Selecting a specific reference ellipsoid to use for a specific area and orienting it to the landscape, defines what is known in Geodesy as a *datum* (note that the plural of datum in geodesy is *datums*, not data!). A datum thus defines an ellipsoid (itself defined by the major and minor semi-axes), an initial location, an initial azimuth (a reference direction to define the direction of north), and the distance between the geoid and the ellipsoid at the initial location. Establishing a datum is the task of geodetic surveyors, and is done in the context of the establishment of national or international geodetic control survey networks. A datum is thus intended to establish a permanent reference surface, although recent advances in survey technology have led many nations to redefine their current datums.

Most datums only attempt to describe a limited portion of the earth (usually on a national or continental scale). For example, the North American Datum (NAD) and the European Datum each describe large portions of the earth, while the Kandawala Datum is used for Sri Lanka alone. Regardless, these are called local datums since they do not try to describe the entire earth. By contrast, we are now seeing the emergence of World Geodetic Systems (such as WGS84) that do try to provide a single smooth reference surface for the entire globe. Such systems are particularly appropriate for measuring systems that do not use gravity as a reference frame, such as Global Positioning Systems (GPS). However, presently they are not very commonly found as a base for mapping. More typically one encounters local datums, of which several hundred are currently in use.

Datums and Geodetic Coordinates

Perhaps the most important thing to bear in mind about datums is that each defines a different concept of geodetic coordinates—latitude and longitude. Thus, in cases where more than one datum exists for a single location, more than one concept of latitude and longitude exists. It can almost be thought of as a philosophical difference. It is common to assume that latitude and longitude are fixed geographic concepts, but they are not. There are several hundred different concepts of latitude and longitude currently in use (one for each datum). It might also be assumed that the differences between them would be small. However, that is not necessarily the case. In North America, a change is being undertaken to convert all mapping to a recently defined datum called NAD83. At Clark University, for example, if one were to measure latitude and longitude according to NAD83 and compare it to the ground position of the same coordinates in the previous system, NAD27, the difference is in excess of 40 meters! Other locations in the US experience differences in excess of 100 meters. Clearly, combining data from sources measured according to different datums can lead to significant discrepancies.

The possibility that more than one datum will be encountered in a mapping project is actually reasonably high. In recent years, many countries have found the need to replace older datums with newer ones that provide a better fit to local geoidal characteristics. In addition, regional or international projects involving data from a variety of countries are very likely to encounter the presence of multiple datums. As a result, it is imperative to be able to transform the geodetic coordinates of one system to those of another. In TerrSet, the PROJECT module incorporates full datum transformation as a part of its operation.

Cartographic Transformation

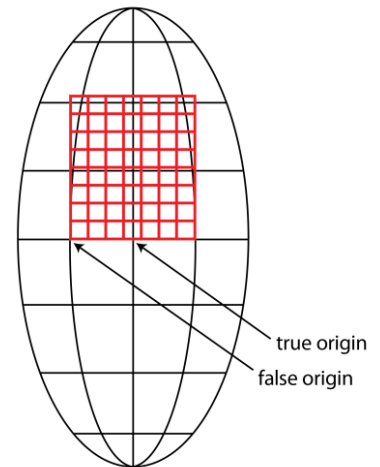
Once a logic has been established for referring to earth locations and a set of measurements has been made, a means of storing and analyzing those positions is required. Traditionally, maps have been the preferred medium for both storage and analysis, while today that format has been supplemented by digital storage and analysis. Both, however, share a common trait—they are most commonly flat! Just as flat maps are a more manageable medium than map globes, plane coordinates are a more workable medium than spherical (ellipsoidal) coordinates for digital applications. As a result, surveyed locations are commonly transformed to a plane grid referencing system before use.

Projection

The process of transforming spheroidal geodetic coordinates to plane coordinate positions is known as projection, and falls traditionally within the realm of cartography. Originally, the concern was only with a one-way projection of geodetic coordinates to the plane coordinates of a map sheet. With the advent of GIS, however, this concern has now broadened to include the need to undertake transformations in both directions in order to develop a unified database incorporating maps that are all brought to a common projection. Thus, for example, a database developed on a Transverse Mercator projection might need to incorporate direct survey data in geodetic coordinates along with map data in several different projections. *Back projecting* digitized data from an existing projection to geodetic coordinates and subsequently using a *forward projection* to bring the data to the final projection is thus a very common activity in GIS. The PROJECT module in TerrSet supports both kinds of transformation.

Projection of spheroidal positions to a flat plane simply cannot be done without some (and oftentimes considerable) distortion. The stretching and warping required to make the transformation work leads to continuous differences in scale over the face of the map, which in turn leads to errors in distance, area and angular relationships. However, while distortion is inevitable, it is possible to engineer that distortion such that errors are minimized, and certain geometrical qualities are maintained.

Of importance to GIS are the family of projections known as *conformal* or *orthomorphic*. These are projections in which the distortion is engineered in such a manner that angular relationships are correctly preserved in the near vicinity of any location. Traditionally, the field of surveying has relied upon the measurement of angular relationships (using instruments such as a theodolite) in order to measure accurate distances over irregular terrain. As a result, conformal projections have been important for the correct transfer of field measurements to a map base, or for the integration of new surveys with existing map data. For the same reasons, conformal projections are also essential to many navigation procedures. As a consequence, virtually all topographic map bases are produced on conformal projections, which form the basis for all of the major grid referencing systems in use today.



Grid Referencing Systems

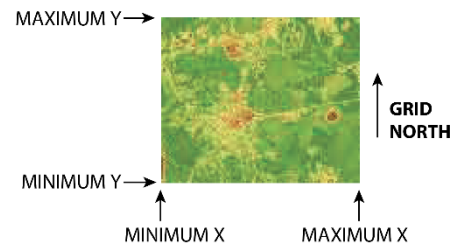
A grid referencing system can be thought of very simply as a systematic way in which the plane coordinates of the map sheet can be related back to the geodetic coordinates of measured earth positions. Clearly, a grid referencing system requires a projection (most commonly a conformal one). It also requires the definition of a plane Cartesian coordinate system to be superimposed on top of that projection. This requires the identification of an initial position that can be used to orient the grid to the projection, much like an initial position is used to orient a datum to the geoid. This initial position is called the *true origin* of the grid and is commonly located at the position where distortion is least severe in the projection. Then, like the process of orienting a datum, a direction is established to represent *grid north*. Most commonly, this will coincide with the direction of true north at the origin. However, because of distortion, it is impossible for true north and grid north to coincide over many other locations.

Once the grid has been oriented to the origin and true north, a numbering system and units of measure are determined. For example, the UTM (Universal Transverse Mercator) system uses the equator and the central meridian of a 6-degree wide zone as the true origin for the northern hemisphere. The point is then given arbitrary coordinates of 500,000 meters east and 0 meters north. This then gives a *false origin* 500 kilometers to the west of the true origin (see figure above). In other words, the false origin marks the location where the numbering system is 0 in both axes. In TerrSet, the numbering logic of a grid referencing system is always given by specifying the latitude and longitude of the true origin and the arbitrary coordinates that exist at that point (in this example, 500,000 E and 0 N).

Georeferencing in TerrSet

The logic that TerrSet uses for georeferencing is quite simple and is based on the issues previously described.

All geographic files are assumed to be stored according to a grid reference system where grid north is aligned with the edges of the raster image or vector file. As a result, the minimum X of a raster image is always the left-hand edge, the maximum Y is the top edge, and so on (see figure). The georeferencing properties of a TerrSet coverage (accessible through the Metadata utility in TerrSet Explorer) include an entry specifying the reference system used by that file when referring to geographic locations. The particulars of that reference system (e.g., projection, datum, origin, etc.) are then contained in a *reference system parameter file* (.ref) (see below). Whenever a projection or datum transformation is required, the PROJECT module refers to this information to control the transformation process.



Every grid reference system must have a reference system parameter file. The only exception to this is a system identified by the keyword "plane". Any coverage that indicates a *plane* coordinate referencing system is understood to use an arbitrary plane system for which geodetic and projection parameters are unknown, and for which a reference system parameter file is not provided. A coverage in a plane reference system cannot be projected.

Over 500 reference system parameter files are supplied with TerrSet. However, .ref files can be created for any grid referencing system using the Metadata utility in TerrSet Explorer or Edit modules. Details on the structure of this simple ASCII text file format are provided below and in the Help System.

For simplicity, geodetic coordinates (latlong) are recognized as a special type of grid referencing system. The true and false origins are identical and occur at 0 degrees of longitude and 0 degrees of latitude. Units are in decimal degrees or radians. Longitudes west of the prime (Greenwich) meridian and latitudes south of the equator are expressed as negative numbers. Thus, for example, Clark University has approximate coordinates of -71.80,+42.27 (71°48' W, 42°16' N). Although geodetic coordinates are truly spheroidal, they are logically treated here as a plane coordinate system. Thus, they are implicitly projected according to a Plate Carrée projection,³² although the projection will actually be listed as "none".

Be aware that there are many possible interpretations of the concept of latitude and longitude, depending upon the datum in use. A single .ref file has been supplied (called LATLONG) for the case where the datum is WGS84. This should be copied and modified for other interpretations.

In TerrSet, the registration point for referencing raster images is in the lower left corner of any cell. Thus, for example, with an image of 10 columns by 5 rows, using a plane reference system of cells 10 meters by 10 meters, the lower left-most corner of the lower left-most cell has the coordinates 0,0 while the upper right-most corner of the upper right-most cell has the position 100, 50. Note also that this lower left-most cell is considered to be in column 0 and row 4 while the upper right-most cell is in column 9, row 0.

TerrSet image and vector documentation files contain several fields that describe the reference system of the file. The *min X*, *max X*, *min Y*, and *max Y* entries give the edge coordinates of an image or bounding rectangle of a vector file in reference system coordinates. For image files, the number of *rows* and *columns* is also given.

The reference system name (*ref. system*) identifies the reference system of the file, and may be "plane", "latlong", or a specific reference system described by a reference system parameter file. In the latter case, the *ref. system* entry in the documentation file should match exactly the name of the corresponding reference system parameter file, without the .ref extension.

³² Note, however, that geodetic coordinates are not considered to be identical to a system based on the Plate Carrée since no new numbering scheme has been undertaken. PROJECT supports transformations based on the Plate Carrée as well, where it is understood that a true plane coordinate system with its own special numbering scheme has been superimposed onto the projected geodetic coordinates.

The *reference units* entry indicates the unit of measure used in the reference coordinate system (e.g., meters).

The *unit distance* refers to the ground distance spanned by a distance of one unit (measured in reference units) in the reference system. Thus, for example, if the hypothetical image in figure above had a unit distance of 2.0 and reference units in meters, it would imply that as we move one in the reference coordinates (e.g., from X=99 to X=100) we actually move 2.0 meters on the ground. Simply think of the unit distance parameter as a multiplier that should be applied to all coordinates to yield the reference units indicated. ***The unit distance will be 1.0 in most cases.*** However, the unit distance should be specified as a value other than 1.0 whenever you have reference units that are not among the standard units supported (meters, feet, miles, kilometers, degrees or radians). For example, a data file where the coordinates were measured in minutes of arc (1/60 of a degree) should have reference units set to degrees and the unit distance set to 0.016667.

In TerrSet, as you move the cursor over a displayed image, the row and column position and the X and Y coordinates are indicated on the status bar at the bottom of the screen. Vector files can also be overlaid (using TerrSet Explorer or Composer) onto the raster image, provided they have the same reference system. *There is no need for the bounding rectangle of the vector file to match that of the raster image.* TerrSet will correctly overlay the portion of the vector file that overlaps the displayed raster image. Onscreen digitizing will always output coordinates in the same reference system as the raster image (or vector layer) displayed.

Reference System Parameter Files

As previously indicated, TerrSet comes with over 400 reference system parameter files. These include one for geodetic coordinates (latitude/longitude) using the WGS84 datum, 120 for the UTM system (one each for the 60 zones, for both the northern and southern hemispheres) using the WGS84 datum, 32 based on the Gauss-Kruger projection, 40 for the UTM system covering North America using NAD27 and NAD83, and 253 for all US State Plane Coordinate (SPC) systems based on the Lambert Conformal Conic and Transverse Mercator projections. ***Appendix 3: Supplied Reference System Parameter Files*** lists each of these files by geographic location. During program installation, these supplied files are installed in the \Georef subfolder of your TerrSet program folder. The contents of these files may be viewed with the Metadata utility in TerrSet Explorer or Edit.

Where .ref Files are Stored

As indicated above, supplied .ref files are stored in a subfolder of the TerrSet program folder called \Georef. If your TerrSet program folder is C:\Program Files (x86)\TerrSet2020, then the supplied reference system parameter files are in C:\Program Files (x86)\TerrSet\Georef. When you create new reference system parameter files, we recommend that you store them in this folder, always adding to your master library of reference files. However, reference system parameter files may be stored anywhere. When a reference system parameter file is given without its path in a dialog box or macro, TerrSet always looks first in the Working Folder, followed by the Resource folders in the order they are named in the current project. If the file is not found, TerrSet will next look in the \Georef subfolder.

Creating New .ref Files

Although TerrSet supplies more than 400 reference system parameter files, many users will need to create new .ref files to suit the needs of specific systems. There are two options available to do this:

1. For cases where a different version of the UTM system is required, the module UTMREF can be used. This will simply require that you enter the zone number, the hemisphere (northern or southern), and the name of the datum along with its associated Molodensky constants. The constants can be found in ***Appendix 2: Datum Parameters*** and are used by PROJECT to undertake datum transformations. If you enter a datum that is not included, you will need to supply the name of the reference ellipsoid and the length of its major and minor semi-axes. These constants for many reference ellipsoids are given in ***Appendix 1: Reference Ellipsoids***.

2. For all other cases, TerrSet Explorer or Edit can be used to create the appropriate file. Often the easiest procedure is to use the copy function in TerrSet Explorer to duplicate an existing .ref file, and then modify it using the Metadata utility in TerrSet Explorer or Edit. The section below indicates the structure and contents of a .ref file.

The .ref File Structure

Like all TerrSet documentation files, .ref files are simple ASCII text files that can be modified by means of any ASCII text editor.

The .ref file type follows the conventions of other documentation files by having the first 14 characters of each line devoted to a field description. Here is an example of a reference system parameter file named US27TM19:

```
ref. system      :  Universal Transverse Mercator Zone 19

projection       :  Transverse Mercator

datum           :  NAD27

delta WGS84      :  -8 160 176

ellipsoid        :  Clarke 1866

major s-ax       :  6378206.5

minor s-ax       :  6356584

origin long      :  -69

origin lat       :  0

origin X         :  500000

origin Y         :  0

scale fac        :  1

units           :  m

parameters      :  0
```

The first line is simply a title and has no analytical use. The second line indicates the projection. In the current release of TerrSet, the following projections are supported:

Mercator

Transverse Mercator

Gauss-Kruger (Gauss-Krueger spelling also accepted.)

Lambert Conformal Conic

Plate Carrée
Hammer Aitoff
Lambert North Polar Azimuthal Equal Area
Lambert South Polar Azimuthal Equal Area
Lambert Transverse Azimuthal Equal Area
Lambert Oblique Azimuthal Equal Area
Cylindrical Equal Area
North Polar Stereographic
South Polar Stereographic
Transverse Stereographic
Oblique Stereographic
Albers Equal Area Conic
Normal Aspect Cylindrical Equal Area
Sinusoidal
none (i.e., geodetic coordinates)

Note that each of the names above are keywords and must appear exactly in this form to be understood by the PROJECT module.

The next line lists the geodetic datum. The text here is informational only, except for the keywords NAD27 and NAD83. These latter two cases are the keywords for the two existing implementations of the North American datum. When PROJECT encounters these keywords as elements of the input and output systems, it uses a special (and more precise) datum transformation technique. For all other cases, the Molodensky transform is used.

The next line lists the differences between the center (i.e., earth center) of the datum in use and that of the WGS84 datum (World Geodetic System, 1984). The values express, in meters, the differences in the X, Y and Z axes respectively, relative to WGS84. These are known as the Molodensky constants. **Appendix 2: Datum Parameters** contains a list of these constants for most datums worldwide. They are used by PROJECT to undertake datum transformations except in the case of conversions between NAD27 and NAD83. These conversions use the US National Geodetic Survey's NADCON procedure.

The next three lines give information about the reference ellipsoid. The line listing the name of the ellipsoid is for informational purposes only. The following two lines describe, in meters, the major and minor semi-axes of the ellipsoid used by the datum. These entries are used analytically by the PROJECT module. **Appendix 1: Ellipsoid Parameters** contains a list of ellipsoids and the lengths of their semi-axes.

The next four lines describe the origin and numbering system of the reference system. The *origin long* and *origin lat* entries indicate the longitude and latitude of the true origin. The *origin X* and *origin Y* entries then indicate what coordinate values exist *at that location* in the reference system being described.

The *scale fac* entry indicates the scale factor to be applied at the center of the projection. A typical value would be 1.0, although it is equally permissible to indicate "na" (an abbreviation for "not applicable"). The most likely case in which users will encounter a need to specify a value other than 1.0 is with the UTM system and other systems based on the Transverse Mercator projection. With the UTM system, the *scale fac* should read 0.9996.

The *units* entry duplicates the units information found in raster and vector documentation files and is included here for confirmation by the system. Therefore, in both reference and documentation files, *units* should be considered as analytical entries.

Finally, the *parameters* entry indicates the number of special parameters included with the .ref file information. Of the projections currently supported, only the Lambert Conformal Conic and Alber's Equal Area Conic require special parameters. All others should have a value of 0 for this entry. In cases using either the Lambert Conformal Conic or Alber's Equal Area Conic projections, the *parameters* entry should read 2 and then include the latitudes (in decimal degrees) of the two standard lines (lines of no distortion)

on the following two lines. Here, for example, is the .ref file for the Massachusetts State Plane Coordinate System, Zone 1 (SPC83MA1), based on the NAD83 datum:

```
ref. system      :  Massachusetts State Plane Coordinate System Mainland Zone

projection       :  Lambert Conformal Conic

datum           :  NAD83

delta WGS84      :  0 0 0

ellipsoid        :  GRS80

major s-ax       :  6378137

minor s-ax       :  6356752.31

origin long      :  -71.5

origin lat       :  41

origin X         :  200000

origin Y         :  750000

scale fac        :  na

units           :  m

parameters       :  2

stand ln 1       :  41.72

stand ln 2       :  42.68
```

Note that the latitude of the "lower" standard line (i.e., that which is nearer the equator) is listed first and that of the latitudinally "higher" standard line is listed second. Future additions of other projections may dictate the need for additional special parameters. However, in this version, only those reference systems based on the Lambert Conformal Conic and Alber's Equal Area Conic projections require special parameters.

Projection and Datum Transformations

Projection transformations for all supported projections are undertaken using ellipsoidal formulas accurate to 2 cm on the ground. As TerrSet uses double precision floating point numbers (15 significant figures) for the representation of coordinate data, this

precision is maintained in all results. However, be aware that the precision of exported coordinates depends on the numeric precision used by the export format being used.

For datum transformations, the primary method used is the Molodensky transformation described in DMA (1987). Whenever the keywords "NAD27" and "NAD83" are encountered in a transformation, however, the more precise US National Geodetic Survey NADCON procedure is used (see Dewhurst, 1990). *This latter procedure should **only** be used within the continental US.* For other regions covered by the North American Datum, do not use the NAD27 or NAD83 keywords, but rather some other spelling in the datum field, and indicate the correct Molodensky transformation constants from *Appendix 2: Datum Parameters*.³³

Algorithms Used by PROJECT

Forward and backward ellipsoidal projection formulae were taken from Snyder (1987). For datum transformations, the Molodensky transform procedure (DMA, 1987) is used. However, for the specific case of NAD27/NAD83 transformations within the continental US, the US National Geodetic Survey's NADCON procedure is used. This procedure is described in Dewhurst (1990) and is accompanied by relevant data files. Briefly, the procedure involves a matrix of corrections to longitude and latitude that are accessed as a two-dimensional look-up table. Final correction values are then interpolated between the four nearest values in the matrix using a bilinear interpolation. To facilitate this, the NADCON data files for the continental US were converted into TerrSet images. These images are contained in the GEOREF subdirectory under the TerrSet program directory. There are two of these files, NADUSLON and NADUSLAT, for the corrections to longitude and latitude respectively.

Further Reading

Geodesy is not a familiar topic for many in GIS. However, as one can see from the material in this chapter, it is essential to effective database development. The following readings provide a good overview of the issues involved:

Burkard, R.K., (1964) *Geodesy for the Layman*, (St. Louis, MO: USAF Aeronautical Chart and Information Center).

Smith, J.R., (1988) *Basic Geodesy*, (Rancho Cordova, CA: Landmark Enterprises).

For projections, there are few texts that can match the scope and approachability of:

Maling, D.H., (1973) *Coordinate Systems and Map Projections*, (London: George Phillip and Son).

Snyder, J.P., (1987) *Map Projections: A Working Manual*. U.S. Geological Survey Professional Paper 1395, (Washington, DC: US Government Printing Office).

Finally, with respect to the procedures used for datum transformations, please refer to:

DMA, (1987) *Department of Defense World Geodetic System 1984: Its Definition and Relationships with Local Geodetic Datums*, DMA TR 8350.2, (Washington, DC: The Defense Mapping Agency).

Dewhurst, W.T., (1990) *NADCON: The Application of Minimum Curvature-Derived Surfaces in the Transformation of Positional Data from the North American Datum of 1927 to the North American Datum of 1983*, NOAA Technical Memorandum NOS NGS-50, (Rockville, MD: National Geodetic Information Center).

³³ Note that while a set of constants exist for the North American Datum as a whole, Molodensky constants also exist for more specific portions of the regions covered to enhance the precision of the transformation.

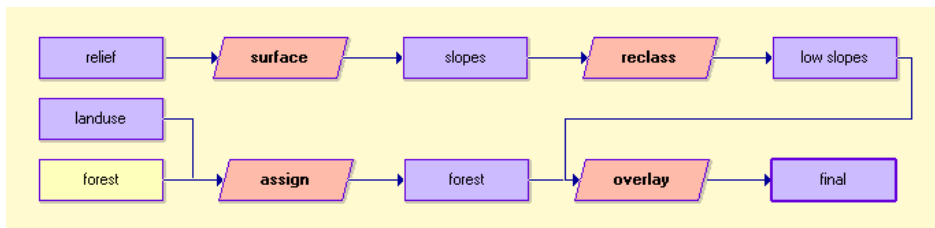
Model Deployment Tools

One of the most fundamental roles for GIS is in the development, testing and utilization of models—suitability models, land change models, soil erosion models, urban growth models, and the like. TerrSet provides an extensive set of tools for creating and deploying models that can either use existing TerrSet toolsets, or allow the more skilled programmer to deploy and integrate their own tools.

The most fundamental is Macro Modeler, a graphical modeling environment that combines the strengths of extensive capabilities and ease of use. Spatial Decision Modeler, similar in its implementation to Macro Modeler, is a graphical modeling environment just for decision support utilizing the main decision support tools found in TerrSet. For simpler equation-based modeling using raster layers, Image Calculator provides rapid equation entry using a familiar calculator interface. A third facility is a macro scripting (.iml) language, used with Run Macro, that is provided largely for legacy applications (this was the original form of modeling tool provided in an early releases of our software). Finally, there is the IDRISI API, an industry standard COM interface that provides access to the internals of the TerrSet system for the most demanding applications and interface development. The API requires a COM-compliant programming environment such as Visual C++, Delphi, Python, or Visual Basic.

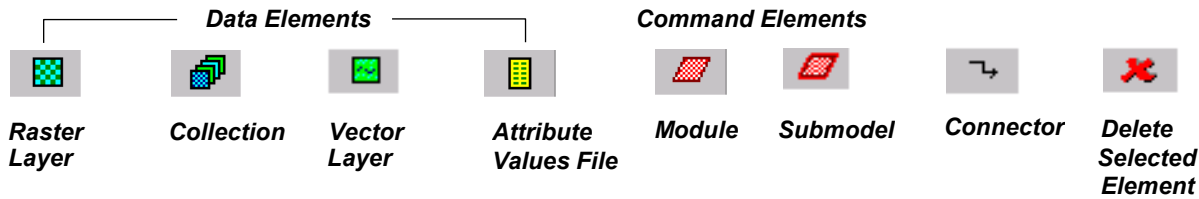
Macro Modeler

Macro Modeler is a graphic environment in which you may assemble and run multi-step analyses. Input files, such as raster images, vector layers, and attribute values files, are linked with TerrSet modules that in turn link to output data files. The result is a graphic model, much like the cartographic models described in the Introductory GIS Exercises of the **Tutorial**. A model may be as simple or complex as desired. The figure below shows a simple suitability mapping model.



Model Construction

To work with Macro Modeler, either click on its icon on the tool bar, or select it from the Idrisi GIS Analysis|Modeling Deployment Tools Menu. This yields a special workspace in the form of a graphic page along with a separate menu and tool bar. Constructing a model involves placing symbols for data files and modules on the graphic page, then linking these model elements with connectors. To place an element, click on its icon (shown in the figure below) or choose the element type from the Macro Modeler menu, then choose the specific file or operation from the supplied pick list. (Note that not all TerrSet modules are available for use in the Macro Modeler.) To connect elements, click on the connector icon, then click on one of the elements and drag the cursor onto the other element and release. It is necessary to connect elements in the proper order because the Macro Modeler assumes that the process flows *from* the element you first clicked *to* the element upon which you released the cursor. To delete any element, first click on that element, then click the delete icon or press the delete key on the keyboard.



Whenever a module is placed in the model, it is placed with the appropriate output data element for that operation already linked. Default temporary filenames are assigned to output files. Right-click on a data file symbol to change the filename. Long filenames are left-justified. To see the entire filename, pause the cursor over the element graphic to cause a balloon to appear with the entire element name. Filenames are stored without paths. Input files may exist in any folder of the current project (specified with TerrSet Explorer) while intermediate and final output files are always written to the Working Folder.

To specify the parameters for the module, right-click on the module symbol. This opens the module parameters dialog. Click on any parameter to see and choose other options. Access the Help System for detailed information about the various options by clicking the Help button on the parameters dialog.

Submodels are user-constructed models that are subsequently encapsulated into a single command element. When a model is saved as a submodel, a parameters dialog is created for the submodel in which the user provides captions for all the model inputs and outputs. A submodel consists of a submodel parameter file (.ims) and its macro model (.imm) file. When a submodel is placed as a command element in a model, the Macro Modeler knows how many and what types of input files are required to run the submodel and what types of output files should be produced. For example, you might create a submodel that stretches a single image then exports it to a JPG image. The model would include both STRETCH and JPGIDRIS, but the submodel would only require the input image and the output image. The user may change the input and output data elements of a submodel through the parameters dialog, but settings for module commands that are part of a submodel must be changed by opening the model from which the submodel was created, making the changes, then resaving the submodel. Submodels not only simplify multi-step processes, but are important elements when models have sections that should loop. This is discussed in further detail below.

Models are saved to an IDRISI Macro Model file (.imm). This file preserves all aspects of the model, including the graphic layout. The file may be opened and modified with the Macro Modeler. The graphic description of the model may be copied (as a .bmp file) to the operating system clipboard then pasted into other word processing and graphics software. The graphic may also be printed. The toolbar icons for Macro Modeler file management are shown in the figure below.



Models may be run at any stage of completion. The Macro Modeler first checks to see if any of the output images already exist. If so, it shows a message indicating which file exists and asks if you wish to continue. Answering Yes (or Yes to All) will cause the existing output file to be deleted before the model runs. Answering No will prevent the Macro Modeler from running the model. As the model runs, the module element that is currently being processed turns green. To stop a model at any point while it is running, click the stop icon. The model will finish the process it is currently doing and will then stop.

The last output created by the model will be automatically displayed. Intermediate images may be displayed by clicking on the data layer symbol of the desired file, then on the display icon on the Macro Modeler toolbar. The metadata may be accessed for any data layer by clicking the data layer symbol, then on the Metadata icon on the Macro Modeler toolbar. Toolbar icons for these functions

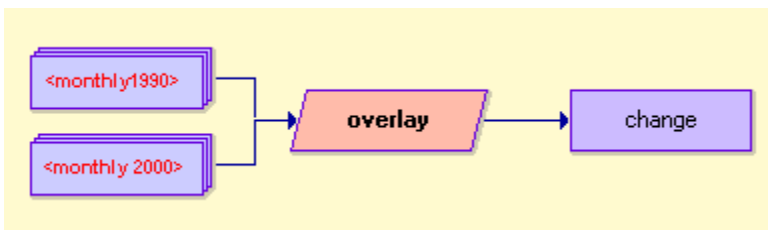
are shown in the figure below.



DynaGroups

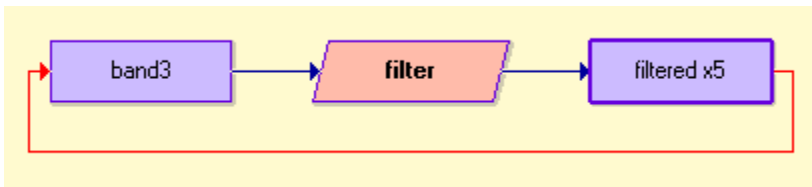
Two distinctive capabilities of the Macro Modeler are its use of DynaGroups to facilitate running the same model on multiple data layers and DynaLinks to construct iterative models in which the output of a process becomes an input for the next iteration of the process.

A DynaGroup is a raster group file or time series file that is used in the Macro Modeler so each member of the group is used to produce an output. Contrast this with the use of a group file as a regular data input to the Macro Modeler in which a group file used as input produces a single output as with the module COUNT. More than one DynaGroup may be used in a process. The module OVERLAY, for example, requires two raster images as input and produces a single raster image output. If DynaGroups are used as the two inputs, then Macro Modeler will first verify that the same number of members is in each group, then it will run OVERLAY using corresponding images from the two group files. If a single raster layer is used as one input and a DynaGroup is used as the other input to OVERLAY, then the modeler will run OVERLAY with the single raster image and each group member in turn. In all cases the number of output images is equal to the number of members in the DynaGroup file. Several options for naming the output files are described in the Help System. In addition, a group file of the output files is automatically constructed. In the model shown below, two DynaGroups are used as input to an overlay operation. The output will be the number of raster images contained in the input DynaGroups and a raster group file.



DynaLinks

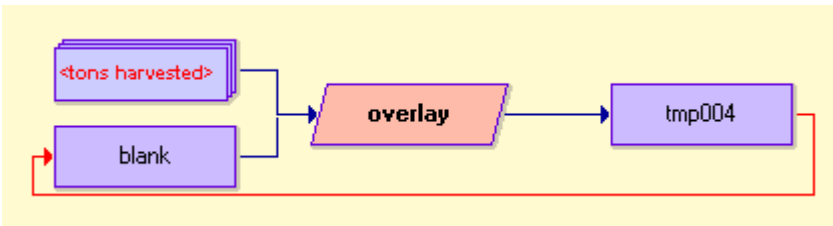
DynaLinks are used to run a model iteratively. An output data element is joined to an initial input data element with a DynaLink. This indicates that when the model performs the second iteration, the linked output filename should be substituted in place of the linked input filename. A DynaLink may only be connected to an initial data layer, i.e., one which has no connector joining to its left-hand side. Macro Modeler asks how many iterations are to be performed and whether each terminal output of the model or only the terminal output of the final iteration should be displayed. The final output of a model that includes a DynaLink is a single raster layer. The model shown below uses a DynaLink. When the model is run, the Macro Modeler asks for the desired number of iterations. Let us assume that in this case we want 5 iterations. The input file called Band3 is filtered to create an output image that is called filtered x5_01. On the second iteration, the file filtered x5_01 is filtered to produce the file filtered x5_02. This file is then substituted back as the input for the next filter operation and so forth. At the end of the 5th iteration, the final image is named filtered x5. Intermediate images (e.g., filtered x5_03) are not automatically deleted until the model is run again.



Note that the entire model is run during each iteration, even if the substitution occurs near the end of the model. To run just a portion of a model iteratively, use a submodel that contains a DynaLink (see below).

Using DynaGroups, DynaLinks and Submodels Together

When DynaGroups and DynaLinks are used in the same model, the number of members in the DynaGroup defines the number of iterations to be run. Each iteration of the model feeds in a new DynaGroup member. The output of a model that contains both elements is a single data layer. This figure shows a simple model in which a group of raster files is summed through the use of a DynaGroup and a DynaLink. In the first iteration, the initial layer BLANK (which is simply an image with the value 0 in every cell) is used with the first DynaGroup member in an Overlay Addition operation. This produces the temporary file TMP004. TMP004 is then substituted back into the place of BLANK for the second iteration. The number of iterations equals the number of images in the DynaGroup.



To iterate or loop only a portion of the model, create a submodel for the portion that should loop. When you run the model, a dialog box for each submodel will appear asking for the number of iterations for that submodel.

Image Calculator

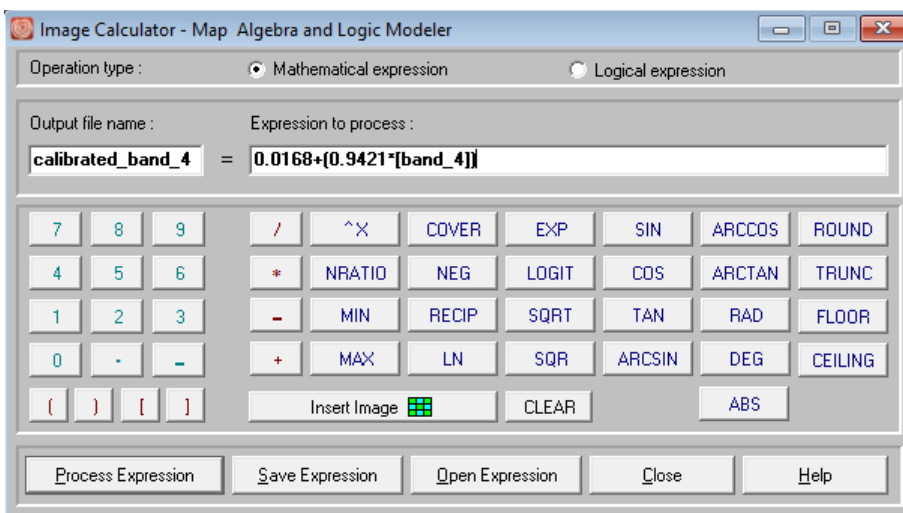


Image Calculator is a calculator tool for quickly evaluating equations using raster image layers. Equations can be saved and edited at a later time. It is extremely easy to use since it works exactly like a scientific calculator and offers the functionality typically

associated with calculators. The only special requirement is that all image files must be specified using square brackets as delimiters. For instance, in the example shown in the figure above, a calibrated image band is created by multiplying its values by a gain of 0.9421 and then adding an offset of 0.0168. It can be accessed through its icon on the tool bar, or from several menu entries: under the Mathematical Operators, Database Query, and Model Deployment Tools sections of the GIS Analysis menu.

Command Line Macros

Macro Modeler and Image Calculator provide very direct and simple interfaces for creating models. However, command line macro scripts are also supported, primarily to support legacy applications from early versions. A command line macro is an ASCII file containing the module names and parameters for the sequence of commands to be performed. It provides no automatic batch capability (though you may quickly copy, paste and edit to achieve this) nor looping.

Specific instructions on the macro command format for each module can be found in the Help System. These instructions can then be placed into an ASCII file with an ".iml" (an acronym for "IDRISI Macro Language") extension. Typically, this will be done with the Edit module. Choosing to save as type Macro file will automatically add the correct filename extension.

Note that some modules do not have a macro command version. These are typically modules that do not produce a resulting file (e.g., Terrset Explorer) or modules that require interaction from the user (e.g., Edit), or the vertical applications. In the menu, any module written in all upper-case letters may be used in a macro.

TerrSet records all the commands you execute in a text file located in the Working Folder. This file is called a LOG file. The commands are recorded in a similar format to the macro command format. It may sometimes be more efficient to edit a LOG file to have the macro format rather than typing in macro commands from scratch. To do so, open the LOG file in Edit (idrisi32.log is the most recent log file, followed by idrisi32.lo2, .lo3, .lo4 and .lo5) and alter it to have the macro file format. You will need to add an "x" between the module name and the rest of the parameters. Save it as a macro file.

Note also that whether from a dialog box or a macro, the command line used to generate each output image is recorded in that image's Lineage field in its documentation file. This may be viewed with the Metadata utility and may be copied and pasted into a macro file using the CTL+C keyboard sequence to copy highlighted text in Metadata and the CTL+V keyboard sequence to paste it into the macro file in Edit.

The Run Macro dialog box, under the GIS Analysis|Model Deployment Tools menu, will ask for the name of the macro to execute along with any command line parameters that should be passed to the macro file itself. This latter box can be left blank if there are none. Parameters would only be needed if the macro contained variables (e.g., %1, %2).

Each line in a macro is completed in sequence because the resulting image of one command is often used as input to a later command. In addition, if more than one macro is launched, the first macro launched will be completed before the second is begun.

Macro File Structure

IDRISI Macro files support the following syntax.

1. TerrSet modules in command line mode. The following is an example of a valid line in an IML file:

```
OVERLAY x 3 * soilsuit * slopsuit * suitland
```

The x after the module name indicates that command line mode is invoked. *All parameters after the x are separated by asterisks.* Short filenames may be given, as in the example above. In this case, the macro will look for the appropriate file

type first in the Working Folder, then in each Resource Folder in the order in which they are listed in the current project file. Long filenames, as well as full paths, may also be given in the command line, e.g.:

```
OVERLAY × 3 * c:\data\soil.rst * c:\data\slopes.rst * c:\output\suitable_land.rst
```

2. The macro processor supports command line variables using the %# format (familiar to DOS users). For example, the above line could be modified as such:

```
OVERLAY × 3 * %1 * %2 * %3
```

In this modification, %1, %2 and %3 are replaced with the first three command line parameters specified in the Macro Parameters text box of the Macro dialog box.

3. External Applications. Any non-TerrSet application can be called from an IML file. The syntax is as follows:

```
CALL [Application Name] [Command Line Parameters]
```

where [Application Name] should specify the full path to the application. For example, to call an application named SOILEROD.EXE in a folder named MODELS on drive C, and pass it the name of an input file named SOILTYPE.RST, the command would be:

```
CALL c:\models\soilerod.exe soiltype.rst
```

4. Other IML files. It is possible to execute other IML files from within a main IML file by using the BRANCH command. The syntax for such an operation is:

```
BRANCH [IML File Name] [Command Line Parameters]
```

The IML filename is the short name of the IML file to call (i.e., without extension). Command line parameters are optional.

The IDRISI API

TerrSet has been designed as an OLE Automation Server using COM Object technology (i.e., a COM Server). As a consequence, it is possible to use high-level development languages, such as Delphi, C++, Python, Visual Basic, as macro languages for controlling the operation of TerrSet. In addition, you can create sophisticated OLE Automation Controller applications that have complete control over the operation of TerrSet. Thus, you would use the API in instances where you wish to:

- create complex models that require the extensive control structures of a high-level computer language (such as many dynamic models);
- create custom applications, perhaps to be distributed to others;
- create custom interfaces.

The OLE Automation Server feature of TerrSet is automatically registered with Windows when TerrSet is first run on your system. Thus, if you have installed TerrSet and run it at least once, you automatically have access to the full API. The IDRISI OLE Automation server provides a wide range of functions for controlling TerrSet, including running modules, displaying files, and manipulating TerrSet environment variables. With visual programming environments such as Delphi, Visual C++ and Visual Basic, access to these functions is very easy. Specific instructions as well as a complete reference manual can be accessed from the Help menu (TerrSet COM API Documentation).

While the API provides a wide range of tools, most users only require a few - typically the Run Module operation that can run any module, the Display File operator (for construction of automatically displayed outputs), and a couple of procedures for accessing the project folder structure.

Instructions are also provided in the API help file (under the Help menu) for how to add your own modules to the TerrSet menu system so that you can develop applications and incorporate them as direct extensions to TerrSet itself.

CHAPTER FOUR

IDRISI GIS ANALYSIS

Of the eight major toolsets in the TerrSet System, none is more fundamental than the IDRISI GIS Analysis toolset. The modules found in this toolset provide the foundation for analyzing raster data. These modules are broken down into eight menu headings, highlighting the powerful functionality of each. These include:

- Database Query
- Mathematical Operators
- Distance Operators
- Context Operators
- Statistics
- Decision Support
- Change/Time Series
- Surface Analysis

The sections below highlight some of the basic GIS analysis tools in TerrSet in context of major GIS operations.

Database Query

Perhaps the most fundamental of analytical operations undertaken in GIS is simple database query, in which we ask questions of the database and examine the results as a map. With a spatial database, two types of questions may be posed—"What locations have this attribute?" and "What is the attribute at this location?" The first is known as *query by attribute*, while the second is called *query by location*.

Query by Attribute

Query by attribute may be performed several ways, depending on the geography of the layers. If you are working with a single geography (e.g. farm fields, provinces) defined by a vector file for which you have multiple attributes in a database, the database query may be accomplished entirely in Database Workshop using an SQL filter. The results may then be linked to a vector file for display or they may be assigned to a raster feature definition image for subsequent display.

For example, if you had a map of the countries of the world, and multiple attributes for each country stored in a database, then you could perform a query such as, "Find all the countries where the median per capita annual income is less than \$5000, but the literacy rate is higher than 60%." The query conditions could be used in an SQL filter in Database Workshop, and the result linked for display to the original vector feature definition file in DISPLAY Launcher. If a new raster image file should be made, the attributes may be assigned to a raster feature definition image from Database Workshop.

However, if the geographies of the attributes of interest are not the same, or if the attributes exist only as image layers, then two steps are involved. First, the features meeting the conditions specified are selected in each layer. This normally involves the use of RECLASS or ASSIGN. Then those selected data are used in an overlay operation, provided through the module OVERLAY, to find the locations that meet all the conditions. (Both steps may be carried out with a single command in Image Calculator, but behind the interface, the individual steps are still carried out in sequence.)

For example, you might ask, "Where are all the locations that have residential land use and are within a half mile of the primary path of planes taking off from the proposed airport?" In this case, the geography of land use and that of the flight paths for the airport are not the same. A Boolean image (zeros and ones only) would be made for each condition using RECLASS or ASSIGN, then these would be combined in an OVERLAY multiply operation. The resulting image would have the value of one only where both conditions are found:

Input Image 1	x	Input Image 2	= Output Image
0		0	0
1		0	0
0		1	0
1		1	1

Reclassification and overlay are fundamental to query by attribute in GIS. In TerrSet, RECLASS and ASSIGN are the tools used to perform database queries on single attributes, and may be used to produce Boolean images either directly or through the Image Calculator.

While RECLASS and ASSIGN may be used to produce similar results, there are several important differences between these modules. Even in cases where either may be used, generally one will be easier to use than the other. The choice will become more apparent as you become familiar with the characteristics of the two modules.

RECLASS works on an image file. Original values may be specified as individual values, or as ranges of values. This information is entered in the RECLASS dialog box. Any values left out of the specified reclassification ranges will remain unchanged. Also, there is an option in RECLASS, User-defined individual reclass, which operates similar to the ASSIGN module and eliminates the step of creating a values file. Also, like ASSIGN, any unspecified values become zero.

With ASSIGN, a feature definition image file and an attribute values file are required. The latter is commonly created with Edit or imported from a spreadsheet or statistical software package. The data values in the feature definition image must be byte or integer. However, the new value to be assigned may be byte, integer or real. Both old and new values must be specified as single numbers, not as ranges. The old and new values are entered in a values file, rather than in the ASSIGN dialog box. Any original values not specified in the values file will automatically be assigned the new value zero in the output image.

Whenever the query involves more than one attribute, it is necessary to use OVERLAY. (Again, the user may choose to use OVERLAY directly, or through Image Calculator.) For example, to find all agricultural land on soil type 6 requires that we first isolate soil type 6 as a Boolean image from the soils layer, and the agricultural land as a Boolean image from the land use layer. These two Boolean images are then overlaid, using the multiplication operation, to find all cases where it is soil type 6 AND agricultural.

Similarly, the maximum option of OVERLAY may be used to produce the Boolean OR result:


Input Image 1	MAX	Input Image 2	=	Output Image
0.00		0.00		0.00
1.00		0.00		1.00
0.00		1.00		1.00
1.00		1.00		1.00

All of the logical operations can be achieved similarly through simple operations on Boolean images. For example, the Boolean XOR (exclusive OR) operation can be performed with an addition operation, followed by a reclassification of all values not equal to 1 to 0. In developing models that require this kind of logic, it is often helpful to construct a table such as the ones above in order to determine the type of operations needed.

In Image Calculator, these analyses are built as logical expressions. While Image Calculator often provides a faster and easier interface, there are advantages, particularly to those new to GIS, to using the modules directly and performing each step individually. Doing so allows each step in the process to be evaluated so that errors in logic may be detected more easily. Using the modules individually also allows the user to become more familiar with their operation, facilitating the use of these modules outside the limits of database query.

Query by Location

Query by location is most easily accomplished in TerrSet with the Cursor Inquiry tool in the Display System. With your cursor placed on the location in question, click the left mouse button. The underlying data value for that location will be displayed on the screen.

Query by location can be extended to include query across multiple raster files by activating the Identify tool (click on the Identify icon  in the toolbar). A left-click on the map will bring up information about the pixel value at that location for all the layers in the map composition. Similarly, by linking a vector file to an associated database in Database Workshop, you can query by location by clicking on a feature to bring up all the linked database field values for the queried object.

Other tools for database query by location include PROFILE, QUERY and EXTRACT. Each of these gives results based on the attributes found at the location of input features.

Database Workshop

Database Workshop is TerrSet's relational database manager, and lies at the heart of TerrSet's support for layer collections that link vector feature definition files to database tables. TerrSet uses the Microsoft ADO and Access Jet Engine as the basis for Database Workshop. With this facility, one can undertake a wide variety of database operations. However, more importantly, one can interact directly with linked-table collections: database queries can be shown immediately on the associated map layer, and map layer queries can be directly linked to the data table. In addition, database field values can be assigned or extracted from raster layers. Each of these is discussed below.

Working with Linked-Table Collections

A linked-table collection consists of a vector feature definition file, a database table and a link file associating the two. The collection is defined in Database Workshop from the Establish Display Link menu entry under the Query menu.¹ The link file contains information about the vector file, database file, table of the database file (a database file may have several tables), and link field for the collection. In a linked-table collection each field (column) in the database, linked to the geographic definition of the features in the vector file, becomes a map layer. These can each be displayed using DISPLAY Launcher by selecting the layer of interest from below the collection filename, or by typing in the full "dot-logic" name of the layer. Database Workshop offers several additional ways to examine these data. Once a display link is made, one can either query the features in a linked map layer to highlight the records in the database or select a record in the database to highlight that feature in the vector map layer.

Launching Database Workshop

To launch Database Workshop, either click its icon on the toolbar or select its entry in the File|Data Entry or GIS Analysis|Database Query menus. When launched, if the selected layer of the map window with focus is from a linked-table collection, Database Workshop will automatically open that table. Otherwise, use the File/Open menu option on the Database Workshop menu to select the desired database file and table.

Displaying Layers from Database Workshop



Simply click the mouse into any record (row) of the field (column) you wish to view and then click the Database Workshop Display icon from the Database Workshop toolbar. The selected field will then be displayed using autoscaling and the default symbol file.

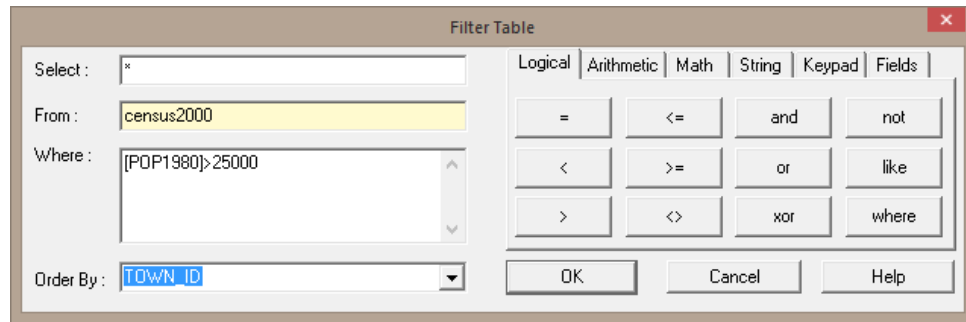


Database Query using an SQL Filter

Database query by attribute is accomplished in Database Workshop by *filtering* the database. This is simply the identification of which records have attributes that meet our query (i.e., filter) criteria. To query the active database table, click on the Filter Table icon or choose the Filter Table option from the Query menu. This opens the SQL Filter dialog which provides a simple interface to the construction of a Structured Query Language (SQL) statement.

The *Select* option at the top of the filter dialog specifies which fields to display in the result. The default asterisk indicates all fields and is fine in most instances. To specify a subset of fields, type their names into this input box separated by commas. Remember that all field names require square brackets around them if they contain spaces in the names. (To avoid ambiguity, it is a good habit to place square brackets around every field name.)

¹ For more about collections, see the chapter Map Layers, Raster Group Files, Vector Collections and Data Structures.



The *Where* input box is where the main part of the filter is constructed. The tabbed options to the right facilitate the placement of filter elements. Note also that any valid SQL clause can be typed in; you are not restricted to the options shown on the tabs.

The *Order By* input box is optional. It simply causes the results of the query to be sorted according to the field chosen.

Clicking OK causes the filter to be executed. Database Workshop will then show only the records that meet the filter criteria. The records that do not meet the criteria are still in the database, but they are hidden. (Remove the filter to restore the full database—see below.)

▪ **Mapping the Filtered Records**

When a filter is executed, Database Workshop checks all open map windows to see if any contain an active layer that is linked to the database that was filtered. If so, it will automatically display the results of the query as a Boolean map with features that meet the filter criteria shown in red and all others shown in black.

Removing the Filter

To remove any filter, choose the Remove Filter icon from the toolbar or choose the option from the Query menu.

Query by Location

When a map window contains an active layer (the one highlighted in Composer) linked to a database, if you click on a feature in the map display, Database Workshop will automatically locate the corresponding record in the database and highlight that value.

Other Database Operations

Calculating Field Values

In addition to querying the database, it is sometimes necessary to create new fields, either through importing external values or calculating new values from existing fields. For example, one might calculate a new field of population density values based on existing fields of population and area. The Calculate Field Values option of the Query menu (also accessed by clicking its icon on the Database Workshop toolbar) produces an SQL dialog area similar to that of the SQL Filter. In this case, it facilitates the construction of an SQL UPDATE SET operation to calculate new values for a field as a function of a mathematical or logical equation. In the SET input box, select the field to be calculated. Then enter the equation into the main input box after the "=" sign using the tabbed options as an aid. As with Filter, any valid SQL clause can be entered—you are not restricted to the options specified in the tabbed control.

Advanced SQL Queries across relational tables

An Advanced SQL editor is also available under the Query menu. It can be used to make more complicated SQL commands in the active table and across tables in the active database file. Queries can also be saved to a text file. Consult with an SQL text for advanced commands.

Finding Specific Records



The Find Next option of the Query Menu (also accessed by clicking the Find Next icon on the Database Workshop toolbar) provides a simple way to search for records. The "=" option looks for the next exact match while the "like" option looks for approximate matches.

Sorting



To sort the records of the database according to the values of a particular field, click the mouse into any record of that field then click either the ascending or descending sort button on the Database Workshop toolbar.

Entering or Modifying Data

You will specifically need to enter Edit Mode before any cell value in the database can be entered or modified. This guards against accidental changes to the data values. Enter edit mode by choosing the option from the Edit menu, or by clicking onto the Edit Mode status icon. The grid changes color when you are in edit mode. Several toolbar options are disabled until edit mode is turned off. You should therefore exit edit mode as soon as you have finished entering data. To do so, choose the option under the Edit menu, or click the Edit Mode status button.

Modifying the Table Structure

The table structure can be modified (i.e., add, rename or remove fields; add or delete records) from the Edit menu. However, this cannot be done if other "users" have access to the table. Any map windows that are linked to this database are considered to be users. Thus, you will need to close all of these map windows before the table structure can be altered.

Assigning Data to and Extracting Data from Raster Layers

Database Workshop provides a very simple means of assigning field data to a raster layer, or extracting data from a raster layer into a database field. To assign field data to a raster layer, use the Export|Field to Raster Image command from the File menu in Database Workshop or right-click in the desired field and select "to Raster Image". A link must be established, and row and column information need to be specified. By default, the X and Y coordinates in the new raster image will be taken from the linked vector file. To extract data from a raster image, use the Import/Raster Image command from the File menu in Database Workshop. A raster feature definition file will need to be specified representing the ID's in the raster feature definition image that was used in the extraction. These identifiers must match one field (the link field) in the database.

Assigning Data to and Importing Data from Vector Layers

Vector files can be either created or imported from the File menu in Database Workshop. A display link must first be established. Then, to import vector files, simply select the File|Import|Table|from Vector File command from the menu. A new table will be added

to the open database containing the values in the vector file. A new ID field will also be created. To export a field to a vector file, simply highlight a value in the field to export and select the File|Export|Field|to Vector File command from the menu (or right-click in the field and choose “to Vector File”. The vector features created will be based on the linked vector file specified in the vector collection file.

Export and Import

Database Workshop also provides selected import and export options under the File|Import|Table from External File menu. Default formats supported are xBase, Microsoft Excel, comma delimited (.csv), tab delimited text files (.txt), attribute values files (.avl).

Decision Support

With rapid increases in population and continuing expectations of growth in the standard of living, pressures on natural resource use have become intense. For the resource manager, the task of effective resource allocation has thus become especially difficult. Clear choices are few and the increasing use of more marginal lands puts one face-to-face with a broad range of uncertainties. Add to this a very dynamic environment subject to substantial and complex impacts from human intervention, and one has the ingredients for a decision making process that is dominated by uncertainty and consequent risk for the decision maker.

For many years, considerable interest has been focused on the use of GIS as a decision support system. For some, this role consists of simply informing the decision making process. However, it is more likely in the realm of resource allocation that the greatest contribution can be made.

The research staff at the Clark Labs has been specifically concerned with the use of GIS as a direct extension of the human decision making process—most particularly in the context of resource allocation decisions. However, our initial investigations into this area indicated that the tools available for this type of analysis were remarkably poor. Despite strong developments in the field of Decision Science, little of this had made a substantial impact on the development of software tools. And yet, at the same time, there was clear interest on the part of a growing contingency of researchers in the GIS field to incorporate some of these developments into the GIS arena. As a consequence, in the early 1990s, we embarked on a project, in conjunction with the United Nations Institute for Training and Research (UNITAR), to research the subject and to develop a suite of software tools for resource allocation. These were first released with Version 4.1 of the MS-DOS version of IDRISI, with a concentration on procedures for Multi-Criteria and Multi-Objective decision making—an area that can broadly be termed *Decision Strategy Analysis*. Since then, we have continued this development, most particularly in the area of *Uncertainty Management*.

Uncertainty is not simply a problem with data. Rather, it is an inherent characteristic of the decision making process itself. Given the increasing pressures that are being placed on the resource allocation process, we need to recognize uncertainty not as a flaw to be regretted and perhaps ignored, but as a fact of the decision making process that needs to be understood and accommodated. Uncertainty Management thus lies at the very heart of effective decision making and constitutes a very special role for the software systems that support GIS. The following discussion is thus presented in two parts: Decision Strategy Analysis and Uncertainty Management.

Introduction to Decision Support²

² The introductory material in this chapter is adapted from Eastman, J.R., 1993. Decision Theory and GIS, *Proceedings, Africa GIS '93*, UNITAR, Geneva. (out of print).

Decision Theory is concerned with the logic by which one arrives at a choice between alternatives. What those alternatives are varies from problem to problem. They might be alternative actions, alternative hypotheses about a phenomenon, alternative objects to include in a set, and so on. In the context of GIS, it is useful to distinguish between policy decisions and resource allocation decisions. The latter involves decisions that directly affect the utilization of resources (e.g., land) while the former is only intended to influence the decision behavior of others who will in turn make resource commitments. GIS has considerable potential in both arenas.

In the context of policy decisions, GIS is most commonly used to inform the decision maker. However, it also has potential (almost entirely unrealized at this time) as a process modeling tool, in which the spatial effects of predicted decision behavior might be simulated. Simulation modeling, particularly of the spatial nature of socio-economic issues and their relation to nature, it is to be expected that GIS will continue to play an increasingly sophisticated role.

Resource allocation decisions are also prime candidates for analysis with a GIS. Indeed, land evaluation and allocation is one of the most fundamental activities of resource development (FAO, 1976). With the advent of GIS, we now have the opportunity for a more explicitly reasoned land evaluation process. However, without procedures and tools for the development of decision rules and the predictive modeling of expected outcomes, this opportunity will largely go unrealized. GIS has been slow to address the needs of decision makers and to cope with the problems of uncertainty that lead to decision risk. In an attempt to address these issues, the Clark Labs has worked in close collaboration with the United Nations Institute for Training and Research (UNITAR) to develop a set of decision support tools for the TerrSet software system.

Although there is now fairly extensive literature on decision making in the Management Science, Operations Research and Regional Science fields (sometimes linked together under the single name *Decision Science*), there is unfortunately a broadly divergent use of terminology (e.g., see Rosenthal, 1985). Accordingly, we have adopted the following set of operational definitions which we feel are in keeping with the thrust of the Decision Science literature and which are expressive of the GIS decision making context.

Definitions

Decision

A decision is a choice between alternatives. The alternatives may represent different courses of action, different hypotheses about the character of a feature, different classifications, and so on. We call this set of alternatives the *decision frame*. Thus, for example, the decision frame for a zoning problem might be [commercial residential industrial]. The decision frame, however, should be distinguished from the individuals to which the decision is being applied. We call this the *candidate set*. For example, extending the zoning example above, the set of all locations (pixels) in the image that will be zoned is the candidate set. Finally, a *decision set* is that set of all individuals that are assigned a specific alternative from the decision frame. Thus, for example, all pixels assigned to the residential zone constitute one decision set. Similarly, those belonging to the commercial zone constitute another. Therefore, another definition of a decision would be to consider it the act of assigning an individual to a decision set. Alternatively, it can be thought of as a choice of alternative characterizations for an individual.

Criterion

A criterion is some basis for a decision that can be measured and evaluated. It is the evidence upon which an individual can be assigned to a decision set. Criteria can be of two kinds, factors and constraints, and can pertain either to attributes of the individual or to an entire decision set.

Factors

A factor is a criterion that enhances or detracts from the suitability of a specific alternative for the activity under consideration. It is therefore most commonly measured on a continuous scale. For example, a forestry company may determine that the steeper the slope, the more costly it is to transport wood. As a result, better areas for logging would be those on shallow slopes — the shallower the better. Factors are also known as *decision variables* in the mathematical programming literature (see Feiring, 1986) and *structural variables* in the linear goal programming literature (see Ignizio, 1985).

Constraints

A constraint serves to limit the alternatives under consideration. A good example of a constraint would be the exclusion from development of areas designated as wildlife reserves. Another might be the stipulation that no development may proceed on slopes exceeding a 30% gradient. In many cases, constraints will be expressed in the form of a Boolean (logical) map: areas excluded from consideration being coded with a 0 and those open for consideration being coded with a 1. However, in some instances, the constraint will be expressed as some characteristic that the decision set must possess. For example, we might require that the total area of lands selected for development be no less than 5000 hectares, or that the decision set consist of a single contiguous area. Constraints such as these are often called *goals* (Ignizio, 1985) or *targets* (Rosenthal, 1985). Regardless, both forms of constraints have the same ultimate meaning—to limit the alternatives under consideration.

Although factors and constraints are commonly viewed as very different forms of criteria, material will be presented later in this chapter which shows these commonly held perspectives simply to be special cases of a continuum of variation in the degree to which criteria tradeoff in their influence over the solution, and in the degree of conservativeness in risk (or alternatively, pessimism or optimism) that one wishes to introduce in the decision strategy chosen. Thus, the very hard constraints illustrated above will be seen to be the *crisp* extremes of a more general class of *fuzzy* criteria that encompasses all of these possibilities. Indeed, it will be shown that continuous criteria (which we typically think of as factors) can serve as *soft* constraints when tradeoff is eliminated. In ecosystems analysis and land suitability assessment, this kind of factor is called a *limiting factor*, which is clearly a kind of constraint.

Decision Rule

The procedure by which criteria are selected and combined to arrive at a particular evaluation, and by which evaluations are compared and acted upon, is known as a decision rule. A decision rule might be as simple as a threshold applied to a single criterion (such as, all regions with slopes less than 35% will be zoned as suitable for development) or it may be as complex as one involving the comparison of several multi-criteria evaluations.

Decision rules typically contain procedures for combining criteria into a single composite index and a statement of how alternatives are to be compared using this index. For example, we might define a composite suitability map for agriculture based on a weighted linear combination of information on soils, slope, and distance from market. The rule might further state that the best 5000 hectares are to be selected. This could be achieved by choosing that set of raster cells, totaling 5000 hectares, in which the sum of suitabilities is maximized. It could equally be achieved by rank ordering the cells and taking enough of the highest ranked cells to produce a total of 5000 hectares. The former might be called a *choice function* (known as an *objective function* or *performance index* in the mathematical programming literature—see Diamond and Wright, 1989) while the latter might be called a *choice heuristic*.

Choice Function

Choice functions provide a mathematical means of comparing alternatives. Since they involve some form of optimization (such as maximizing or minimizing some measurable characteristic), they theoretically require that each alternative be evaluated in turn. However, in some instances, techniques do exist to limit the evaluation only to likely alternatives. For example, the Simplex Method in linear programming (see Feiring, 1986) is specifically designed to avoid unnecessary evaluations.

Choice Heuristic

Choice heuristics specify a procedure to be followed rather than a function to be evaluated. In some cases, they will produce an identical result to a choice function (such as the ranking example above), while in other cases they may simply provide a close approximation. Choice heuristics are commonly used because they are often simpler to understand and easier to implement.

Objective

Decision rules are structured in the context of a specific objective. The nature of that objective, and how it is viewed by the decision makers (i.e., their motives) will serve as a strong guiding force in the development of a specific decision rule. An objective is thus a *perspective* that serves to guide the structuring of decision rules.³ For example, we may have the stated objective to determine areas suitable for timber harvesting. However, our perspective may be one that tries to minimize the impact of harvesting on recreational uses in the area. The choice of criteria to be used and the weights to be assigned to them would thus be quite different from that of a group whose primary concern was profit maximization. Objectives are thus very much concerned with issues of motive and social perspective.

Evaluation

The actual process of applying the decision rule is called evaluation.

Multi-Criteria Evaluations

To meet a specific objective, it is frequently the case that several criteria will need to be evaluated. Such a procedure is called *Multi-Criteria Evaluation* (Voogd, 1983; Carver, 1991). Another term that is sometimes encountered for this is *modeling*. However, this term is avoided here since the manner in which the criteria are combined is very much influenced by the objective of the decision.

Multi-criteria evaluation (MCE) is most commonly achieved by one of two procedures. The first involves *Boolean overlay* whereby all criteria are reduced to logical statements of suitability and then combined by means of one or more logical operators such as intersection (AND) and union (OR). The second is known as *Weighted Linear Combination* (WLC) wherein continuous criteria (*factors*) are standardized to a common numeric range, and then combined by means of a weighted average. The result is a continuous mapping of suitability that may then be masked by one or more Boolean *constraints* to accommodate qualitative criteria, and finally thresholded to yield a final decision.

While these two procedures are well established in GIS, they frequently lead to different results, as they make very different statements about how criteria should be evaluated. In the case of Boolean evaluation, a very extreme form of decision making is used. If the criteria are combined with a logical AND (the intersection operator), a location must meet *every* criterion for it to be included in the decision set. If even a single criterion fails to be met, the location will be excluded. Such a procedure is essentially risk-averse, and selects locations based on the most cautious strategy possible—a location succeeds in being chosen only if its worst quality (and therefore all qualities) passes the test. On the other hand, if a logical OR (union) is used, the opposite applies—a location will be included in the decision set even if only a single criterion passes the test. This is thus a gambling strategy, with (presumably) substantial risk involved.

³ It is important to note here that we are using a somewhat broader definition of the term objective than would be found in the goal programming literature (see Ignizio, 1985). In goal programming, the term *objective* is synonymous with the term *objective function* in mathematical programming and *choice function* used here.

Now compare these strategies with that represented by weighted linear combination (WLC). With WLC, criteria are permitted to tradeoff their qualities. A very poor quality can be compensated for by having several very favorable qualities. This operator represents neither an AND nor an OR—it lies somewhere in between these extremes. It is neither risk averse nor risk taking.

For reasons that have largely to do with the ease with which these approaches can be implemented, the Boolean strategy dominates vector approaches to MCE, while WLC dominates solutions in raster systems. But clearly neither is better—they simply represent two very different outlooks on the decision process—what can be called a decision strategy. TerrSet also includes a third option for multi-criteria evaluation, known as an Ordered Weighted Average (OWA) (Eastman and Jiang, 1996). This method offers a complete spectrum of decision strategies along the primary dimensions of degree of tradeoff involved and degree of risk in the solution.

Multi-Objective Evaluations

While many decisions we make are prompted by a single objective, it also happens that we need to make decisions that satisfy several objectives. A multi-objective problem is encountered whenever we have two candidate sets (i.e., sets of entities) that share members. These objectives may be complementary or conflicting in nature (Carver, 1991: 322).

▪ *Complementary Objectives*

With complementary or non-conflicting objectives, land areas may satisfy more than one objective, i.e., an individual pixel can belong to more than one decision set. Desirable areas will thus be those which serve these objectives together in some specified manner. For example, we might wish to allocate a certain amount of land for combined recreation and wildlife preservation uses. Optimal areas would thus be those that satisfy both of these objectives to the maximum degree possible.

▪ *Conflicting Objectives*

With conflicting objectives, competition occurs for the available land since it can be used for one objective or the other, but not both. For example, we may need to resolve the problem of allocating land for timber harvesting and wildlife preservation. Clearly the two cannot coexist. Exactly how they compete, and on what basis one will win out over the other, will depend upon the nature of the decision rule that is developed.

In cases of complementary objectives, multi-objective decisions can often be solved through a *hierarchical* extension of the multi-criteria evaluation process. For example, we might assign a weight to each of the objectives and use these, along with the suitability maps developed for each, to combine them into a single suitability map. This would indicate the degree to which areas meet all of the objectives considered (see Voogd, 1983). However, with conflicting objectives the procedure is more involved.

With conflicting objectives, it is sometimes possible to rank order the objectives and reach a *prioritized* solution (Rosenthal, 1985). In these cases, the needs of higher ranked objectives are satisfied before those of lower ranked objectives are dealt with. However, this is often not possible, and the most common solution for conflicting objectives is the development of a *compromise* solution. Undoubtedly the most commonly employed techniques for resolving conflicting objectives are those involving optimization of a choice function such as mathematical programming (Fiering, 1986) or goal programming (Ignizio, 1985). In both, the concern is to develop an allocation of the land that maximizes or minimizes an objective function subject to a series of constraints.

Uncertainty and Risk

Clearly, information is vital to the process of decision making. However, we rarely have perfect information. This leads to uncertainty, of which two sources can be identified: *database* and *decision rule uncertainty*.

Database Uncertainty

Database uncertainty is that which resides in our assessments of the criteria which are enumerated in the decision rule. Measurement error is the primary source of such uncertainty. For example, a slope of 35% may represent an important threshold. However, because of the manner in which slopes are determined, there may be some uncertainty about whether a slope that was measured as 34% really *is* 34%. While we may have considerable confidence that it is most likely *around* 34%, we may also need to admit that there is some finite probability that it is as high as 36%. Our expression of database uncertainty is likely to rely upon probability theory.

Decision Rule Uncertainty

Decision rule uncertainty is that which arises from the manner in which criteria are combined and evaluated to reach a decision. A very simple form of decision rule uncertainty is that which relates to parameters or thresholds used in the decision rule. A more complex issue is that which relates to the very structure of the decision rule itself. This is sometimes called *specification error* (Alonso, 1968), because of uncertainties that arise in specifying the relationship between criteria (as a model) such that adequate evidence is available for the proper evaluation of the hypotheses under investigation.

▪ **Decision Rule Uncertainty and Direct Evidence: Fuzzy versus Crisp Sets**

A key issue in decision rule uncertainty is that of establishing the relationship between the evidence and the decision set. In most cases, we are able to establish a direct relationship between the two, in the sense that we can define the decision set by measurable attributes that its members should possess. In some cases these attributes are crisp and unambiguous. For example, we might define those sewer lines in need of replacement as those of a particular material and age. However, quite frequently the attributes they possess are fuzzy rather than crisp. For example, we might define suitable areas for timber logging as those forested areas that have gentle slopes and are near to a road. What is a gentle slope? If we specify that a slope is gentle if it has a gradient of less than 5%, does this mean that a slope of 5.0001% is not gentle? Clearly there is no sharp boundary here. Such classes are called *fuzzy sets* (Zadeh, 1965) and are typically defined by a set membership function. Thus we might decide that any slope less than 2% is unquestionably gentle, and that any slope greater than 10% is unquestionably steep, but that membership in the gentle set gradually falls from 1.0 at a 2% gradient to 0.0 at a 10% gradient. A slope of 5% might then be considered to have a membership value of only 0.7 in the set called "gentle." A similar group of considerations also surround the concept of being "near" to a road.

Fuzzy sets are extremely common in the decision problems faced with GIS. They represent a form of uncertainty, but it is not measurement uncertainty. The issue of what constitutes a shallow slope is over and above the issue of whether a measured slope is actually what is recorded. It is a form of uncertainty that lies at the very heart of the concept of factors previously developed. *The continuous factors of multi-criteria decision making are thus fuzzy set membership functions, whereas Boolean constraints are crisp set membership functions.* But it should be recognized that the terms factor and constraint imply more than fuzzy or crisp membership functions. Rather, these terms give some meaning also to the manner in which they are aggregated with other information.

▪ **Decision Rule Uncertainty and Indirect Evidence: Bayes versus Dempster Shafer**

Not all evidence can be directly related to the decision set. In some instances we only have an indirect relationship between the two. In this case, we may set up what can be called a *belief function* of the degree to which evidence implies the membership in the decision set. Two important tools for accomplishing this are Bayesian Probability Theory and Dempster-Shafer Theory of Evidence. These will be dealt with at more length later in this chapter in Part B on Uncertainty Management.

Decision Risk

Decision Risk may be understood as the likelihood that the decision made will be wrong.⁴ Risk arises as a result of uncertainty, and its assessment thus requires a combination of uncertainty estimates from the various sources involved (database and decision rule uncertainty) and procedures, such as Bayesian Probability theory, through which it can be determined. Again, this topic will be discussed more thoroughly in Part B of this chapter.

A Typology of Decisions

Given these definitions, it is possible to set out a very broad typology of decisions as illustrated in the figure below.

	Single Criterion	Multi-Criteria
Single Objective		
Multi-Objective		

Decisions may be characterized as *single-* or *multi-objective* in nature, based on either a *single criterion* or *multiple criteria*. While one is occasionally concerned with single criterion problems, most problems approached with a GIS are multi-criteria in nature. For example, we might wish to identify areas of concern for soil erosion based on slope, landuse, soil type and the like. In these instances, our concern lies with how to combine these criteria to arrive at a composite decision. Therefore, the first major area of concern in GIS with regard to Decision Theory is Multi-Criteria Evaluation.

Most commonly, we deal with decision problems of this nature from a single perspective. However, in many instances, the problem is multi-objective in nature (Diamond and Wright, 1988). Multi-objective problems arise whenever the same resources belong to more than one candidate set. Thus, for example, a paper company might include all forest areas in its candidate set for consideration of logging areas, while a conservation group may include forest areas in a larger candidate set of natural areas to be protected. Any attempt, therefore, to reconcile their potential claims to this common set of resources presents a multi-objective decision problem.

Despite the prevalence of multi-objective problems, current GIS software is severely lacking in techniques to deal with this kind of decision. To date, most examples of multi-objective decision procedures in the literature have dealt with the problem using linear programming optimization (e.g., Janssen and Rietveld 1990; Carver, 1991; Campbell et. al., 1992; Wright et. al., 1983). However, in most cases, these have been treated as choice problems between a limited number (e.g., less than 20) of candidate sites previously isolated in a vector system. The volume of data associated with raster applications (where each pixel is a choice alternative) clearly overwhelms the computational capabilities of today's computing environment. In addition, the terminology and procedures of linear programming are unknown to most decision makers and are complex and unintuitive by nature. As a consequence, the second major area of Decision Theory of importance to GIS is Multi-Objective Land Allocation. Here, the focus will be on a simple decision heuristic appropriate to the special needs of raster GIS.

⁴ Note that different fields of science define risk in different ways. For example, some disciplines modify the definition given here to include a measure of the cost or consequences of a wrong decision (thus allowing for a direct relationship to cost/benefit analysis). The procedures developed in TerrSet do not preclude such an extension. We have tried here to present a fairly simple perspective that can be used as a building block for more specific interpretations.

Multi-Criteria Decision Making in GIS

As indicated earlier, the primary issue in multi-criteria evaluation is concerned with how to combine the information from several criteria to form a single index of evaluation. In the case of Boolean criteria (constraints), the solution usually lies in the union (logical OR) or intersection (logical AND) of conditions. However, for continuous factors, a weighted linear combination (Voogd, 1983: 120) is most commonly used. With a weighted linear combination, factors are combined by applying a weight to each followed by a summation of the results to yield a suitability map, i.e.:

$$S = \sum w_i x_i \quad \text{where} \quad S = \text{suitability}$$

w_i = weight of factor i

x_i = criterion score of factor i

This procedure is not unfamiliar in GIS and has a form very similar to the nature of a regression equation. In cases where Boolean constraints also apply, the procedure can be modified by multiplying the suitability calculated from the factors by the product of the constraints, i.e.:

$$S = \sum w_i x_i \times \Pi c_j \quad \text{where} \quad c_j = \text{criterion score of constraint } j$$

Π = product

All GIS software systems provide the basic tools for evaluating such a model. In addition, in TerrSet, a special module named MCE has been developed to facilitate this process. However, the MCE module also offers a special procedure called an Ordered Weighted Average that greatly extends the decision strategy options available. The procedure will be discussed more fully in the section on Evaluation below. For now, however, the primary issues relate to the standardization of criterion scores and the development of the weights.

Criterion Scores

Because of the different scales upon which criteria are measured, it is necessary that factors be standardized⁵ before combination using the formulas above, and that they be transformed, if necessary, such that all factors maps are positively correlated with suitability.⁶ Voogd (1983: 77-84) reviews a variety of procedures for standardization, typically using the minimum and maximum values as scaling points. The simplest is a linear scaling such as:

⁵ In using the term *standardization*, we have adopted the terminology of Voogd (1983), even though this process should more properly be called *normalization*.

⁶ Thus, for example, if locations near to a road were more advantageous for industrial siting than those far away, a distance map would need to be transformed into one expressing proximity.

$$x_i = \frac{(R_i - R_{\min})}{(R_{\max} - R_{\min})} \times \text{standardized_range} \quad \text{where } R = \text{raw score}$$

However, if we recognize that continuous factors are really fuzzy sets, we easily recognize this as just one of many possible set membership functions. In TerrSet, the module named FUZZY is provided for the standardization of factors using a whole range of fuzzy set membership functions. The module is quick and easy to use and provides the option of standardizing factors to either a 0-1 real number scale or a 0-255 byte scale. This former option is generally used and will be assumed in the discussion that follows. Importantly, the higher value of the standardized scale must represent the case of being *more* likely to belong to the decision set.

A critical issue in the standardization of factors is the choice of the end points at which set membership reaches either 0.0 or 1.0. Our own research has suggested that blindly using a linear scaling (or indeed any other scaling) between the minimum and maximum values of the image is ill-advised. In setting these critical points for the set membership function, it is important to consider their inherent meaning. Thus, for example, if we feel that industrial development should be placed as far away from a nature reserve as possible, it would be dangerous to implement this without careful consideration. Taken literally, if the map were to cover a range of perhaps 100 km from the reserve, then the farthest point away from the reserve would be given a value of 1.0. Using a linear function, then, a location 5 km from the reserve would have a standardized value of only 0.05. And yet it may be that the primary issue was noise and minor disturbance from local citizens, for which a distance of only 5 kilometers would have been equally as good as being 100 km away. Thus the standardized score should really have been 1.0. If an MCE were undertaken using the blind linear scaling, locations in the range of a few 10s of km would have been severely devalued when in fact they might have been quite good. In this case, the recommended critical points for the scaling should have been 0 and 5 km. In developing standardized factors using FUZZY, then, careful consideration should be given to the inherent meaning of the end points chosen.

Criterion Weights

A wide variety of techniques exist for the development of weights. In very simple cases, assigning criteria weights may be accomplished by dividing 1.0 among the criteria. (It is sometimes useful for people to think about "spending" one dollar, for example, among the criteria). However, when the number of criteria is more than a few, and the considerations are many, it becomes quite difficult to make weight evaluations on the set as a whole. Breaking the information down into simple pairwise comparisons in which only two criteria need be considered at a time can greatly facilitate the weighting process, and will likely produce a more robust set of criteria weights. A pairwise comparison method has the added advantages of providing an organized structure for group discussions, and helping the decision making group hone in on areas of agreement and disagreement in setting criterion weights.

The technique described here and implemented in TerrSet is that of pairwise comparisons developed by Saaty (1977) in the context of a decision making process known as the Analytical Hierarchy Process (AHP). The first introduction of this technique to a GIS application was that of Rao et. al. (1991), although the procedure was developed outside the GIS software using a variety of analytical resources.

In the procedure for Multi-Criteria Evaluation using a weighted linear combination outlined above, it is necessary that the weights sum to one. In Saaty's technique, weights of this nature can be derived by taking the principal eigenvector of a square reciprocal matrix of pairwise comparisons between the criteria. The comparisons concern the relative importance of the two criteria involved in determining suitability for the stated objective. Ratings are provided on a 9-point continuous scale. For example, if one felt that proximity to roads was very strongly more important than slope gradient in determining suitability for industrial siting, one would enter a 7 on this scale. If the inverse were the case (slope gradient was very strongly more important than proximity to roads), one would enter 1/7.

1/9	1/7	1/5	1/3	1	3	5	7	9
extremely	very strongly	strongly	moderately	equally	moderately	strongly	very strongly	extremely
<i>less important</i>					<i>more important</i>			

The Continuous Rating Scale

In developing the weights, an individual or group compares every possible pairing and enters the ratings into a pairwise comparison matrix. Since the matrix is symmetrical, only the lower triangular half actually needs to be filled in. The remaining cells are then simply the reciprocals of the lower triangular half (for example, since the rating of slope gradient relative to town proximity is 4, the rating of town proximity relative to slope gradient will be 1/4). Note that where empirical evidence exists about the relative efficacy of a pair of factors, this evidence can also be used.

	Road Proximity	Town Proximity	Slope Gradient	Small Holder Settlement	Distance from Park
Road Proximity	1				
Town Proximity	1/3	1			
Slope Gradient	1	4	1		
Small Holder Set.	1/7	2	1/7	1	
Distance from Park	1/2	2	1/2	4	1

An example of a pairwise comparison matrix for assessing the comparative importance of five factors to industrial development suitability.

The procedure then requires that the principal eigenvector of the pairwise comparison matrix be computed to produce a *best fit* set of weights (table below). If no procedure is available to do this, a good approximation to this result can be achieved by calculating the weights with each column and then averaging over all columns. For example, if we take the first column of figures, they sum to 2.98. Dividing each of the entries in the first column by 2.98 yields weights of 0.34, 0.11, 0.34, 0.05, and 0.17 (compare to the values in the table below). Repeating this for each column and averaging the weights over the columns usually gives a good approximation to the values calculated by the principal eigenvector. In the case of TerrSet, however, a special module named WEIGHT has been developed

to calculate the principal eigenvector directly. Note that these weights will sum to one, as is required by the weighted linear combination procedure.

RoadProx	0.33
TownProx	0.08
Slope	0.34
SmallHold	0.07
ParkDist	0.18

Weights derived by calculating the principal eigenvector of the pairwise comparison matrix.

Since the complete pairwise comparison matrix contains multiple paths by which the relative importance of criteria can be assessed, it is also possible to determine the degree of consistency that has been used in developing the ratings. Saaty (1977) indicates the procedure by which an index of consistency, known as a *consistency ratio*, can be produced. The consistency ratio (CR) indicates the probability that the matrix ratings were randomly generated. Saaty indicates that matrices with CR ratings greater than 0.10 should be re-evaluated. In addition to the overall consistency ratio, it is also possible to analyze the matrix to determine where the inconsistencies arise. This has also been developed as part of the WEIGHT module in TerrSet.

Evaluation

Once the criteria maps (factors and constraints) have been developed, an evaluation (or aggregation) stage is undertaken to combine the information from the various factors and constraints. The MCE module offers three logics for the evaluation/aggregation of multiple criteria: Boolean intersection, weighted linear combination (WLC), and the ordered weighted average (OWA).

MCE and Boolean Intersection

The most simplistic type of aggregation is the Boolean intersection or logical AND. This method is used only when factor maps have been strictly classified into Boolean suitable/unsuitable images with values 1 and 0. The evaluation is simply the multiplication of all the images.

MCE and Weighted Linear Combination

The derivation of criterion (or factor) weights is described above. The weighted linear combination (WLC) aggregation method multiplies each standardized factor map (i.e., each raster cell within each map) by its factor weight and then sums the results. Since the set of factor weights for an evaluation must sum to one, the resulting suitability map will have the same range of values as the standardized factor maps that were used. This result is then multiplied by each of the constraints in turn to "mask out" unsuitable areas. All these steps could be done using either a combination of SCALAR and OVERLAY, or by using the Image Calculator. However, the module MCE is designed to facilitate the process.

The WLC option in the MCE module requires that you specify the number of criteria (both constraints and factors), their names, and the weights to be applied to the factors. All factors must be standardized to a byte (0-255) range. (If you have factors in real format, then use one of the options other than MCE mentioned above.) The output is a *suitability* map masked by the specified constraints.

MCE and the Ordered Weighted Average

In its use and implementation, the ordered weighted average approach is not unlike WLC. The dialog box for the OWA option is almost identical to that of WLC, with the exception that a second set of weights appears. This second set of weights, the *order weights*, controls the manner in which the weighted factors are aggregated (Eastman and Jiang, 1996; Yager, 1988). Indeed, WLC turns out to be just one variant of the OWA technique. To introduce the OWA technique, let's first review WLC in terms of two new concepts: tradeoff and risk.

▪ Tradeoff

Factor weights are weights that apply to specific factors, i.e., all the pixels of a particular factor image receive the same factor weight. They indicate the relative degree of importance each factor plays in determining the suitability for an objective. In the case of WLC the weight given to each factor also determines how it will tradeoff relative to other factors. For example, a factor with a high factor weight can tradeoff or compensate for poor scores on other factors, even if the unweighted suitability score for that highly-weighted factor is not particularly good. In contrast, a factor with a high suitability score but a small factor weight can only weakly compensate for poor scores on other factors. The factor weights determine how factors tradeoff but, as described below, order weights determine the *overall* level of tradeoff allowed.

▪ Risk

Boolean approaches are extreme functions that result either in very risk-averse solutions when the AND operator is used or in risk-taking solutions when the OR operator is used.⁷ In the former, a high aggregate suitability score for a given location (pixel) is only possible if *all* factors have high scores. In the latter, a high score in any factor will yield a high aggregate score, even if all the other factors have very low scores. The AND operation may be usefully described as the *minimum*, since the minimum score for any pixel determines the final aggregate score. Similarly, the OR operation may be called the *maximum*, since the maximum score for any pixel determines the final aggregate score. The AND solution is *risk-averse* because we can be sure that the score for every factor is at least as good as the final aggregate score. The OR solution is *risk-taking* because the final aggregate score only tells us about the suitability score for the single most suitable factor.

The WLC approach is an averaging technique that softens the hard decisions of the Boolean approach and avoids the extremes. In fact, given a continuum of risk from minimum to maximum, WLC falls exactly in the middle; it is neither risk-averse nor risk-taking.

▪ Order Weights, Tradeoff and Risk

The use of order weights allows for aggregation solutions that fall anywhere along the risk continuum between AND and OR. Order weights are quite different from factor weights. They do not apply to any specific factor. Rather, they are applied on a pixel-by-pixel basis to factor scores as determined by their rank ordering across factors at each location (pixel). Order weight 1 is assigned to the lowest-ranked factor for that pixel (i.e., the factor with the lowest score), order weight 2 to the next higher-ranked factor for that pixel, and so forth. Thus, it is possible that a single order weight could be applied to pixels from any of the various factors depending upon their relative rank order.

To examine how order weights alter MCE results by controlling levels of tradeoff and risk, let us consider the case where factor weights are equal for three factors A, B, and C. (Holding factor weights equal will make clearer the effect of the order weights.) Consider a single pixel with factor scores A= 187, B=174, and C=201. The factor weights for each of the factors is 0.33. When ranked from minimum value to maximum value, the order of these factors for this pixel is [B,A,C]. For this pixel, factor B will be assigned order weight 1, A order weight 2 and C order weight 3.

⁷ The logic of the Boolean AND and OR is implemented with fuzzy sets as the minimum and maximum. Thus, as we are considering continuous factor scores rather than Boolean 0-1 images in this discussion, the logical AND is evaluated as the minimum value for a pixel across all factors and the logical OR is evaluated as the maximum value for a pixel across all factors.

Below is a table with thirteen sets of order weights that have been applied to this set of factor scores [174,187,201]. Each set yields a different MCE result even though the factor scores and the factor weights are the same in each case.

Order Weights			
Min (1)	(2)	Max (3)	Result
1.00	0.00	0.00	174
0.90	0.10	0.00	175
0.80	0.20	0.00	177
0.70	0.20	0.10	179
0.50	0.30	0.20	183
0.40	0.30	0.30	186
0.33	0.33	0.33	187
0.30	0.30	0.40	189
0.20	0.30	0.50	191
0.10	0.20	0.70	196
0.00	0.20	0.80	198
0.00	0.10	0.90	200
0.00	0.00	1.00	201

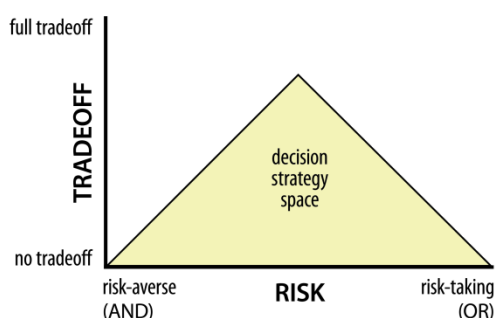
The first set of order weights in the table is [1, 0, 0]. The weight of factor B (the factor with the minimum value in the set [B, A, C]) will receive all possible weight while factors A and C will be given no weight at all. Such a set of order weights make irrelevant the factor weights. Indeed, the order weights have altered the evaluation such that no tradeoff is possible. As can be seen in the table, this has the effect of applying a minimum operator to the factors, thus producing the traditional intersection operator (AND) of fuzzy sets.

Similarly, the last set of order weights $[0, 0, 1]$ has the effect of a maximum operator, the traditional union operator (OR) of fuzzy sets. Again, there is no tradeoff and the factor weights are not employed.

Another important example from the table is where the order weights are equal, $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$. Here all ranked positions get the same weight; this makes tradeoff fully possible and locates the analysis exactly midway between AND and OR. Equal order weights produce the same result as WLC.

In all three cases, the order weights have determined not only the level of tradeoff but have situated the analysis on a continuum from (risk-averse, minimum, AND) to (risk-taking, maximum, OR).

As seen in the table, the order weights in the OWA option of MCE are not restricted to these three possibilities, but instead can be assigned any combination of values that sum to 1.0. Any assignment of order weights results in a decision rule that falls somewhere in a triangular decision strategy space that is defined by the dimensions of risk and tradeoff as shown in the figure below.



Whether most of the order weight is assigned to the left, right or center of the order weights determines the position in the risk dimension. The logical AND operator is the most risk-averse combination and the logical OR is the most risk-taking combination. When order weights are predominantly assigned to the lower-ranked factors, there is greater risk aversion (more of an AND approach). When order weights are more dominant for the higher-ranked factors, there is greater risk taking (more of an OR approach). As discussed above, equal order weights yield a solution at the middle of the risk axis.

The degree of tradeoff is governed by the relative distribution of order weights between the ranked factors. Thus, if the sum of the order weights is evenly spread between the factors, there is strong tradeoff, whereas if all the weight is assigned to a single factor rank, there is no tradeoff. (It may be

helpful to think of this in terms of a graph of the order weights, with rank order on the X axis and the order weight value on the Y axis. If the graph has a sharp peak, there is little tradeoff. If the graph is relatively flat, there is strong tradeoff.)

Thus, as seen from the table, the order weights of $[0.5, 0.3, 0.2]$ would indicate a strong (but not perfect) degree of risk aversion (because weights are skewed to the risk-averse side of the risk axis) and some degree of tradeoff (because the weights are spread out over all three ranks). Weights of $[0, 1, 0]$, however, would imply neither risk aversion nor acceptance (exactly in the middle of the risk axis), and no tradeoff (because all the weight is assigned to a single rank).

The OWA method is particularly interesting because it provides this continuum of aggregation procedures. At one extreme (the logical AND), each criterion is considered necessary (but not sufficient on its own) for inclusion in the decision set. At the other extreme (the logical OR), each criterion is sufficient on its own to support inclusion in the decision set without modification by other factors. The position of the weighted linear combination operator halfway between these extremes is therefore not surprising. This operator considers criteria as neither necessary nor sufficient—strong support for inclusion in the decision set by one criterion can be equally balanced by correspondingly low support by another. It thus offers full tradeoff.

Using OWA

Given this introduction, it is worth considering how one would use the OWA option of MCE. Some guidelines are as follows:

1. Divide your criteria into three groups: hard constraints, factors that should not tradeoff, and factors that should tradeoff. For example, factors with monetary implications typically tradeoff, while those associated with some safety concern typically do not.

2. If you find that you have factors that both tradeoff and do not tradeoff, separate their consideration into two stages of analysis. In the first, aggregate the factors that tradeoff using the OWA option. You can govern the degree of tradeoff by manipulating the order weights. Then use the result of the first stage as a new factor that is included in the analysis of those that do not tradeoff.
3. If you run an analysis with absolutely no tradeoff, the factor weights have no real meaning and can be set to any value.

Completing the Evaluation

Once a suitability map has been prepared, it is common to decide, as a final step, which cells should belong to the set that meets a particular land allocation area target (the decision set). For example, having developed a map of suitability for industrial development, we may then wish to determine which areas constitute the *best* 5000 hectares that may be allocated. Oddly, this is an area where most raster systems have difficulty achieving an exact solution. One solution would be to use a choice function where that set of cells is chosen which maximizes the sum of suitabilities. However, the number of combinations that would need to be evaluated is prohibitive in a raster GIS. As a result, we chose to use a simple choice heuristic—to select the top-ranked cells that are required to meet the area target. In TerrSet, a module named TOPRANK is available that allows the selection of top-ranked cells. The user can specify the percentage of top-ranked cells to extract or can extract the best cells up to a specified cost (as defined by a separate price image). However, an even more flexible option is to use the MOLA module.

The name MOLA derives from its primary use in solving Multi-Objective Land Allocation problems (as will be described below). However, it can also be used to solve single objective allocations. Here you can choose to select the best lands up to a specified area target or the best within a specified budget. You can also choose whether the cells should be contiguous or not. If you choose contiguous, you can also indicate whether they should all be in a single contiguous cluster or several separate but internally contiguous clusters. You can also indicate the degree of compactness associated with contiguous clusters by specifying the minimum span that must occur either in the horizontal or vertical directions.

Multi-Objective Decision Making in GIS

Multi-objective decisions are so common in environmental management that it is surprising that specific tools to address them have not yet been further developed within GIS. The few examples one finds in the literature tend to concentrate on the use of mathematical programming tools outside the GIS, or are restricted to cases of complementary objectives.

Complementary Objectives

As indicated earlier, the case of complementary objectives can be dealt with quite simply by means of a hierarchical extension of the multi-criteria evaluation process (e.g., Carver, 1991). Here a set of suitability maps, each derived in the context of a specific objective, serve as the factors for a new evaluation in which the objectives are themselves weighted and combined by linear summation. Since the logic which underlies this is multiple use, it also makes sense to multiply the result by all constraints associated with the component objectives.

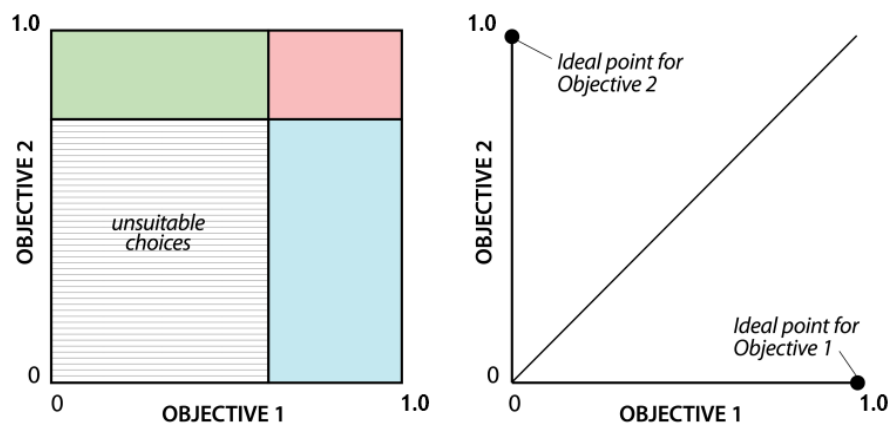
Conflicting Objectives

With conflicting objectives, land can be allocated to one objective but not more than one (although hybrid models might combine complementary and conflicting objectives). As was indicated earlier, one possible solution lies with a prioritization of objectives (Rosenthal, 1985). After the objectives have been ordered according to priority, the needs of higher priority objectives are satisfied (through the selection of top-ranked cells to meet areal goals) before those of lower priority ones. This is done by successively satisfying the needs of higher priority objectives and then removing (as a new constraint) areas taken by that objective from

consideration by all remaining objectives. A prioritized solution is easily achieved with the use of the TOPRANK module. However, instances are rare where a *prioritized* solution makes sense. More often a *compromise* solution is required.

As noted earlier, compromise solutions to the multi-objective problem have most commonly been approached through the use of mathematical programming tools outside GIS (e.g., Diamond and Wright, 1988; Janssen and Rietveld, 1990; Campbell, et. al., 1992). Mathematical programming solutions (such as linear or integer programming) can work quite well in instances where only a small number of alternatives are being addressed. However, in the case of raster GIS, the massive data sets involved will typically exceed present-day computing power. In addition, the concepts and methodology of linear and integer programming are not particularly approachable to a broad range of decision makers. As a result, we have sought a solution to the problem of multi-objective land allocation under conditions of conflicting objectives such that large raster datasets may be handled using procedures that have an immediate intuitive appeal.

The procedure we have developed is an extension of the decision heuristic used for the allocation of land with single objective problems. This is best illustrated by the diagram in the figure below. Each of the suitability maps may be thought of as an axis in a multi-dimensional space. Here we consider only two objectives for purposes of simple explanation. However, any number of objectives can be used.



Every raster cell in the image can be located within this decision space according to its suitability level on each of the objectives. To find the best x hectares of land for Objective 1, we simply need to move a decision line down from the top (i.e., far right) of the Objective 1 suitability axis until enough of the best raster cells are captured to meet our area target. We can do the same with the Objective 2 suitability axis to capture the best y hectares of land for it. As can be seen in the figure above, this partitions the decision space into four regions—areas best for Objective 1 (blue) and not suitable for Objective 2, areas best for Objective 2 (green) and not suitable for Objective 1, areas not suitable for either, and areas judged best for both (red). The latter represents areas of conflict.

To resolve these areas of conflict, a simple partitioning of the affected cells is used. As can be seen in the figure above, the decision space can also be partitioned into two further regions: those closer to the *ideal point* for Objective 1 and those closer to that for Objective 2. The ideal point represents the best possible case—a cell that is maximally suited for one objective and minimally suited for anything else. To resolve the conflict zone, the line that divides these two regions is overlaid onto it and cells are then allocated to their closest ideal point. Since the conflict region will be divided between the objectives, both objectives will be short on achieving their area goals. As a result, the process will be repeated with the decision lines being lowered for both objectives to gain more territory. The process of resolving conflicts and lowering the decision lines is iteratively repeated until the exact area targets are achieved.

It should be noted that a 45-degree line between a pair of objectives assumes that they are given equal weight in the resolution of conflicts. However, unequal weighting can be given. Unequal weighting has the effect of changing the angle of this dividing line. In fact, the tangent of that angle is equal to the ratio of the weights assigned to those objectives.

As a result of the above considerations, the module named MOLA (Multi-Objective Land Allocation) was developed to undertake the compromise solution to the multi-objective problem. As indicated earlier, MOLA is capable of single-objective solutions, but its primary purpose is for the solution to conflicting multi-objective problems. MOLA requires the names of the objectives and their relative weights, the names of the suitability maps for each, and the amount of area that should be allocated to each (or alternatively the maximum budget to expend in acquiring the best land using a separate land price map). It then iteratively performs a first stage allocation for each objective separately, checks for conflicts in the allocations, and then resolves conflicts based on a minimum-distance-to-ideal-point rule using the weighted ranks. At the end of each iteration, progress towards the area or budget goals are assessed and new parameters are set. The process continues until all goals are met. As with single-objective solutions, MOLA also allows you to indicate whether the allocations should be contiguous and to control their compactness by specifying a minimum span.

A Worked Example

To illustrate these multi-criteria/multi-objective procedures, we will consider the following example of developing a zoning map to regulate expansion of the carpet industry (one of the largest and most rapidly growing industries in Nepal) within agricultural areas of the Kathmandu Valley of Nepal. The problem is to zone 1500 hectares of current agricultural land outside the ring road of Kathmandu for further expansion of the carpet industry. In addition, 6000 hectares will be zoned for special protection of agriculture. The problem clearly falls into the realm of multi-objective/multi-criteria decision problems. In this case, we have two objectives: to protect lands that are best for agriculture, and at the same time find other lands that are best suited for the carpet industry. Since land can be allocated to only one of these uses at any one time, the objectives must be viewed as conflicting (i.e., they may potentially compete for the same lands). Furthermore, the evaluation of each of these objectives can be seen to require multiple criteria.

In the illustration that follows, a solution to the multi-objective/multi-criteria problem is presented as developed with a group of Nepalese government officials as part of an advanced seminar in GIS.⁸ While the scenario was developed purely for the purpose of demonstrating the techniques used, and while the result does not represent an actual policy decision, it is one that incorporates substantial field work and the perspectives of knowledgeable decision makers. The procedure follows a logic in which each of the two objectives is first dealt with as a separate multi-criteria evaluation problem. The result consists of two separate suitability maps (one for each objective) which are then compared to arrive at a single solution that balances the needs of the two competing objectives.

1. Solving the Single Objective Multi-Criteria Evaluations

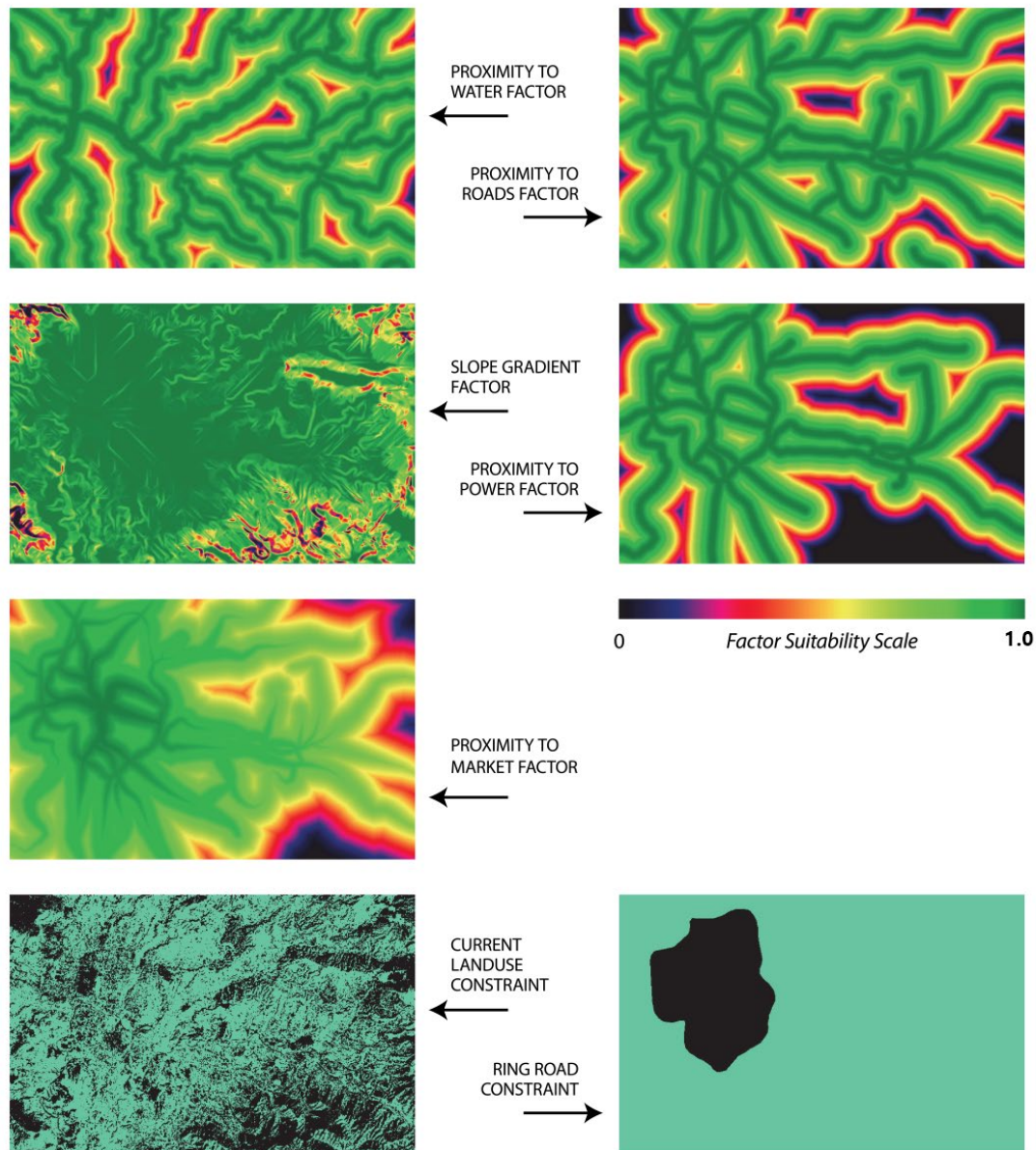
▪ 1.1 Establishing the Criteria: Factors and Constraints

The decision making group identified five factors as being relevant to the siting of the carpet industry: proximity to water (for use in dyeing and the washing of carpets), proximity to roads (to minimize road construction costs), proximity to power, proximity to the market, and slope gradient. For agriculture, they identified three of the same factors: proximity to water (for irrigation), proximity to market, and slope gradient, as well as a fourth factor, soil capability. In both cases, they identified the same constraints: the allocation would be limited to areas outside the ring road surrounding Kathmandu, land currently in some form of agricultural use, and slope gradients less than 100%. For factor images, distance to water, road and power lines was calculated based on the physical distance, and the proximity to market was developed as a cost distance surface (accounting for variable road class frictions).

▪ 1.2 Standardizing the Factors

⁸ The seminar was hosted by UNITAR at the International Center for Integrated Mountain Development (ICIMOD) in Nepal, September 28-October 2, 1992.

Each of the constraints was developed as a Boolean map while the factors were standardized using the module FUZZY so that the results represent fuzzy membership in the decision set. For example, for the carpet industry allocation, the proximity to water factor map was standardized using a *sigmoidal monotonically decreasing* fuzzy membership function with control points at 10 and 700 meters. Thus, areas less than 10 meters were assigned a set membership of 1.0, those between 10 and 700 meters were assigned a value which progressively decreased from 1.0 to 0 in the manner of an s-shaped curve, and those beyond 700 meters to a river were considered to be too far away (i.e., they were assigned a value of 0). The figure below illustrates the standardized results of all five factors and the constraints for the carpet industry allocation.



Carpet Industry Factors and Constraints

▪ 1.3 Establishing the Factor Weights

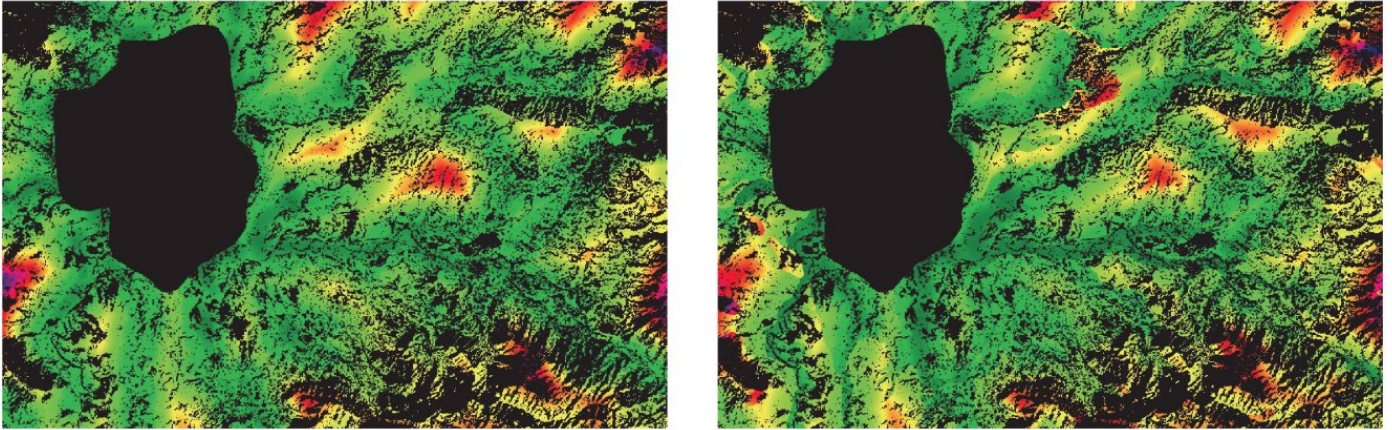
The next stage was to establish a set of weights for each of the factors. In the nature of a focus group, the GIS analyst worked with the decision makers as a group to fill out a pairwise comparison matrix. Each decision maker was asked in turn to estimate a rating and then to indicate why he or she assigned the rating. The group would then be asked if they agreed. Further discussion would ensue, often with suggestions for different ratings. Ultimately, if another person made a strong case for a different rating that seemed to have broad support, the original person who provided the rating would be asked if he/she were willing to change (the final decision would in fact rest with the original rater). Consensus was not difficult to achieve using this procedure. It has been found through repeated experimentation with this technique that the only cases where strong disagreement arose were cases in which a new variable was eventually identified as needing to be incorporated. This is perhaps the greatest value of the pairwise comparison technique—it is very effective in uncovering overlooked criteria and reaching a consensus on weights through direct participation by decision makers.

Once the pairwise comparison matrices were filled, the WEIGHT module was used to identify inconsistencies and develop the best fit weights. The table below shows the factor weights evaluated for the suitability for carpet industry development.

FACTOR	FACTOR WEIGHT
Proximity to Water	0.51
Proximity to Roads	0.05
Proximity to Power	0.25
Accessibility to Market	0.16
Low Slopes	0.03

▪ 1.4 Undertaking the Multi-Criteria Evaluation

Once the weights were established, the module MCE (for Multi-Criteria Evaluation) was used to combine the factors and constraints in the form of a weighted linear combination (WLC option). The procedure is optimized for speed and has the effect of multiplying each factor by its weight, adding the results and then successively multiplying the result by each of the constraints. Since the weights sum to 1.0, the resulting suitability maps have a range from 0-1. The figure below shows the result of separate multi-criteria evaluations to derive suitability maps for the carpet and agricultural industries.

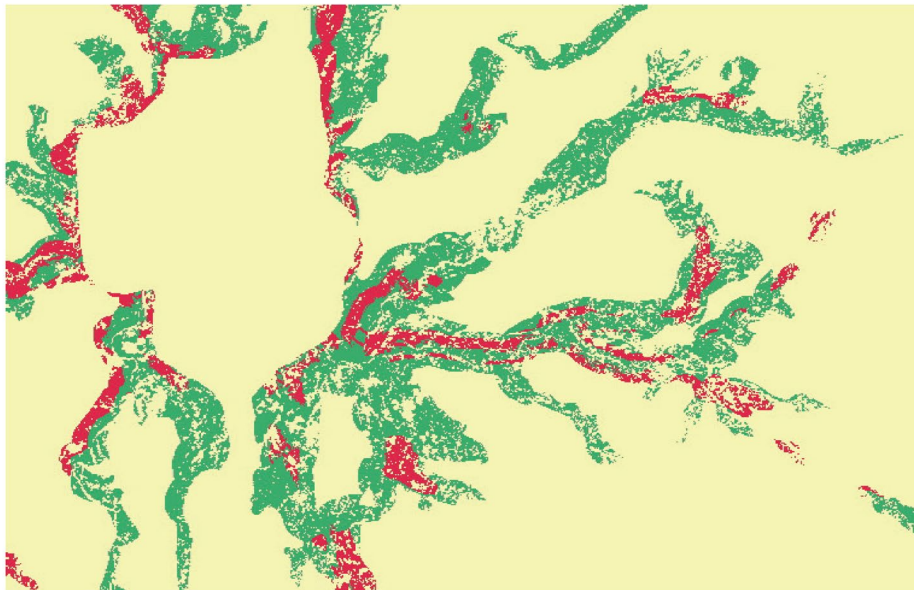


Composite Suitability images for Carpet Industry (left) and Agriculture (right). Suitability scale corresponds to that in previous figure.

2. Solving the Multi-Objective Land Allocation Problem

Once the multi-criteria suitability maps have been created for each objective, the multi-objective decision problem can be approached using the MOLA module.

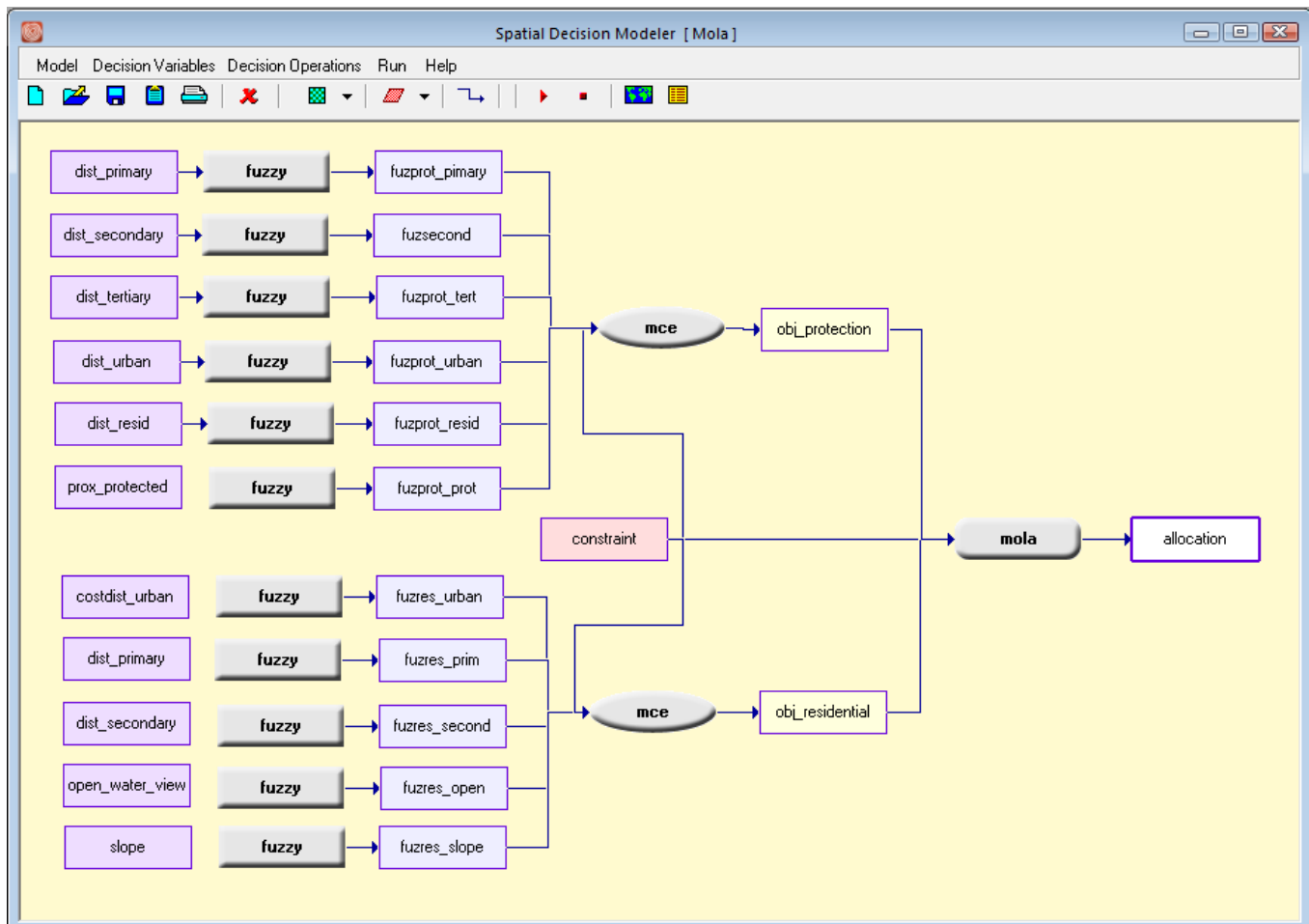
MOLA requires the names of the suitability maps, the names of objectives, the relative weight to assign to each, and the area to be allocated to each. The module then undertakes the iterative procedure of allocating the best ranked cells to each objective according to the areal goals, looking for conflicts, and resolving conflicts based on the weighed minimum-distance-to-ideal-point logic. The figure below shows the final result for a non-contiguous allocation, achieved after 6 iterations. Note, however, that MOLA also provides the option for a contiguous solution and provides the ability to control its compactness.



Final allocation to the carpet industry (red) and agriculture (green) objectives

Spatial Decision Modeler

The Spatial Dependence Modeler (SDM) is a graphical environment for decision support and is at the heart of this integrated environmental decision support system. It provides a graphical interface for developing decision models that can resolve complex resource allocation decisions, using many of the outputs produced by the other tools in the TerrSet system. Spatial Decision Modeler uses the language and the logic developed around the TerrSet decision support tools, including the development of factors and constraints with tools such as FUZZY and RECLASS, the combination of factors to produce suitability maps with the tool MCE, and the further combination of multiple objectives with the MOLA tool. MOLA includes additional options for forcing contiguity and compactness in the allocation result, as well as a new non-spatial budget constraint option.



A Closing Comment

The decision support tools provided in TerrSet are still under active development. We therefore welcome written comments and observations to further improve the modules and enhance their application in real-world situations.

References / Further Reading

- Alonso, W., 1968. Predicting Best with Imperfect Data, *Journal of the American Institute of Planners*, 34: 248-255.
- Carver, S.J., 1991. Integrating Multi-Criteria Evaluation with Geographical Information Systems, *International Journal of Geographical Information Systems* 5(3): 321-339.
- Campbell, J.C., Radke, J., Gless, J.T. and Wirtshafter, R.M., 1992. An Application of Linear Programming and Geographic Information Systems: Cropland Allocation in Antigua, *Environment and Planning A*, 24: 535-549.
- Diamond, J.T. and Wright, J.R., 1988. Design of an Integrated Spatial Information System for Multiobjective Land-Use Planning, *Environment and Planning B: Planning and Design*, 15: 205-214.
- Diamond, J.T. and Wright, J.R., 1989. Efficient Land Allocation, *Journal of Urban Planning and Development*, 115(2): 81-96.
- Eastman, J.R., 1996. Uncertainty and Decision Risk in Multi-Criteria Evaluation: Implications for GIS Software Design, Proceedings, UN University International Institute for Software Technology Expert Group Workshop on Software Technology for Agenda'21: Decision Support Systems, February 26-March 8.
- Eastman, J.R., and Jiang, H., 1996. Fuzzy Measures in Multi-Criteria Evaluation, *Proceedings, Second International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Studies*, May 21-23, Fort Collins, Colorado, 527-534.
- Eastman, J.R., Jin, W., Kyem, P.A.K., and Toledano, J., 1995. Raster Procedures for Multi-Criteria/Multi-Objective Decisions, *Photogrammetric Engineering and Remote Sensing*, 61(5): 539-547.
- Eastman, J.R., Kyem, P.A.K., and Toledano, J., 1993. A Procedure for Multi-Objective Decision Making in GIS Under Conditions of Competing Objectives, *Proceedings, EGIS'93*, 438-447.
- Eastman, J.R., Kyem, P.A.K., Toledano, J. and Jin, W., 1993. *GIS and Decision Making*, Explorations in Geographic Information System Technology, 4, UNITAR, Geneva.
- FAO, 1976. *A Framework for Land Evaluation*, Soils Bulletin 32. Food and Agricultural Organization of the United Nations, Rome.
- Feiring, B.R., 1986. Linear Programming: An Introduction, *Quantitative Applications in the Social Sciences*, Vol. 60, Sage Publications, London.
- Honea, R.B., Hake, K.A., and Durfee, R.C., 1991. Incorporating GISs into Decision Support Systems: Where Have We Come From and Where Do We Need to Go? In: M. Heit and A. Shortreid (eds.), *GIS Applications in Natural Resources*. GIS World, Inc., Fort Collins, Colorado.
- Ignizio, J.P., 1985. Introduction to Linear Goal Programming, *Quantitative Applications in the Social Sciences*, Vol. 56, Sage Publications, London.
- Janssen, R. and Rietveld, P., 1990. Multicriteria Analysis and Geographical Information Systems: An Application to Agricultural Land Use in the Netherlands. In: H.J. Scholten and J.C.H. Stillwell, (eds.), *Geographical Information Systems for Urban and Regional Planning*: 129-139. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Rao, M., Sastry, S.V.C., Yadav, P.D., Kharod, K., Pathan, S.K., Dhinwa, P.S., Majumdar, K.L., Sampat Kumar, D., Patkar, V.N. and Phatak, V.K., 1991. *A Weighted Index Model for Urban Suitability Assessment—A GIS Approach*. Bombay Metropolitan Regional Development Authority, Bombay, India.
- Rosenthal, R.E., 1985. Concepts, Theory and Techniques: Principles of Multiobjective Optimization. *Decision Sciences*, 16(2): 133-152.
- Saaty, T.L., 1977. A Scaling Method for Priorities in Hierarchical Structures. *J. Math. Psychology*, 15: 234-281.

Voogd, H., 1983. Multicriteria Evaluation for Urban and Regional Planning. Pion, Ltd., London.

Wright, J., ReVelle, C. and Cohon, J., 1983. A Multiobjective Integer Programming Model for the Land Acquisition Problem. *Regional Science and Urban Economics*, 13: 31-53.

Zadeh, L.A., 1965. Fuzzy Sets. *Information and Control*, 8: 338-353.

Uncertainty Management

Uncertainty is inevitable in the decision making process. In the GIS community, the issue of uncertainty has received a considerable amount of interest (see Goodchild and Gopal, 1989), however, attention has focused particularly on measurement error: the expression of error (Burrough, 1986; Lee et. al., 1987; Maling, 1989; Stoms, 1987), error assessment (Congalton, 1991), error propagation (Burrough, 1986), and the reporting of data quality (Moellering et. al., 1988; Slonecker and Tosta, 1992). There has also been considerable interest in other forms of uncertainty such as that expressed by fuzzy sets (e.g., Fisher, 1991). However, there has been less attention paid to how these uncertainties combine to affect the decision process and decision risk.

As the field becomes more conversant in the understanding and handling of uncertainty and its relationship to decision risk, it is inevitable that we will see a movement of GIS away from the *hard* decisions of traditional GIS (where it is assumed that the database and models are perfect) to procedures dominated by *soft* decisions. Given a knowledge of uncertainties in the database and uncertainties in the decision rule, it is possible to change the hard Boolean results of traditional GIS decisions into soft probabilistic results—to talk not of whether an area does or does not have a problem with soil erosion, but of the *likelihood* that it has a problem with soil erosion; not of whether an area is suitable or not for land allocation, but of the degree to which it is suitable. This would then allow a final hard decision to be developed based on the level of risk one is willing to assume. Thus, for example, one might decide to send an agricultural extension team to visit only those farms where the likelihood (or possibility) of a soil erosion problem exceeds 70%.

The movement to soft decision rules will require, in part, the development of uncertainty management capabilities in GIS. It requires data structures to carry uncertainty information and a revision of existing routines to assess and propagate error information. It also requires new procedures for analyzing different kinds of uncertainty and their effects on decision making. In TerrSet, a variety of procedures are available for this task.

A Typology of Uncertainty

Uncertainty includes any known or unknown error, ambiguity or variation in both the database and the decision rule. Thus, uncertainty may arise from such elements as measurement error, inherent variability, instability, conceptual ambiguity, over-abstraction, or simple ignorance of important model parameters.

Considering the decision making process as a set membership problem is a useful perspective from which to understand the source and role of uncertainty in decision making. As previously defined, a *decision frame* contains all the alternatives (or hypotheses) under consideration, and *evidence* is that information through which set membership of a location in the *decision set* (the set of chosen alternatives) can be evaluated. Thus, the decision making process contains three basic elements within which uncertainty can occur—the evidence, the decision set, and the relation that associates the two.

Uncertainty in the Evidence

In examining evidence to decide which elements of the candidate set belong to the set of alternatives to be chosen (the decision set), one evaluates the qualities and characteristics of those entities as represented in the database. However, there is a significant concern here with measurement error and how it propagates through a decision rule. This kind of uncertainty is usually represented by an RMS (root mean square) error in the case of quantitative data, or proportional error in the case of qualitative data, and relies upon classical probability theory and statistical inference for its assessment and propagation.

Uncertainty in the Relation

The second basic element of a decision is the specification of the relationship between the evidence and the decision set. Uncertainty arises here from at least three sources.

1. The first is in cases where the *definition* of a criterion (as opposed to its measurement) is subject to uncertainty. Sets with clearly defined attributes are known as *crisp* sets and are subject to the logic of classical sets. Thus, for example, the set of areas that would be inundated by a rise in sea level is clearly defined. Disregarding measurement error, if an area is lower than the projected level of the sea, it is unambiguously a member of the set. However, not all sets are so clearly defined. Consider, for example, the set of areas with steep slopes. What constitutes a steep slope? If we specify that a slope is steep if it has a gradient of 10% or more, does this mean that a slope of 9.99999% is not steep? Clearly there is no sharp boundary here. Such sets are called *fuzzy* sets (Zadeh, 1965) and are typically defined by a set membership function, as will be discussed further below. Although recognition of the concept of fuzzy sets is somewhat new in GIS, it is increasingly clear that such sets are prevalent (if not dominant) in land allocation decisions.
2. The second case where uncertainty arises is in cases where the evidence does not directly and perfectly imply the decision set under consideration. In the examples of inundated lands or steep slopes, there is a direct relationship between the evidence and the set under consideration. However, there are also cases where only indirect and imperfect evidence can be cited. For example, we may have knowledge that water bodies absorb infrared radiation. Thus we might use the evidence of low infrared reflectance in a remotely sensed image as a statement of the belief that the area is occupied by deep open water. However, this is only a belief since other materials also absorb infrared radiation.

Statements of belief in the degree to which evidence implies set membership are very similar in character to fuzzy set membership functions. However, they are not definitions of the set itself, but simply statements of the degree to which the evidence suggests the presence of the set (however defined). Thus the logic of fuzzy sets is not appropriate here, but rather, that of *Bayes* and *Dempster-Shafer* theory.

3. The third area where uncertainty can occur in specifying the relation between the evidence and the decision set is most often called *model specification error* (Alonso, 1968). In some instances, decisions may be based on a single criterion, but commonly several criteria are required to define the decision set. Thus, for example, one might define areas suitable for development as being those on shallow slopes and near to roads. Two issues here would be of concern: are these criteria adequate to define suitable areas, and have we properly aggregated the evidence from these criteria? If set membership indicated by slopes is 0.6 and proximity to roads is 0.7, what is the membership in the decision set? Is it the 0.42 of probabilities, the 0.6 of fuzzy sets, the 0.78 of Bayes, the 0.88 of Dempster-Shafer, or the 0.65 of linear combination? Further, how well does this aggregated value truly predict the degree to which the alternative under consideration truly belongs to the decision set? Clearly the construction of the decision rule can have an enormous impact on the set membership value deduced.

Uncertainty in the Decision Set

The final area of concern with respect to uncertainty in the decision process concerns the final set deduced. As outlined above, the process of developing the decision set consists of converting the evidence for each criterion into an elementary set statement, and then aggregating those statements into a single outcome that incorporates all of the criteria considered. Clearly, uncertainty here is some aggregate of the uncertainties which arose in acquiring the evidence and in specifying the relationship between that evidence and the decision set. However, in the presence of uncertainty about the degree to which any candidate belongs to the final set (as a result of the evidence gathered or its implications about set membership), some further action is required in order to develop the final set—a threshold of uncertainty will need to be established to determine which alternatives will be judged to belong to the decision set. To do so thus logically implies some likelihood that the decision made will be wrong—a concept that can best be described as *decision risk*. For example, given a group of locations for which the likelihood of being below a projected new sea level has been assessed, the final decision about which locations will be assumed to ultimately flood will be solved by establishing a threshold of likelihood. Clearly this threshold is best set in the context of decision risk.

In the remainder of this chapter, a set of tools in TerrSet will be explored for the management of uncertainty that arises in the evidence (database uncertainty) and in specifying the relation between that evidence and the decision set (decision rule uncertainty). In addition, in each of these two sections, consideration will be given to the problem of making a definitive judgment in the context of uncertainty, and thus the accommodation of decision risk.

Database Uncertainty and Decision Risk

An assessment of measurement error and an analysis of its propagation through data models combining different data layers is an essential aspect of uncertainty management. In this section, we examine procedures available in TerrSet for error assessment and propagation, and very importantly, procedures for evaluating the effects of this error on the decision process through a consideration of decision risk.

Error Assessment

The assessment of measurement error is normally achieved by selecting a sample of sites to visit on the ground, remeasuring the attribute at those locations using some more accurate instrument, and then comparing the new measurements to those in the data layer. To assist this procedure, TerrSet provides the SAMPLE and ERRMAT modules.

SAMPLE has the ability to lay out a sample of points (in vector format) according to a random, systematic or stratified random scheme. The latter is usually preferred since it combines the best qualities of the other two—the unbiased character of the random sampling scheme with the even geographic coverage of the systematic scheme.

The size of the sample (n) to be used is determined by multiplying an estimate of the standard error of the evaluation statistic being calculated by the square of the standard score (z) required for the desired level of confidence (e.g., 1.96 for 95% confidence), and dividing the result by the square of the desired confidence interval (e) (e.g., 0.01 for $\pm 10\%$). For estimates of the sample size required for estimating an RMS error, this formula simplifies to:

$$n = \frac{z^2 s^2}{2e^2} \quad \text{where } s \text{ is the estimated RMS}$$

For estimates of the proportional error in categorical data, the formula becomes:

$$n = \frac{z^2 pq}{e^2} \quad \text{where } p \text{ is the estimated proportional error and } q = (1 - p)$$

Note that the term *stratified* in stratified random means that it is spatially stratified according to a systematic division of the area into rectangular regions. In cases where some other stratification is desired, and/or where the region to be sampled is not rectangular in shape, the following procedure can be used:

1. Determine the area of the stratum or irregular region using the AREA module and divide it by the area of the total image. This will indicate the proportional area of the stratum or irregular region.
2. Divide the desired sample size by the proportional area. This will indicate a new (and larger) sample size that will be required to ensure that the desired number of sample points will fall within the area of interest.
3. Run SAMPLE with the new sample size and use only those points that fall within the area of interest.

Once the ground truth has been undertaken at the sample points, the characteristic error can be assessed. In the case of assessing quantitative data using RMS, the standard formula for RMS derived from a sample can be used:

$$RMS = \sqrt{\frac{\sum(x_i - t)^2}{n - 1}}$$

where x_i = measurement

t = true value

However, in the case of qualitative data, an error matrix should be used to assess the relationship between mapped categories and true values. To facilitate this process, the ERRMAT module can be used. ERRMAT requires two input files: the original categorical image (e.g., a landuse map) and a second image containing the true categories. This *truth* map is typically in the form of a map dominated by zeros (the background) with isolated cells indicating the positions of sample points with their true values. Using these data, ERRMAT outputs an error matrix and summary statistics.

The error matrix produced by ERRMAT contains a tabulation of the number of sample points found in each possible combination of true and mapped categories. The table illustrates the basic error matrix output. As can be seen, tabulations along the diagonal represent cases where the mapped category matched the true value. Off-diagonal tabulations represent errors and are tabulated as totals in the margins. The error marginals represent the proportional error by category, with the total proportional error appearing in the bottom-right corner of the table. Proportional errors along the bottom of the graph are called *errors of omission* while those along the right-hand edge are called *errors of commission*. The former represents cases where sample points of a particular category were found to be mapped as something different, while the latter includes cases where locations mapped as a particular category were found to be truly something else. Careful analysis of these data allows not only an assessment of the amount of error, but also of where it occurs and how it might be remedied. For example, it is typical to look at errors of omission as a basis for judging the adequacy of the mapping, and the errors of commission as a means of determining how to fix the map to increase the accuracy.

		True				Total	error	<i>errors of commission</i>
		Conifers	Mixed	Deciduous	Water			
Mapped	Conifers	24	0	0	3	27	0.11	
	Mixed	3	36	16	0	55	0.35	
	Deciduous	0	0	28	0	28	0.00	
	Water	2	0	0	14	16	0.12	
	Total	29	36	44	17	126		

error	0.17	0.00	0.36	0.18	0.19
--------------	-------------	-------------	-------------	-------------	-------------

An Error Matrix

In addition to the basic error matrix, ERRMAT also reports the overall and per category Kappa Index of Agreement (KIA) values. The Kappa Index of Agreement is similar to a proportional accuracy figure (and thus the complement of proportional error), except that it adjusts for chance agreement.

Error Propagation

When uncertainty exists in data layers, that error will propagate through any analysis and combine with the error from other sources. Specific formulas do exist for the expected error propagation arising from typical GIS mathematical operations (such as those involved with SCALAR and OVERLAY). **Appendix 4** contains a representative set of such formulae. In addition, TerrSet contains two modules that under certain circumstances will propagate error information automatically using such procedures. The first is the MCE module described earlier in this chapter while the second is SURFACE.

If all the input factors presented to MCE contain error (RMS) information in the *value error* field of their documentation files, MCE will determine the propagated output error and place it in the documentation file of the result. However, bear in mind that it makes two *large* assumptions—first, that there is no correlation between the factors, and second, that there is no uncertainty in the weights since that uncertainty has been resolved through deriving a consensus. If these assumptions are not valid, a new assessment should be derived using a Monte Carlo procedure as described further below.

In the case of SURFACE, error information will also be propagated when deriving slopes from a digital elevation model where the RMS error has been entered in the *value error* field of its documentation file.

Despite the availability of propagation formulas, it is generally difficult to apply this approach to error propagation because:

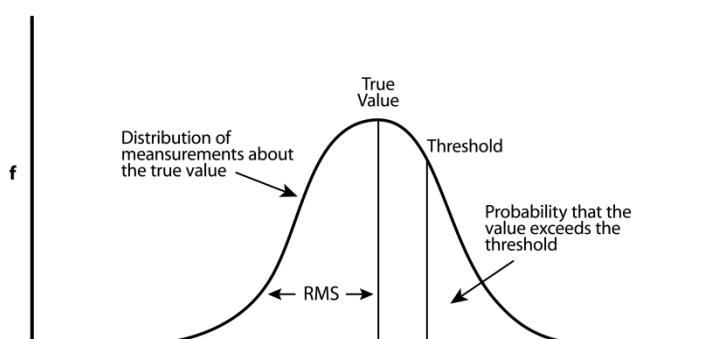
1. propagation is strongly affected by intercorrelation between variables and the correlation may not always be known at the outset;
2. only a limited number of formulas are currently available, and many GIS operations have unknown propagation characteristics.

As a result, we have provided in TerrSet the tools for a more general approach called *Monte Carlo Simulation*.

Monte Carlo Simulation

In the analysis of propagation error through Monte Carlo Simulation, we simulate the effects of error in each of the data layers to

assess how it propagates through the analysis. In practice, the analysis is run twice—first in the normal fashion, and then a second time using data layers containing the simulated error. By comparing the two results, the effects of the error can be gauged—the only reason they differ is because of the error introduced. Typically, HISTO would be used to examine the distribution of these errors as portrayed in a difference image produced with OVERLAY. With a normally distributed result, the standard deviation of



this difference image can be used as a good indicator of the final RMS.⁹

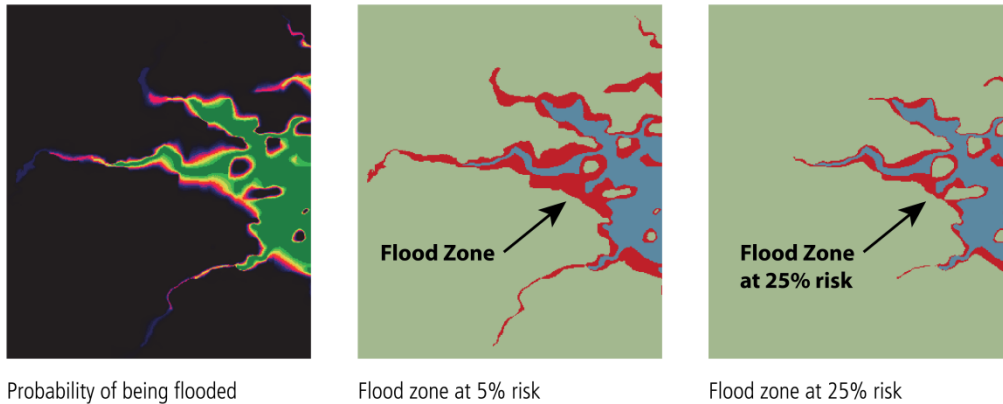
The tool that is used to introduce the simulated error is **RANDOM**. **RANDOM** creates images with random values according to a rectilinear, normal or lognormal model. For normal and lognormal distribution, the RMS error can be either one uniform value for the entire image, or be defined by an image that has spatially varied values. For categorical data, the rectilinear model outputs integer values that can be used as category codes. For quantitative data, all models can generate real numbers. For example, to add simulated error for a digital elevation model with an RMS error of 3 meters, **RANDOM** would be used to generate a surface using a normal model with a mean of 0 and a standard deviation of 3. This image would then be *added* to the digital elevation model. Note that the result is not meant to have any specific claim to reality—just that it contains error of the same nature as that believed to exist in the original.

Database Uncertainty and Decision Risk

Given an estimate of measurement error and an analysis of how it has propagated through the decision rule, the **PCLASS** module can be used to determine a final decision in full recognition of the decision risk that these uncertainties present. **PCLASS** evaluates the likelihood that the data value in any raster cell exceeds or is exceeded by a specified threshold. **PCLASS** assumes a random model of measurement error, characterized by a Root Mean Square (RMS) error statement. In the **TerrSet** system, the metadata for each raster image contains a field where error in the attribute values can be stated, either as an RMS for quantitative data, or as a proportional error for qualitative data. **PCLASS** uses the RMS recorded for a quantitative image to evaluate the probability that each value in the image lies either above or below a specified threshold. It does so by measuring the area delineated by that threshold under a normal curve with a standard deviation equal to the RMS (figure above). The result is a probability map as is illustrated in the figure below, expressing the likelihood that each area belongs to the decision set.

With **PCLASS** we have the *soft* equivalent of a *hard* **RECLASS** operation. For example, consider the case of finding areas that will be inundated by a rise in sea level as a result of global warming. Traditionally, this would be evaluated by reclassifying the heights in a digital elevation model into two groups—those below the projected sea level and those above. With **PCLASS**, however, recognition is made of the inherent error in the measurement of heights so that the output map is not a *hard* Boolean map of zeros and ones, but a *soft* probability map that ranges continuously from zero to one. In the first figure below, for example, is an illustration of the output from **PCLASS** after evaluating the probability that heights are less than a new projected sea level of 1.9 meters above the current level in Boston Harbor in the USA. Given this continuous probability map, a final decision can be made by reclassifying the probability map according to the level of decision risk one is willing to assume. The second two figures, for example, show the difference between the original coastline and that associated with the new sea level while accepting only a 5% chance of being wrong compared to that of accepting a 25% chance. Clearly, the Digital Elevation Model (DEM) used in this assessment is not very precise. However, this illustrates the fact that even poor data can be used effectively if we know how poor they are.

⁹ Monte Carlo Simulation relies upon the use of a very large set of simulations to derive its characterizations. In cases such as this where each cell provides a new simulation, the total composite of cells can provide such a large sample. Results are improved by repeated runs of such an analysis and an averaging of results.



Decision Rule Uncertainty

In the *Typology of Uncertainty* presented earlier, the second major element of uncertainty that was identified (after measurement error) was that in specifying the relationship between the evidence and the final decision set—an aspect that can broadly be termed *decision rule uncertainty*. This is an area where much further research is required. However, the TerrSet system does include an extensive set of tools to facilitate the assessment and propagation (or aggregation in this context) of this form of uncertainty.

All of these tools are concerned with the uncertainty inherent in establishing whether an entity belongs in the final decision set, and thus fall into a general category of uncertain set membership expression, known as a *fuzzy measure*. The term *fuzzy measure* (not to be confused with the more specific instance of a fuzzy set) refers to any set function which is monotonic with respect to set membership (Dubois and Prade, 1982). Notable examples of fuzzy measures include Bayesian *probabilities*, the *beliefs* and *plausibilities* of Dempster-Shafer theory, and the *possibilities* of fuzzy sets.

A common trait of fuzzy measures is that they follow DeMorgan's Law in the construction of the intersection and union operators (Bonissone and Decker, 1986), and thereby, the basic rules of uncertainty propagation in the aggregation of evidence. DeMorgan's Law establishes a triangular relationship between the intersection, union and negation operators such that :

$$T(a, b) = \sim S(\sim a, \sim b)$$

where T = Intersection (AND) = T – Norm

where S = Union (OR) = T – CoNorm

and \cong Negation (NOT)

The intersection operators in this context are known as *triangular norms*, or simply *T-Norms*, while the union operators are known as *triangular co-norms*, or *T-CoNorms*.

A T-Norm can be defined as (Yager, 1988):

a mapping $T: [0,1] \times [0,1] \rightarrow [0,1]$ such that:

$$T(a, b) = T(b, a) \quad (\text{commutative})$$

$$T(a, b) \geq T(c, d) \text{ if } a \geq c \text{ and } b \geq d \quad (\text{monotonic})$$

$$T(a, T(b, c)) = T(T(a, b), c) \quad (\text{associative})$$

$$T(1, a) = a$$

Some examples of T-Norms include:

$$\min(a, b) \quad (\text{the intersection operator of fuzzy sets})$$

$$a \times b \quad (\text{the intersection operator of probabilities})$$

$$1 - \min(1, ((1 - a)^p + (1 - b)^p)^{\frac{1}{p}}) \quad (\text{for } p \geq 1)$$

$$\max(0, a + b - 1)$$

Conversely, a T-CoNorm is defined as:

a mapping $S: [0,1] \times [0,1] \rightarrow [0,1]$ such that:

$$S(a, b) = T(b, a) \quad (\text{commutative})$$

$$S(a, b) \geq S(c, d) \text{ if } a \geq c \text{ and } b \geq d \quad (\text{monotonic})$$

$$S(a, S(b, c)) = S(S(a, b), c) \quad (\text{associative})$$

$$S(0, a) = a$$

Some examples of T-CoNorms include:

$$\max(a, b) \quad (\text{the union operator of fuzzy sets})$$

$$a + b - a \times b \quad (\text{the union operator of probabilities})$$

$$\min\left(1, (a^p + b^p)^{\frac{1}{p}}\right) \quad (\text{for } p \geq 1)$$

$$\min(1, a + b)$$

These examples show that a very wide range of operations are available for fuzzy measure aggregation, and therefore, criteria aggregation in decision making processes. Among the different operators, the most extreme (in the sense that they yield the most extreme numeric results upon aggregation) are the *minimum* T-Norm operator and the *maximum* T-CoNorm operator. These

operators also have special significance as they are the most used aggregation operators for fuzzy sets. Furthermore, they have been shown by Yager (1988) to represent the extreme ends of a continuum of related aggregation operators that can be produced through the operation of an ordered weighted average. As was indicated in the section **Decision Support**, this continuum also includes the traditional weighted linear combination operator that is commonly encountered in GIS. However, the important issue here is not that a family of aggregation operators is correct or better than another, but simply that different expressions of decision rule uncertainty require different aggregation procedures.

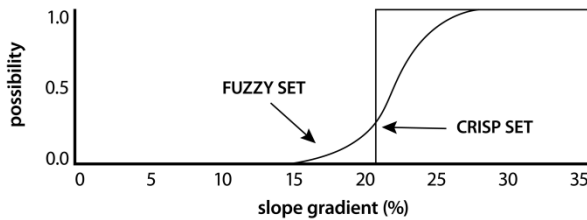
Currently, three major logics are in use for the expression of decision rule uncertainty, all of which are represented in the TerrSet module set: fuzzy set theory, Bayesian statistics, and Dempster-Shafer theory. Each is distinct and has its own very different set of T-Norm/T-CoNorm operators. However, the context in which one uses one as opposed to another is not always clear. In part, this results from the fact that decision rules may involve more than one form of uncertainty. However, this also results from a lack of research within the GIS field on the context in which each should be used. That said, here are some general guidelines that can be used:

- Decision problems that can be cast in the framework of suitability mapping can effectively be handled by the logic of fuzzy sets. This procedure has been covered in detail under the section on Multi-Criteria Evaluation in the **Decision Support section**. For example, if we define suitability in terms of a set of continuous factors (distance from roads, slope, etc.), the expression of suitability is continuous. There is no clear separation between areas that are suitable and those that are not. Many (if not most) GIS resource allocation problems fall into this category, and thus belong in the realm of fuzzy sets.
- The presence of *fuzziness*, in the sense of ambiguity, does not always imply that the problem lies in the realm of fuzzy sets. For example, measurement uncertainty associated with a crisp set can lead to a set membership function that is essentially identical in character to that of a fuzzy set. Rather, the distinguishing characteristic of a fuzzy set is that the set is itself inherently ambiguous. For example, if one considers the case of deciding on whether an area will be flooded as the result of the construction of a dam, some uncertainty will exist because of error in the elevation model. If one assumes a random error model, and spatial independence of errors, then a graph of the probability of being inundated against reported height in the database will assume an s-shaped cumulative normal curve, much like the typical membership function of a fuzzy set. However, the set itself is not ambiguous—it is crisp. It is the measure of elevation that is in doubt.
- The presence of *fuzziness*, in the sense of inconclusiveness, generally falls into the realm of Bayesian probability theory or its variant known as Dempster-Shafer theory. The problem here is that of indirect evidence—that the evidence at hand does not allow one to directly assess set membership, but rather to infer it with some degree of uncertainty. In their prototypical form, however, both logics are concerned with the substantiation of crisp sets—it is the strength of the relationship between the evidence and the decision set that is in doubt. A classic example here is the case of the supervised classification procedure in the analysis of remotely sensed imagery. Using training site data, a Bayesian classifier (i.e., decision engine) establishes a statistical relationship between evidence and the decision set (in the form of a conditional probability density function). It is this established, but uncertain, relationship that allows one to infer the degree of membership of a pixel in the decision set.
- Despite their common heritage, the aggregation of evidence using Bayes and Dempster-Shafer can yield remarkably different results. The primary difference between the two is characterized by the role of the absence of evidence. Bayes considers the absence of evidence in support of a particular hypothesis to therefore constitute evidence in support of alternative hypotheses, whereas Dempster-Shafer does not. Thus, despite the fact that both consider the hypotheses in the decision frame to be exhaustive, Dempster-Shafer recognizes the concept of ignorance while Bayes does not. A further difference is that the Bayesian approach combines evidence that is conditioned upon the hypothesis in the decision set (i.e., it is based on training data), while Dempster-Shafer theory aggregates evidence derived from independent sources.

Despite these broad guidelines, the complete implementation of these logics is often difficult because their theoretical development has been restricted to prototypical contexts. For example, fuzzy set theory expresses ambiguity in set membership in the form of a membership function. However, it does not address the issue of uncertainty in the form of the membership function itself. How, for example, does one aggregate evidence in the context of indirect evidence *and* an ambiguous decision set? Clearly there is much to be

learned here. As a start, the following section begins to address the issues for each of these major forms for the expression of uncertainty.

Fuzzy Sets

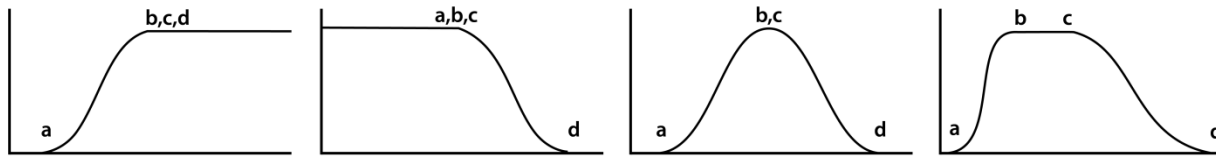


Fuzzy sets are sets (or classes) without sharp boundaries; that is, the transition between membership and nonmembership of a location in the set is gradual (Zadeh, 1965; Schmucker, 1982). A fuzzy set is characterized by a fuzzy membership grade (also called a *possibility*) that ranges from 0.0 to 1.0, indicating a continuous increase from nonmembership to complete membership. For example, in evaluating whether a slope is steep, we may define a fuzzy membership function such that a slope of 10% has a membership of 0, and a slope of 25% has a membership of 1.0. Between 10% and 25%, the fuzzy membership of a slope gradually increases on the scale from 0 to 1 (see figure). This contrasts with the classic *crisp* set which has distinct boundaries. However, a crisp set can also be seen as a special case of fuzzy set where fuzzy membership changes instantaneously from 0 or 1.

Fuzzy set theory provides a rich mathematical basis for understanding decision problems and for constructing decision rules in criteria evaluation and combination. In use, the FUZZY module in TerrSet is designed for the construction of Fuzzy set membership functions, while the OWA option of the MCE module offers a range of appropriate aggregation operators. FUZZY offers four types of membership function:

1. **Sigmoidal:** The *sigmoidal* ("s-shaped") membership function is perhaps the most commonly used function in fuzzy set theory. It is produced here using a cosine function as described in the Help System. In use, FUZZY requires the positions (along the X axis) of 4 *inflection points* governing the shape of the curve. These are indicated in the figure below as points *a*, *b*, *c* and *d*, and represent the inflection points as the membership function rises above 0, approaches 1, falls below 1 again, and finally approaches 0. The right-most function of figure shows all four inflection points as distinct. However, this same function can take different forms. Beginning at the left, the *monotonically increasing* function shape rises from 0 to 1 then never falls. The previously mentioned concept of steep slopes is a good example here where the first inflection point *a* would be 10%, and the second *b* would be 25%. Since it never falls again, inflection points *c* and *d* would be given the same value as *b* (FUZZY understands this convention). However, the FUZZY interface facilitates data input in this case by requesting values only for inflection points *a* and *b*. The second curve shows a *monotonically decreasing* function that begins at 1 then falls and stays at 0. In this case where the membership function starts at 1 and falls to 0 but never rises, *a* and *b* would be given identical values to *c* (the point at which it begins to fall), and *d* would be given the value of the point at which it reaches 0. The FUZZY interface only requires inflection points *c* and *d* for this type of function. The last two functions shown are termed *symmetric* as they rise then fall again. In the case where the function rises and then immediately falls (the third curve in the figure below), points *b* and *c* take on the same value. Finally, where it rises, stays at 1 for a while, and then falls, all four values are distinct. In both cases, the FUZZY interface requires input of all four inflection points. Note that there is no requirement of geometric symmetry for symmetric functions, only that the curves rise then fall again. It is quite likely that the shape of the curve between *a* and *b* and the shape between *c* and *d* would be different, as illustrated in the right-most

curve in the figure below.



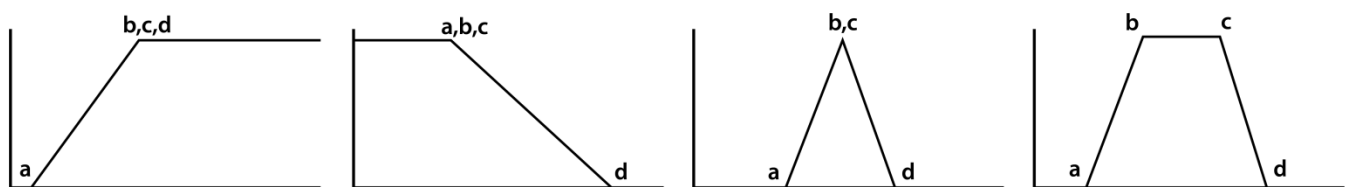
Sigmoidal Membership Function

2. **J-Shaped:** The J-Shaped function is also quite common, although in most cases it would seem that a sigmoidal function would be better. The figure below shows the different possibilities of J-shaped functions and the positions of the inflection points. *It should be pointed out that with the J-shaped function, the function approaches 0 but only reaches it at infinity. Thus the inflection points a and d indicate the points at which the function reaches 0.5 rather than 0.*



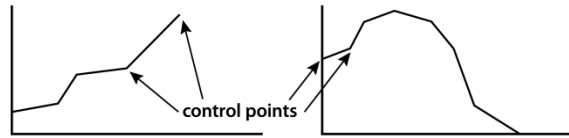
J-Shaped Membership Function

3. **Linear:** The figure below shows the linear function and its variants, along with the position of the inflection points. This function is used extensively in electronic devices advertising fuzzy set logic, in part because of its simplicity, but also in part because of the need to monitor output from essentially linear sensors.



Linear Membership Function

4. **User-defined:** When the relationship between the value and fuzzy membership does not follow any of the above three function types, the user-defined function is most applicable. An unlimited number of control points may be used in this function to define the fuzzy membership curve. The fuzzy membership between any two control points is linearly interpolated, as in the figure below.



User-Defined Membership Function

In Multi-Criteria Evaluation, fuzzy set membership is used in the standardization of criteria. Exactly which function should be used will depend on the understanding of the relationship between the criterion and the decision set, and on the availability of information to infer fuzzy membership. In most cases, either the sigmoidal or linear functions will be sufficient.

Bayesian Probability Theory

When complete information is available or assumed, the primary tool for the evaluation of the relationship between the indirect evidence and the decision set is Bayesian Probability theory. Bayesian Probability theory is an extension of Classical Probability theory which allows us to combine new *evidence* about an hypothesis along with prior knowledge to arrive at an estimate of the likelihood that the hypothesis is true. The basis for this is Bayes' Theorem which states that (in the notation of probability theory):

$$\text{where: } p(h|e) = \frac{p(e|h) \cdot p(h)}{\sum_i (e|h) \cdot p(h_i)}$$

$p(h|e)$ = the probability of the hypothesis being true given the evidence (posterior probability)

$p(e|h)$ = the probability of finding that evidence given the hypothesis being true

$p(e|h)$ = the probability of the hypothesis being true regardless of the evidence (prior probability)

For those unfamiliar with probability theory, this formula may seem intimidating. However, it is actually quite simple. The simplest case is when we have only two hypotheses to choose from—an hypothesis h and its complement $\sim h$ (that h is **not** true), the probabilities of which are represented by $p(h)$ and $p(\sim h)$, respectively. For example, is an area going to be flooded or is it not? The first question to consider is whether we have any prior knowledge that leads us to the probability that one or the other is true. This is called an *a priori probability*. If we do not, then the hypotheses are assumed to be equally probable.

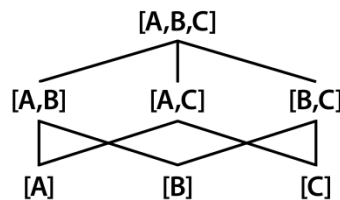
The term $p(e|h)$ expresses the probability that we would find the evidence we have if the hypothesis being evaluated were true. It is known as a *conditional* probability, and is assessed on the basis of finding areas in which we know the hypothesis to be true and gathering data to evaluate the probability that the evidence we have is consistent with this hypothesis. We will refer to this as ground truth data even though it may be assessed on theoretical grounds or by means of a simulation.

The term $p(h|e)$ is a *posterior* probability created after prior knowledge and evidence for the hypothesis are combined. By incorporating extra information about the hypotheses, the probability for each hypothesis is modified to reflect the new information. It is the assumption of Bayes' Theorem that complete information is achievable, and thus the only reason that we do not have an accurate probability assessment is a lack of evidence. By adding more evidence to the prior knowledge, theoretically one could reach a true probability assessment for all the hypotheses.

Dempster-Shafer Theory

Dempster-Shafer theory, an extension of Bayesian probability theory, allows for the expression of ignorance in uncertainty management (Gordon and Shortliffe, 1985; Lee et al., 1987). The basic assumptions of Dempster-Shafer theory are that ignorance exists in the body of knowledge, and that belief for a hypothesis is not necessarily the complement of belief for its negation.

First, Dempster-Shafer theory defines hypotheses in a hierarchical structure (figure below) developed from a basic set of hypotheses that form the *frame of discernment*.¹⁰ For example, if the frame of discernment includes three basic hypotheses: {A, B, C}, the structure of hypotheses for which Dempster-Shafer will accept evidence includes all possible combinations, [A], [B], [C], [A, B], [A, C], [B, C], and [A, B, C]. The first three are called singleton hypotheses as each contains only one basic element. The rest are non-singleton hypotheses containing more than one basic element. Dempster-Shafer recognizes these hierarchical combinations because it often happens that the evidence we have supports some combinations of hypotheses without the ability to further distinguish the subsets. For example, we may wish to include classes of [deciduous] and [conifer] in a landcover classification, and find that evidence from a black and white aerial photograph can distinguish forest from non-forested areas, but not the type of forest. In this case we may use this evidence as support for the hierarchical combination [deciduous, coniferous]. Clearly this represents a statement of uncertainty. However, it also provides valuable information that will be used to advantage by the Dempster-Shafer procedure in any statement of belief about these hypotheses.



Hierarchical Structure of the Subsets in the Whole Set [A,B,C]

In expressing commitment to any of these hypotheses, Dempster-Shafer theory recognizes six important concepts: *basic probability assignment (BPA)*, *ignorance*, *belief*, *disbelief*, *plausibility*, and *belief interval*.

A basic probability assignment (BPA) represents the support that a piece of evidence provides for one of these hypotheses and not its proper subsets. Thus a BPA for [A, B] represents that mass of support for [A,B], but not [A] or [B]—i.e., that degree of support for some indistinguishable combination of [A] and [B]. This is usually symbolized with the letter "m" (for *mass*), e.g.:

$$m(A, B) = \text{basis probability assignment to } [A, B]$$

¹⁰ The *frame of discernment* in Dempster-Shafer theory has essentially the same meaning as the term *decision frame* as used in this paper—i.e., the set of alternative hypotheses or classes that can be substantiated or assigned to entities. Dempster-Shafer considers these hypotheses to be exhaustive. Thus, statements of support for any hierarchical combination of classes represents a degree of inability to commit to one of the singleton hypotheses in the frame of discernment. However, in practice, Dempster-Shafer does treat these hierarchical combinations as additional hypotheses. In addition, in a GIS and Remote Sensing context, there may be good reason to treat some unresolvable commitment to one of these hierarchical combinations as truly evidence of an independent class/hypothesis to which entities might be assigned. For example, with a frame of discernment that includes [forest] and [wetland], the presence of commitment to a [forest wetland] combination may in fact represent the presence of a "forested wetland" class that cannot be resolved by attaining better evidence. As a result, we recognize here that the analyst may wish to consider the decision frame as containing all of the hierarchical combinations, and not just the more limited set of singletons that forms the Dempster-Shafer frame of discernment. This does not violate the logic of Dempster-Shafer, since we are simply making the post-analysis judgement that certain combinations represent new classes and thus may form a decision set.

The basic probability assignment for a given hypothesis may be derived from subjective judgment or empirical data. Since a BPA is a fuzzy measure, the FUZZY module can also be used in TerrSet to develop a BPA from a given data set.

The sum of all BPAs will always equal 1.0. Thus, the BPA for the ultimate superset ([A, B, C] in this example) will equal the complement of the sum of all other BPAs. This quantity thus represents *ignorance*—the inability to commit to any degree of differentiation between the elements in the frame of discernment.

Belief represents the total support for an hypothesis, and will be drawn from the BPAs for all subsets of that hypothesis, i.e.:

$$BEL(X) = \sum m(Y) \quad \text{when } Y \subseteq X$$

Thus, the belief in [A, B] will be calculated as the sum of the BPAs for [A, B], [A], and [B]. In this example, belief represents the probability that an entity is A or B. Note that in the case of singleton hypotheses, the basic probability assignment and belief are identical.

In contrast to belief, *plausibility* represents the degree to which an hypothesis cannot be disbelieved. Unlike the case in Bayesian probability theory, *disbelief* is not automatically the complement of belief, but rather, represents the degree of support for all hypotheses that do not intersect with that hypothesis. Thus:

$$PL(X) = 1 - BEL(\sim X) \quad \text{where } \sim X = \text{not } X$$

$$\text{thus } PL(X) = \sum m(Y) \quad \text{when } Y \cap X \neq \Phi$$

Interpreting these constructs, we can say that while belief represents the degree of hard evidence in support of an hypothesis, plausibility indicates the degree to which the conditions *appear to be right* for that hypothesis, even though hard evidence is lacking. For each hypothesis, then, belief is the lower boundary of our commitment to that hypothesis, and plausibility represents the upper boundary. The range between the two is called the *belief interval*, and represents the degree of uncertainty in establishing the presence or absence of that hypothesis. As a result, areas with a high belief interval are those in which new evidence will supply the greatest degree of information. Dempster-Shafer is thus very useful in establishing the value of information and in designing a data gathering strategy that is most effective in reducing uncertainty.

Compared with Bayesian probability theory, it is apparent that Dempster-Shafer theory is better able to handle uncertainty that involves ignorance. In Bayesian probability theory only singleton hypotheses are recognized and are assumed to be exhaustive (i.e., they must sum to 1.0). Thus, ignorance is not recognized, and a lack of evidence for a hypothesis therefore constitutes evidence against that hypothesis. These requirements and assumptions are often not warranted in real-world decision situations. For example, in establishing the habitat range for a particular bird species, evidence in the form of reported sightings might be used. However, the absence of a sighting at a location does not necessarily imply that the species was not present. It may simply indicate that there was no observer present, or that the observer failed to see a bird that was present. In cases such as this, Dempster-Shafer theory is appropriate (Gordon and Shortliffe, 1985; Srinivasan and Richards, 1990).

Dempster-Shafer Aggregation Operators

The full hierarchy of hypotheses and the BPAs associated with each represent a state of knowledge that can be added to at any time. In aggregating probability statements from different sources of evidence, Dempster-Shafer employs the following rule of combination:

$$m(Z) = \frac{\sum m_1(X) \times m_2(Y)}{1 - \sum m_1(X) \times m_2(Y)} \quad \begin{array}{l} \text{when } (X \cap Y) = Z \\ \text{when } (X \cap Y) = \Phi \end{array}$$

$$\text{If } \sum m_1(X) \times m_2(Y) = 0 \text{ for } (X \cap Y) = \Phi$$

$$m(z) = \sum m_1(X) \times m_2(Y) = 0 \text{ for } (X \cap Y) = Z$$

The final belief, plausibility, and belief interval for each of the hypotheses can then be calculated based on the basic probability assignment calculated using the above equations. Ignorance for the whole set can also be derived. In most cases, after adding new evidence, the ignorance is reduced.

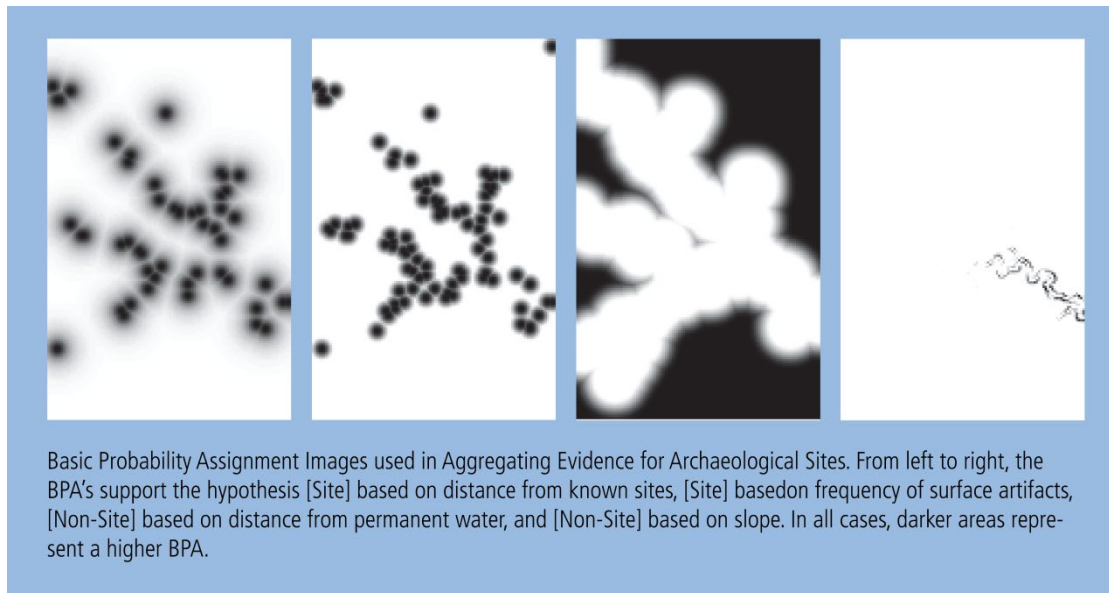
Working with Dempster-Shafer Theory: Belief

In TerrSet, the Belief module can be used to implement the Dempster-Shafer logic. Belief constructs and stores the current state of knowledge for the full hierarchy of hypotheses formed from a frame of discernment. In addition, it has the ability to aggregate new evidence with that knowledge to create a new state of knowledge, that may be queried in the form of map output for the belief, plausibility or belief interval associated with any hypothesis.

Belief first requires that the basic elements in the frame of discernment be defined. As soon as the basic elements are entered, all hypotheses in the hierarchical structure will be created in the hypothesis list. For each line of evidence entered, basic probability assignment images (in the form of real number images with a 0 - 1 range) are required with an indication of their supported hypothesis. The BUILD KNOWLEDGE BASE item in the ANALYSIS menu then incorporates this new evidence by recalculating the state of knowledge using the Dempster-Shafer rule of combination, from which summary images in the form of belief, plausibility or belief interval statements for each hypothesis can be selected. All the information entered can be saved in a knowledge base file for later use when more evidence is obtained.

The Dempster-Shafer rule of combination provides an important approach to aggregating indirect evidence and incomplete information. Consider, for example, the problem of estimating where an archaeological site of a particular culture might be found. The decision frame includes two basic elements, [site] and [non-site].¹¹ Four pieces of evidence are used: the locations of known sites, the frequency of surface artifacts (such as pottery shards), proximity to permanent water, and slopes. The first may be seen as direct evidence (at the exact positions of the sites themselves) for areas that have known archaeological sites. However, what we are concerned about are the areas that do not have a site, for which the known sites do not provide direct information. Therefore, the evidence is largely indirect. For areas that are close to the existing sites, one could believe the likelihood for the presence of another site would be higher. Thus the FUZZY module is used to transform a map of distance from known sites into an image of probability (a basic probability assignment image in support of the [site] hypothesis). The frequency of surface artifacts is also used as evidence in support of the [site] hypothesis. The distance from permanent water and slope images, however, have been used as disbelief images (see note 1 under "Using Belief" below). They therefore have both been scaled to a 0-1 range using FUZZY to provide support for the [non-site] hypothesis. The figure below shows these basic probability assignment images.

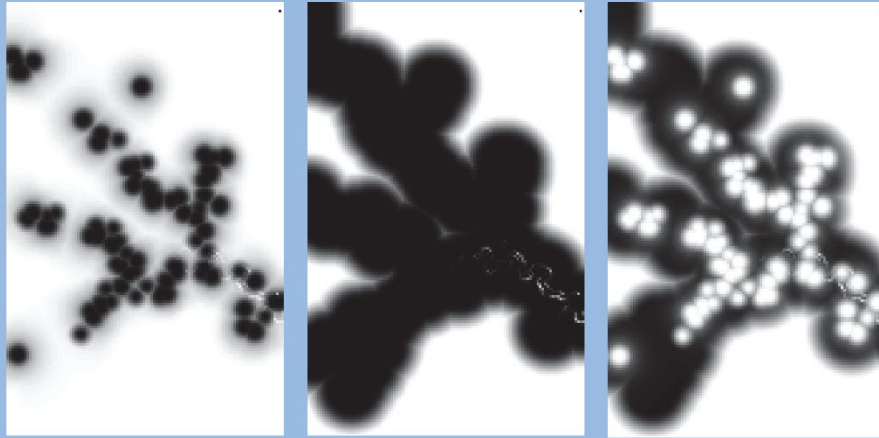
¹¹ The total number of hypotheses that Dempster-Shafer generates in the full hierarchy is $2^n - 1$. Implicitly, there is an extra hypothesis that is the null set, which is assumed by Dempster-Shafer to be automatically false. Thus in this example, the [non-site] hypothesis is not the null set, nor is it automatically assumed by Dempster-Shafer. In this example it was entered as a positive hypothesis, and member of the frame of discernment.



The module Belief combines information from all four sources and has been used to produce belief, plausibility and belief interval images for the [site] hypothesis as illustrated in the figure above. The belief interval image is particularly interesting in that it shows us where we have substantial uncertainty. Further sampling of evidence in these areas might prove profitable since the conditions support the plausibility of a site, even though concrete evidence is poor.

Using Belief

1. You may find it difficult to decide whether a particular piece of evidence should be used to support the belief of an hypothesis or, alternatively, the complement of that image should be used to support its disbelief. The latter is actually a statement in support of the plausibility of an hypothesis, but not its belief, and is very common in GIS. For example, in the case above, proximity to permanent water was treated as a distance image in support of disbelief in the possibility of a site. The reason for this is that if one were near to water there is no reason to believe that a site would or would not be present, but if one were far from water, there is excellent reason to assume that a site could not have existed. In deciding how to treat lines of evidence, consider carefully whether the data provide true evidence in support of an hypothesis, or simply support for its plausibility (i.e., the inability to deny its possibility).
2. To enter a disbelief, indicate that the evidence supports the collection of all hypotheses that do not include the one of concern. In the archaeology example, distance from water was entered as evidence for [non-site]. In a case with three hypotheses [A, B, C], to indicate that a particular line of evidence supports the disbelief in A, you would indicate that it provides support for [B, C].
3. For each line of evidence that is incorporated using Belief, make sure that you enter all of the hypotheses that a particular piece of evidence supports in one run. The reason for this is that Belief needs to undertake some internal calculations related to ignorance, and thus it needs to know also about the hypotheses for which that evidence does not add support. You only need to enter a basic probability assignment image if the evidence supports the hypothesis to some degree larger than zero. For the hypotheses that the evidence does not support, the module assumes 0 probability.
4. For each line of evidence, the basic probability assignment images must be real number images with a range that does not exceed 0-1.



Belief (left), Plausibility (middle) and Belief Interval (right) images for the presence of archaeological sites after Dempster-Shafer combination of evidence.

Decision Rule Uncertainty and Decision Risk

In the context of measurement error, it is a fairly straightforward matter to relate uncertainty to decision risk. In TerrSet, the PCLASS module achieves this based on the logic of classical sets (as was discussed earlier). However, as we move from the strong frequentist interpretation of probability associated with measurement error, to the more indirect relationship of Bayesian and Dempster-Shafer beliefs, to the quite independently established concept of fuzzy sets, we move further and further away from the ability to establish risk in any absolute sense (Eastman, 1996). Indeed, with a decision based on fuzzy sets, we can establish that the inclusion of an alternative is less risky than another, but not what the actual risk is. Thus, instead of calculating *absolute risk*, we need to be able to establish *relative risk*.

The concept of relative risk is one that is quite familiar. For example, in evaluating a group of candidates for employment, we might examine a number of quantifiable criteria—grades, rating charts, years of experience, etc.—that can permit the candidates to be ranked. We then attempt to hire the best ranked individuals on the assumption that they will perform well. However, there is no absolute scale by which to understand the likelihood that they will achieve the goals we set. In a similar manner, the RANK module in TerrSet can be used to rank the suitabilities achieved through a multi-criteria aggregation procedure. This result can then be divided by the maximum rank to produce an image of relative risk. This result can then be thresholded to extract a specific percentage of the best (i.e., least risky) solutions available. The importance of this solution is that it can be applied to any decision surface regardless of the nature of the uncertainties involved.

Anisotropic Cost Analysis

Cost surface modeling is now a familiar feature of many raster geographic information systems. In developing a cost surface, one accounts for the cost of moving through space, where costs are a function of both the standard (or *base*) costs associated with movement, and also of frictions and forces that impede or facilitate that movement.

Isotropic Costs

Isotropic cost surface modeling is accomplished in TerrSet with the COST module. Given input images of a set of features from which cost distances should be calculated and the frictions that affect movement, COST outputs a cost surface that expresses costs of movement in terms of distance equivalents. Thus, for example, if a cell contains a value of 100, it simply expresses that the cost of moving from the nearest starting feature (target) to that point is the equivalent of moving over 100 cells at the base cost. It could equally arise from traveling over 100 cells with a relative friction (i.e., relative to the friction associated with the base cost) of 1, or 50 cells with frictions of 2, or 1 cell with a relative friction of 100.

Anisotropic Costs

With the COST module, frictions have identical effect in any direction. It doesn't matter how you move through a cell—its friction will be the same. We can call such a friction *isotropic* since it is equal in all directions. However, it is not very difficult to imagine *anisotropic* frictions—frictional elements that have different effects in different directions. Take, for example, the case of slopes. If we imagine the costs of walking (perhaps in calories per hour at normal walking speed), then slopes will affect that cost differently in different directions. Traveling upslope will cause that friction to act full-force; traveling perpendicularly across the slope will have no effect at all; and traveling downslope will act as a force that reduces the cost. Traditional cost analysis cannot accommodate such an effect.

Anisotropic Cost Modules in TerrSet

In TerrSet, four modules are supplied for the modeling of anisotropic costs. Anisotropic cost analysis is still a very new area of analysis, and we therefore encourage users to send information, in writing, on their applications and experiences using these modules.

At the core of the set are two different modules for the analysis of anisotropic costs, VARCOST and DISPERSE, and two support modules for the modeling of forces and frictions that affect those costs, RESULTANT and DECOMP.

VARCOST models the effects of anisotropic frictions on the movement of phenomena that have their own motive force. The example just given of walking in the presence of slopes is an excellent example, and one that is perfectly modeled by VARCOST. DISPERSE, on the other hand, models the movement of phenomena that have no motive force of their own, but which are acted upon by anisotropic forces to disperse them over time. A good example of this would be a point-source pollution problem such as a chemical spill on land. Upon absorption into the soil, the contaminant would move preferentially with ground water under the force of gravity according to the hydraulic gradient. The resulting pattern of movement would look plume-like because of the decreasing probability of movement as one moved in a direction away from the maximum gradient (slope). DISPERSE and VARCOST are thus quite similar in concept, except in the nature of how forces and frictions change in response to changes in the direction of movement. This we call the anisotropic function, as will be discussed below. However, to understand such functions, it is useful to review the distinction between forces and frictions in the modeling of costs.

Forces and Frictions

In cost modeling, forces and frictions are not inherently different. In all of the cost modeling procedures—COST, VARCOST and DISPERSE—frictions are expressed as relative frictions using the base cost as a reference. Thus, for example, if it takes 350 calories to walk along flat ground, and 700 calories to walk across more rugged terrain at equal speed, we would indicate that rugged terrain has a friction of 2. However, if we were to walk down a slope such that our energy expended was only 175 calories, then we would express that as a friction of 0.5. But what are frictions less than 1? They are, in fact, forces. To retain consistency, all relative frictions

in TerrSet are expressed as values greater than 1 and relative forces are expressed as values less than 1. Thus, if we were concerned with wind forces and we had a base force of 10 km/hour, a wind of 30 km/hour would be specified as a relative force of 0.33.

With anisotropic cost modeling, a single image cannot describe the nature of forces and frictions acting differently in different directions. Rather, a pair of images is required—one describing the magnitude of forces and frictions, expressed as relative quantities exactly as indicated above, and the other describing the direction of those forces and frictions, expressed as azimuths.¹² These magnitude/direction image pairs thus describe a field of force/friction vectors which, along with the anisotropic function discussed below, can be used to determine the force or friction in any direction at any point. The term *force/friction image pair* refers to a magnitude image and its corresponding direction image for either forces (used with DISPERSE) or frictions (used with VARCOST).

It is important to understand the nature of the direction images required for both VARCOST and DISPERSE. With VARCOST, the *friction* direction image must represent the direction of movement that would incur the greatest cost to movement. For example, if you are modeling the movement of a person walking across a landscape and the frictions encountered are due to slopes (going uphill is difficult, going downhill is easy), then the values in the friction direction image should be azimuths from north that point uphill.

With DISPERSE, the *force* direction image must represent the direction in which the force acts most strongly. For example, if you are modeling the dispersion of a liquid spill over a landscape (flowing easily downhill, flowing with great difficulty uphill), then the values in the force direction image should be azimuths from north that point downhill.

In the use of VARCOST and DISPERSE, a single anisotropic force/friction vector image pair is specified. Since analyses may involve a number of different forces acting simultaneously, a pair of modules has been supplied to allow the combination of forces or frictions. The first of these is RESULTANT. RESULTANT takes the information from two force/friction image pairs to produce a new force/friction image pair expressing the resultant vector produced by their combined action. Thus, RESULTANT can be used to successively combine forces and frictions to produce a single magnitude/direction image pair to be used as input to VARCOST or DISPERSE.

The second module that can be used to manipulate force/friction image pairs is DECOMP. DECOMP can decompose a force/friction image pair into its X and Y component images (i.e., the force/friction in X and the force/friction in Y). It can also recombine X and Y force/friction components into magnitude and direction image pairs. Thus DECOMP could be used to duplicate the action of RESULTANT.¹³ However, a quite different and important use of DECOMP is with the interpolation of force/friction vectors. If one takes the example of winds, it is not possible to interpolate the data at point locations to produce an image, since routines such as TREND and INTERPOL cannot tell that the difference between 355 and 0 degrees is the same as between 0 and 5. However, if a raster image pair of the point force/friction data is constructed and then decomposed into X and Y components (using DECOMP), these component images can be interpolated (e.g., with TREND) and then recombined into a force/friction pair using DECOMP.

Anisotropic Functions

With force/friction image pairs, one has an indication of both the magnitude and direction with which forces and frictions act. However, what is the interpretation of direction? If a force is said to act at 45° (northeast), does this mean it acts fully at 45° and not at all at 44°? The answer to this is not easily determined and it ultimately depends upon the application. If one takes the earlier example of walking against slopes of varying degrees, the force/friction image describes only the direction and magnitude of the steepest descending slope. If one faced directly into the slope one would feel the full force of the friction (i.e., effective friction = stated friction). Facing directly away from the slope (i.e., pointing downslope), the friction would be transformed into a force to the fullest possible extent (i.e., effective friction = 1/(stated friction)). Between the two, intermediate values would occur. Moving

¹² Azimuths express directions in degrees, clockwise from north. In IDRISI, it is also permissible to express an azimuth with the value of -1 to indicate that no direction is defined.

¹³ To undertake a process similar to RESULTANT, DECOMP is used to decompose all force/friction image pairs acting upon an area into their X and Y components. These X and Y component images are then added to yield a resulting X and Y pair. The recombination option of DECOMP is then used with these to produce a resultant magnitude/direction image pair.

progressively in a direction farther away from the maximum friction, the friction would progressively decrease until one reached 90°. At 90°, the effect of the slope would be neutralized (effective friction = 1). Then as one moves past 90° towards the opposite direction, frictions would become forces progressively increasing to the extreme at 180°.

This variation in the effective friction/force as a function of direction is here called the anisotropic function. With VARCOST, the following default function is used:

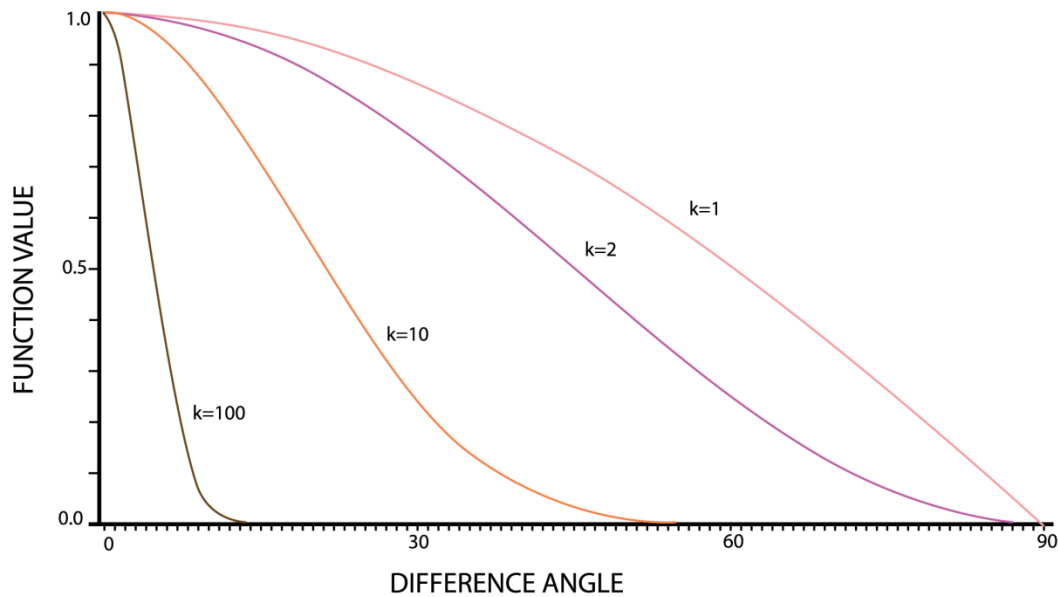
$$\text{effective_friction} = \text{stated_friction}^f$$

$$\text{where } f = \cos^k \alpha$$

and k = a user – defined coefficient

and α = difference angle

The difference angle in this formula measures the angle between the direction being considered and the direction *from which* frictions are acting (or equivalently, the direction *to which* forces are acting). The figure indicates the nature of this function for various exponents (k) for difference angles from 0 to 90°.



You will note in the figure that the exponent k makes the function increasingly direction-specific. At its limit, an extremely high exponent would have the effect of causing the friction to act fully at 0°, to become a fully acting force at 180°, and to be neutralized at all other angles. The default anisotropic function returns negative values for all difference angles from 90° to 270° regardless of the exponent used (i.e., negative cosine values, when raised to odd *or even* exponents, return negative values for the function). Hence, these angles always yield effective friction values that are less than one (i.e., act as forces).

We have not presumed that this function will be appropriate in all circumstances. As a result, we have provided the option of entering a user-defined function. The procedure for doing so is quite simple— VARCOST has the ability to read a data file of function values for difference angles from 0-360° in increments of 0.05°. The format for this file is indicated in the VARCOST module description in the Help System. The important thing to remember, however, is that with VARCOST, the values of that function represent an exponent as follows:

$$\text{effective_friction} = \text{stated_friction}^f$$

where f = a user – defined function

With DISPERSE, the same general logic applies to its operation except that the anisotropic function is different:

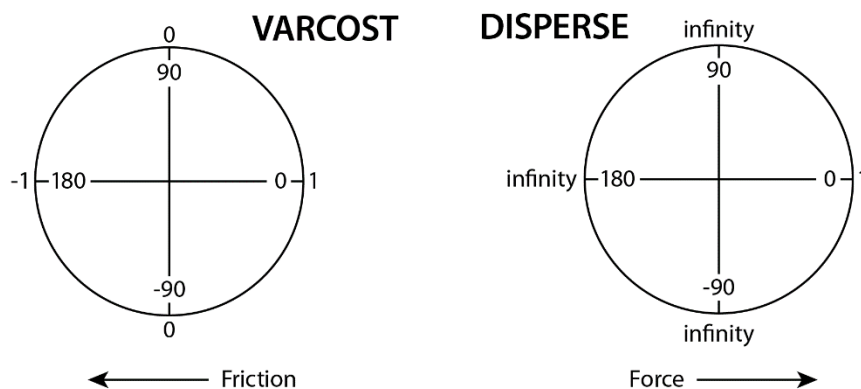
$$\text{effective_friction} = \text{stated_friction} \times f$$

where $f = 1/\cos^k \alpha$

and k = a user – defined coefficient

and α = difference angle

The effect of this function is to modify frictions such that they have full effect at an angle of 0° with progressive increases in friction until they reach infinity at 90° . The function is designed so that effective frictions remain at infinity for all difference angles greater than 90° . The figure shows the values returned by the default functions of f , illustrating this difference between the functions of VARCOST and DISPERSE.



Like VARCOST, DISPERSE also allows the entry of a user-defined function. The procedure is identical, allowing for the reading of a data file containing function values for difference angles from 0 - 360° in increments of 0.05° . The format for this file is indicated in the DISPERSE module description in the Help System. Unlike VARCOST, however, the values of that function represent a multiplier (rather than an exponent) as follows:

$$\text{effective_friction} = \text{stated_friction} \times f$$

where f = a user – defined function

Applications of VARCOST and DISPERSE

VARCOST and DISPERSE have proven useful in a variety of circumstances. VARCOST is a direct extension of the logic of the COST module (i.e., as a means of gauging the effects of frictions and forces on the costs of movement through space, with the special additional capability to moderate frictional effects with varying directions of movement through cells). One might use VARCOST, for example, along with ALLOCATE, to assign villages to rural health centers where the costs of travel on foot are accommodated given landuse types (an isotropic friction) and slopes (an anisotropic friction).

DISPERSE is useful in cases where the phenomenon under study has no motive force of its own, but moves due to forces that act upon it. Potential applications might include point source pollution studies, forest and rangeland fire modeling, and possibly oil spill monitoring and projection.

We encourage users to share with us their experiences using these modules and how they might be changed or augmented to facilitate such studies. We would also welcome the submission of user-defined anisotropic functions that meet the needs of special applications and might be useful to a broader user group.

Surface Interpolation

In GIS, we often want to combine information from several layers in analyses. If we only know the values of a selection of points and these sample points do not coincide between the layers, then such analyses would be impossible. Even if the sample points do coincide, we often want to describe a process for *all* the locations within a study area, not just for selected points. In addition, we need full surfaces because many processes modeled in GIS act continuously over a surface, with the value at one location being dependent upon neighboring values.

Any GIS layer, whether raster or vector, that describes *all* locations in a study area might be called a surface. However, in surface analysis, we are particularly interested in those surfaces where the attributes are quantitative and vary continuously over space. A raster Digital Elevation Model (DEM), for instance, is such a surface. Other example surfaces might describe NDVI, population density, or temperature. In these types of surfaces, each pixel may have a different value than its neighbors.

A landcover map, however, would not be considered a surface by this definition. The values are qualitative, and they also do not vary continuously over the map. Another example of an image that does not fit this particular surface definition would be a population image where the population values are assigned uniformly to census units. In this case, the data are quantitative, yet they do not vary continuously over space. Indeed, change in values is present only at the borders of the census units.

No GIS surface layer can match reality at every scale. Thus the term *model* is often applied to surface images. The use of this term indicates a distinction between the surface as represented digitally and the actual surface it describes. It also indicates that different models may exist for the same phenomenon. The choice of which model to use depends upon many things, including the application, accuracy requirements, and availability of data.

It is normally impossible to measure the value of an attribute for every pixel in an image. (An exception is a satellite image, which measures average reflectance for every pixel.) More often, one needs to fill in the gaps between sample data points to create a full surface. This process is called interpolation. TerrSet offers several options for interpolation which are discussed in this chapter. Further technical information about these modules may be found in the Help System.

Surface Interpolation

The choice of interpolation technique depends on what type of surface model you hope to produce and what data are available. In this section, the techniques available in TerrSet are organized according to input sample data type—points or lines. A description of the algorithm used and the general characteristics of the techniques are given. For a more theoretical treatment of the characteristics of surface models produced by particular interpolation techniques, consult the references provided at the end of this chapter.

Interpolation techniques may be described as global or local. A global interpolator derives the surface model by considering all the data points at once. The resulting surface gives a "best fit" for the entire sample data set, but may provide a very poor fit in particular locations. A local interpolator, on the other hand, calculates new values for unknown pixels by using the values of known pixels that are nearby. Interpolators may define "nearby" in various ways. Many allow the user to determine how large an area or how many of the nearest sample data points should be considered in deriving interpolated values.

Interpolation techniques are also classified as exact or inexact. An exact interpolation technique always retains the original values of the sample data points in the resulting surface, while an inexact interpolator may assign new values to known data points.

Interpolation from Point Data

Trend Surface Analysis

Trend surfaces are typically used to determine whether spatial trends exist in a data set, rather than to create a surface model to be used in further analyses. Trend surfaces may also be used to describe and remove broad trends from data sets so more local influences may be better understood. Because the resulting surface is an ideal mathematical model, it is very smooth and is free from local detail.

In TerrSet, the module TREND is used to produce a trend surface image from sample data points. TREND is a global interpolator since it calculates a surface that gives the best fit, overall, to the entire set of known data points. TREND is also an inexact interpolator. The values at known data points may be modified to correspond to the best fit surface for the entire data set.

TREND fits up to a 9th order polynomial surface model to the input point data set. To visualize how TREND works, we will use an example of temperature data at several weather stations. The linear surface model is flat (i.e., a plane). Imagine the temperature data as points floating above a table top. The height of each point above the table top depends on its temperature. Now imagine a flat piece of paper positioned above the table. Without bending it at all, one adjusts the tilt and height of the paper in such a way that the sum of the distances between it and every point are minimized. Some points would fall above the plane of the paper and some below. Indeed, it is possible that no points would actually fall on the paper itself. However, the overall separation between the model (the plane) and the sample data points is minimized. Every pixel in the study area could then be assigned the temperature that corresponds to the height of the paper at that pixel location.

One could use the same example to visualize the quadratic and cubic trend surface models. However, in these cases, you would be allowed to bend the paper (but not crease it). The quadratic surface allows for broad bends in the paper while the cubic allows even more complex bending.

TREND operates much like this analogy except a polynomial formula describing the ideal surface model replaces the paper. This formula is used to derive values for all pixels in the image. In addition to the interpolated surface produced, TREND reports (as a percentage) how well the chosen model fits the input points. TREND also reports the F-ratio and degrees of freedom, which may be used to test if the modeled trend is significantly different from zero (i.e., no trend at all).

Thiessen or Voronoi Tessellation

The term *tessellation* means to break an area into pieces or tiles. With a Thiessen tessellation, the study area is divided into regions around the sample data points such that every pixel in the study area is assigned to (and takes on the value of) the data point to which it is closest.

Because it produces a tiled rather than a continuous surface, this interpolation technique is seldom used to produce a surface model. More commonly it is used to identify the *zones of influence* for a set of data points.

Suppose a set of new health centers were proposed for a rural area and its inhabitants needed to be assigned to their *closest* facility. If Euclidean distance was used as the definition of *closest*, then THIESSEN would provide the desired result. Zones of influence that are based on more complex variables than Euclidean distance may also be defined in TerrSet using the COST and ALLOCATE modules in sequence. In the same example, if shortest travel time rather than shortest euclidean distance defined *closest*, then COST would be used to develop a travel-time surface (incorporating information about road types, paths, etc.) and ALLOCATE would be used to assign each pixel to its nearest facility in terms of shortest travel time.

Distance-Weighted Average

The distance-weighted average preserves sample data values and is therefore an exact interpolation technique. In TerrSet, it is available in the module INTERPOL (also known as IDW, for Inverse Distance Weighting).

The user may choose to use this technique either as a global or a local interpolator. In the global case, all sample data points are used in calculating all the new interpolated values. In the local case, only a subset of the points is used. This can be controlled by specifying either the number of nearest control points to use, or a search radius within which all included control points are used. The local option is generally recommended, unless data points are very uniformly distributed and the user wants a smoother result.

With either the global or local implementation, the user can define how the influence of a known point varies with distance to the unknown point. The idea is that the attribute of an interpolated pixel should be most similar to that of its closest known data point, a bit less similar to that of its next closest known data point, and so on. Most commonly, the function used is the inverse square of distance ($1/d^2$, where d is distance).

For every pixel to be interpolated, the distance to every sample point to be used is determined and the inverse square of the distance is computed. Each sample point attribute is multiplied by its respective inverse square distance term and all these values are summed. This sum is then divided by the sum of the inverse square distance terms to produce the interpolated value.

The user may choose to use an exponent other than 2 in the function. Using an exponent greater than 2 causes the influence of the closest sample data points to have relatively more weight in deriving the new attribute. Using an exponent of 1 would cause the data points to have more equal influence on the new attribute value.

The distance-weighted average will produce a smooth surface in which the minimum and maximum values occur at sample data points. In areas far from data points, the surface will tend toward the local average value, where *local* is determined by the search radius. The distribution of known data points greatly influences the utility of this interpolation technique. It works best when sample data are many and are fairly evenly distributed.

Triangulated Irregular Networks

A Triangulated Irregular Network, or TIN, is a vector data structure. The sample data points become the vertices of a set of triangular facets that completely cover the study area. In TerrSet, the TIN is generated and then used to create a continuous raster surface model. The section **Triangulated Irregular Networks and Surface Generation** is devoted to this set of procedures.

Kriging and Simulation

Continuous surfaces can also be derived from point data using geostatistical techniques. Various kriging options are offered in TerrSet through three interfaces to the Gstat¹⁴ software package: Spatial Dependence Modeler, Model Fitting, and Kriging and Simulation. Like the techniques offered in INTERPOL, kriging methods may be used either as global or local interpolators. However, the local implementation is most often used. Kriging preserves sample data values and is therefore an exact interpolator. Simulation does not preserve sample data values, making it an inexact interpolator.

The main difference between kriging methods and a simple distance-weighted average is that they allow the user great flexibility in defining the model to be used in the interpolation for a particular data set. These customized models are better able to account for changes in spatial dependence across the study area. Spatial dependence is simply the idea that points that are closer together have more similar values than points that are further apart. Kriging recognizes that this tendency to be similar to nearby points is not restricted to a Euclidean distance relationship and may exhibit many different patterns.

The kriging procedure produces, in addition to the interpolated surface, a second image of variance. The variance image provides, for each pixel, information about how well the interpolated value fits the overall model that was defined by the user. The variance image may thereby be used as a diagnostic tool to refine the model. The goal is to develop a model with an even distribution of variance that is as close as possible to zero.

Kriging produces a smooth surface. Simulation, on the other hand, incorporates per-pixel variability into the interpolation and thereby produces a rough surface. Typically, hundreds of such surfaces are generated and summarized for use in process modeling.

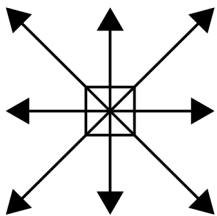
The geostatistical tools provided through TerrSet's interfaces to Gstat are discussed in greater detail in the section **Geostatistics**.

Interpolation from Isoline Data

Sometimes surfaces are created from isoline data. An isoline is a line of equal value. Elevation contours are one example of isolines. Isolines are rarely field measurements; they are more likely the result of digitizing paper maps. One must be aware that the methods involved in creating the isolines may have already included some sort of interpolation. Subsequent interpolation between isolines adds other types of error.

Linear Interpolation from Isolines

A linear interpolation between isolines is available in TerrSet through both the INTERCON and TIN modules. The INTERCON procedure is older and is maintained only for backward compatibility. TIN is described in the next section, while INTERCON is described here. The isolines must first be rasterized, with the attributes of the pixels representing isolines equal to the isoline value. It is also possible to add points of known value prior to interpolation. It is perhaps more useful, however, to add in lines that define ridges, hill crests or other such break features that are not described by the original isoline data set.



¹⁴ Gstat, © Edzer Pebesma, is licensed freeware available from GNU. See the Help System for more details.

In the interpolation, four lines are drawn through a pixel to be interpolated, as shown in the figure. The lines are extended until they intersect with a pixel of known value in each direction. The slope along each of the four lines is calculated by using the attributes of the intersected pixels and their X,Y coordinates. (Slope is simply the change in attribute from one end of the line to the other, divided by the length of the line.) The line with the greatest slope is chosen and is used to interpolate the unknown pixel value.¹⁵ The value at the location of the pixel to be interpolated is calculated based on the attribute values of the intersected pixels, the slope of the line, and the X,Y position of the pixel to be interpolated. This process is carried out for all unknown pixels.

Choice of resolution when the isolines are rasterized is crucial. If the resolution is too coarse, more than one line may rasterize into a single pixel. In this case, only the latter value is retained and a poor interpolation will result. It is recommended that one set the initial resolution to be equal or less than the distance between the closest isolines. A coarser resolution surface can be generated after the initial interpolation using RESAMPLE or CONTRACT. Note that one can easily produce a surface with more apparent detail than is actually present in the isoline data. Del Barrio et al (1992) present a quantitative method for determining a resolution that captures the optimum information level achievable given the characteristics of the input isoline data.

Linear interpolation from isolines may produce some obvious and undesirable artifacts in the resulting surface. A histogram of a surface produced by this interpolation technique tends to show a "scaloped" shape, with histogram peaks at the input isoline values. In addition, star-shaped artifacts may be present, particularly at peaks in the surface. These characteristics can be mitigated to some degree (but not removed) by applying a mean filter (with the FILTER module). Finally, hill tops and valley bottoms will be flat with the value of the enclosing contour. In many cases, if isoline data are available, the constrained and optimized TIN method described below will produce a better surface model.

INTERCON is an exact interpolator, since isolines retain their values. It could also be termed a local interpolator, though the isolines used to interpolate any particular pixel may be quite distant from that pixel.

Constrained Triangulated Irregular Networks

As discussed above, triangulated irregular networks may be generated from point data. In addition, the TerrSet TIN module allows for input of isoline data for TIN creation. In doing so, the TIN can be constrained so no triangular facet edge crosses an isoline. This forces the triangulation to preserve the character of the surface as defined by the isolines. A TIN developed from isolines can also be optimized to better model features such as hill tops and valley bottoms. Once the TIN is developed, it may be used to generate a raster surface model with the module TINSURF.

All the steps involved in this process are detailed in the chapter **Triangulated Irregular Networks and Surface Generation** which follows shortly.

Choosing a Surface Model

No single surface generation method is better than others in the abstract. The relative merit of any method depends upon the characteristics of the input sample data and the context in which the surface model will be used. The precision of sample point measurements, as well as the frequency and distribution of sample points relative to the needed scale of variation, influence the choice of interpolation technique to apply to those data. In addition, the scale of the processes to be modeled is key in guiding the creation of an interpolated surface model. Surface shape (e.g., convexity, concavity) and level of local variation are often key aspects of process models, where the value or events in one pixel influence those of the neighboring pixels. It is not unusual to develop several surface models and use each in turn to assess the sensitivity of an analysis to the type of surface generation techniques used.

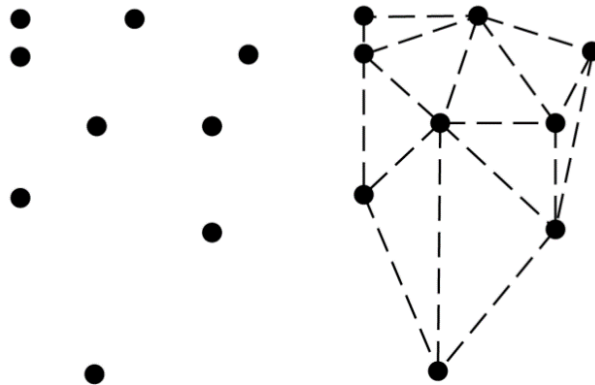
¹⁵ The line of greatest slope is used to avoid flat lines that result when a line intersects the same isoline on both ends. This is quite common with topographic maps and would lead to an abundance of flat areas in the interpolated surface.

References / Further Reading

- Blaszczynski, J., 1997. Landform Characterization With Geographic Information Systems, *Photogrammetric Engineering and Remote Sensing*, 63(2): 183-191.
- Burrough, P., and McDonnell, R., 1998. *Principles of Geographical Information Systems*, 98-161, Oxford University Press, London.
- del Barrio, G., Bernardo, A., and Diez, C., 1992. The Choice of Cell Size in Digital Terrain Models: An Objective Method, Conference on Methods of Hydrologic Comparison, Oxford, UK, September 29-October 20.
- Desmet, J., 1997. Effects of Interpolation Errors on the Analysis of DEMs, *Earth Surface Processes and Landforms*, 22: 563-580.
- Lam, N., 1983. Spatial Interpolation Methods: A Review, *The American Cartographer*, 10(2): 129-149

Triangulated Irregular Networks and Surface Generation

Triangulated Irregular Networks (TINs) are the most commonly-used structure for modeling continuous surfaces using a vector data model. They are also important to raster systems because they may be used to generate raster surface models, such as DEMs. With triangulation, data points with known attribute values (e.g., elevation) are used as the vertices (i.e., corner points) of a generated set of triangles. The result is a triangular tessellation of the entire area that falls within the outer boundary of the data points (known as the *convex hull*). The figure below illustrates a triangulation from a set of data points.

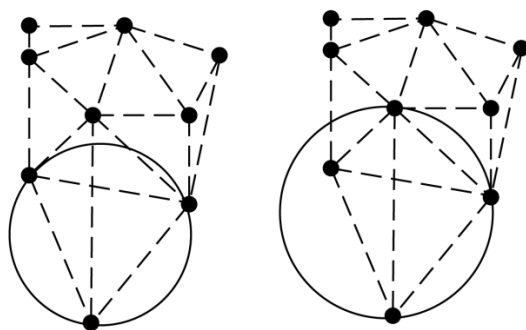


A set of data points (left) and a triangulation of those data points (right).

There are many different methods of triangulation. The Delaunay triangulation process is most commonly used in TIN modeling and is that which is used by TerrSet. A Delaunay triangulation is defined by three criteria: 1) a circle passing through the three points of any triangle (i.e., its circumcircle) does not contain any other data point in its interior, 2) no triangles overlap, and 3) there are no gaps in the triangulated surface. The figure below shows examples of Delaunay and non-Delaunay triangulations.

A natural result of the Delaunay triangulation process is that the minimum angle in any triangle is maximized. This property is used by the TerrSet algorithm in constructing the TIN. The number of triangles (Nt) that make up a Delaunay TIN is $Nt=2(N-1)-Nh$, and the number of edges (Ne) is $Ne=3(N-1)-Nh$, where N is the number of data points, and Nh is the number of points in the convex hull.

TerrSet includes options for using either true point data or vertex points extracted from isolines¹⁶ as input for TIN generation. The TIN module also offers options to use non-constrained or constrained triangulation, to optimize the TIN by removing “tunnel” and “bridge” edges, and to generate a raster surface from the TIN by calling the module TINSURF. Modules are also available for preparing TIN input data. These are all discussed in detail below.



Delaunay triangulation (left) and non-Delaunay triangulation (right). The shaded triangle doesn't meet the empty circumcircle criterion.

In TerrSet, the TIN file structure consists of a vector line file (containing the triangle edges) and an associated ASCII TIN file (containing information indicating which points make up each triangle). File structure details may be found in the Help System.

Preparing TIN Input Data

Points

Normally there will be little data preparation necessary when point data is used to create a TIN. In some cases it may be desirable to reduce the number of points to be used in the triangulation. For example, if the number and density of the points exceeds the required accuracy of the TIN, the user may choose to remove points since fewer points will lead to faster processing of the TIN. The module GENERALIZATION offers this point-thinning capability.

If point data are used as input to TIN generation, only the non-constrained triangulation option, described below, is available. If isoline data are used as input to TIN generation, both the non-constrained and constrained options are available and a better TIN result can be expected. If a raster surface is the desired final output of input point data, the INTERPOL module and the TerrSet interfaces to Gstat offer alternatives to TIN/TINSURF. (See the chapters **Surface Analysis** and **Geostatistics**.)

¹⁶ In this chapter, the term isoline refers to any line representing a constant attribute value. Elevation contours are one example of isolines.

Lines

When an isoline file is used as input to TIN, only the vertices¹⁷ that make up the lines are used in the triangulation. It may be useful to examine the density of the vertices in the isoline file prior to generating the TIN. The module GENERALIZATION may be used to extract the vertices of a line file to a vector point file for visualization.

It may be desirable to add points along the lines if points are so far apart they create long straight-line segments that result in large TIN facets. Point thinning along lines is also sometimes desirable, particularly with isoline data that was digitized in stream mode. In this case, the number of points in the lines may be much greater than that necessary for the desired resolution of the TIN, and thus will only serve to slow down the TIN generation process. The module TINPREP performs along-line point addition or thinning. Other line generalization options are also available in the GENERALIZATION module.

If line data are used to create a TIN, both the non-constrained and the constrained triangulation options are available. The differences between these options are described below. If a raster surface is the desired final output of input isoline data, the module INTERCON offers an alternative to TIN/TINSURF. However, the latter normally produces a superior result.

Command Summary

Following is a list and brief description of the modules mentioned in this section.

GENERALIZATION thins or "generalizes" point vector data, extracts the vertices (points) from a line vector to a point vector file, and generalizes line vector data..

INTERPOL interpolates a raster surface from point data.

TerrSet interfaces to Gstat provide geostatistical tools that can be used to create a raster surface from point data.

TINPREP adds or thins vertices along vector lines.

INTERCON interpolates a raster surface from rasterized isoline data.

TIN creates a TIN from point or line vector data. TINSURF may be automatically called from the TIN dialog if a raster surface output is desired.

TINSURF creates a raster surface from an existing TIN.

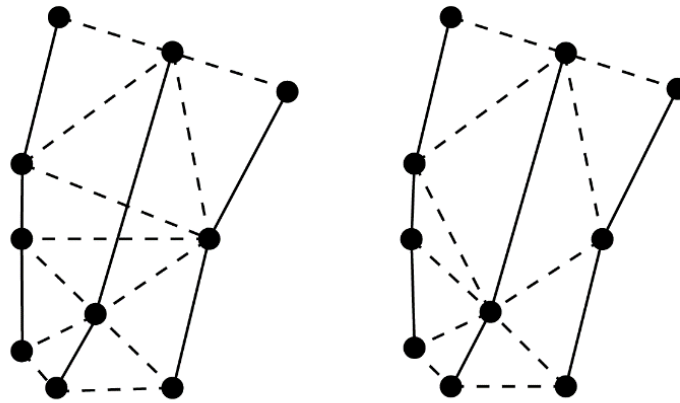
Non-Constrained and Constrained TINs

The non-constrained Delaunay triangulation is described in the Introduction section above and is implemented in the TerrSet TIN module using an algorithm designed for speed of processing. First, the set of input points (or isoline vertices) are divided into sections. Then each of the sections is triangulated. The resulting "mini-TINs" are then merged together. A local optimization procedure is always implemented during the merging process to maximize the minimum angles and thus satisfy Delaunay criteria for the triangulation.

A constrained Delaunay triangulation is an extension of the non-constrained triangulation described above, with additional conditions applied to the selection of triangle vertices. In TerrSet, the constrained Delaunay triangulation uses isolines as non-crossing break-line constraints to control the triangulation process. This process ensures that triangle edges do not cross isolines and

¹⁷ In this context, the term "vertices" refers to all the points that make up a line, including the beginning and ending points.

that the resulting TIN model is consistent with the original isoline data. Not all triangles will necessarily meet the Delaunay criteria when the constrained triangulation is used.

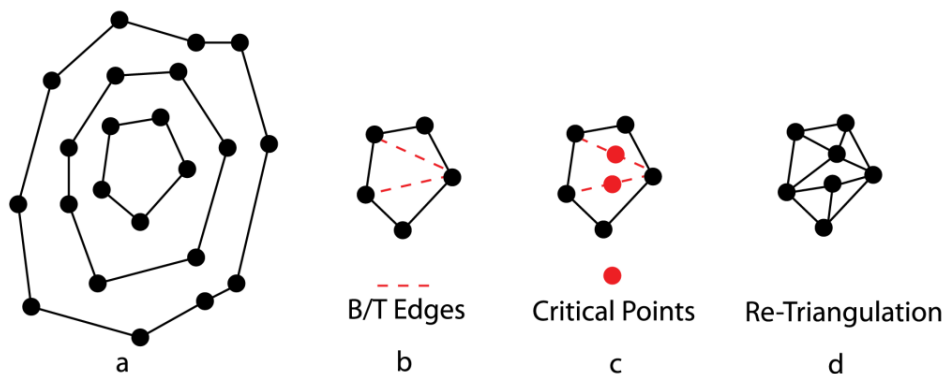


Unconstrained (left) and constrained (right) Delaunay triangulations. Solid lines represent isolines.

In TerrSet, the constrained TIN is created in a two-step process. First, a non-constrained triangulation is completed. Then triangle edges are checked for isoline intersections. When such an intersection is encountered, a local optimization routine is again run until no isoline intersections remain.

The figure below shows constrained and unconstrained TINs created from the same set of isoline vertex data points.

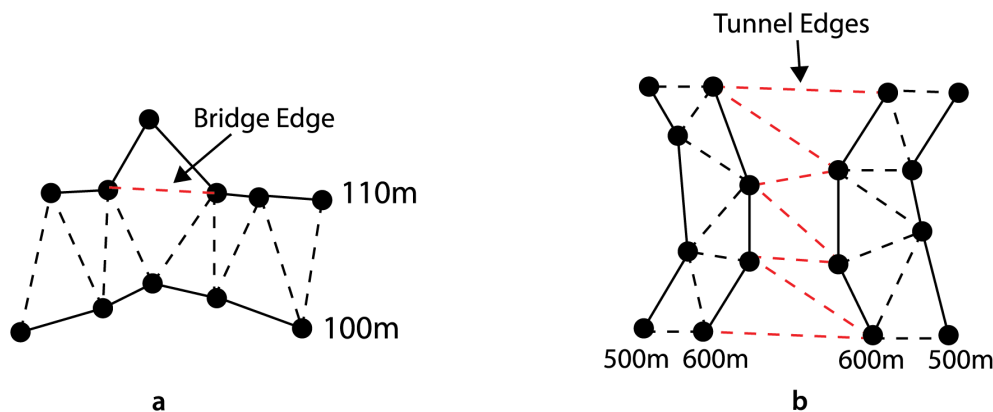
Removing TIN “Bridge” and “Tunnel” Edges



a) Contours at the top of a hill; b) triangulation of highest contour, with B/T edge identified; c) placement of critical points on B/T edges; d) re-triangulation

Contour lines at the top of a hill are shown in the figure below (a). In (b), the highest contour is shown along with the resulting triangles created within it when a constrained TIN is generated. Because all three of the points for all of the triangles have the same elevation, the top of the hill is perfectly flat in the TIN model. Our experience with actual terrain tells us that the true surface is probably not flat, but rather rises above the TIN facets. The edges of the TIN facets that lie below the true surface in this case are examples of what are called “tunnel edges”. These are identified in (b). A tunnel edge is any triangle edge that lies below the true surface. Similarly, if the contours of (a) represented a valley bottom or depression, the TIN facets shown in (b) would describe a flat surface that is higher than the true surface. The edges of the TIN facets that lie above the true surface would then be termed “bridge edges”.

Bridge and tunnel (B/T) edges are not restricted to hill tops and depression bottoms. They can also occur along slopes, particularly where isolines are undulating, and along ridges or channels. Two such examples are shown in the figure below.



a) Contours at a stream; b) contours at a “saddle” feature. Contours are shown with solid line, constrained triangle edge with dashed lines. B/T edge are shown in red.

To optimize a TIN, B/T edges may be removed. B/T edge removal could technically be performed on an unconstrained TIN, but this is not recommended and is not allowed in TerrSet. An optimal TIN will be generated if isolines are used as the original input for TIN generation, the constrained triangulation is used, and B/T edges are removed.

While many of the concepts of this section are illustrated with elevation data, the procedures are not limited to such data.

Bridge and Tunnel Edge Removal and TIN Adjustment

The TerrSet TIN module includes an option to create a TIN with all B/T edges removed. This option is only available if isoline data and the constrained triangulation option are used. First, a normal TIN is created from the vector input data. Then, all of the B/T edges in the TIN are identified. In TerrSet, a B/T edge is defined as any triangle edge with endpoints of the same attribute, where these endpoints are not neighboring points on an isoline.

New points, termed *critical points*, are created at the midpoints of the B/T edges (figure c below). The areas around the critical points are then re-triangulated (figure d above). When a B/T edge is shared by two triangles, four new triangles result. When a B/T edge is part of the TIN boundary, and is thus used by only one triangle, two new triangles result.

Once the critical points have been placed and the triangulation has been adjusted, the next step is to assign appropriate attribute values (e.g., elevations) to these new points.

Attribute Interpolation for the Critical Points

In TerrSet, the recommended method for determining the attribute of a critical point uses a parabolic shape. The parabola, as a second-order non-linear polynomial method, was chosen because it combines computational simplicity and a shape that is compatible with most topographic surfaces.¹⁸ Before describing the mathematical details and the algorithm of the calculation of critical point values, we will use an illustration to think through the general logic.

▪ *General Logic*

Let us assume that the contours of figure a above describe a hill and that the hilltop beyond the highest contour has a somewhat rounded peak. Given this, we could imagine fitting a parabolic surface (like an inverted, U-shaped bowl) to the top of the hill. The particular parabolic surface we would choose would depend on the shape of the nearby terrain. If slopes were gentle leading up to the highest contour, then we would choose a surface with gently sloping sides and a wide top. But if slopes were quite steep, we would choose a surface with more vertical sides and a narrower top. Once a surface was chosen, all critical points on the tunnel edges at the top of the hill could be projected onto the parabolic surface. They could then each be assigned the elevation of the surface at their location.

The actual implementation of the interpolation differs from the general logic described above in that two-dimensional parabolas are used rather than parabolic surfaces. Up to eight parabolas, corresponding to eight directions, are fit through each critical point location. An attribute for the critical point is derived for each parabola, and the final attribute value assigned to the point is their average. Details of the process are given below.

▪ *Calculating the Critical Point Attribute*

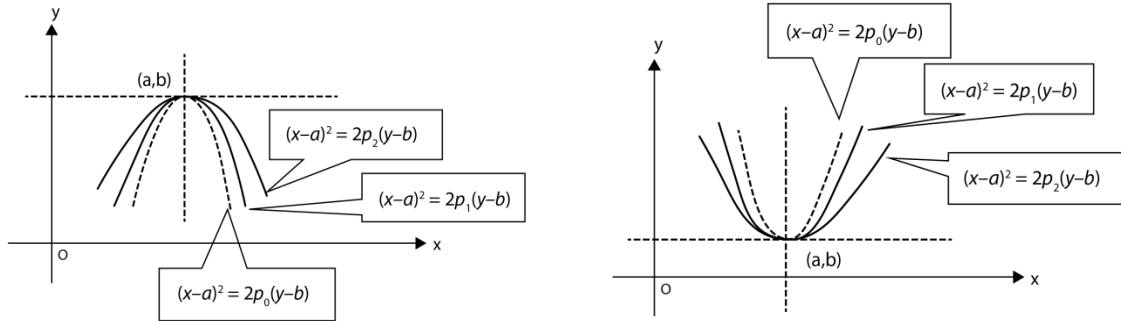
A parabola is defined by the following equation:

$$(X - a)^2 = 2p(Y - b)$$

Where the point (a,b) defines the center (top or bottom) point of the parabola and the parameter p defines the steepness of the shape. When p is positive, the parabola is U-shaped. When p is negative, the parabola is inverted. The larger the absolute value of p, the wider the parabola.

The figure below shows several parabolas and their equations.

¹⁸ Although the parabolic algorithm is recommended, linear and optimized linear options are also available as critical point interpolation methods in the module TIN. In the example of the hilltop, a sharp peak would be modeled by the linear method in contrast to the rounded peak of the parabolic method. The optimized linear method uses a linear interpolation unless slopes in all eight directions (see the discussion of the parabolic interpolation) are zero, in which case it uses the parabolic.



Example parabolas and their equations. On the left, p is negative. On the right, p is positive.

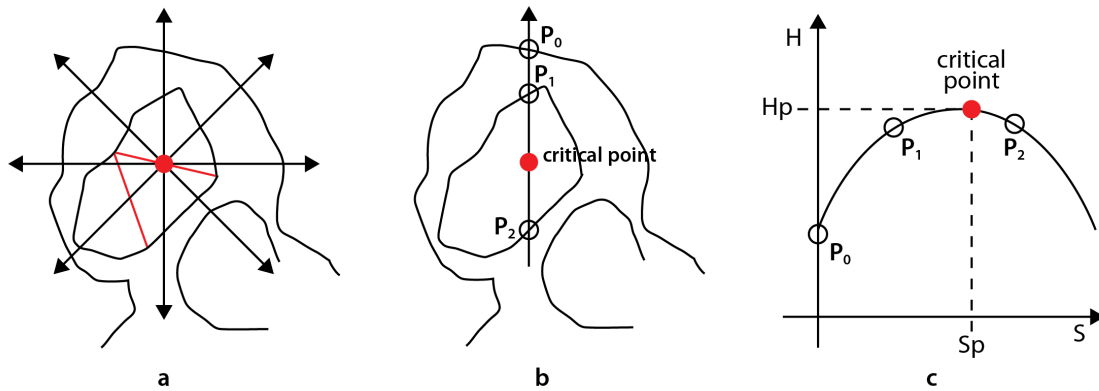
To translate the general parabolic equation to the critical point attribute interpolation problem, we re-label the axes in the above figure from X,Y to S,H where S represents distance from the origin (o) and H represents the attribute value (e.g., elevation) from the origin (o). (The origin is defined by the location and attribute of the *original point* as described below.) In the example of a critical point on a tunnel edge at the top of a hill, the plane of the parabola is a cross section of the hill.

To define a parabola for a critical point, three points with known coordinates and attributes that lie on that same parabola must be found.¹⁹ Up to eight parabolas, each defined by three points, are developed for each critical point.

For each critical point, a search process is undertaken to find intersections with isolines in each of eight directions, as shown in the figure below. If two intersections are found in each direction, then eight parabolas can be defined. Each is defined by three points, with two points taken from one direction from the critical point and the other one taken from the opposite direction. In (b) in the figure, the intersection points for one search direction, points P_0 , P_1 and P_2 , are used to define the parabola shown in (c). The point that lies between two intersections from the critical point is always termed the *original point* and is labeled P_0 . This point is set at $S=0$, so distances (S) to all other points are measured from this original point. P_1 lies between the critical point and the original point, and P_2 lies on the opposite side of the critical point.

¹⁹ Any parabola can be defined once three points on it are known. Three equations (one for each point) can be written as below. For each, the distance (S) from that point to the origin and the attribute (H) are known. The simultaneous equations can then be solved for a , b , and p .

$$(S_0 - a)^2 = 2p(H_0 - b) \quad (S_1 - a)^2 = 2p(H_1 - b) \quad (S_2 - a)^2 = 2p(H_2 - b)$$



a) Eight-direction search for isoline intersections for one critical point; b) intersection points for one direction; c) parabola derived from intersection points. Attribute (h_p) for the critical point can be found, given the critical point's distance (S_p) from P.

If three intersections are not found for a particular parabola (e.g., at the edge of a coverage), then it is undefined and the number of parabolas used to interpolate the attribute value for that critical point will be fewer than eight.

For each defined parabola, the attribute value of any point on the parabola can be found by entering its distance from the original point into the parabolic equation. The following equation can be used to calculate the attribute of a critical point for one of its parabolas:²⁰

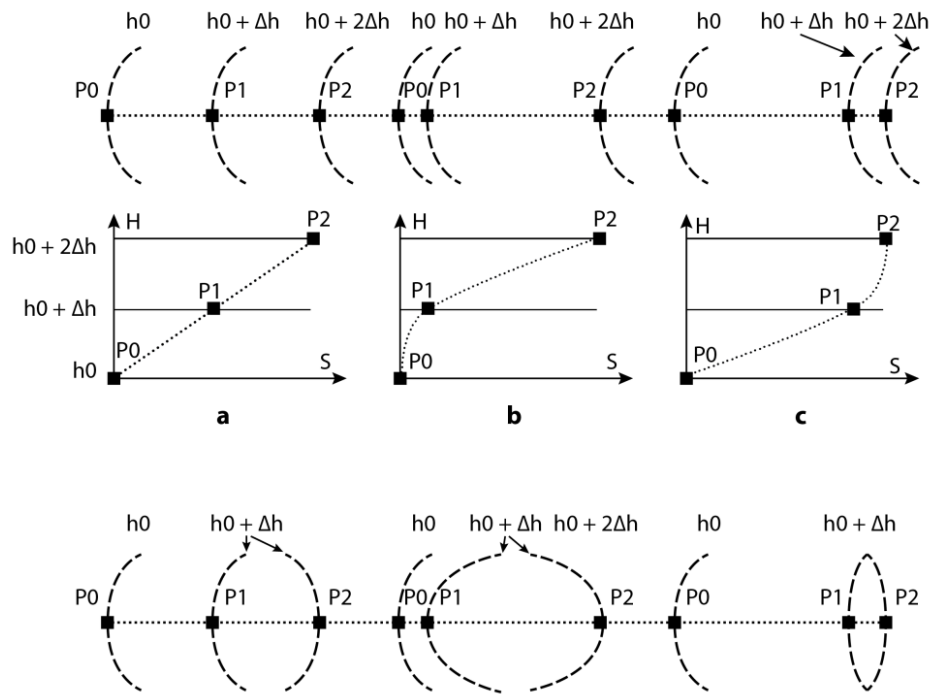
$$H = \sum_{i=0}^2 h_i \cdot \prod_{j=0, j \neq i}^2 \left[\frac{(S_{point} - S_j)}{(S_i - S_j)} \right]$$

Where h_i , $i = 0, 1, 2$ are attribute values of the three intersection points, P_0 , P_1 and P_2 ; S_i , S_j , $i, j = 0, 1, 2$ represent the distances from the original point to the intersection points, and S_{point} represents the distance from the original point to the critical point. According to the above definitions of the intersection points (b in figure below), we know $S_0 = 0$, while $S_1 = P_1P_0$, and $S_2 = P_2P_0$.

For each parabola, the attribute value at the position of the critical point is calculated in this manner. The final attribute value that is assigned to the critical point is the average of all valid interpolated values (invalid cases are discussed below).

The figure below shows several examples of cases in which B/T edges would be identified and a new value for the critical points placed on their midpoints would be interpolated. In each figure, only one search direction is illustrated. In the figures, a, b and c are examples of cases where critical points occur along slopes while d, e and f are cases where critical points occur on hill tops. For cases in which the attribute value of the critical point is lower than those of the surrounding isolines, the curves would be inverted.

²⁰ The equation incorporates the derivation of the parabolic parameters a, b, and p.



Six examples of parabolic critical point interpolation. The upper part of each example illustrates the map view of the isolines (dashed) and the intersection points (P0, P1 and P2) for one search direction (dotted line) for a critical point. The lower part shows the parabola for that set of intersection points. The attribute for the critical point (which is always between P1 and P2) can be found by plotting the critical point on the parabolic curve at its distance (S) from P0.

▪ Invalid Cases

There are two extreme circumstances in which the parabolic interpolation procedure is invalid:

1. If all three intersection points have the same attribute value, the three points are not used for interpolation. An interpolated value for the critical point is therefore not calculated for this direction. The attribute value assigned would be an average of the other interpolated values.
2. If the interpolated value is greater than (in the case of tunnel edges) or less than (in the case of bridge edges) the value of the next expected contour, then the critical point is assigned the value of the next expected contour.²¹ The nature of contour maps requires such a limitation.

²¹ The algorithm uses the local contour interval for each critical point, so isoline data with variable contour intervals do not pose a problem.

Outputs of TIN

The outputs of the TIN module are a vector line file defining the triangle edges, an ASCII .TIN file containing the topological information for the triangulation and, if B/T edge removal was used, a point vector file of the critical points that were added. All these pieces except for the triangle edge vector file, in addition to the original vector data file, are used by the TINSURF module to create a raster surface from the TIN.

Generating a Raster Surface from a TIN

A raster surface may be generated from the TIN at the time the TIN is created or may be created from an existing TIN file later. The TINSURF module creates the raster surface. Its dialog asks only for the TIN file as input. However, the TIN file stores the name of the original vector file used to create the TIN as well as whether B/T edge removal was used. If the TIN is the result of B/T edge removal, then TINSURF also requires the critical point vector file. Therefore you should not delete, move or rename any of these files prior to creating the raster surface.

For each raster pixel in the output image, an attribute value is calculated. This calculation is based on the positions and attributes of the three vertex points of the triangular facet within which the pixel center falls and the position of the pixel center.²² The logic is as follows:

1. Solve the following set of simultaneous equations for A, B and C:

$$H_1 = Ax_1 + By_1 + C$$

$$H_2 = Ax_2 + By_2 + C$$

$$H_3 = Ax_3 + By_3 + C$$

where $H_{1,2,3}$ are the attribute values (e.g., elevations) of the three triangle facet vertices and $(x,y)_{1,2,3}$ are their reference system coordinates.

2. Given A, B and C, as derived above, solve the following for H_p :

$$H_p = Ax_p + By_p + C$$

where H_p is the attribute of the pixel and $(x,y)_p$ is the reference system coordinate of the pixel center.

3. Assign the pixel the attribute value H_p .

The algorithm proceeds on a facet-by-facet basis, so the derivation of A, B, and C in step 1 is carried out only once for all the pixels that fall within a single facet.

Raster Surface Optimization

For optimal generation of a raster surface from a TIN model, care should be taken in preparing the data used to create the TIN. If isoline data is used, the isolines should not cross. The distribution of points in the input vector file should be evaluated visually and adjusted, if necessary, by thinning or adding points. If point attribute values are available at peaks and valleys in the study area,

²² Each pixel center will fall in only one TIN facet, but a single facet may contain several pixel center points.

adding these to the input data will reduce bridge and tunnel edge effects and will enhance the quality of the resulting TIN and the subsequent raster surface.

A TIN will cover only the area inside the convex hull of the data points. This may present a problem if the original vector data does not cover the entire study area. The areas outside the convex hull will not be covered by triangles in the TIN and will be assigned a background value in the resulting raster surface. An option to add corner points is available on the TIN dialog to help mitigate this problem for the corners of the image. However, there may still be areas outside the convex hull even when corner points are added. If possible, it is recommended that the vector point or isoline data used to create the TIN extend beyond the limits of the desired raster study area. Then specify the final raster bounding coordinates in TINSURF. This will produce a TIN that covers the entire rectangular study area and a raster surface that contains no background values.

Further Reading

Lee J., 1991. Comparison of Existing Methods for Building Triangular Irregular Network Models of Terrain From Grid Digital Elevation Models, *International Journal of Geographic Information Systems*, 3: 267-285.

Tsai, V. J. D., 1993. Delaunay Triangulations in TIN Creation: an Overview and a Linear-time Algorithm, *International Journal of Geographic Information Systems*, 6: 501-512.

Zhu, H., Eastman, J. R., and Schneider, K., 1999. Constrained Delaunay Triangulation and TIN Optimization Using Contour Data, *Proceedings of the Thirteenth International Conference on Applied Geologic Remote Sensing*, 2: 373-380, Vancouver, British Columbia, Canada.

Geostatistics

Geostatistics provides tools for the exploration and statistical characterization of sample point data. It also provides a number of techniques for interpolating surfaces from such data. *Ordinary kriging* is the most well-known of these. While the techniques originated with scientists working in the mining industry, a broader audience has been found in those fields in which both data values and their locations are considered analytically important.

Several interpolation techniques were introduced in the section **Surface Interpolation**. Geostatistical techniques are distinct from these in that they provide GIS analysts with the ability to incorporate information about patterns of spatial continuity into the interpolation model as well as to produce surfaces that include elements of local variation. The methods allow for a high degree of user flexibility in detecting and defining structures that describe the nature of a data set. Indeed, a set of structures can be nested, each describing a particular aspect of the data set.

With this flexibility, however, also comes some risk. From the same data set, it is possible to produce many surfaces—all very different, and all seemingly reasonable representations of reality. The new user is encouraged to enter into geostatistics deliberately and with some caution. An understanding of, and respect for, the underlying assumptions of these techniques is essential if the results are to provide meaningful information to any analysis.

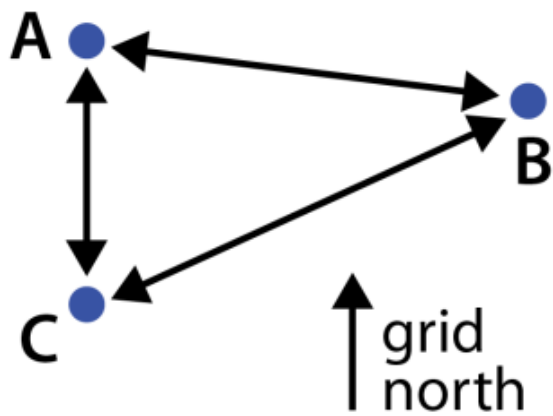
This chapter presents a very brief overview of the geostatistical capabilities offered through TerrSet interfaces to Gstat.²³ For more complete and theoretical treatments of geostatistics, consult the references listed at the end of this chapter. The **Tutorial** includes an extensive exercise illustrating the use of the geostatistical tools available in TerrSet.

²³ IDRISI provides a graphical user interface to Gstat, a program for geostatistical modeling, prediction and simulation written by Edzer J. Pebesma (Department of Physical Geography, Utrecht University). Gstat is freely available under the GNU General Public License from www.gstat.org. Clark Labs' modifications of the Gstat code are available from the downloads section of the Clark Labs Web site at www.clarklabs.org.

Spatial Continuity

The underlying notion that fuels geostatistical methods is quite simple. For continuously varying phenomena (e.g., elevation, rainfall), locations that are close together in space are more likely to have similar values than those that are further apart. This tendency to be most similar to one's nearest neighbors is quantified in geography through measures of spatial autocorrelation and continuity. In geostatistics, the complement of continuity, variability, is more often the focus of analysis.

The first task in using geostatistical techniques to create surfaces is to describe as completely as possible the nature of the spatial variability present in the sample data. Spatial variability is assessed in terms of distance and direction. The analysis is carried out on *pairs* of sample data points. Every data point is paired with every other data point. Each pair may be characterized by its *separation distance* (the Euclidean distance between the two points) and its *separation direction* (the azimuth in degrees of the direction from one point to the other).²⁴ The sample data point set shown below would produce pairs characterized as shown in the table following.



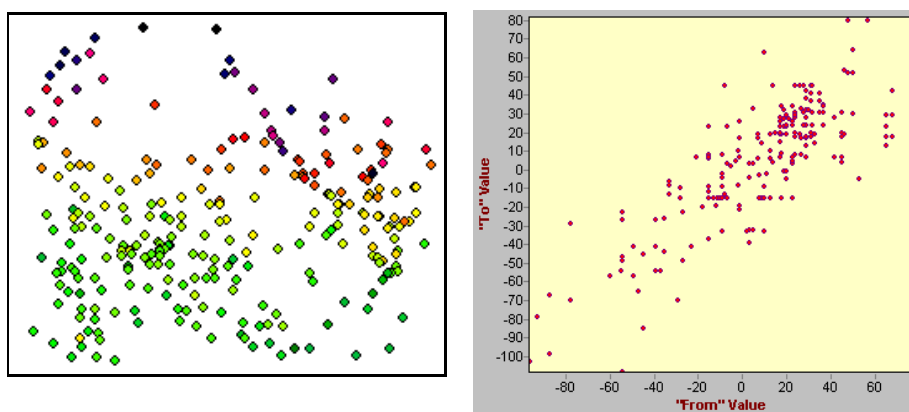
Pair	SeparationDistance	Separation Direction
AB	80 m	100
AC	50 m	0

²⁴ The points in a pair are identified as the *from* point and the *to* point. No pair is repeated.

The distance measure is typically referred to in units of *lags*, where the length of a lag (i.e., the lag distance or lag interval) is set by the user. In specifying a particular lag during the analysis, the user is limiting the pairs under consideration to those that fall within the range of distances defined by the lag. If the lag were defined as 20 meters, for example, an analysis of data at the third lag would include only those data pairs with separation distances of 40 to 60 meters.

Direction is measured in degrees, clockwise from grid north. As with distance, direction is typically specified as a range rather than a single azimuth.

The h-scatterplot is used as a visualization technique for exploring the variability in the sample data pairs. In the h-scatterplot, the X axis represents the attribute at one point of the pair (the from point) and the Y axis represents that same attribute at the other point of the pair (the to point). The h-scatterplot may be used to plot all of the pairs, but is more often restricted to a selection of pairs based on a certain lag and/or direction. The figure shows the spatial distribution of 250 rainfall sample points from a 1000 km² area. These points were paired and data pairs that are within 1 lag (0-1 km) and for all directions are plotted in the h-scatterplot shown in the figure below.



The h-scatterplot is typically used to get a sense of what aspects of the data pair distribution are influencing the summary of variability for a particular lag. H-scatterplots are interpreted by assessing the dispersion of the points. For example, if the pairs were perfectly linearly correlated (i.e., no variability at this separation and direction), then all the points would fall along a line. A very diffuse point pattern in the h-scatterplot indicates high variability for the given ranges of distance and direction. The h-scatterplot is available through the Spatial Dependence Modeler interface.

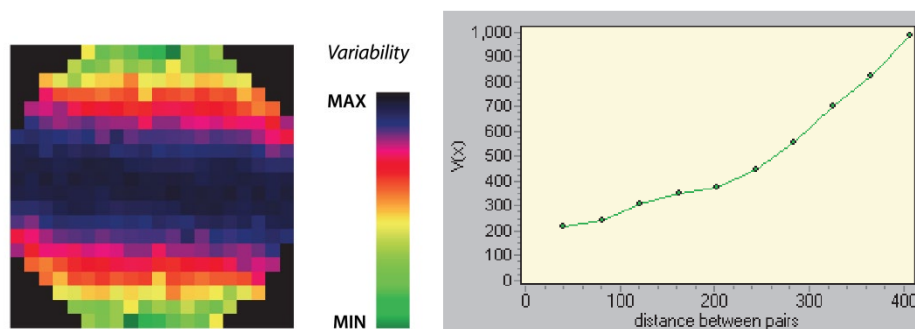
The semivariogram is another tool for exploring and describing spatial variability and is also available through the Spatial Dependence Modeler interface. The semivariogram summarizes the variability information of the h-scatterplots and may be presented both as a surface graph and a directional graph. The surface graph shows the average variability in all directions at different lags. The center position in the graph, called the *origin*, represents zero lags. The lags increase from the center toward the edges. The direction is represented in the surface graph with grid north directly up from the center pixel, 90 degrees directly to the right, and so on.²⁵ The magnitude of variability is represented by color using the default palette. Low values are shown in darker

²⁵ Note that some geostatistical software plot zero degrees to the right rather than the top of the surface graph.

colors and higher values in brighter colors. When one moves the cursor over the surface graph, its location, in terms of direction and distance from the origin, is shown at the bottom of the graph.

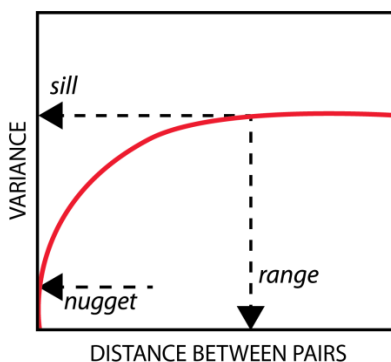
A surface graph semivariogram of the same sample rainfall points from the above figure is shown below. The lag distance is set to 1 km. One can readily see that in the West-East direction, there is low variability among the pairs across all lags. It appears that the direction of minimum variability (i.e., maximum continuity) is approximately 95 (and 275) degrees. We would expect data points that are separated from each other in this direction to have attributes that are more similar than data points separated by the same distance but in a different direction.

The other graphic form of the semivariogram is the directional graph, as shown alongside the surface graph. It is used to develop the structures that describe the patterns of variability in the data. In the directional graph, a single summary point is plotted for each lag. The X-axis shows the separation distance, labeled in reference units (e.g., km), while the Y-axis shows the average variability for the sample data pairs that fall within each lag. All pairs may be considered regardless of direction (an omnidirectional plot), or the plot may be restricted to pairs from a particular range of directions.



Usually one begins with plotting an omnidirectional semivariogram. From the omnidirectional graph, one may gain insight into the overall variability of the data. The user then may create several plots, using different directions and lag distances, to gain a better understanding of the structure of the data set.

The structure of the data may be described by four parameters: the *sill*, the *range*, the *nugget* and *anisotropy*. The first three are labeled in the figure below. In most cases involving environmental data, spatial variability between sample pairs increases as the separation distance increases. Eventually, the variability reaches a plateau where an increase in separation distance between pairs no longer increases the variability between them, i.e., there is no spatial dependence at this and larger distances. The variance value at which the curve reaches the plateau is called the *sill*. The total separation distance from the lowest variance to the sill is known as the *range*. The range signifies the distance beyond which sample data should not be considered in the interpolation process when selecting points that define a local neighborhood.



The *nugget* refers to the variance at a separation distance of zero, i.e., the Y-intercept of the curve that is fit to the data. In theory, we would expect this to be zero. However, noise or uncertainty in the sample data may produce variability that is not spatially dependent and this will result in a non-zero value, or a *nugget effect*. A nugget structure increases the variability uniformly across the entire graph because it is not related to distance or direction of separation.

The fourth parameter that defines the structure is the anisotropy of the data set. The transition of spatial continuity may be equal in all directions, i.e., variation is dependent on the separation distance only. This is known as an *isotropic model*. A model fit to any direction is good for all directions. In most environmental data sets, however, variability is not isotropic. The point data illustrated above, for example, exhibits a minimum direction of variability in the West-East direction. In any other direction, variability increases more rapidly at the same separation distance. This type of data requires an *anisotropic model*. Anisotropy is described by directional axes of minimum and maximum continuity. To determine the parameters to be used, the user views directional semivariograms for multiple directions.

In kriging and simulation interpolation processes, structures that describe the pattern of spatial variability represented by directional semivariograms are used to determine the influence of spatial dependence on neighborhoods of sample points selected to predict unknown points. The structures influence how their attributes should be weighted when combined to produce an interpolated value. Semivariograms, however, because they are based on the inherent incompleteness of sample data, need smoother curves that define the shape of the spatial variability across all separation distances. Using ancillary information and the semivariograms, mathematical functions are combined to delineate a smooth curve of spatial variability. At this stage, a nugget structure, and sills, ranges, and anisotropies of additional structures are defined for the smooth curve. The Model Fitting interface offers several mathematical functions that may be used to design a curve for the spatial variability. Those functions that do not plateau at large separation distances, such as the linear and the power functions, are termed non-transitional. Those that do reach a plateau, such as the gaussian and exponential functions, are called transitional functions.

Together, the nugget structure, and the sills, ranges, and anisotropies of additional structures mathematically define a nested model of spatial variability. This is used when locally deriving weights for the attributes of sample data within the neighborhood of a location to be interpolated. Using the Spatial Dependence Modeler interface, one unearths a pattern of spatial variability through the plotting of many variograms until a representative semivariogram can be determined. Through the Model Fitting interface, the user fits a mathematical curve described by sills, ranges, a nugget, anisotropy and selected functions to the detected spatial variability. This curve is used to derive the weights applied to locally selected samples during the interpolation by kriging or conditional simulation.

Semivariograms are statistical measures that assume the input sample data are normally distributed and that local neighborhood means and standard deviations show no trends. Each sample data set must be assessed for conformity to these assumptions. Transformations of the data, editing of the data set, and the selection of different statistical estimators of spatial variability are all used to cope with data sets that diverge from the assumptions.

The ability to identify true spatial variability in a data set depends to a great extent on ancillary knowledge of the underlying phenomenon measured. This detection process can also be improved with the inclusion of other attribute data. The crossvariogram, like the semivariogram, plots variability along distances of joint datasets and uses one set of data to help explain and improve the description of variability in another. For example, when interpolating a rainfall surface from point rainfall data, incorporating a highly correlated variable such as elevation could help improve the estimation of rainfall. In such a case where the correlation is known, sampled elevation data could be used to help in the prediction of a rainfall surface, especially in those areas where rainfall sampling is sparse.

The semivariogram and another method, the robust estimator of the semivariogram, are the measures of variability that are used for the final fitting of a variability model to be used with the data set. They are also the only estimators of variability used by TerrSet for kriging and simulation. However, other methods for detecting spatial contiguity are available through the Spatial Dependence Modeler interface. These include the correlogram, the cross-correlogram, the covariogram, and the cross-covariogram.

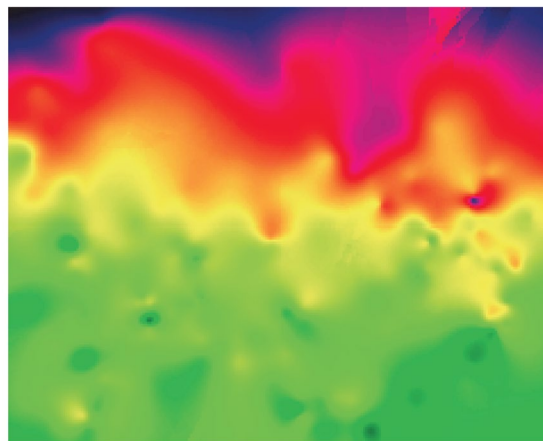
Kriging and Conditional Simulation

The Kriging and Simulation interface utilizes the model developed in the Spatial Dependence Modeler and Model Fitting interfaces to interpolate a surface. The model is used to derive spatial continuity information that will define how sample data will be weighted when combined to produce values for unknown points. The weights associated with sample points are determined by direction and distance to other known points, as well as the number and character of data points in a user-defined local neighborhood.

With ordinary kriging, the variance of the errors of the fit of the model is minimized. Thus it is known as a Best Linear Unbiased Estimator (B.L.U.E.).

By fitting a smooth model of spatial variability to the sample data and by minimizing the error of the fit to the sample data, kriging tends to underestimate low values and overestimate large values. Kriging minimizes the error produced by the differences in the fit of the spatial continuity to each local neighborhood. In so doing, it produces a smooth surface.

The surface shown in the figure below was produced using kriging with the sample precipitation points shown in the point data above.



The goal of kriging is to reduce the degree of variance error in the estimation across the surface. The variance error is a measure of the accuracy of the fit of the model and neighborhood parameters to the sample data, not the actual measured surface. One can only interpret this information in terms of knowledge about how well the sample data represents the actual surface. The more uniform the fit of the spatial model, the more likely it is good. The variance error is used to identify problems in the sample data, in the model parameters, and in the definition of the local neighborhood. It is not a measure of surface accuracy.

In TerrSet, two tools are available to assess the fit of the model to the sample data. First, the cross-validation tool iteratively removes a sample data point and interpolates a new value for the location. A table is produced to show the difference between the predicted attributes and the known attributes at those locations. Second, a variance image is produced that shows the spatial variation of uncertainty as a result of the fitted model. The variance image provides information to assist in identifying the problem areas where the relationship between the fitted model and the sample data points is poor.

Cokriging is an extension of kriging that uses a second set of points of different attributes to assist in the prediction process. The two attributes must be highly correlated with each other to derive any benefit. The description of spatial variability of the added variable can be used in the interpolation process, particularly in areas where the original sample points are sparse.

In conditional simulation, a non-spatially dependent element of variability is added to the model previously developed. The variability of each interpolated point is used to randomly choose another estimate. The resulting surface maintains the spatial variability as defined by the semivariogram model, but also represents pixel-by-pixel variability. The resulting surface is not smooth. Typically many of these surfaces (perhaps hundreds) are produced, each representing one model of reality. The surfaces differ from each other because of the random selection of estimates. Conditional simulation is best suited for developing multiple representations of a surface that may serve as inputs to a Monte Carlo analysis of a process model.

Summary

Geostatistics provides a large collection of tools for exploring and understanding the nature of a data set. Rather than simply seeking to produce a visually-pleasing interpolated surface, one engages in geostatistical analysis with the foremost purpose of understanding why various methods produce particular and different results. Interpretation of the information presented through the various techniques is dependent upon knowledge of other data characteristics and the actual surface. While spatial variability measures themselves are relatively simple descriptive statistics, understanding how they may be used with data sets that diverge from ideal assumptions requires practice and experience.

References / Further Reading

Geostatistical analysis is a well developed field and much literature is available. The brief list that follows should provide a good introduction to geostatistical exploration for those who already have a good command of statistics.

Burrough, P., and McDonnell, R., 1998. *Principles of Geographical Information Systems*, 98-161, Oxford University Press, Oxford.

Cressie, N., 1991. *Statistics for Spatial Data*, John Wiley and Sons, Inc., New York.

Cressie, N., and Hawkins, D., 1980. Robust Estimation of the Variogram, *Journal International Association of Mathematical Geology*, 12:115-125.

Deutsch, C., and Journel, A., 1998. *GSLIB Geostatistical Software Library and User's Guide, 2nd Edition*, Oxford University Press, Oxford.

Goovaerts, P., 1997. *Geostatistics for Natural Resources Evaluation*, Oxford University Press, Oxford.

Issaks, E., and Srivastava, R., 1989. *Applied Geostatistics*, Oxford University Press, Oxford.

Journel, A., and Huijbregts, C., 1978. *Mining Geostatistics*, Academic Press, New York.

Myers, J., 1997. *Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping*, Van Nostrand Reinhold, New York.

Pebesma, E., 1991-1998. Gstat, GNU Software Foundation.

Pebesma, E., and Wesseling, C., 1998. Gstat: A Program for Geostatistical Modelling, Prediction and Simulation, *Computers and Geosciences*, 24(1): 17-31.

Soares, A., Gómez-Hernandez, J., and Froidevaux, R., eds., 1997. *geoENVI – Geostatistics for Environmental Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands .

Solow, A., and Ratik, S., 1994. Conditional Simulation and the Value of Information, In: *Geostatistics for the Next Century*, R. Simitrakopoulos (ed.), Kluwer Academic Publishers, Dordrecht, The Netherlands.

CHAPTER FIVE

IDRISI IMAGE PROCESSING

The TerrSet System includes the industry leading toolset for image processing of remotely sensed data, the IDRISI Image Processing Toolset. The tools include state-of-the-art machine learning classifiers, including neural network classifiers and classification tree analysis. The tools available are broken down into 10 menu headings. These include:

- **Restoration**
- **Enhancement**
- **Transformation**
- **Fourier Analysis**
- **Signature Development**
- **Hard Classifiers**
- **Soft Classifiers / Mixture Analysis**
- **Segmentation Classifiers**
- ***Hyperspectral Image Analysis***
- ***Accuracy Assessment***

Image Restoration

Image restoration is broken down into two broad sub-areas: radiometric restoration and geometric restoration. Radiometric restoration is concerned with the fidelity of the readings of electromagnetic energy by the sensor while geometric restoration is concerned with the spatial fidelity of images. However, while the word restoration suggests a return to conditions that once existed, the reality is that image restoration is concerned with establishing measurement conditions that probably never existed—the measurement of radiometric characteristics under perfect and invariant conditions on an abstract geographic reference ellipsoid. Radiometric restoration is thus concerned with issues such as atmospheric haze, sensor calibration, topographic influences on illumination, system noise, and so on. Geometric restorations are less of a burden to the general data user because most geometric restorations are already completed by the imagery distributor. All of these rectifications can be achieved using existing IDRISI Image Processing modules. However, IDRISI also provides a series of imagery-specific modules that streamline many of these operations into a single click solution (as will be described at the end of the section on radiometric restoration).

Radiometric Restoration

Sensor Calibration

The detectors on sensors can vary between instruments (such as on successive satellites in a series, such as the NOAA TIROS-N weather satellites) and within an instrument over time or over the face of the image if multiple detectors are used (as is commonly the case). Sensor calibration is thus concerned with ensuring uniformity of output across the face of the image, and across time.

Radiance Calibration

Pixel values in satellite imagery typically express the amount of radiant energy received at the sensor in the form of uncalibrated relative values simply called Digital Numbers (DN). Sometimes these DN are referred to as the brightness values. For many (perhaps most) applications in remote sensing (such as classification of a single-date image using supervised classification), it is not necessary to convert these values. However, conversion of DN to absolute radiance values or relative surface reflectance values is a necessary procedure for comparative analysis of several images taken by different sensors (for example, Landsat-2 versus Landsat-5). Since each sensor has its own calibration parameters used in recording the DN values, the same DN values in two images taken by two different sensors may actually represent two different radiance values.

Usually, detectors are calibrated so that there is a linear relationship between DN and spectral radiance. This linear function is typically described by three parameters: the range of DN values in the image, and the lowest (L_{min}) and highest (L_{max}) radiances measured by a detector over the spectral bandwidth of the channel. L_{min} is the spectral radiance corresponding to the minimum DN value (usually 0). L_{max} is the radiance corresponding to the maximum DN (varies). Not only each sensor, but each band within the same sensor, has its own L_{min} and L_{max} . The information about sensor calibration parameters (L_{min} and L_{max}) is usually supplied with the data or is available elsewhere.¹ The equation² relating DN in remotely sensed data to radiance is:

$$L = \left(\frac{L_{max} - L_{min}}{255} \right) DN + L_{min}$$

where L is the radiance expressed in $Wm^{-2}sr^{-1}$

Alternatively, the calibration of the sensor may be expressed in the form of an offset and gain. In this case, radiance can be calculated as:

$$L = \text{Offset} + (\text{Gain} \times DN)$$

Note that it is also possible to convert between an Offset/Gain specification and L_{min}/L_{max} as follows:

$$\text{Offset} = L_{min}$$

and

$$\text{Gain} = \frac{L_{max} - L_{min}}{\text{MaxDN}}$$

¹ The Landsat satellites calibration parameters can be found in: EOSAT Landsat Technical Notes No. 1, August 1986, or the Landsat Data User's Handbook.

² For an explanation of radiance computation from DN, you may wish to consult: Lillesand, T.M., R.W. Kiefer and J.W. Chipman, 2004. *Remote Sensing and Image Interpretation*. Fifth Edition. John Wiley and Sons.

or alternatively:

$$L_{min} = \text{Offset}$$

$$L_{max} = (\text{Gain} \times \text{MaxDN}) + L_{min}$$

Either the CALIBRATE or RADIANCE modules can be used to convert raw DN values to calibrated radiances. The RADIANCE module has the most extensive options. It contains a lookup table of L_{min} and L_{max} for both the MSS and TM sensors on Landsat 1-5. For other satellite systems, it permits the user to enter system-specific L_{min}/L_{max} , or Offset/Gain values. CALIBRATE is more specifically geared towards brightness level matching, but does allow for adjustment to a specific offset and gain. In either case, special care must be taken that the calibration coefficients are correctly matched to the output units desired. The most common expression of radiance is in $\text{mWcm}^{-2}\text{sr}^{-1}\text{mm}^{-1}$ (i.e., milliWatts per square centimeter per steradian per micron). However, it is also common to encounter $\text{Wm}^{-2}\text{sr}^{-1}\text{mm}^{-1}$ (i.e., watts per square meter per steradian per micron).

Band Striping

Striping or banding is systematic noise in an image that results from variation in the response of the individual detectors used for a particular band. This usually happens when a detector goes out of adjustment and produces readings that are consistently much higher or lower than the other detectors for the same band. In the case of MSS data, there are 6 detectors per band which scan in a horizontal direction. If one of the detectors is miscalibrated, then horizontal banding occurs repetitively on each 6th line. Similarly, in the case of TM data with 16 detectors per band, each 16th line in the image will be affected. Multispectral SPOT data have a pushbroom scanner with 3000 detectors for each band, one detector for each pixel in a row. Since detectors in a pushbroom scanner are arranged in a line perpendicular to the satellite orbit track, miscalibration in SPOT detectors produces vertical banding. Since the SPOT satellite has an individual detector for each column of data, there is no repetitive striping pattern in the image.

The procedure that corrects the values in the bad scan lines is called *destriping*. It involves the calculation of the mean (or median) and standard deviation for the entire image and then for each detector separately. Some software packages offer an option for applying a mask to the image to exclude certain areas from these calculations (for example, clouds and cloud shadows should be excluded). Also, sometimes only a portion of the image, usually a homogeneous area such as a body of water, is used for these calculations. Then, depending on the algorithm employed by a software system, one of the following adjustments is usually made:

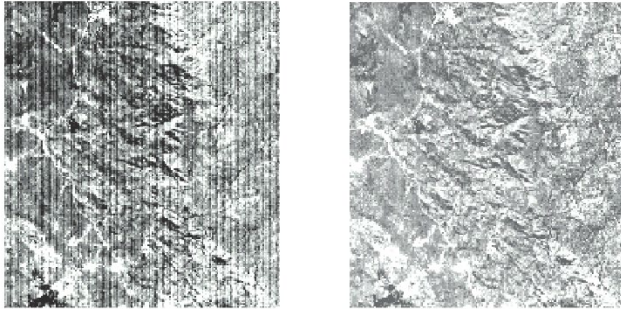
1. The output from each detector is scaled to match the mean and standard deviation of the entire image. In this case, the value of each pixel in the image is altered.
2. The output from the problem detector is scaled to resemble the mean and standard deviation of the other detectors. In this case, the values of the pixels in normal data lines are not altered.

The module DESTRIPE employs the first method. Results of this transformation are shown in the figures below.

If satellite data are acquired from a distributor already fully georeferenced, then radiometric correction via DESTRIPE is no longer possible. In this case, we suggest a different methodology. One can run an Unstandardized Principal Components Analysis on the collection of input bands. The last few components usually represent less than 1 percent of the total information available and tend to hold information relevant to striping. If these components are removed completely and the rest of the components re-assembled, the improvement can be dramatic as the striping effect can even disappear. To re-assemble component images, it is necessary to save the table information reporting the eigenvectors for each component. In this table, the rows are ordered according to the band number and the column eigenvectors reading from left to right represent the set of transformation coefficients required to linearly transform the bands to produce the components. Similarly, each row represents the coefficients of the reverse transformation from the components back to the original bands. Multiplying each component image by its corresponding eigenvector element for a particular band and summing the weighted components together reproduces the original band of information. If the noise components are

simply dropped from the equation, it is possible to compute the new bands, free of these effects. This can be achieved very quickly using the Image Calculator facility.

The section on **Fourier Analysis** details how that technique may also be used to remove striping or noise from satellite imagery.



Mosaicing

Mosaicing refers to the process of matching the radiometric characteristics of a set of images that fit together to produce a larger composite. The MOSAIC module facilitates this process. The basic logic is to equalize the means and variances of recorded values across the set of images, based on an analysis of comparative values in overlap areas. The first image specified acts as the primary, to which all other images are adjusted.

Atmospheric Correction

The atmosphere can affect the nature of remotely sensed images in a number of ways. At the molecular level, atmospheric gases cause Rayleigh scattering that progressively affects shorter wavelengths (causing, for example, the sky to appear blue). Further, major atmospheric components such as oxygen, carbon dioxide, ozone and water vapor (particularly these latter two) cause absorption of energy at selected wavelengths. Aerosol particulates (an aerosol is a gaseous suspension of fine solid or liquid particles) are the primary determinant of haze, and introduce a largely non-selective (i.e., affecting all wavelength equally) Mie scattering. Atmospheric effects can be substantial (see figure below). Thus, remote sensing specialists have worked towards the modeling and correction of these effects. TerrSet offers several approaches to atmospheric correction, with the most sophisticated being the module ATMOSC.

Dark Object Subtraction Model

The effect of haze is usually a relatively uniform elevation in spectral values in the visible bands of energy. One means of reducing haze in imagery is to look for values in areas of known zero reflectance, such as deep water. Any value above zero in these areas is likely to represent an overall increase in values across the image and can be subtracted easily from all values in the individual band using SCALAR. However, ATMOSC also offers a Dark Object Subtraction model with the added benefit that it compensates for variations in solar output according to the time of year and the solar elevation angle. To do this, it requires the same estimate of the Dn of haze (e.g., the Dn of deep clear lakes), the date and time of the image, the central wavelength of the image band, the sun elevation, and radiance conversion parameters. These additional parameters are normally included with the documentation for remotely sensed images.

Cos(t) Model

One of the difficulties with atmospheric correction is that the data necessary for a full accommodation are often not available. The Cos(t) model was developed by Chavez (1996) as a technique for approximation that works well in these instances. It is also available in the ATMOSC module and incorporates all of the elements of the Dark Object Subtraction model (for haze removal) plus a procedure for estimating the effects of absorption by atmospheric gases and Rayleigh scattering. It requires no additional parameters over the Dark Object Subtraction model and estimates these additional elements based on the cosine of the solar zenith angle (90 - solar elevation).

Full Correction Model

The full model is the most demanding in terms of data requirements. In addition to the parameters required for the Dark Object Subtraction and Cos(t) models, it requires an estimate of the optical thickness of the atmosphere (the Help System for ATMOSC gives guidelines for this) and the spectral diffuse sky irradiance (the downwelling diffuse sky irradiance at the wavelength in question arising from scattering—see Forster (1984) and Turner and Spencer (1972). In cases where this is unknown, the default value of 0 can be used.

Apparent Reflectance Model

ATMOSC offers a fourth model known as the Apparent Reflectance Model. It is rarely used since it does very little accommodation to atmospheric effect (it only accommodates the sun elevation, and thus the effective thickness of the atmosphere). It is included, however, as a means of converting Dn into approximate reflectance values.

An Alternative Haze Removal Strategy

Another effective method for reducing haze involves the application of Principal Components Analysis . The PCA module separates a collection of bands into statistically separate components. The method of removing a component is described in the *Noise Effects* section (this chapter) for the removal of image striping. We suggest using this method to reduce any other atmospheric effects as well. The figures below are Landsat TM Band 1 images before restoration and after. The area is in north-central Vietnam at a time with such heavy amounts of haze relative to ground reflectance as to make sensor banding also very apparent. Principal Components Analysis was applied on the seven original bands. The 6th and 7th components, which comprised less than 0.002 percent of the total information carried across the bands, were dropped when the components were reverse-transformed into the new band. (The reverse-transformation process is detailed above in the section on Band Striping.) Notice that not only do the haze and banding disappear in the second image, but also what appears to be clouds is greatly reduced.



Topographic Effects

*Topographic effect*³ is defined simply as the difference in radiance values from inclined surfaces compared to horizontal ones. The interaction of the angle and azimuth of the sun's rays with slopes and aspects produce topographic effects resulting in variable illumination. Images are often taken in the early morning hours or late afternoon, when the effect of sun angle on slope illumination can be extreme. In mountainous environments, reflectances of slopes facing away from the sun are considerably lower than the overall reflectance or mean of an image area. In extreme terrain conditions, some areas may be shadowed to the extent that meaningful information is lost altogether.

Shadowing and scattering effects exaggerate the difference in reflectance information coming from similar earth materials. The signature of the same landcover type on opposite facing slopes may not only have a different mean and variance, but may even have non-overlapping reflectance ranges. In the classification process, the highly variable relationship between slope, landcover, and sun angle can lead to a highly exaggerated number of reflectance groups that make final interpretation of data layers more costly, difficult, and time consuming. Even when landcovers are not the same on opposite sides of the mountain (which is often the case since slope and aspect help determine the cover type present), variable illumination nonetheless makes it difficult to derive biomass indices or perform other comparisons between landcover classes.

Several techniques for mitigating topographic effect have evolved in recent years. However, many tend to be only appropriate for the specific environment in which they were developed, or they require high-detail ancillary data that is often unavailable. The three most accessible techniques used are band ratioing, partitioning an image into separate areas for classification, and illumination modeling based on a DEM. More sophisticated techniques (which are not discussed here) involve the modeling of such illumination effects as backscattering and indirect diffusion effects.

Band Ratioing

In band ratioing, one band image is divided by another.

$$\frac{BandA}{BandB}$$

The resulting output image is then linearly stretched back to the 0 to 255 value range, and used for image classification. Band ratioing is based on the principle that terrain variations (in slope and aspect) cause variations in illumination that are consistent across different wavelengths. Thus in an area with a uniform landcover, the relative reflectance of one band to another will be the same regardless of slope and aspect variations. Band ratioing is the simplest technique to implement to mitigate topographic effect. It will not be very effective, however, when the variance in signature ranges is highly compressed. This commonly occurs in extreme shadowing conditions.

Image Partitioning

Image partitioning works from the simple assumption that since different areas within an image are affected differently by illumination effects resulting from slope and aspect, these distinct areas should be classified separately. Using a digital elevation model to produce mask images, the bands are subdivided according to different elevation, slope, and aspect categories. These sub-scenes are then classified separately and the results recombined after classification. Like band ratioing, the technique is simple and intuitive, but is effective only under the right conditions. Such partitioning works best where landcover conditions are stratified environmentally. Otherwise there is the potential to create hundreds of meaningless clusters when using an unsupervised classification or to misclassify pixels when applying supervised techniques.

The thresholds set for slope, aspect, and elevation are dependent upon the known sun angle and azimuth. Without solar information, the significant thresholds of topographic effect may be imprecisely determined by analyzing the shadow effect visually. The

³ The Topographic Effects section is condensed from the "Mitigating Topographic Effects in Satellite Imagery" exercise in Schneider and Robbins, 1995. *UNITAR Explorations in GIS Technology, Volume 5, GIS and Mountain Environments*, UNITAR, Geneva, also available from the Clark Labs.

registration of the DEM to the satellite data must be as precise as possible, otherwise the inexact nature of thresholds will further increase the possibility of less meaningful classifications.

Illumination Modeling

The analytical tools associated with most raster GIS software systems offer a very effective technique for modeling illumination effects. The steps, using TerrSet, are as follows:

1. Use HISTO to calculate the mean of the image band to be corrected.
2. Using a Digital Elevation Model (DEM) for the image area, use the HILLSHADE (SURFACE) module to create a map of analytical hillshading. This will be the model of illumination effects for all bands and simply needs to be calibrated for use.
3. Use REGRESS to calculate the linear relationship between the hillshading map and the image to be corrected. Use the image as the dependent variable and the hillshading as the independent variable.
4. Use CALIBRATE to apply the offset (the intercept of the regression equation) and gain (the slope of the regression equation) to the hillshading map. The result is a model of the terrain-induced illumination component.
5. Use Image Calculator to subtract the result of the previous step from the original image and then add the mean calculated in the first step. The result is a reasonable estimate of what the image would have looked like if the ground had been flat.

Noise

Noise in images occurs because of any number of mechanical or electronic interferences in the sensing system that lead to transmission errors. Noise either degrades the recorded signal or it virtually eliminates all radiometric information. Noise can be systematic, such as the periodic malfunctioning of a detector, resulting in striping or banding in the imagery. Or it can be more random in character, causing radiometric variations described as having a "salt-and-pepper" appearance. In RADAR imagery, "speckle" occurs because the signal's interaction with certain object edges or buildings produces a highly elevated recording, which when frequent, has a similar effect as "salt-and-pepper" noise.

Scan Line Drop Out

Scan line drop out occurs when temporary signal loss from specific detectors causes a complete loss of data for affected lines. In TerrSet this problem can be solved by a sequence of steps. First, reclassify the affected band in RECLASS to create a Boolean mask image in which the pixels where drop-lines are present have a value of 1 and the rest have a value of zero. Then, for horizontal drop lines, run a user-defined 3 x 3 FILTER across the original band containing the following kernel values:

0	0.5	0
0	0	0
0	0.5	0

This will have the effect of assigning to each pixel the average of the values in the scanlines above and below. If the drop line is vertical, simply rotate the filter kernel values by 90 degrees. Then, use OVERLAY to multiply the mask and the filtered image together. This creates a result in which filtered values only appear in those locations where scan lines were lost. OVERLAY this result on the original band using the COVER operation, which will cause these values to be placed only where the data were lost.

"Salt-and-Pepper" Noise

Random noise often produces values that are abnormally high or low relative to surrounding pixel values. Given the assumption that noisy reflectance values show abrupt changes in reflectance from pixel to pixel, it is possible to use filtering operations to replace these values with another value generated from the interpretation of surrounding pixels. The FILTER module provides several options for this purpose. A 3 x 3 or 5 x 5 median filter commonly are applied. The noisy pixels are replaced by the median value selected from the neighbors of the specified window. Because all pixels are processed by median filtering, some detail and edges may be lost. This is especially problematic with RADAR imagery because of the particularly high level of speckle that can occur. Therefore, we have included an Adaptive Box filter that is an extension of the common Lee filter. The Adaptive Box filter determines locally within a specified window (3 x 3, 5 x 5, or 7 x 7) the mean and the min/max value range based on a user-specified standard deviation. If the center window value is outside the user-specified range, then it is assumed to be noise and the value is replaced by an average of the surrounding neighbors. You may choose the option of replacing the value with a zero. The filter also allows the user to specify a minimum threshold variance in order to protect pixels in areas of very low variation. See the Help System of the FILTER module to learn more about this highly flexible approach.

Imagery-Specific All-in-One Procedures

There are a growing series of imagery-specific modules in the IDRISI Image Processing toolset that not only import the data for use with TerrSet, but also apply various radiometric corrections. The LANDSAT, SENTINEL and DIGITALGLOBE modules all offer these abilities. Options include the import of raw Dn, Top-of-Atmosphere (TOA) radiances, TOA reflectances, and Bottom-of-Atmosphere (BOA) reflectances using either a Dark Object Subtraction (DOS) or COS(t) approximation.

Geometric Restoration

As stated in the chapter **Introduction to Remote Sensing and Image Processing**, most elements of geometric restoration associated with image capture are corrected by the distributors of the imagery, most importantly *skew correction* and *scanner distortion correction*. Distributors also sell imagery already *georeferenced*. Georeferencing is not only a restoration technique but a method of reorienting the data to satisfy the specific desires and project requirements of the data user. As such, it is particularly important that georeferenced imagery meets the data user's standards and registers well with other data in the same projection and referencing system.

It is our experience that even if one's standards are not very stringent for the particular imaging task at hand, it is well worth the time one takes to georeference the imagery oneself rather than having the distributor do so. This is true for a number of reasons. First, certain radiometric corrections become more difficult (if not impossible) to perform if the data are already georeferenced. Of particular concern is the ability to reduce the effects of banding, scan line drop, and topographic effects on illumination. If the geometric orientation of the effects is altered, then standard restoration techniques are rendered useless. Given that the severity of these effects is not usually known prior to receiving the data, georeferencing the data oneself is important in maintaining control over the image restoration process.

Another reason to georeference imagery oneself is to gain more control over the spatial uncertainties produced by the georeferencing process. Only then is it possible to know how many control points are used, where they are located, what the quality of each is individually, and what the most satisfying combination of control points is to select. The RESAMPLE module provides the user with significant control over this process. The user may freely drop and add points and evaluate the effects on the overall RMS error and the individual residuals of points as they are fitted to a new equation. This interactive evaluation is especially important if significant rubbersheeting is required to warp the data to fit the projection needs of one's area.

See the section on **Georeferencing** for a broader discussion of this issue. See also the exercise on Georeferencing in the tutorial for a worked example of how the RESAMPLE module is used in TerrSet to georegister a satellite image.

References

- Chavez, P.S., (1996) "Image-Based Atmospheric Corrections - Revisited and Improved", *Photogrammetric Engineering and Remote Sensing*, 62, 9, 1025-1036.
- Forster, B.C., (1984) "Derivation of atmospheric correction procedures for Landsat MSS with particular reference to urban data", *International Journal of Remote Sensing*, 5, 5, 799-817.
- Lillesand, T.M. and R.W. Kiefer, (1994) *Remote Sensing and Image Interpretation*. Third Edition. John Wiley and Sons.
- Turner, R.E., and Spencer, M.M., (1972) "Atmospheric Model for Correction of Spacecraft Data", *Proceedings, Eighth International Symposium on Remote Sensing of the Environment*, Vol. II, 895-934.

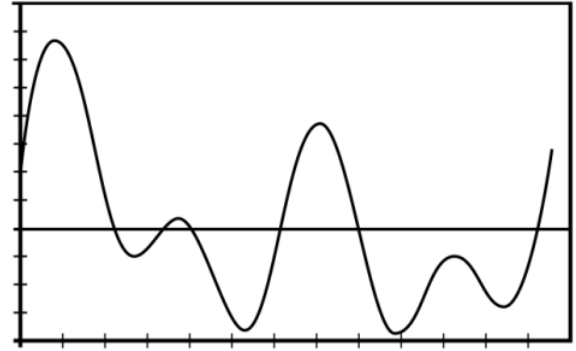
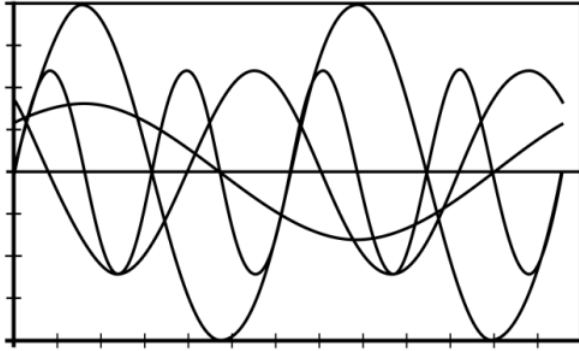
Fourier Analysis

Fourier Analysis is a signal/image decomposition technique that attempts to describe the underlying structure of an image as a combination of simpler elements. Specifically, it attempts to describe the signal/image as a composite of simple sine waves. Thus the intent of Fourier Analysis is somewhat similar to that of Principal Components Analysis—to break the image down into its structural components for the purpose of analyzing and modifying those components before eventual reconstruction into an enhanced form. While Fourier Analysis has application in a number of fields ranging from optics to electronics, in the context of image processing, it is most often used for noise removal.

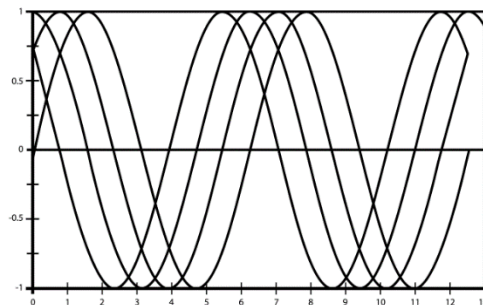
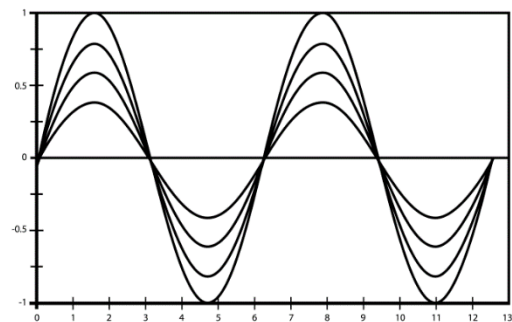
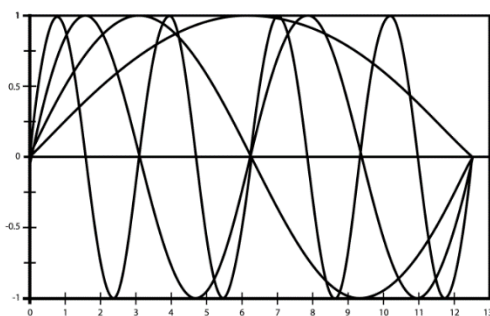
The Logic of Fourier Analysis

It is unfortunate that the mathematical treatment of Fourier Analysis makes it conceptually inaccessible to many. Since there are ample treatments of Fourier Analysis from a mathematical perspective, the description offered here is intended as a more conceptual treatment.

Any image can be conceptually understood as a complex wave form. For example, if one were to graph the grey levels along any row or column, they would form the character of a complex wave. For a two dimensional image, the logical extension of this would be a surface, like the surface of an ocean or lake. Imagine dropping a stone into a pool of water—a simple sine wave pattern would be formed with a wave length dependent upon the size of the stone. Now imagine dropping a whole group of stones of varying size and at varying locations. At some locations the waves would cancel each other out while at other locations they would reinforce each other, leading to even higher amplitude waves. The surface would thus exhibit a complex wave pattern that was ultimately created by a set of very simple wave forms. The figures illustrate this effect. The figure on the left shows a series of sine waves of varying frequency, amplitude, and phase (these terms will be explained below). The figure on the right shows the complex wave form that would result from the combination of these waves.



Fourier Analysis uses this logic in reverse. It starts with a complex wave form and assumes that this is the result of the additive effects of simple sine waves of varying *frequency*, *amplitude* and *phase*. Frequency refers to the number of complete wavelengths that occur over a specified distance. The upper left figure below shows a series of sine waves that differ only in their frequency. Amplitude refers to the height or strength of the wave. The upper right figure below shows a series of sine waves that vary only in amplitude. Finally, phase refers to the offset of the first wave crest from origin. The bottom figure below shows a series of waves that vary only in phase. In the decomposition of digital images, a finite set of waves are assumed, ranging from the lowest frequency wave which completes a single wavelength over the extent of the image in X and Y, to one that completes 2 wavelengths over that extent, to 3 wavelengths, and so on, up to the highest frequency wave with a wavelength equal to twice the pixel resolution (known as the *Nyquist frequency*). The task of the Fourier Analysis process is then simply one of estimating the phase and amplitudes of these waves.



How Fourier Analysis Works

The easiest way to understand how Fourier Analysis works is to use an analogy. In the presence of a sound (a complex wave form), the string of a guitar (or any other stringed instrument) will vibrate, or resonate, if the sound contains the same note (frequency). The strength of that sympathetic vibration will be a function of the amplitude of that note within the original sound. In essence, this is how Fourier Analysis works—it "listens" for the degree of resonance of a set of specific frequencies with the sound (image) being analyzed. It is like placing a harp into the presence of a sound, and gauging the degree to which each string resonates.

The process of testing for the presence of varying frequencies is achieved by multiplying the complex wave by the corresponding amplitude of the sine wave being evaluated, and summing the results. The resulting sum is the resonance at that frequency. The only problem, however, is phase. What if the wave in question is present, but our test wave is out of phase? In the worst case they are exactly out of phase, so that the peaks in the test frequency are balanced by troughs in the complex wave—the net effect is that they will cancel each other out, leading one to believe that the wave is not present at all.

The answer to the problem of phase is to test the complex wave against two versions of the same frequency wave, exactly out of phase with each other. This can easily be done by testing both a sine wave and a cosine wave of the same frequency (since sines and cosines are identical except for being exactly out of phase). In this way, if there is little resonance with the sine wave because of a problem of phase, it is guaranteed to resonate with the cosine wave.⁴

Interpreting the Mathematical Expression

Given the discussion above, the formula for the Fourier Series is not so difficult to understand. Considering the one-dimensional case, the complex function over x can be described as the sum of sine and cosine components as follows:

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos(k\omega x) + b_k \sin(k\omega x))$$

where $f(x)$ is the value of the function at position x , ω is equal to $2\pi/T$ where T is the period (the length of the series), and k is the harmonic.⁵ The coefficients a and b are determined by means of the Fourier Transform.

The Fourier Transform itself makes use of Euler's Formula :

$$e^{-i2\pi ux} = \cos 2\pi ux - i \sin 2\pi ux$$

where i^2 is -1 , leading to the following formulation for the case of discrete data:

$$F(u) = \frac{1}{N} \sum f(x) e^{-ik\omega x}$$

and an inverse formula of:

$$f(x) = \sum F(u) e^{ik\omega x}$$

It is not critical that these mathematical expressions be fully understood in order to make productive use of the Fourier Transform. However, it can be appreciated from the above that:

⁴ The use of a sine/cosine pair is identical in concept to describing locations using a pair of coordinates— X and Y . In both cases, the reference pair are known as basis vectors. In plane two-dimensional space, any location can be defined by its X and Y coordinates. Similarly, in the frequency domain, any wave can be defined by its sine and cosine coordinates.

⁵ The term "harmonic" used here refers to the relation of quantities whose reciprocals are in arithmetic progression (e.g., $1, 1/2, 1/3, 1/4$, etc.); or to points, lines, functions, etc. involving such a relation.

these formulas express the forward and inverse transforms for one-dimensional data. Simple extensions make these applicable to two-dimensional data.

The implementation of the Fourier Transform uses complex-valued inputs and outputs. A complex number is one with both real and imaginary parts of the form $a + bi$, where $i^2 = -1$. In the case considered here, for input into the FOURIER module, the image grey-level values make up the real part, while the imaginary part is set to zero (this is done automatically by the FOURIER module) since it doesn't exist. Thus the input to the Fourier Transform is a single image.

the resulting output of the forward transform thus consists of two images—a real and an imaginary part. The real part expresses the cosine component of the Fourier series while the imaginary part expresses the sine component. These are the amplitudes a and b of the cosine and sine components expressed in the first formula in this section. From these, the amplitude and phase of the wave can be determined as follows:

$$\text{Amplitude} = \sqrt{a^2 + b^2}$$

$$\text{Phase} = \tan^{-1}\left(\frac{b}{a}\right)$$

Together, the real and imaginary parts express the frequency spectrum of an image. While both the amplitude and phase can be readily calculated from these parts (Image Calculator in TerrSet can be used to do this), neither is commonly used. More commonly, the power spectrum is calculated and the phase is ignored. The power spectrum is simply the square of the amplitude. However, for purposes of visual examination, it is commonly expressed as a power spectrum image as follows:

$$\text{PowerSpectrum} = \ln(1 + \text{amplitude}^2)$$

This is the formula used in FOURIER. Thus the forward transform produced by the FOURIER module yields three outputs—a real part image, an imaginary part image, and a power image. These are commonly referred to as frequency domain images (as opposed to the original image which expresses the spatial domain).

The primary intent of the Fourier Transform is to examine the spectrum and modify its characteristics before reconstructing the original by means of the inverse transform. Examination of the spectrum is done with the power image. However, modifications are implemented on the real and imaginary parts. The FILTERFQ, DRAWFILT and FREQDIST modules can be used to create a variety of filters to be applied to the real and imaginary parts of the frequency domain.

The FOURIER module can also be used to compute the inverse transform. In this case, it is necessary to supply both the real and imaginary parts of the frequency domain. The result is a real image—the imaginary part is assumed to be zero, and is discarded.

The actual implementation of the Fourier Transform in FOURIER is by means of the Fast Fourier Transform (FFT) procedure. This algorithm is comparatively very fast, but requires that the image size (in X and Y) be a power of 2 (e.g., 2, 4, 8, 16, 32, 64, 128, etc.). In cases where the image is some other size, the edges can be padded with zeros to make the image conform to the proper dimensions. The ZEROPAD module facilitates this process in TerrSet.

Fourier Analysis assumes that the image itself is periodic to infinity. Thus it is assumed that the image repeats itself endlessly in X and Y. The effect is similar to bending the image in both X and Y such that the last column touches the first and the last row touches the first. If the grey values at these extremes are quite different, their juxtaposition will lead to spurious high frequency components in the transform. This can be mitigated by zero padding the edges. Zero padding is therefore often quite desirable.

Using Fourier Analysis in TerrSet

Gathering together and extending the information above, the following procedures are typical of Fourier Analysis in TerrSet.

1. Prepare the image for analysis. Both the rows and columns must be powers of 2. However, it is not necessary that they be the same. Thus, for example, an original image of 200 columns and 500 rows would need to be padded out to be an image of 256 columns and 512 rows. Use the ZEROPAD module in TerrSet to do this. If the original image already has rows and columns that are a power of 2, this step can be omitted.
2. Run FOURIER with the forward transform using the image prepared in Step 1 as the input. The image can contain byte, integer or real data—the data type is not important. FOURIER will produce three outputs—a real part image, an imaginary part image, and a power spectrum image. The last of these is intended for visual analysis while the first two are used for numeric analysis.
3. Examine the power spectrum image and design a filter (a topic covered at greater length in the next section). To create the filter, use either the FREQDIST module (followed by either RECLASS or FUZZY), the FILTERFQ module, or the DRAWFILT module.
4. Apply the filter to the real and imaginary part images created in Step 2. This is typically done through multiplication using the OVERLAY module.
5. Run FOURIER again and use the modified real and imaginary parts as input to the inverse transform. The result is the reconstructed image. Note that this output is always a real number image which may be converted to either byte or integer form if desired (and the range of values permits). The original image that was submitted to the forward transform must also be supplied. This image provides reference system information. If zero padding was added prior to the forward transform, then that image (with the zero padding) should be given here as the original image.
6. If zero padding was added, use WINDOW to extract the image to a new file.

Interpreting Frequency Domain Images

When first encountered, frequency domain images appear very strange. Indeed it is hard to believe that they contain all the information required to completely restore the full spatial domain image. However, the pixels in the real and imaginary part images contain a complete record of the sine waves that will form the image when combined. In these images as well as the power spectrum image, the position of each pixel in relation to the center cell indicates the frequency of the wave, while the pixel value indicates the amplitude.

For purposes of visual analysis, the power spectrum image is always used since it contains a visually enhanced record of the amplitude of the component waves.⁶ Within this image, each pixel represents a different wave frequency, with the sole exception of the central pixel located at $(\text{columns}/2)$ and $(\text{rows}/2 - 1)$. This pixel represents a frequency of 0—an expression of the average grey level of the image, similar in concept to an intercept in regression analysis.

The pixels to the immediate right $((\text{columns}/2) + 1)$ and above $(\text{rows}/2)$ represent the lowest frequency (longest wavelength) waves in the decomposition, with a frequency of one complete wave over the entire extent of the image, i.e., a frequency of $(1/(nd))$ where n =number of rows or columns, and d =pixel resolution. Thus with an image of 512 columns by 512 rows, and 30 meter cells, the pixel to the immediate right or above the center represents a frequency of $(1/(nd)) = (1/(512 \times 30)) = 1/15,360$ meters. Similarly, the second-most pixel to the right or above the center represents a frequency of $(2/(nd)) = (2/(512 \times 30)) = 1/7,680$ meters. Likewise, the third-most pixel to the right or above the center represents a frequency of $(3/(nd)) = (3/(512 \times 30)) = 1/5,120$ meters. This logic continues to the right-most and top-most edges, which would have a frequency of $(255/(nd)) = (255/(512 \times 30)) = 1/60.24$ meters. This latter value is only one short of the limiting frequency of $(256/(nd)) = (256/(512 \times 30)) = 1/60$ meters. This limiting frequency is the *Nyquist frequency*, and represents the shortest wavelength that can be described by pixels at a given resolution. In this example, the shortest wave repeats every 60 meters.

⁶ The phase information is not important for visual analysis, and is lost in the production of the power spectrum image. However, all mathematical manipulations are undertaken on the real and imaginary part images, which together contain complete amplitude and phase information.

The upper-right quadrant describes waves with positive frequencies in X and Y. All other quadrants contain at least one dimension that describes negative waves. For example, the lower-left quadrant describes frequencies that are negative in both X and Y. Negative frequencies are a consequence of the fact that Fourier Analysis assumes that the information analyzed is infinitely periodic. This is achieved by imagining that the original image could be bent into a cylinder shape in both X and Y so that the first and last columns adjoin one another and similarly that the top and last rows adjoin one another. Note also that the upper-right and lower-left quadrants are mirror images of each other as are the upper-left and lower-right quadrants. This symmetry arises from the fact that the input data are real number and not complex number images.

Finally, note that the first column and the last row (and the last column and the first row) represent the amplitude and power of waves at the Nyquist frequency for both positive and negative frequencies—i.e., using the example above, both $(256/(nd)) = (256/(512 \times 30)) = 1/60$ meters and $(-256/(nd)) = (-256/(512 \times 30)) = -1/60$ meters. The reason why these represent both the positive and negative frequencies relates to the cylindrical folding indicated above that is necessary to create an infinitely periodic form.

Given the above logic to the structure of the amplitude and power spectrum image s, several key points can be appreciated:

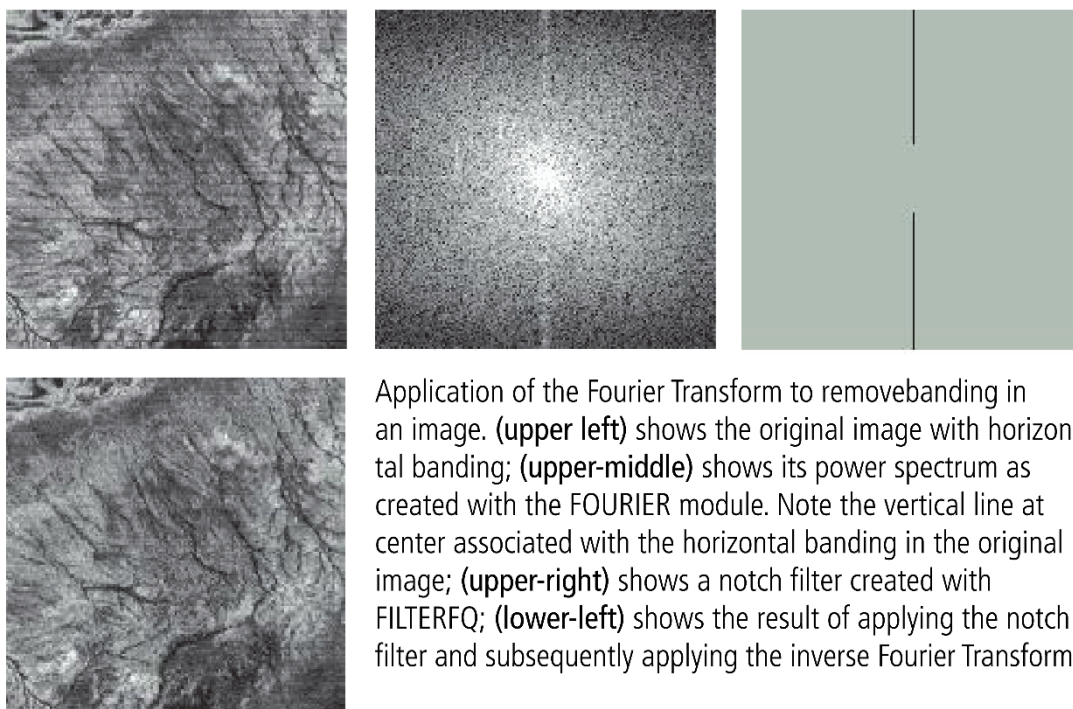
Amplitude and power spectrum images have a character that is radial about the central point representing zero frequency.

Noise elements are readily apparent as aberrant point or line features in the power spectrum. Linear noise elements will appear at a 90 degree angle in the power spectrum image to their spatial domain direction, e.g., vertical striping in the original image will produce horizontal elements in the power spectrum image.

These noise elements can be filtered out by reducing their amplitudes to 0 in the frequency domain and then doing the inverse transform.

Frequency Domain Filter Design

The upper left figure below shows an example of an image with severe horizontal banding, while the upper middle figure shows its power spectrum created with FOURIER . The upper right figure shows a notch filter created with FILTERFQ . Both the real and imaginary parts of the frequency domain transform were then multiplied by this filter in order to block out these wavelengths (reduce their amplitudes to 0). Finally, the bottom figure shows the result of applying the reverse transform on the modified real and imaginary part images.



Application of the Fourier Transform to remove banding in an image. (upper left) shows the original image with horizontal banding; (upper-middle) shows its power spectrum as created with the FOURIER module. Note the vertical line at center associated with the horizontal banding in the original image; (upper-right) shows a notch filter created with FILTERFQ; (lower-left) shows the result of applying the notch filter and subsequently applying the inverse Fourier Transform.

TerrSet supplies a variety of facilities for developing frequency domain filters. One would most commonly use FILTERFQ, which offers 26 filters, each of which can be controlled for the specific characteristics of the data being manipulated as well as for the specific purpose of the filter. The next most commonly used module is DRAWFILT, an interactive filter development tool in which one literally draws the areas (i.e., frequencies) to be removed. Finally, for even further flexibility, TerrSet offers a module named FREQDIST. As the name suggests, this module creates a frequency distance image (as measured from the center point of the power spectrum). This can then be used as the basic input to a variety of filter shaping tools such as RECLASS or FUZZY. The frequency distance image can also be submitted to SURFACE to create an aspect image which can then be shaped by RECLASS or FUZZY to create directional filters.

Regardless of how the filter is created, however, application of that filter is achieved the same way in all cases—by simple multiplication using OVERLAY with both the real and imaginary part images. As it turns out, multiplication in the frequency domain is the equivalent of convolution in the spatial domain.

References

Good references for frequency domain filtering include:

Gonzalez, R.C., and Woods, R.E., 1992. *Digital Image Processing*, Addison-Wesley, Reading, Massachusetts.

Mather, P., 1987. *Computer Processing of Remotely Sensed Images*, John Wiley and Sons, New York.

Jensen, J.R., 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective*, Prentice Hall, Upper Saddle River, NJ.

Classification of Remotely Sensed Imagery

This section we will cover the general logic and strategy to be used in working with the classification modules in the IDRISI Image Processing system. Detailed notes on the use of each module can be found in the Help System. Furthermore, examples in the form of exercises can be found in the online tutorial. Also, there is an Appendix: Introduction to Remote Sensing and Image Processing that provides fundamental concepts for undertaking image processing analysis.

Supervised Versus Unsupervised Classification

Classification is the process of developing interpreted maps from remotely sensed images. As a consequence, classification is perhaps the most important aspect of image processing to GIS. Traditionally, classification was achieved by visual interpretation of features and the manual delineation of their boundaries. However, with the advent of computers and digital imagery, attention has focused on the use of computer-assisted interpretation. Although the human eye still brings a superior set of capabilities to the classification process, the speed and consistency of digital procedures make them very attractive. As a consequence, the majority of classification projects today make use of digital classification procedures, guided by human interpretation.

As indicated in the **Introduction to Remote Sensing and Image Processing** chapter, there are two basic approaches to the classification process: supervised and unsupervised classification. With supervised classification, one provides a statistical description of the manner in which expected landcover classes should appear in the imagery, and then a procedure (known as a *classifier*) is used to evaluate the likelihood that each pixel belongs to one of these classes. With unsupervised classification, a very different approach is used. Here another type of classifier is used to uncover commonly occurring and distinctive reflectance patterns in the imagery, on the assumption that these represent major landcover classes. The analyst then determines the identity of each class by a combination of experience and ground truth (i.e., visiting the study area and observing the actual cover types).

In both of these cases, the process of classification can be seen as one of determining the set to which each pixel belongs. In the case of supervised classification, the sets are known (or assumed to be known) before the process is begun. Classification is thus a decision making process based upon available information. With unsupervised classification, however, the classes are unknown at the outset. Thus, the process is really one of segmentation rather than decision making per se.

Spectral Response Patterns versus Signatures

As explained in the **Introduction to Remote Sensing and Image Processing** chapter, each type of material interacts with electromagnetic energy by either reflecting, absorbing or transmitting it, with the exact nature of that interaction varying from one wavelength to the next—a pattern known as a *Spectral Response Pattern* (SRP). The basis for classification is thus to find some area of the electromagnetic spectrum in which the nature of that interaction is distinctively different from that of other materials that occur in the image. Many refer to this as a *signature*—a spectral response pattern that is characteristic of that material. However, in practice, the determination of consistently distinctive signatures is difficult to achieve for the following reasons:

- most vegetation types do not have consistent spectral response patterns—phenological changes throughout the growing season can lead to highly variable signatures.
- changes in illumination (because of slope or the time of year) and moisture variations can also lead to significantly different spectral response patterns.
- most landcover consist of mixtures of elementary features that are sensed as single pixels. For example, a row crop such as maize actually contains a mixture of plant and soil as sensed by a satellite. Likewise, a pixel may contain a mixture of conifers and deciduous species in a forest area.

- for a given sensor, there is no guarantee that the wavelengths in which it senses will be the same as those in which a material is most distinctive. Currently, multispectral sensors examine several very important areas of the spectrum, particularly for the differentiation of vegetation. However, the usable areas not examined far outnumber those that are, and many of the wavelengths that could potentially distinguish many rock types, for example, are not typically examined.

As a result of these problems, there has been a strong orientation within the remote sensing community to develop signatures with reference to specific examples within the image to be classified rather than relying on the use of more general libraries of characteristic spectral response patterns. These very specific examples are called *training sites*, named thus because they are used to train the classifier on what to look for. By choosing examples from within the image itself (usually confirmed by a ground truth visit), one develops signatures that are specific to the wavelengths available. One also avoids the problems of variations in both solar zenith angle and stage of the growing season. One can also choose examples that are characteristic of the various cover class mixtures that exist.

Despite this very pragmatic approach to the classification process, it remains very much a decision problem. We ask the process to create a definitive classification in the presence of considerable variation. For example, despite differences in growth stage, soil background and the presence of intercropping, we ask the process to distill all variations of maize cropping into a single maize class.

Recently, however, interest has focused on relaxing this traditional approach in two areas, both strongly represented in the TerrSet system. The first is the development of *soft* classifiers, while the second extends the logic of multispectral sensing to *hyperspectral* sensing.

Hard Versus Soft Classifiers

Traditional classifiers can be called *hard* classifiers since they yield a hard decision about the identity of each pixel. In contrast, *soft* classifiers express the degree to which a pixel belongs to each of the classes being considered. Thus, for example, rather than deciding that a pixel is either deciduous or coniferous forest, it might indicate that its membership grade in the deciduous class is 0.43 and coniferous is 0.57 (which a hard classifier would conclude is coniferous). One of the motivations for using a soft classifier is to determine the mixture of landcover classes present. If we could assume that these two classes were the only ones present, it might be reasonable to conclude that the pixel contains 43% deciduous cover and 57% coniferous. Such a conclusion is known as *sub-pixel classification*.

A second motivation for the use of a soft classifier is to measure and report the strength of evidence in support of the best conclusion that can be made. TerrSet introduces special soft classifiers that allow us to determine, for example, that evidence for deciduous is present to a level of 0.26, for coniferous to 0.19 and some unknown type to 0.55. This would immediately suggest that while the pixel has some similarities to our training sites for these two classes, it really belongs to some type that we have not yet identified.

A third motivation for the use of soft classifiers concerns the use of GIS data layers and models to supplement the information used to reach a final decision. For example, one might extract a mapping of the probability that each pixel belongs to a residential landcover class from the spectral data. Then a GIS data layer of roads might be used to develop a mapping of distance from roads, from which the probability of *not* being residential might be deduced (areas away from roads are unlikely to be residential). These two lines of evidence can then be combined to produce a stronger statement of the probability that this class exists. A final hard decision can subsequently be achieved by submitting the individual class membership statements to an appropriate *hardener*—a decision procedure that chooses the most likely alternative.

Multispectral Versus Hyperspectral Classifiers

The second major new development in classifiers is the use of *hyperspectral* data. Most sensors today are termed *multispectral* in that they sense electromagnetic energy in more than one area of the spectrum at once. Each of these areas is called a band and is represented by a single monochrome image. For example, the Landsat Thematic Mapper (TM) sensor system images seven simultaneous bands in the blue (Band 1), green (Band 2), red (Band 3), near infrared (Band 4), middle infrared (Bands 5 and 7) and

thermal infrared (Band 6) wavelength areas. Hyperspectral sensors are really no different in concept except that they image in many narrowly defined bands. For example, the AVIRIS experimental system developed by the Jet Propulsion Laboratory (JPL) images in 224 bands over a somewhat similar wavelength range as the TM sensor. Similarly, the EOS- MODIS system, launched in December 1999, spreads 36 bands over essentially the same range as that covered by the five bands on the corresponding AVHRR system of the NOAA series satellites.

It is tempting to think that more is better—i.e., that the greater number and higher spectral resolution of hyperspectral bands would naturally lead to better classifications. However, this is not necessarily the case. Hyperspectral images are most often highly correlated with other bands of similar wavelength. Thus one must process a substantially increased amount of data (which does affect the level of sophistication of the classifier algorithm) without a corresponding gain of information. The real benefit of hyperspectral imagery is gained from the ability to prospect for very narrow absorption features (spectral regions exhibiting strong absorption from specific materials) at high resolution. This has achieved most notable success in the context of geological applications. For example, recent extraterrestrial missions such as the NASA Mars Surveyor, Galileo, and Cassini missions all carry hyperspectral sensors for the purpose of mineral mapping. The high spectral resolution of these systems provides the ability to measure mineral absorption patterns with high precision, leading to the ability to map both the presence and abundance of surficial materials. Hyperspectral classification is still quite new and effectively experimental in character. TerrSet includes a range of procedures for working with these data.

Supervised Classification

General Logic

There is a consistent logic to all of the supervised classification routines in TerrSet, regardless of whether they are hard or soft classifiers. In addition, there is a basic sequence of operations that must be followed no matter which of the supervised classifiers is used. This sequence is described here. The **Tutorial** manual also contains worked examples of this process.

1. Define Training Sites

The first step in undertaking a supervised classification is to define the areas that will be used as training sites for each landcover class. This is usually done by using the on-screen digitizing feature as outlined in the TerrSet System Overview chapter. You should choose a band with strong contrast (such as a near-infrared band) or a color composite for use in digitizing. Then display that image on the screen (use autoscaling if necessary to gain good contrast) and use the on-screen digitizing feature to create one or more vector files of training site polygons—vector outlines of the training site areas.⁷

Good training site locations are generally those with as pure a sample of the information class as possible. For example, if you were choosing to define a site of deciduous forest, it would be important to choose an area that was not mixed with conifers and that had little soil background or understory vegetation visible. When digitizing training sites you should also avoid including any pixels belonging to adjacent landcover. This will be easiest to achieve if you zoom in on the area before digitizing that site.

⁷ IDRISI offers two procedures for the digitizing of training site polygons. With the default procedure, one digitizes a set of points that form the boundary of the training site polygon. The second procedure creates the polygon by aggregating together all contiguous pixels surrounding a designated point that fall within a specified tolerance of the spectral characteristics of the central pixel. This is called a flood polygon since it is analogous to the concept of water flowing outward from the designated point. You will also note that with this option a maximum distance can be specified to limit how far this growth procedure will spread. Note that the system also allows you to define training sites by means of a raster image. In some instances, it may make sense to define these locations by direct reference to ground locations (such as by means of point locations gathered with a GPS). However, this requires very exact prior georeferencing of the image, and a confidence that positional errors will not include unwanted pixels in the training sites. Some georeferencing procedures also alter image characteristics which may be undesirable. It is for these reasons that classification is commonly undertaken on ungeoreferenced imagery using on-screen digitizing of training sites. The final classified image is then georeferenced at a later stage.

In general, you should aim to digitize enough pixels so that there are *at least* 10 times as many pixels for each training class as there are bands in the image to classify. Thus, for a Landsat TM image with seven bands, you should aim to have at least 70 pixels per training class (more than that is not difficult to achieve and is recommended—the more the better). If this is difficult to achieve with a single site, simply digitize more than one training site for that class. The on-screen digitizing facility requires that you give an integer identifier for each feature. Thus to digitize more than one training site for a landcover class, simply assign the same identifier to each example. It may be helpful to make a list of ID's and their corresponding information classes.

Finally, note that there is no requirement that all training sites be included in a single vector file created with the on-screen digitizing feature. You can create a single vector file for each information class if you wish. This will simply require that you undertake the signature development stage for each of these files. Alternatively, you can join these vector files into a single vector file with CONCAT or rasterize all the vector files into a single raster image and develop the signatures from that single vector file or raster image.

For some machine learning classifiers (such as MLP), it is advantageous to have the training sites be about the same size. This is known as a balanced sample. To facilitate this, the digitize dialog offers the ability to monitor training site areas while you are digitizing. You have the choice of having the units expressed as pixels or ground units.

2. Extract Signatures

After the training site areas have been digitized, the next step will depend on whether a statistical classifier or a machine learning procedure is being used. With a machine learning classifier, the classification stage is tightly bound to a learning stage. With a statistical classifier, it is more common to create a statistical characterization of each informational class. These are called *signatures* in TerrSet. This is achieved with the MAKESIG module.⁸ MAKESIG will ask for the name of the vector or raster file that contains the training sites for one or more informational classes, and the bands to be used in the development of signatures. It will then ask for a name for each of the included classes. These names should be suitable as TerrSet file names since they will be used to create signatures files (.sig extension) for each informational class. If you used more than one vector file to store your training site polygons, run MAKESIG for each of these files. Your goal is to create a SIG file for every informational class.

SIG files contain a variety of information about the landcover classes they describe.⁹ These include the names of the image bands from which the statistical characterization was taken, the minimum, maximum and mean values on each band, and the full variance/covariance matrix associated with that multispectral image band set for that class. To examine the contents of this file in detail, use SIGCOMP. Note also that the SEPSIG module can be used to assess the distinctiveness of your set of signatures.

3. Classify the Image

The third (and sometimes final) step is to classify the image. This can be done with any of the hard or soft classifiers described below. Clearly there are many choices here. However, here are some tips:

- the parallelepiped procedure (the PIPED module) and the MINDIST (minimum distance to means) module are included for pedagogic reasons only. They date from the early days of digital image processing when computing resources were limited. Generally they offer no other advantages than speed.
- when training sites are known to be strong (i.e., well-defined with a large sample size), and are multivariate normal in character, the MAXLIKE procedure is very strong, and provides the advantage of allowing additional information to support the classification (by means of prior probability images). Similarly, when signatures are multivariate normal, the

⁸ The FUZSIG module offers an interesting, but quite different logic for extracting signatures from impure training sites. This is discussed further in the section on fuzzy signatures below. The ENDSIG module also creates signatures for use with the UNMIX classifier.

⁹ Each SIG file also has a corresponding SPF file that contains the actual pixel values used to create the SIG file. It is used only by HISTO in displaying histograms of signatures.

Mahalanobis Typicality (MAHALCLASS) classifier provides a unique feature. While all other classifiers assume that you have complete information about all of the land cover classes in the image, MAHALCLASS does not. It is the only classifier that provides information about the existence of a single class. The *typicality* it maps is an expression from 0-1 that indicates how typical the pixel is of the class it was trained on.

- several classifiers focus on establishing the boundaries between classes in band space (i.e., variable space, where each remotely sensed image band represents a different dimension. These include, in order of sophistication, Minimum Distance to Means (MINDIST), the Fisher/Linear Discriminant Analysis Classifier (FISHER) and the Support Vector Machine (SVM) classifier. The latter of these is a very powerful classifier and provides excellent results with non-normal signatures.
- there is an interesting relationship between the Multinomial Logistic Regression (MULTILOGISTICREG) classifier and the Multi-Layer Perceptron (MLP) neural network. Although the solution mechanics are a bit different, MULTILOGISTICREG is essentially the same as a MLP, but with no hidden layer. The hidden layer in MLP acts somewhat like an intermediate Principal Components. It abstracts a form of latent structure that allows for interesting non-linearities and interaction effects. MLP is a very powerful classifier as long as the number of output classes is small.
- there are many other interesting pairings. SOM is a neural network approach that is very similar to a K-Means (KMEANS) clustering. Similarly, Fuzzy Artmap (FUZZYARTMAP) is very similar in character to a Chain Clustering (CHAINCLUSTER) technique.
- Classification Tree Analysis (CTA) and DECISIONFOREST (an implementation of the random forest algorithm) are clearly related. The former produces a single tree while the latter classifies on the basis of many trees that are derived from different random draws of training data and different random selections of input variables. The former can sometimes be useful in understanding the classification process, but is highly susceptible to overfitting (becoming too sensitive to meaningless details). The latter solves this by working with many different subsets of the available information, but sacrifices transparency in the process. DECISIONFOREST is currently one of the most popular classification procedures in production use.
- for sub-pixel classification (i.e., analysis of mixture components), use one of the UNMIX soft classifier options. The other soft classifiers are primarily used for uncertainty analysis.
- KNN is a k-nearest neighbor classifier. It is a very simple classifier that classifies based on the similarity of pixels to known training examples. There are also some similarities to the RBF neural network that classifies on the basis of similarity to clusters developed through K-Means.

4. Generalization

The fifth stage is optional and frequently omitted. After classification, there may be many cases of isolated pixels that belong to a class that differs from the majority that surround them. This may be an accurate description of reality, but for mapping purposes, a very common post-processing operation is to generalize the image and remove these isolated pixels. This is done by passing a mode filter over the result (using the FILTER module). The mode filter replaces each pixel with the most frequently occurring class within a 3x3 window around each pixel. This effectively removes class patches of one or a few pixels and replaces them with the most common neighboring class. Use this operation with care—it is a generalization that truly alters the classified result.

5. Post-Classification Cleanup

All is fair in love and classification! Sometimes the only solution to achieve a highly accurate classification is to engage in some manual cleanup. The digitize dialog offers the ability to use the vector features in a digitized layer to update the pixels in an existing

classified image. Options are provided to replace existing pixels based on the new digitized areas directly, or to apply various masking options.

6. Accuracy Assessment

The final stage of the classification process usually involves an accuracy assessment. Traditionally this is done by generating a random set of locations (using the stratified random option of the SAMPLE module) to visit on the ground for verification of the true landcover type. A simple values file is then made to record the true landcover class (by its integer index number) for each of these locations. This values file is then used with the vector file of point locations to create a raster image of the true classes found at the locations examined. This raster image is then compared to the classified map using ERRMAT. ERRMAT tabulates the relationship between true landcover classes and the classes as mapped. It also tabulates *errors of omission* and *errors of commission* as well as the overall proportional error.

The size of the sample (n) to be used in accuracy assessment can be estimated using the following formula:

$$n = z^2 pq / e^2$$

where

z is the standard score required for the desired level of confidence (e.g., 1.96 for 95% confidence, 2.58 for 99%, etc.) in the assessment

e is the desired confidence interval (e.g., 0.1 for $\pm 10\%$)

p is the *a priori* estimated proportional error, and

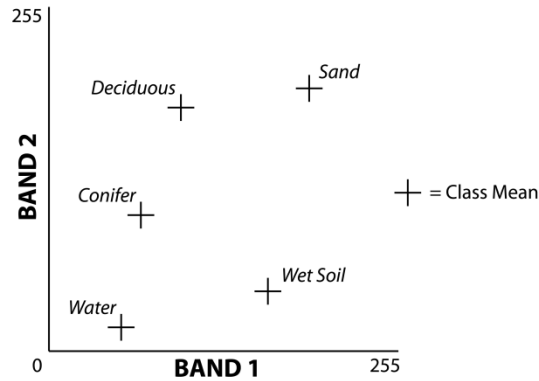
$$q = 1 - p$$

Hard Classifiers

The hard classifiers are so named because they all reach a hard (i.e., unequivocal) decision about the class to which each pixel belongs. They are all based on a logic that describes the expected position of a class (based on training site data) in what is known as *band space*, and then gauging the position of each pixel to be classified in the same band space relative to these class positions. From this perspective, the easiest classifier to understand is the MINDIST procedure.

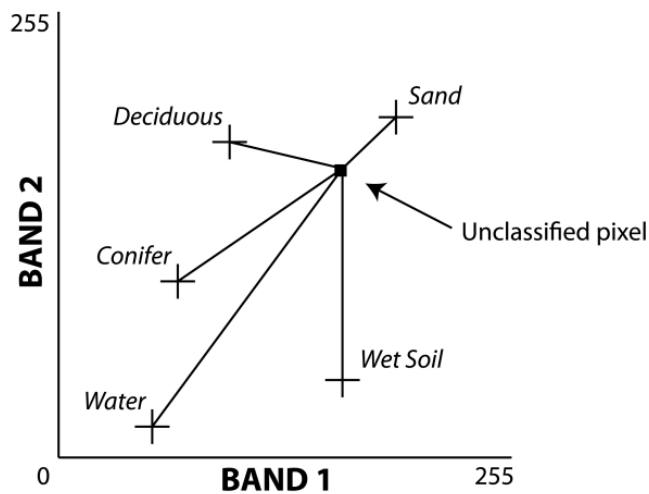
Minimum-Distance-to-Means

The MINDIST module implements a *Minimum-Distance-to-Means* classifier. Based on training site data, MINDIST characterizes each class by its mean position on each band. For example, if only two bands were to be used, the figure below might characterize the positions of a set of known classes as determined from the training site data.

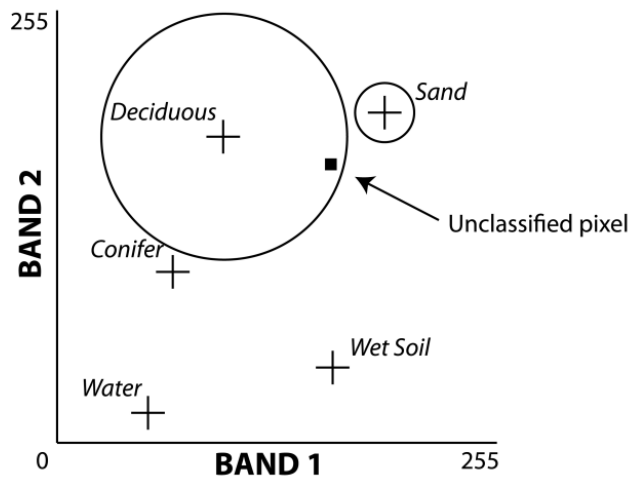


Here each axis indicates reflectance on one of the bands. Thus, using the mean reflectance on these bands as X,Y coordinates, the position of the mean can be placed in this band space. Similarly, the position of any unclassified pixel can also be placed in this space by using its reflectance on the two bands as its coordinates.

To classify an unknown pixel, MINDIST then examines the distance from that pixel to each class and assigns it the identity of the nearest class. For example, the unclassified pixel shown in the figure below would be assigned the "sand" class since this is the class mean to which it is closest.



Despite the simplicity of this approach, it actually performs quite well. It is fast and can employ a maximum distance threshold which allows for any pixels that are unlike any of the given classes to be left unclassified. However, the approach does suffer from problems related to signature variability. By characterizing each class by its mean band reflectances only, it has no knowledge of the fact that some classes are inherently more variable than others. This, in turn, can lead to misclassification. For example, consider the case of a highly variable deciduous class and a very consistent sand class in classifying the unclassified pixel in the figure below.



The circles in this figure illustrate the variability of each of these two classes. If we assume that these circles represent a distance of two standard deviations from the mean, we can see that the pixel lies within the variability range of the deciduous category, and outside that of sand. However, we can also see that it is closer to the mean for sand. In this case, the classifier would misclassify the pixel as sand when it should really be considered to be deciduous forest.

This problem of variability can be overcome if the concept of distance is changed to that of standard scores. This transformation can be accomplished with the following equation:

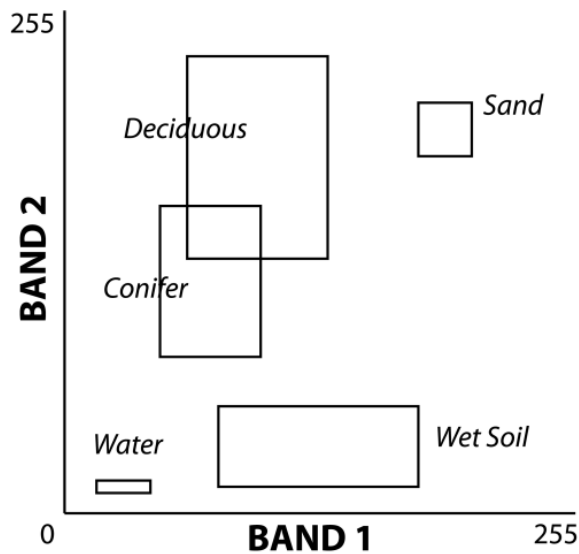
$$\text{standardized distance} = (\text{original distance} - \text{mean}) / \text{standard deviation}$$

The MINDIST procedure in IDRISI Image Processing toolset offers this option of using standardized distances, which is highly recommended. In the example above, the pixel would be correctly classified as deciduous since its standardized distance from the mean for deciduous would be less than 2 (perhaps 1.95 in this illustration), while that for sand would be greater than 2 (probably close to 4 in this illustration).

Our experience with MINDIST has been that it can perform very well when standardized distances are used. Indeed, it often outperforms a maximum likelihood procedure whenever training sites have high variability.

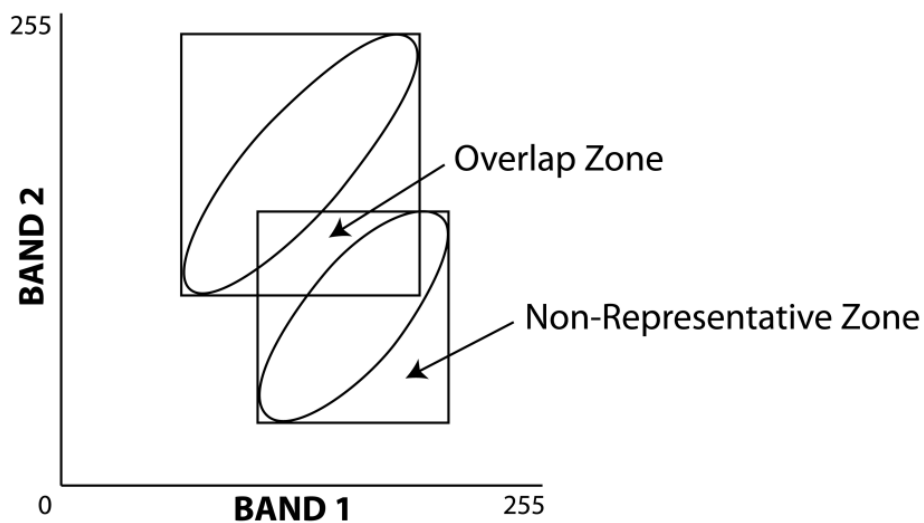
Parallelepiped

The PIPED module implements the parallelepiped procedure for image classification. The parallelepiped procedure characterizes each class by the range of expected values on each band. This range may be defined by the minimum and maximum values found in the training site data for that class, or (more typically) by some standardized range of deviations from the mean (e.g., ± 2 standard deviations). With multispectral image data, these ranges form an enclosed box-like polygon of expected values known as a *parallelepiped*. Unclassified pixels are then given the class of any parallelepiped box they fall within. If a pixel does not fall within any box, it is left unassigned. The figure below illustrates this effect.



This classifier has the advantage of speed and the ability to take into account the differing variability of classes. In addition, the rectangular shape accommodates the fact that variability may be different along different bands. However, the classifier generally performs rather poorly because of the potential for overlap of the parallelepipeds. For example, the conifer and deciduous parallelepipeds overlap in this illustration, leaving a zone of ambiguity in the overlap area. Clearly, any choice of a class for pixels falling within the overlap is arbitrary.

It may seem that the problem of overlapping parallelepipeds would be unlikely. However, they are extremely common because of the fact that image data are often highly correlated between bands. This leads to a cigar-shaped distribution of likely values for a given class that is very poorly approximated by a parallelepiped as shown in the figure below.

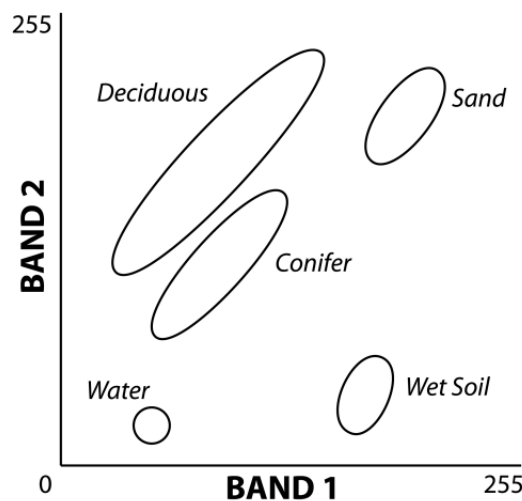


Clearly the MINDIST procedure would not encounter this problem, since the line of separation between these classes would fall in between these two distributions. However, in this context of correlation between bands (which is virtually guaranteed), the parallelepiped procedure produces both zones of overlap and highly non-representative areas that really should not be included in the class. In general, then, the parallelepiped procedure should be avoided, despite the fact that it is the fastest of the supervised classifiers.¹⁰

Maximum Likelihood

To compensate for the main deficiencies of both the Parallelepiped and Minimum-Distance-to-Means procedures, the Maximum Likelihood procedure, provided by the MAXLIKE module in IDRISI Image Processing toolset, is used. The Maximum Likelihood procedure is based on Bayesian probability theory. Using the information from a set of training sites, MAXLIKE uses the mean and variance/covariance data of the signatures to estimate the posterior probability that a pixel belongs to each class.

In many ways, the MAXLIKE procedure is similar to MINDIST with the standardized distance option. The difference is that MAXLIKE accounts for intercorrelation between bands. By incorporating information about the covariance between bands as well as their inherent variance, MAXLIKE produces what can be conceptualized as an elliptical zone of characterization of the signature. In actuality, it calculates the posterior probability of belonging to each class, where the probability is highest at the mean position of the class, and falls off in an elliptical pattern away from the mean, as shown in the figure below.



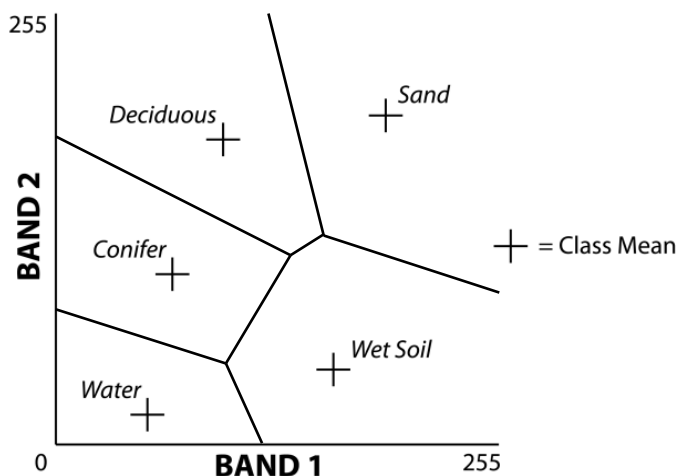
Linear Discriminant Analysis (Fisher Classifier)

The next two classifiers (FISHER and SVM) both aim to segment band space to define regions associated with each class. FISHER does this by estimating the formulas of hyperplanes that intersect band space to create these regions. It does this by conducting a linear discriminant analysis of the training site data to form a set of linear functions that express the degree of support for each class. The assigned class for each pixel is then that class which receives the highest support after evaluation of all functions. These functions have a form similar to that of a multivariate linear regression equation, where the independent variables are the image bands, and the dependent variable is the measure of support. In fact, the equations are calculated such that they maximize the variance between

¹⁰ In the early days of image processing when computing resources were poor, this classifier was commonly used as a quick look classifier because of its speed.

classes and minimize the variance within classes. The number of equations will be equal to the number of bands, each describing a hyperplane of support.

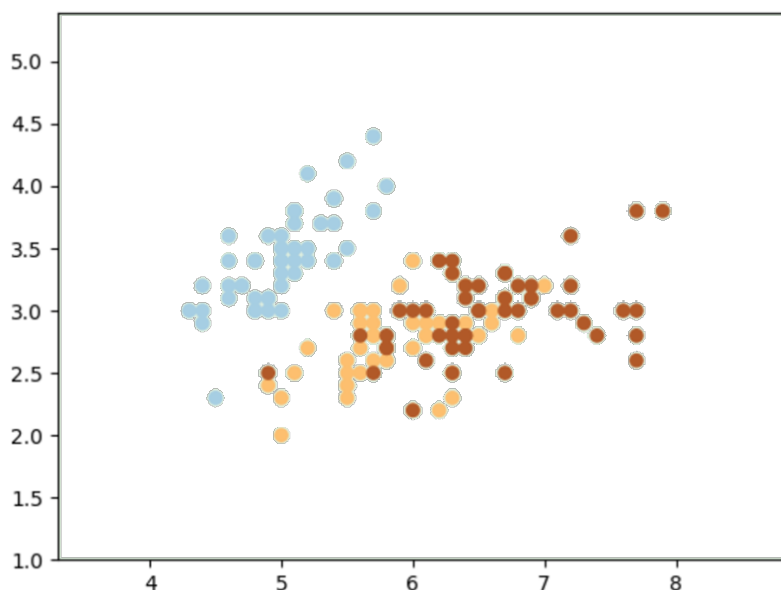
Note that, although it does it in a very different way, the MINDIST classifier is also implicitly setting up a regionalization of band space. In the case of MINDIST it is effectively creating a Voronoi tessellation (Thiessen polygons).



Support Vector

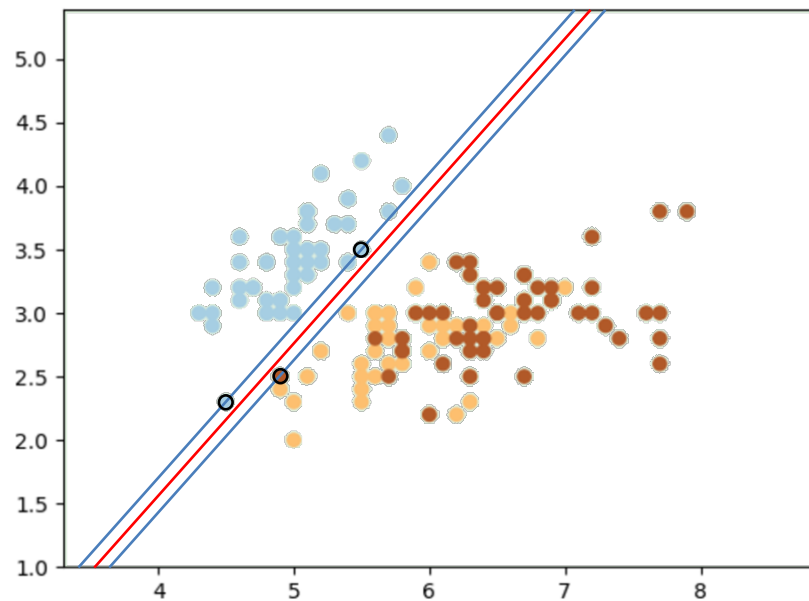
Machine

A Support Vector Machine (SVM) tries to do something similar to FISHER, but with a greater degree of flexibility. The diagram below shows a scatter diagram of the training data for three classes, colored blue, yellow and red. The X and Y axes represent two spectral reflectance bands (actually, it's a well-known dataset about flowers, but we'll pretend they're image bands and land cover classes).

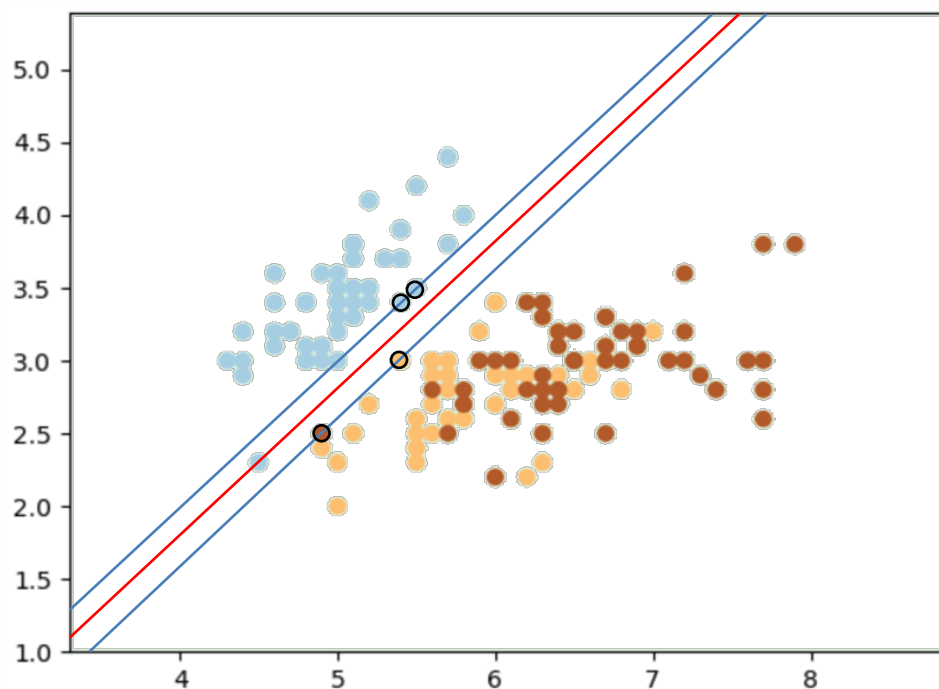


Like FISHER, SVM tries to establish hyperplanes that can separate these classes in band space (see the next diagram). However, unlike FISHER, it tries to establish a hyperplane (the red line) that bisects the widest possible margin (the blue lines) that it can

achieve to separate the classes. In the next diagram, we see the hyperplane and margin separating the blue class from the others. The pixels that touch this margin are known as the support vectors (outlined in black).



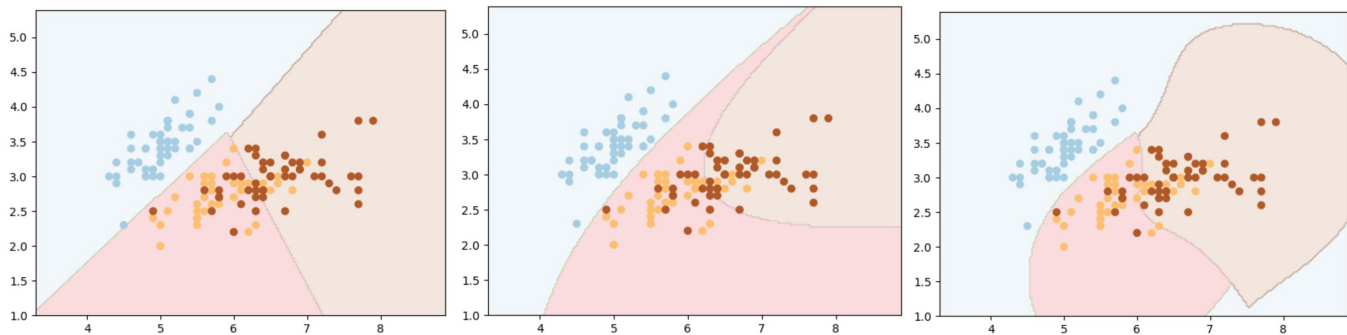
In this example, an even wider margin could be established if we simply ignore the blue pixel with the lowest Y value and consider it to be an outlier.



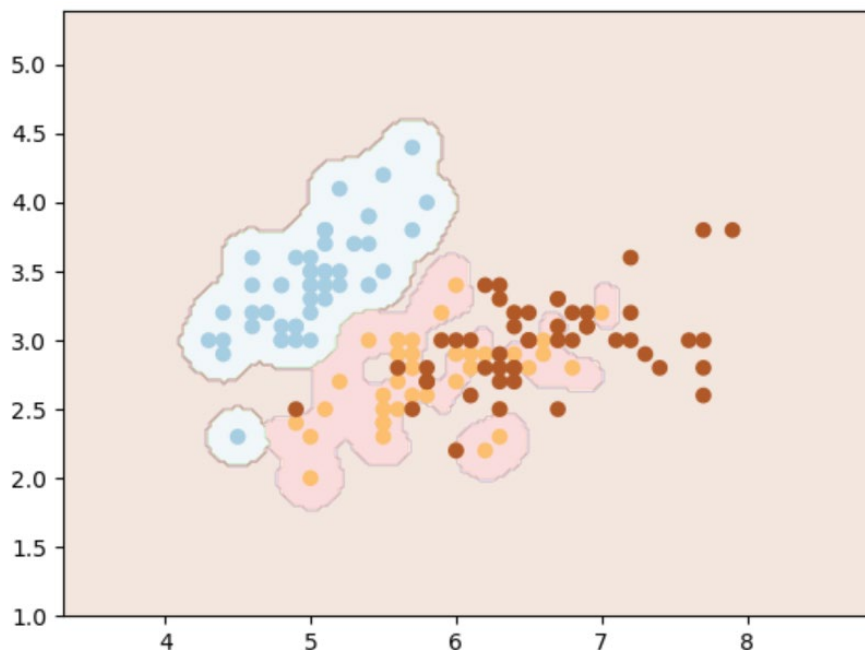
Outliers are not uncommon and may not be very representative of the class. We want the widest margin possible because we're only looking at the training data and we don't know what we'll encounter when we start to classify all the pixels in the image. Wide margins give us the best chance of avoiding misclassification in general (at the cost of misclassifying a few outliers). You can control the degree to which you ignore outliers with the C parameter. C stands for *cost*, and allows you to change the cost of misclassifying

outliers. Technically, it is known as a *regularization* parameter. Use a low value of C (e.g., 1) to ignore outliers and a high C to treat them more as data to respect.

In this example, a linear hyperplane did quite well in separating the blue class from the other two, but it's clearly going to have problems separating the other two. Here is where SVM comes into its own. SVM offers several different kernels that govern the shape of hyperplanes that can be fit. These include linear (left in the diagram below), polynomial (middle) and radial basis function (right) kernels.



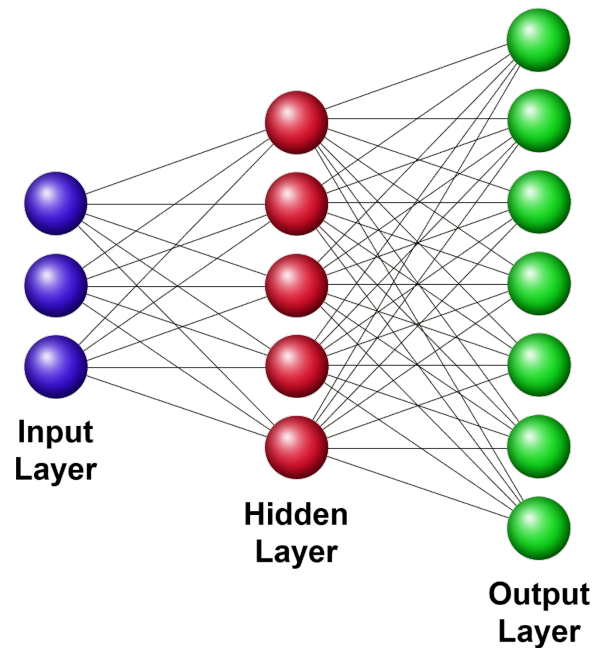
The RBF (radial basis function) is probably the most commonly used, and is the default in TerrSet. It can separate any set of classes, although at a risk of overfitting. Consider this example:



Clearly it did a good job in separating the classes, but how do you know what parameters to use to achieve the desired result? The issue of SVM parameters is complex and beyond the scope of this manual. However, there are many resources on-line. Like Scikit-learn, TerrSet uses LIBSVM (<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>) as its SVM engine so virtually all such resources apply to TerrSet as well. That said, there is typically no way to know the best parameters, so the generally recommended approach is to use the *grid-search* option where TerrSet separates the data into training and validation sets and tests a range of settings.

Neural Networks

TerrSet provides a variety of Artificial Neural Network (ANN) tools, of which probably the most important is the Multi-Layer Perceptron (MLP). MLP conceptualizes the band values for each pixel as a set of input neurons (analogous to sensory neurons) and the output class assignments as output neurons (analogous to motor neurons). Between the two is one or more hidden layers with neurons that are analogous to those in the human cortex. With ANN's these neurons are called *perceptrons*.



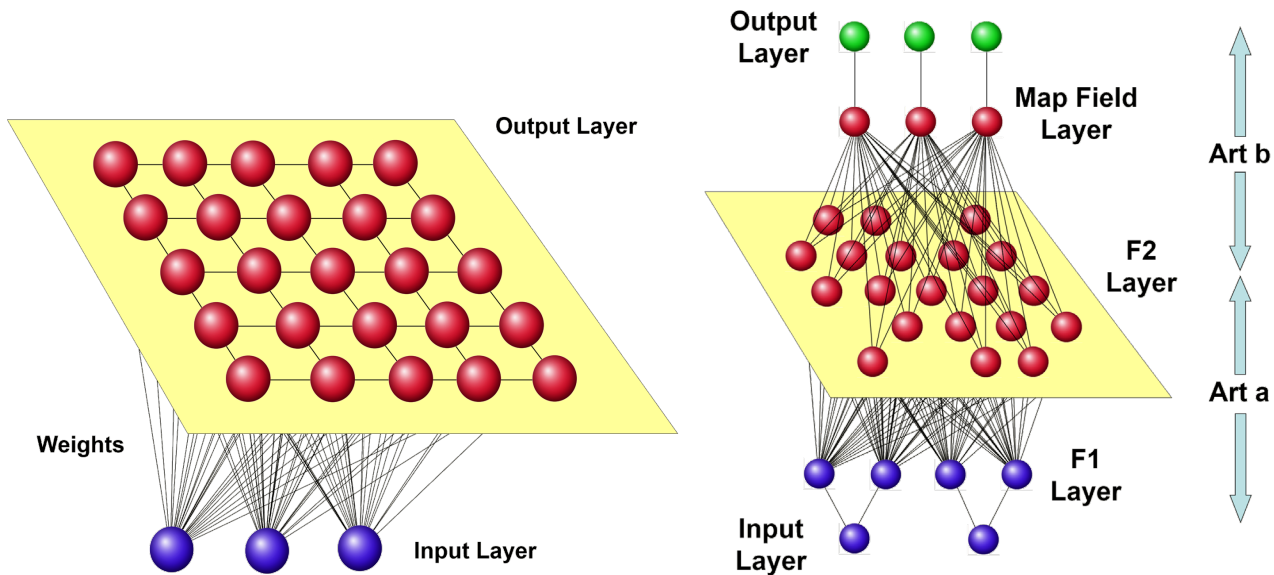
Perceptrons within any of these layers are not connected, while perceptrons between layers are fully connected. Each connection has an associated connection weight. The perceptrons in the input layer gain their values from the pixel reflectances. Those in the hidden and output layers gain their values by multiplying each perceptron value in the previous layer to which it is connected by the appropriate connection weights. These weighted values are then summed, and then the sum is modified by a sigmoidal function that forces the result to be within a range from 0-1. Thus, the perceptrons in the hidden and output layers are actually just multi-criteria evaluation units that perform a weighted linear combination of their connected inputs.

In use, the intention with MLP is that a set of pixel reflectances, when fed through the network and acted on by these perceptrons will result in a value close to 1.0 for the correct land cover class, and values close to 0.0 for the incorrect classes. Thus, the challenge is how to develop the correct connection weights to make this happen. This is done through training. Initially, the connection weights are all set to be random numbers (typically between -1 and +1). Then the reflectances of a training pixel are fed through the network. Since it is a training pixel, MLP knows what the correct answer is. It therefore compares what gets propagated to the output layer with what it should truly be. Clearly, the first time through the network, the error will be substantial. It therefore adjusts the weights a small amount in order to make the error a bit smaller. It then repeats this with the next training site pixel and keeps adjusting the weights until it exhausts the entire training data set. This constitutes one *iteration* (also known as an *epoch*). It then continues adjusting the weights by using the training data again. It is not uncommon to train the network for thousands of iterations. To help you assess when to stop the training, MLP holds back some of the training data for validation and given you an accuracy statement and loss function (RMS error).

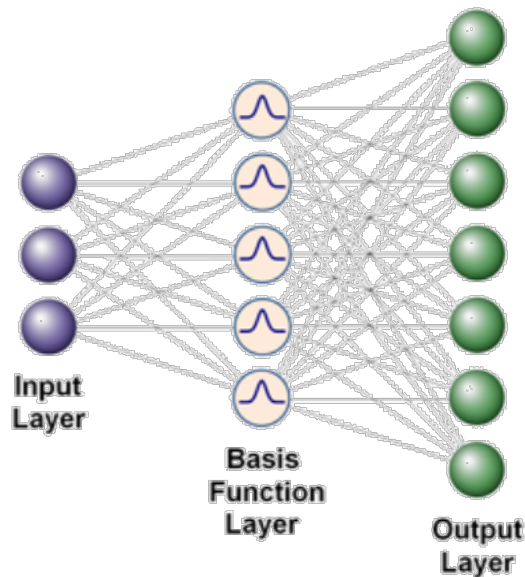
Each of the ANN's in TerrSet operate by learning connection weights in a similar fashion. MLP is probably the most important of these and has exceptional ability in learning complex and non-linear relationships. It is also one of the foundation modeling tools in the Land Change Modeler. Other ANN's include SOM and Fuzzy Artmap. SOM (left in the diagram below) is at it's heart, an unsupervised technique similar to a K-Means (KMEANS in TerrSet) clustering procedure (think of the neurons in the output layer

being like the clusters in a K-Means clustering. It can also be used as a supervised technique after unsupervised training by associating training data with the most similar output neuron.

Fuzzy Artmap is also, at its heart, an unsupervised technique. It is very similar to a chain clustering algorithm (CHAINCLUSTER in TerrSet). It starts with a single neuron and adjusts the weights to try to describe as best as possible, unless a pixel is encountered that would require a very big adjustment. In that case it adds another neuron. It continues in this fashion until it can describe all pixels. In a second stage, a supervised classification can be done by associating neurons with training site classes.

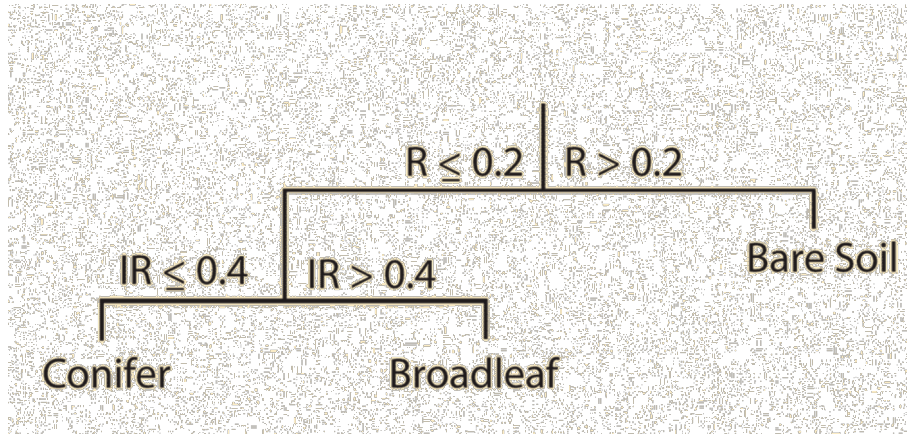


The final ANN supported is a Radial Basis Function NN (RBFNN), which has some similarities to MLP. In this case (diagram below) the hidden layer neurons are determined by K-Means clustering, and the weights between the input layer and the hidden layer describe how similar pixels are to these clusters. In addition, the activation function is a Gaussian that expresses how near (i.e., similar) an input pixels is to the cluster. The second set of connections between the hidden layer and the output layer is achieved with multinomial logistic regression (MULTILOGISTICREG in TerrSet).



Classification Tree Classifiers

TerrSet supports two classification tree classifiers, CTA and DECISIONFOREST. CTA (Classification Tree Analysis) is an implementation of the C4.5 procedure. Based on training data, it develops a decision tree of binary decisions that can ultimately classify pixels correctly. For example, imagine distinguishing between three classes: conifer forest, broadleaf forest and bare soil, using just two bands – red reflectance and infrared reflectance. The tree is upside down, with the root at the top and the leaves at the bottom. Given a pixel, the tree indicates that you should start the red band. If the reflectance is greater than 0.2, then it is bare soil. Otherwise, look at the infrared band. If the IR reflectance is greater than 0.4 it is broadleaf forest, otherwise it is conifer forest.



CTA can learn to classify the training data perfectly. However, it is highly susceptible to overfitting and may perform poorly. One solution is to use *pruning* where some of the terminal leaves are removed, leaving less specific leaves that do contain some impurities. Another, and now the favored approach, is to use multiple trees that are trained on different subsets of the data and to assign classes based on a majority vote. This approach is known as a *Random Forest*.

The implementation of Random Forest in TerrSet is called DECISIONFOREST (because Random Forest is a trademarked name – see https://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm). Note that DECISIONFOREST can also be accessed from the shortcut box by typing DF. Also note that DF can be accessed as a standalone module for the classification of remotely sensed imagery and also as an empirical modeling technique in LCM. This section describes its use for image classification.

Like all of TerrSet's image classifiers, use of DF requires the specification of a training site image and a set of image bands. Based on the number of image bands, DF has an algorithm to determine the size of the subset of variables to be used in determining each split. For example, CTA will consider all of the image bands in determining candidate variables for each split. With DF, it will only consider a random subset of bands. For example, with 7 Landsat bands it will, by default, make a random selection of just 2 bands as it determines each split. It will also work with a different random subset of the training data. Training data not used is then used for validation to estimate what is known as the "Out of Bag" error. Clearly each tree will differ because they develop based on consideration of different variable and different subsets of data. However, with a large number of trees, it is reasoned that the most common classification will be the correct one. Typically, the number of trees developed is 100 or more.

DF is a popular classifier because it is capable of performing well with classes that are not normally distributed. It can even perform well with classes that are multimodal. However, that flexibility comes at the cost of computational time in the classification stage. For example, if the model is developed with 500 trees, then 500 trees need to be traversed in order to classify each pixel.

Soft Classifiers

Unlike hard classifiers, soft classifiers defer making a definitive judgment about the class membership of any pixel in favor of a group of statements about the degree of membership of that pixel in each of the possible classes. Like traditional supervised classification procedures, each uses training site information for the purpose of classifying each image pixel. However, unlike traditional hard classifiers, the output is not a single classified landcover map, but rather a set of images (one per class) that express for each pixel the degree of membership in the class in question. In fact, each expresses the degree to which each pixel belongs to the set identified by a signature according to one of the following set membership metrics:

BAYCLASS	based on Bayesian probability theory ,
BELCLASS	based on Dempster-Shafer theory,
MAHALCLASS	based on Mahalanobis distance,
FUZCLASS	based on Fuzzy Set theory, and
UNMIX	based on the Linear Mixture model.
MULTILOGISTICREG	based on multinomial logistic probabilities.

It is important to recognize that each of these falls into a general category of what are known as *Fuzzy Measures* (Dubois and Prade, 1982) of which Fuzzy Sets is only one instance. Fuzziness can arise for many reasons and not just because a set is itself fuzzy. For example, measurement error can lead to uncertainty about the class membership of a pixel even when the classes (sets) are crisply defined. It is for this reason that we have adopted the term *soft*—it simply recognizes that the class membership of a pixel is frequently uncertain for reasons that are varied in origin.

Image Group Files

Since the output of each of these soft classifiers is a set of images, each also outputs a raster image group file (.rgf). This can be used with cursor inquiry to examine the set membership values for a pixel in each class simultaneously in either numeric or graph form (see the Help System section on Display). Note that the classification uncertainty image described below is also included in each group file produced.

Classification Uncertainty

In addition to these set membership images, each of these soft classifiers outputs an image that expresses the degree of *classification uncertainty* it has about the class membership of any pixel. Classification uncertainty measures the degree to which no class clearly stands out above the others in the assessment of class membership of a pixel. In the case of BAYCLASS, BELCLASS and FUZCLASS, it is calculated as follows:

$$\text{ClassificationUncertainty} = 1 - \frac{\max - \frac{\text{sum}}{n}}{1 - \frac{1}{n}}$$

where

max = the maximum set membership value for that pixel

sum = the sum of the set membership values for that pixel

n = the number of classes (signatures) considered

The logic of this measure is as follows:

The numerator of the second term expresses the difference between the maximum set membership value and the total dispersion of the set membership values over all classes.

The denominator of the second term expresses the extreme case of the difference between a maximum set membership value of 1 (and thus total commitment to a single class) and the total dispersion of that commitment over all classes.

By taking the ratio of these two quantities, one develops a measure that expresses the degree of commitment to a specific class relative to the largest possible commitment that can be made. Classification uncertainty is thus the complement of this ratio.

In spirit, the measure of uncertainty developed here is similar to the entropy measure used in Information Theory. However, it differs in that it is concerned not only with the degree of dispersion of set membership values between classes, but also the total amount of commitment present. Following are some examples that can clarify this concept.

Examples

Assuming a case where three classes are being evaluated, consider those with the following allocations of set membership:

(0.0 0.0 0.0)	Classification Uncertainty = 1.00
(0.0 0.0 0.1)	Classification Uncertainty = 0.90
(0.1 0.1 0.1)	Classification Uncertainty = 1.00
(0.3 0.3 0.3)	Classification Uncertainty = 1.00
(0.6 0.3 0.0)	Classification Uncertainty = 0.55
(0.6 0.3 0.1)	Classification Uncertainty = 0.60
(0.9 0.1 0.0)	Classification Uncertainty = 0.15
(0.9 0.05 0.05)	Classification Uncertainty = 0.15
(1.0 0.0 0.0)	Classification Uncertainty = 0.00

With UNMIX, however, classification uncertainty is measured as the residual error after calculation of the fractions of constituent members. This will be discussed further below.

BAYCLASS and Bayesian Probability Theory

BAYCLASS is a direct extension of the MAXLIKE module. It outputs a separate image to express the posterior probability of belonging to each considered class according to Bayes' Theorem :

$$p(h|e) = \frac{p(e|h) - p(h)}{\sum_i p(e|h_i) - p(h_i)}$$

where :

$p(h e)$	= the probability of the hypothesis being true given the evidence (posterior probability)
$p(e h)$	= the probability of finding that evidence given the hypothesis being true
$p(h)$	= the probability of the hypothesis being true regardless of the evidence (prior probability)

In this context, the variance/covariance matrix derived from training site data is that which allows one to assess the multivariate conditional probability $p(e|h)$. This quantity is then modified by the prior probability of the hypothesis being true and then normalized by the sum of such considerations over all classes. This latter step is important in that it makes the assumption that the classes considered are the only classes that are possible as interpretations for the pixel under consideration. Thus even weak support for a specific interpretation may appear to be strong if it is the strongest of the possible choices given.

This posterior probability $p(h|e)$ is the same quantity that MAXLIKE evaluates to determine the most likely class, and indeed, if the output images of BAYCLASS were to be submitted directly to HARDEN, the result would be identical to that of MAXLIKE. In essence, BAYCLASS is a *confident* classifier. It assumes that the only possible interpretation of a pixel is one of those classes for which training site data have been provided. It therefore admits to no ignorance. As a result, lack of evidence for an alternative hypothesis constitutes support for the hypotheses that remain. In this context, a pixel for which reflectance data only very weakly support a particular class is treated as unequivocally belonging to that class ($p = 1.0$) if no support exists for any other interpretation.

The prime motivation for the use of BAYCLASS is sub-pixel classification —i.e., to determine the extent to which mixed pixels exist in the image and their relative proportions. It is also of interest to observe the underlying basis of the MAXLIKE procedure. However, for *In-Process Classification Assessment* (IPCA), the BELCLASS procedure is generally preferred because of its explicit recognition that some degree of ignorance may surround the classification process.

In the context of mixture analysis, the probabilities of BAYCLASS are interpreted directly as statements of proportional representation. Thus if a pixel has posterior probabilities of belonging to deciduous and conifer of 0.68 and 0.32 respectively, this would be interpreted as evidence that the pixel contains 68% deciduous species and 32% conifers. Note, however, that this requires several important assumptions to be true. First, it requires that the classes for which training site data have been provided are exhaustive (i.e., that there are no other possible interpretations for that pixel). Second, it assumes that the conditional probability distributions $p(e|h)$ do not overlap in the case of pure pixels. In practice, these conditions may be difficult to meet.

In testing at Clark Labs, we have found that while BAYCLASS is effective in determining the constituent members of mixed pixels, it is often not so effective in determining the correct proportions. Rather, we have found that procedures based on the Linear Mixture model (UNMIX) perform considerably better in this respect. However, Linear Spectral Unmixing has its own special limitations. Thus, we favor a hybrid approach using the better qualities of Bayesian decomposition and Linear Spectral Unmixing, as will be discussed below.

BELCLASS and Dempster-Shafer Theory

BELCLASS is probably the most complex of the soft classifier group in its underlying theory. It is based on Dempster-Shafer theory—a variant of Bayesian probability theory that explicitly recognizes the possibility of ignorance. Dempster-Shafer theory is explained more fully in the **Decision Support: Uncertainty Management** chapter. However, a good introduction can be provided by considering the output from BAYCLASS.

Consider a classification where training sites have been developed for the classes [conifer], [deciduous], [grass], [urban], and [water]. If a pixel shows some degree of similarity to [conifer] and to no others, BAYCLASS will assign a value of 1.0 to that class and 0.0 to all others. In fact, it will do so even if the actual support for the [conifer] class is low, because Bayesian probability theory does not recognize the concept of ignorance. It assumes that lack of evidence *for* a hypothesis constitutes evidence *against* that hypothesis. Thus, in this example, the absence of evidence for any combination of [deciduous grass urban water] is therefore interpreted as evidence that it must be [conifer], no matter how weak the direct evidence for [conifer] actually is (so long as it is greater than 0).

In contrast to this, Dempster-Shafer theory does not assume that it has full information but accepts that the state of one's knowledge may be incomplete. The absence of evidence about a hypothesis is treated as just that—lack of evidence. Unlike Bayesian probability theory, it is not assumed that this, therefore, constitutes evidence against that hypothesis. As a consequence, there can be a difference between one's belief in an hypothesis and one's attendant disbelief in that same hypothesis!

In the language of Dempster-Shafer theory, the degree to which evidence provides concrete support for an hypothesis is known as *belief*, and the degree to which the evidence does not refute that hypothesis is known as *plausibility*. The difference between these two is then known as a *belief interval*, which acts as a measure of uncertainty about a specific hypothesis.

Returning to the previous example, if the evidence supports [conifer] to the degree 0.3 and all other classes to the degree 0.0, Bayesian probability would assign a posterior probability of 1.0 to [conifer]. However, Dempster-Shafer theory would assign a belief

of 0.3 to [conifer] and a plausibility of 1.0, yielding a belief interval of 0.7. Furthermore, it would assign a belief of 0.0 to all other classes and a plausibility of 0.7.

If a second piece of evidence were then to be considered, and it was found that this evidence supported [urban] to a degree of 0.6 and gave no support to any other hypothesis, this would affect the hypothesis [conifer] by lowering its plausibility to 0.4. At this point, then, belief in [conifer] is 0.3 and plausibility is 0.4. Thus our uncertainty about this class is very low (0.1).

The combination of evidence is generally somewhat more complex than these contrived examples would suggest and uses a logic known as *Dempster's Rule*. The Belief module implements this rule. However, BELCLASS is less concerned with the combination of evidence than it is with decomposing the evidence to determine the degree of support (expressed as belief or plausibility) for each of the classes for which training data have been supplied.¹¹

In addition to the concepts of belief and plausibility, the logic of Dempster-Shafer theory can also express the degree to which the state of one's knowledge does not distinguish between the hypotheses. This is known as *ignorance*.

Ignorance expresses the incompleteness of one's knowledge as a measure of the degree to which we cannot distinguish between any of the hypotheses. Using the example above, ignorance thus expresses one's commitment to the indistinguishable set of all classes [conifer deciduous grass urban water]—the inability to tell to which class the pixel belongs.

In normal use, Dempster-Shafer theory requires that the hypotheses (classes) under consideration be mutually exclusive and exhaustive. However, in developing BELCLASS, we felt that there was a strong case to be made for non-exhaustive categories—that the pixel may indeed belong to some unknown class, for which a training site has not been provided. In order to do this we need to add an additional category to every analysis called [other], and assign any incompleteness in one's knowledge to the indistinguishable set of all possible classes (including this added one)—e.g., [conifer deciduous grass urban water other]. This yields a result which is consistent with Dempster-Shafer theory, but which recognizes the possibility that there may be classes present about which we have no knowledge.

Defining ignorance as a commitment to the indistinguishable set of all classes suggests that it may have some relationship to the classification uncertainty image produced by BAYCLASS. The classification uncertainty image of BAYCLASS expresses the uncertainty the classifier has in assigning class membership to a pixel. Uncertainty is highest whenever there is no class that clearly stands out above the others in the assessment of class membership for a pixel. We have found that in the context of BELCLASS, this measure of uncertainty is almost identical to that of Dempster-Shafer ignorance. As a result, we have modified the output of the classification uncertainty image of BELCLASS slightly so that it outputs true Dempster-Shafer ignorance.¹²

The operation of BELCLASS is essentially identical to that of BAYCLASS and thus also MAXLIKE. Two choices of output are given: beliefs or plausibilities. In either case, a separate image of belief or plausibility is produced for each class. In addition, a classification uncertainty image is produced which can be interpreted in the same manner as the classification uncertainty image produced by all of the soft classifiers (but which is truly a measure of Dempster-Shafer ignorance).

The prime motivation for the use of BELCLASS is to check for the quality of one's training site data and the possible presence of unknown classes during *In-Process Classification Assessment*. In cases where one believes that one or more unknown classes exist (and thus that some portion of total ignorance arises because of this presence of an unknown class), the BELCLASS routine should be used. BELCLASS does this by implicitly adding an [other] class to the set of classes being considered. This is a theoretical concession to the mechanics of the BELCLASS process and will not be directly encountered by the user.

Comparing the output of BELCLASS with BAYCLASS, you will notice a major difference. Looking at the images produced by BAYCLASS, it will appear as if your training sites are strong. BAYCLASS is a very confident classifier (perhaps overly confident) since it assumes no ignorance. BELCLASS, however, appears to be a very reserved classifier. Here we see a result in which all of the

¹¹ The logic of the decomposition process is detailed in the module description for BELCLASS in the on-line Help System.

¹² The filename for the classification uncertainty image in BELCLASS is composed by concatenating the prefix supplied by the user and the letter string "clu".

uncertainties in our information become apparent. It does not presume to have full information, but explicitly recognizes the possibility that one or more unknown classes may exist.

BELCALC and MAXSET

In BELCLASS, concern is directed to the degree of membership that each pixel exhibits for each of the classes for which training data have been provided. However, the logic of Dempster-Shafer theory recognizes a whole hierarchy of classes, made up of the indistinguishable combinations of these basic classes. For example, given basic classes (called *singletons*) of [conifer] [deciduous] [grass], Dempster-Shafer theory recognizes the existence of all of the following classes:¹³

- [conifer]
- [deciduous]
- [grass]
- [conifer deciduous]
- [conifer grass]
- [deciduous grass]
- [conifer deciduous grass]

Dempster-Shafer theory also allows one to make two different kinds of assessments about each of these classes. The first is clearly belief. The second is known as a *Basic Probability Assignment* (BPA). Both require further explanation.

When evidence provides some degree of commitment to one of these non-singleton classes and not to any of its constituents separately, that expression of commitment is known as a Basic Probability Assignment (BPA). The BPA of a non-singleton class thus represents the degree of support for the presence of one or more of its constituents, but without the ability to tell which.

Given this understanding of a BPA, belief in a non-singleton class is then calculated as the sum of BPAs for that class and all sub-classes. For example, to calculate the belief in the class [conifer deciduous], you would add the BPAs for [conifer deciduous], [conifer] and [deciduous]. Belief is thus a broader concept than a BPA. It represents the total commitment to all members of a set combined.

In the context of remote sensing, these non-singleton classes are of interest in that they represent mixtures, and thus might be used for a more detailed examination of sub-pixel classification. However, the sheer number of such classes makes it impractical to have a software module that outputs all possible classes. For a set of n singleton classes, the total number of classes in the entire hierarchy is $(2^n - 1)$. Thus, in a case where 16 landcover classes are under consideration, 65,535 classes are included in the full hierarchy—over two terabytes of output for a full Landsat scene!

As it turns out, however, only a small number of these non-singleton classes contain any significant information in a typical application. These can very effectively be determined by running the MAXSET module. MAXSET is a hard classifier that assigns to each pixel the class with the greatest degree of commitment from the full Dempster-Shafer class hierarchy. The significance of these mixtures can further be determined by running the AREA module on the MAXSET result. Then BELCALC can be used to calculate the mixture BPA that underlies the MAXSET result.

Belief

Both BELCLASS and BELCALC deconstruct the evidence to infer belief and plausibility for each class. One of the motivations for doing so is that it allows the user to combine ancillary information with that determined from the reflectance data. New evidence can

¹³ Clearly these are sets of classes. However, since evidence may support one of these sets without further distinction about which members of the set are supported, the set itself can be thought of as a class.

be combined with existing knowledge with the Belief module. Belief is described more fully in the chapter on **Decision Support: Uncertainty Management**, and is used in exactly the same manner for the data described here.

MULTILOGISTICREG and Multinomial Logistic Regression

Logistic Regression is used for regression problems where the dependent variable is dichotomous. This kind of variable violates the assumptions of the linear regression model. Logistic regression solves the problem by using the data to estimate the probability that the dependent variable is 1. This probability distribution follows a logistic curve. Multinomial Logistic Regression estimates these probabilities for multiple classes. Thus, the MULTILOGISTIC module can be used to estimate the probabilities of each class (soft classifications). It also produces a hard output by assigning the class with the highest probability. Clearly, this module is outputting probabilities, just like BAYCLASS. However, the results are slightly different. Those interested in exploring these differences might wish to consult *John Hogland, Nedret Billor & Nathaniel Anderson (2013) Comparison of standard maximum likelihood classification and polytomous logistic regression used in remote sensing, European Journal of Remote Sensing, 46:1, 623-640, DOI: 10.5721/EuJRS20134637*.

FUZCLASS and Fuzzy Set Theory

The third soft classifier is FUZCLASS. As the name suggests, this classifier is based on the underlying logic of Fuzzy Sets. Just as BAYCLASS and BELCLASS are based on the fundamental logic of MAXLIKE, FUZCLASS is based on the underlying logic of MINDIST—i.e., fuzzy set membership is determined from the distance of pixels from signature means as determined by MAKESIG.

There are two important parameters that need to be set when using FUZCLASS. The first is the z-score distance where fuzzy membership becomes zero. The logic of this is as follows.

It is assumed that any pixel at the same location in band space as the class mean (as determined by running MAKESIG) has a membership grade of 1.0. Then as we move away from this position, the fuzzy set membership grade progressively decreases until it eventually reaches zero at the distance specified. This distance is specified as a standard score (z-score) to facilitate its interpretation. Thus, specifying a distance of 1.96 would force 5% of the data cells to have a fuzzy membership of 0, while 2.58 would force 1% to have a value of 0.

The second required parameter setting is whether or not the membership values should be normalized. Normalization makes the assumption (like BAYCLASS) that the classes are exhaustive, and thus that the membership values for all classes for a single pixel must sum to 1.0. This is strictly required to generate true fuzzy set membership grades. However, as a counterpart to BELCLASS, the option is provided for the calculation of un-normalized values. As was suggested earlier, this is particularly important in the context of In-Process Classification Assessment for evaluation of a MINDIST-based classification.

UNMIX and the Linear Mixture Model

The Linear Mixture Model assumes that the mixture of materials within a pixel will lead to an aggregate signature that is an area-weighted average of the signatures of the constituent classes. Thus if two parent materials (called end members in the language of Linear Spectral Unmixing) had signatures of 24, 132, 86 and 56, 144, 98 on three bands, a 50/50 mixture of the two should yield a signature of 40, 138, 92. Using this simple model, it is possible to estimate the proportions of end member constituents within each pixel by solving a set of simultaneous equations. For example, if we were to encounter a pixel with the signature 32, 135, 89, and assumed that the pixel contained a mixture of the two end members mentioned we could set up the following set of equations to solve:

$$f_1(24) + f_2(56) = 32$$

$$f_1(132) + f_2(144) = 135$$

$$f_1(86) + f_2(98) = 89$$

where f_1 and f_2 represent the fractions (proportions) of the two end members. Such a system of simultaneous equations can be solved using matrix algebra to yield a best fit estimate of f_1 and f_2 (0.75 and 0.25 in this example). In addition, the sum of squared residuals between the fitted signature values and the actual values can be used as a measure of the uncertainty in the fit.

The primary limitation of this approach is that the number of end members cannot exceed the number of bands. This can be a severe limitation in the case of SPOT imagery, but of little consequence with hyperspectral imagery. The IDRISI Image Processing toolset thus offers three approaches (in UNMIX) to Linear Spectral Unmixing:

The standard linear spectral unmixing approach (as indicated above) for cases where sufficient bands are available.

A probability guided option for cases where insufficient bands exist. Although the total number of possible end members may be large, the number that coexist within a single pixel is typically small (e.g., 2-3). This approach thus uses a first stage based on the BAYCLASS module to determine the most likely constituents (up to the number of bands), with a second stage linear spectral unmixing to determine their fractions. Experiments at Clark Labs have shown this to produce excellent results.

An exhaustive search option for cases where insufficient bands exist. In this instance, one specifies the number of constituents to consider (up to the total number of bands). It then tests all possible combinations of that many end members and reports the fractions of that combination with the lowest sum of squared residuals. This approach is considerably slower than the other options. In addition, experiments at Clark Labs have shown that this approach yields inferior results to the probability guided procedure in cases where the end member signatures are drawn from training sites. Pure end member signatures, created with ENDSIG, should be used.

End member signatures are specified using standard signature (.sig) files, created using either MAKESIG or ENDSIG. The former is used in the case where end members are derived from training sites, while the latter is used in cases where pure end member values are known (such as from a spectral library). Note that only the mean value on each band is used from these signature files—variance/covariance data are ignored.

Accommodating Ambiguous (Fuzzy) Signatures in Supervised Classification

All the discussions to this point have assumed that the signature data were gathered from pure examples of each class. However, it sometimes happens that this is impossible. For example, it might be difficult to find a pure and uniform stand of white pine forest, because differences in tree spacing permit differing levels of the understory material (perhaps a mixture of soil, dead needles and ferns) to show through the canopy. From the perspective of a single pixel, therefore, there are different grades of membership in the white pine class, ranging from 0.0, where no white pine trees are present within the pixel to 1.0 where a dense closed canopy of white pine exists (such as a plantation). Thus, even though the white pine set (class) is itself inherently crisp, our gathering of data by pixels that span several to many meters across forces our detection of that set to be necessarily fuzzy in character.

Wang (1990) has cited examples of the above problem to postulate that the logic of class membership decision problems in remote sensing is that of Fuzzy Sets. However, as indicated earlier in this chapter, while this problem truly belongs in the realm of Fuzzy Measures, it is not strictly one of Fuzzy Sets in most cases. For example, the white pine class mentioned earlier is not a fuzzy set—it is unambiguous. It is only our difficulty in detecting it free of other cover types that leads to ambiguity. The problem is thus one of imprecision that can best be handled by a Bayesian or Dempster-Shafer procedure (which is, in fact, ultimately the procedure used by Wang, 1990).

As pointed out earlier, ambiguity can arise from a variety of sources, and not just fuzzy sets. However, the special case of resolution and mixed pixels is one that is commonly encountered in the classification process. As a result, it is not unusual that even the best signatures have some degree of inter-class mixing. In such cases, it is desirable to use a procedure for signature development that can recognize this ambiguity. Wang (1990) has proposed an interesting procedure for signature development in this context that we have implemented in a module named FUZSIG. As the name suggests, the module is a variant of MAKESIG for the special case of

ambiguous (i.e., fuzzy) training site data. However, we use the association with fuzzy in the more general sense of fuzzy measures. The logic of the procedure is built around the concept of mixtures and should be restricted to instances where the source of the fuzziness is accommodated.

The use of FUZSIG requires that a specific sequence of operations be followed.

1. Define Training Sites

This stage proceeds much the same as usual: training sites are digitized using the on-screen digitizing facility. However, there is no requirement that training sites be as homogeneous as possible—only that the relative proportions of cover types within each training site pixel can be estimated.

2. Rasterize the Training Sites

Although this stage is not strictly necessary, it can make the next stage much easier if the training sites are collected into a single raster image. This is done by running INITIAL to create a blank byte binary image (i.e., one initialized with zeros), and then rasterizing the digitized polygons with RASTERVECTOR.

3. Create Fuzzy Partition Matrix in Database Workshop

The next step is to create a *fuzzy partition matrix* in Database Workshop. A fuzzy partition matrix indicates the membership grades of each training site in each class. To do this, first set up a database with an integer identifier field and one field (column) per information class in 4-byte real number format. Then put numeric identifiers in the ID field corresponding to each of the training sites (or training site groups if more than one polygon is used per class). For N classes and M training sites, then, an $N \times M$ matrix is formed.

The next step is to fill out the fuzzy partition matrix with values to indicate the membership grades of each training site (or training site group) in the candidate classes. This is best filled out by working across the columns of each row in turn. Since each row represents a training site (or training site group), estimate the proportions of each class that occur within the training site and enter that value (as a real number from 0.0 to 1.0). In this context, the numbers should add to 1.0 along each row, but typically will not along each column.

Once the fuzzy partition matrix is complete, use the Export as Values File option from the Database Workshop File menu to create a series of values files, one for each class. Next create a series of raster images expressing the membership grades for each class. To do so, in The IDRISI Image Processing toolset use ASSIGN to assign each values file to the training site raster image created in the previous step (this is the feature definition image for ASSIGN). Name the output image for each class using a name that is a concatenation of the letters "fz" plus the name of the class. For example, if the class is named "conifer", the output produced with the ASSIGN operation would be named "fzconifer." This "fz" prefix is a requirement for the next stage.

4. Extract Fuzzy Signatures

The next stage is to create the fuzzy signatures. This is done with the FUZSIG module. The operation of FUZSIG is identical to MAKESIG. You will need to specify the name of the file that defines your signatures (if you followed the steps above, this will be the image file you created in Step 2), the number of bands to use in the signature development, the names of the bands and the names of the signatures to be created. It is very important, however, that the signature names you specify coordinate with the names of the fuzzy membership grade images you created in Step 3. Continuing with the previous example, if you specify a signature name of "conifer", it will expect to find an image named "fzconifer" in your working directory. Similarly, a signature named "urban" would be associated with a fuzzy membership grade image named "fzurban."

The output from FUZSIG is a set of signature files (.sig) of identical format to those output from MAKESIG. FUZSIG gives each pixel a weight proportional to its membership grade in the determination of the mean, variance and covariance of each band for each class (see Wang, 1990). Thus a pixel that is predominantly composed of conifers will have a large weight in the determination of the conifer signature, but only a low weight in determining the signature for other constituents.

Because these signature files are of identical format to those produced by MAKESIG, they can be used with any of the classifiers supported by the IDRISI Image Processing toolset, both hard and soft. However, there are some important points to note about these files:

The minimum and maximum reflectances on each band cannot meaningfully be evaluated using the weighting procedure of FUZSIG. As a consequence, the minimum and maximum value recorded are derived only from those pixels where the membership grade for the signature of concern is greater than 0.5. This will typically produce a result that is identical to the output of MAKESIG. However, it is recommended that if the PIPED classifier is to be used with these signatures, the parallelepipeds should be defined by standard deviation units rather than the minimum and maximum data values.

Unlike MAKESIG, FUZSIG does not output a corresponding set of signature pixel files (.spf) to the signature files produced. However, since the signature files are simple ASCII text files, they can be examined in Edit. Their simple structure is explained in the Help System.

Hardeners

Once a soft classifier has been applied to a multispectral image set, the soft results can be re-evaluated to produce a hard classification by using one of the following hardeners from the module HARDEN:

BAYCLASS

Using the results from BAYCLASS, this option determines the class possessing the maximum posterior probability for each cell, given a set of probability images. Up to four levels of abstraction can be produced. The first is the most likely class, just described. The second outputs the class of the second highest posterior probability, and so on, up to the fourth highest probability.

BELCLASS

Using the results from BELCLASS, this option is essentially identical to the BAYCLASS hardener, except that it is designed for use with Dempster-Shafer beliefs.

FUZCLASS

Using the results from FUZCLASS, this option is essentially identical to the BAYCLASS hardener, except that it is designed for use with Fuzzy Sets.

UNMIX

Using the results from UNMIX, this option is essentially identical to the BAYCLASS hardener, except that it is designed for use with the mixture fractions produced by UNMIX.

MAHALCLASS

Using the results from MAHALCLASS, this option is essentially identical to the BAYCLASS hardener, except that it is designed for use with the typicalities produced by MAHALCLASS.

Unsupervised Classification

General Logic

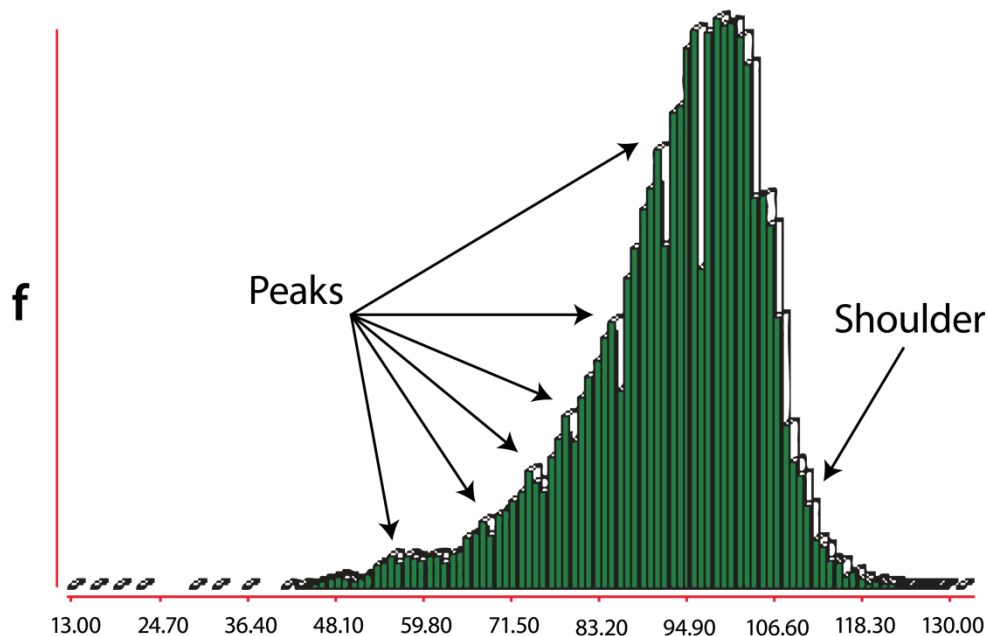
Unsupervised classification techniques share a common intent to uncover the major landcover classes that exist in the image without prior knowledge of what they might be. Generically, such procedures fall into the realm of *cluster analysis*, since they search for clusters of pixels with similar reflectance characteristics in a multi-band image. They are also all generalizations of landcover occurrence since they are concerned with uncovering the *major* landcover classes, and thus tend to ignore those that have very low frequencies of occurrence. However, given these broad commonalities, there is little else that they share in common. There are almost as many approaches to clustering as there are image processing systems on the market. TerrSet is no exception. The primary unsupervised procedure the IDRISI Image Processing toolset offers is unique (CLUSTER). However, TerrSet also offers a commonly used procedure (ISODATA) a variant on this (ISOCLUST). As implemented here, ISOCLUST is really an iterative combination of unsupervised and supervised procedures, as is also the case with the third procedure offered, MAXSET. TerrSet also has a true K-means clustering classifier. TerrSet offers additional supervised modules that also have unsupervised capabilities: SOM and Fuzzy ARTMAP, act in this fashion.

CLUSTER

The CLUSTER module in the IDRISI Image Processing toolset implements a special variant of a Histogram Peak cluster analysis technique (Richards, 1993). The procedure can best be understood from the perspective of a single band. If one had a single band of data, a histogram of the reflectance values on that band would show a number of peaks and valleys. The peaks represent clusters of more frequent values associated with commonly occurring cover types.

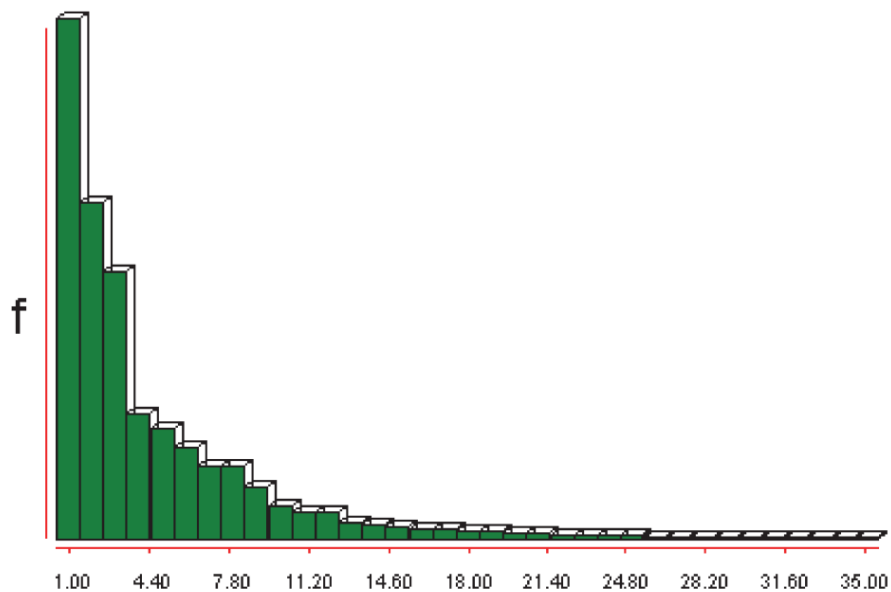
The CLUSTER procedure thus searches for peaks by looking for cases where the frequency is higher than that of its immediate neighbors on either side. In the case of two bands, these peaks would be hills, while for three bands they would be spheres, and so on. The concept can thus be extended to any number of bands. Once the peaks have been located, each pixel in the image can then be assigned to its closest peak, with each such class being labeled as a cluster. It is the analyst's task to then identify the landcover class of each cluster by looking at the cluster image and comparing it to ground features.

CLUSTER offers two levels of generalization. With the broad level of generalization, clusters must occur as distinct peaks in the multi-dimensional histogram as outlined above. However, with the fine level of generalization, CLUSTER also recognizes shoulders in the curve as cluster peaks. Shoulders occur when two adjacent clusters overlap to a significant extent. Peaks and shoulders are identified in the histogram shown in the figure below.



The CLUSTER procedure in TerrSet works with a maximum of seven bands. Experience in using the CLUSTER routine has shown that it is fast and is capable of producing excellent results. However, we have learned that the following sequence of operations is particularly useful.

Run CLUSTER using saturation percentage of 1% with the “fine” generalization level while selecting the “retain all clusters” options. Next, Display a histogram of this image. This histogram shows the frequency of pixels associated with each of the clusters that can be located in the image. Many of these clusters have very small frequencies, and thus are somewhat insignificant. The figure below presents an example of just such a histogram.



As can be seen, there are three clusters that dominate the image. Then there is a sharp break with a second group of strong clusters through to Cluster 12. Then there is a third group that follows until Cluster 25, followed by a small group of very insignificant clusters. Experience suggests then that a good generalization of the data would be to extract the first 12 clusters, with a more detailed analysis focusing on the first 25.

Once the number of clusters to be examined has been determined, run CLUSTER again, but this time choose the "fine generalization" and set the "maximum number of clusters" to the number you determined (e.g., 12).

Display the resulting image with a qualitative color palette and then try to identify each of the clusters in turn. You can use the interactive legend editing option to change the legend caption to record your interpretation. You may also wish to change the color of that category to match a logical color scheme. Also, remember that you may highlight all pixels belonging to a category by holding down the mouse button over a legend category color box.

At the end of the identification process, you may need to combine several categories. For example, the cluster analysis may have uncovered several pavement categories, such as asphalt and concrete, that you may wish to merge into a single category. The simplest way of doing this is to use Edit to create a values file containing the integer reassignments. This file has two columns. The left column should record the original cluster number, while the right column should contain the new category number to which it should be reassigned. After this file has been created, run ASSIGN to assign these new category indices to the original cluster data.

The CLUSTER procedure in the IDRISI Image Processing toolset is fast and remarkably effective in uncovering the basic landcover structure of the image. It can also be used as a preliminary stage to a hybrid unsupervised/supervised process whereby the clusters are used as the training sites to a second classification stage using the MAXLIKE classifier.¹⁴ This has the advantage of allowing the use of a larger number of raw data bands, as well as providing a stronger classification stage of pixels to their most similar cluster. In fact, it is this basic logic that underlies the ISOCLUST procedure described below.

ISOCLUST

The ISOCLUST module is an iterative self-organizing unsupervised classifier based on a concept similar to the well-known ISODATA routine of Ball and Hall (1965) and cluster routines such as the H-means and K-means procedures. The typical logic is as follows:

The user decides on the number of clusters to be uncovered. One is clearly blind in determining this. As a consequence, a common approach is to ask for a large number and then aggregate clusters after interpretation. A more efficient approach to this problem will be offered below, based on the specific implementation in TerrSet.

A set of N clusters is then arbitrarily located in band space. In some systems, these locations are randomly assigned. In most, they are systematically placed within the region of high frequency reflectances.

Pixels are then assigned to their nearest cluster location.

After all pixels have been assigned, a new mean location is computed.

Steps 3 and 4 are iteratively repeated until no significant change in output is produced.

The implementation of this general logic in TerrSet is different in several respects.

After entering the raw image bands to be used, you will be presented with a histogram of clusters that expresses the frequency with which they occur in the image. You should examine this graph and look for significant breaks in the curve. These represent major changes in the generality of the clusters. Specify the number of clusters to be created based on one of these major breaks.

¹⁴ This is possible because MAKESIG can create signatures based on training sites defined by either a vector file or an image. In this case, the image option is used.

The cluster seeding process is actually done with the CLUSTER module in the IDRISI Image Processing toolset. CLUSTER is truly a clustering algorithm (as opposed to a segmentation operation as is true of many so-called clustering routines). This leads to a far more efficient and accurate placement of clusters than either random or systematic placement.

The iterative process makes use of a full Maximum Likelihood procedure. In fact, you will notice it make iterative calls to MAKESIG and MAXLIKE. This provides a very strong cluster assignment procedure.

Because of the efficiency of the seeding step, very few iterations are required to produce a stable result. The default of 3 iterations works well in most instances.

ISODATA

The ISODATA module is a commonly used unsupervised technique that employs the so-called iterative self-organizing data analysis algorithm to partition n-dimensional imagery into a number of clusters according to a specified value. In general ISODATA begins by initializing a number of centroids using given parameters, then assigns each pixel to the cluster whose centroid is the nearest. It then updates the cluster centroids, and splits and merges clusters whenever splitting and merging criteria apply. ISODATA uses a Euclidean distance for calculating the distances between pixels and cluster centroids. The performance of ISODATA depends on the initial estimation of the partition and the parameters specified. See the classification chapter in: Richards, J.A., and X. Jia, 1999. Remote Sensing Digital Image Analysis (New York: Springer), for more detail.

MAXSET

As previously described, MAXSET is a hard classifier that assigns to each pixel the class with the greatest degree of commitment from the full Dempster-Shafer class hierarchy that describes all classes and their hierarchical combination. Although it is run as if it were a supervised classifier (it requires training site data), ultimately it behaves as if it were an unsupervised classifier in that it can assign a pixel to a class for which no exclusive training data have been supplied.

MAXSET is very similar in concept to the PIPED, MINDIST, and MAXLIKE classifiers in that it makes a hard determination of the most likely class to which each pixel belongs according to its own internal logic of operation. MAXSET is different, however, in that it recognizes that the best assignment of a pixel might be to a class that is mixed rather than unique. For example, it might determine that a pixel more likely belongs to a class of mixed conifers and deciduous forest than it does to either conifers or deciduous exclusively. The logic that it uses in doing so is derived from Dempster-Shafer theory, a special variant of Bayesian probability theory that is described more fully in the section on the BELCLASS soft classifier above. Dempster-Shafer theory provides a logic for expressing one's belief in the degree to which an item belongs to a particular set. MAXSET evaluates the degree of support for the membership of every pixel in the hierarchy of sets which includes each of the basic classes plus all possible combinations of classes. Thus, for example, in a case with basic landcover classes A, B and C, MAXSET would evaluate the degree of membership in each of the following classes:

- [A]
- [B]
- [C]
- [A,B]
- [A,C]
- [B,C]
- [A,B,C]

The importance of the supersets (sets with combinations of classes) is that they represent indistinguishable combinations of the basic classes. Thus when evidence supports the combination [A,B], it can be interpreted as support for A, or B, or A and B, but that it is unable to determine which. The reasons for this are basically twofold. Either the evidence is inconclusive, or the indistinguishable

superset really exists. Thus if MAXSET concludes that a pixel belongs to the indistinguishable superset [conifer, deciduous], it may be because the pixel truly belongs to a mixed forest class, or it may simply mean that the training sites chosen for these two classes have not yielded unambiguous signatures.

MAXSET is an excellent starting point for In-Process Classification Assessment (as described above).

Segmentation Classification

The IDRISI Image Processing toolset provides three modules for classification from image segments. Together they provide a hybrid methodology between pixel-based and segment-based classification. The module SEGMENTATION creates an image of segments. The module SEGTRAIN interactively develops training sites and signatures based on the segments from SEGMENTATION. And the module SEGCLASS is a majority rule classifier based on the majority class within a segment. The majority class within a segment is derived from a previously classified image, typically from a pixel-based classifier such as MAXLIKE or MLP. SEGCLASS can improve the accuracy of the pixel-based classification and produce a smoother map-like classification result while preserving the boundaries between segments.

Segmentation is a process by which pixels are grouped that share a homogeneous spectral similarity. The module SEGMENTATION groups adjacent pixels into image segments according to their spectral similarity. Specifically, SEGMENTATION employs a watershed delineation approach to partition input imagery based on their variance. A derived variance image is treated as a surface image allocating pixels to particular segments based on variance similarity. Across space and over all input bands, a moving window assesses this similarity and segments are defined according to a stated similarity threshold. The smaller the threshold, the more homogeneous the segments. A larger threshold will cause a more heterogeneous and generalized segmentation result.

See the Help for SEGMENTATION for complete details on the algorithm.

Hyperspectral Remote Sensing

As indicated in the introduction to this chapter, the IDRISI Image Processing toolset includes a set of routines for working with hyperspectral data. The basic procedures are similar to those used with supervised classification of multispectral data. Signatures are developed for each landcover of interest, then the entire image is classified using information from those signatures. Details of the process are given below.

Importing Hyperspectral Data

TerrSet works with hyperspectral data as a series—i.e., as a collection of independent images that are associated through either a raster image group file (.rgf) or a sensor band file (.sbf). Many hyperspectral images available today are distributed in Band-Interleaved-by-Line (BIL) format. Thus you may need to use GENERICRASTER in the import/export module group to convert these data to TerrSet format. In addition, if the data has come from a UNIX system in 16-bit format, make sure you indicate this within the GENERICRASTER dialog box.

Hyperspectral Signature Development

The signature development stage uses the module HYPERSIG, which creates and displays hyperspectral signature files. Two sources of information may be used with HYPERSIG for developing hyperspectral signatures: training sites (as described in the supervised classification section above) or library spectral curve files.

Image-based Signature Development

The IDRISI Image Processing toolset offers both a supervised and unsupervised procedure for image-based signature development. The former approach is to use a procedure similar to that outlined in earlier sections for supervised classification where training sites are delineated in the image and signatures are developed from their statistical characteristics. Because of the large number of bands involved, both the signature development and classification stages make use of different procedures from those used with multispectral data. In addition, there is a small variation to the delineation of training sites that needs to be introduced. The steps are as follows:

Create a color composite using three of the hyperspectral bands and digitize the training sites.

Rasterize the vector file of training sites by using INITIAL to create a blank image and then using RASTERVECTOR to rasterize the data.

Run HYPERSIG to create a hyperspectral signature file ¹⁵ for each landcover class. Hyperspectral signature files have an .hsg extension, and are similar in intent (but different in structure) to a multispectral signature file (.sig).

Run any of the HYPERSAM , HYPERMIN, HYPERUSP, HYPEROSP, HYPERUNMIX, or HYPERABSORB modules to classify the image.

The unsupervised procedure is somewhat experimental. HYPERAUTOSIG discovers signatures based on the concept of signature power. Users wishing to experiment with this should consult the Help System for specific details.

Library-based Signature Development

The second approach to classifying hyperspectral data relies upon the use of a library of spectral curves associated with specific earth surface materials. These spectral curves are measured with very high precision in a lab setting. These curves typically contain over a thousand readings spaced as finely as 0.01 micrometers over the visible, near and middle-infrared ranges. Clearly there is a great deal of data here. However, there are a number of important issues to consider in using these library curves.

Since the curves are developed in a lab setting, the measurements are taken without an intervening atmosphere. As a consequence, measurements exist for areas in the spectrum where remote sensing has difficulty in obtaining useable imagery. You may therefore find it necessary to remove those bands in which atmospheric attenuation is strong. A simple way to gauge which bands have significant atmospheric attenuation is to run PROFILE and examine the standard deviation of selected features over the set of hyperspectral images. Atmospheric absorption tends to cause a dramatic increase in the variability of feature signatures. To eliminate these bands, edit the sensor band file, described below, to delete their entries and adjust the number of bands at the top of the file accordingly.

Even in bands without severe attenuation, atmospheric effects can cause substantial discrepancies between the spectral curves measured in the lab and those determined from hyperspectral imagery. As a consequence, the hyperspectral modules in TerrSet assume that if you are working with spectral libraries, the imagery has already been atmospherically corrected. The SCREEN module can be used to screen out bands in which atmospheric scattering has caused significant image degradation. The ATMOSC module can then be used on the remaining bands to correct for atmospheric absorption and haze.

¹⁵ The contents of these files are also known as image spectra.

Library spectral curves are available from a number of research sites on the web, such as the United States Geological Survey (USGS) Spectroscopy Lab (<http://speclab.cr.usgs.gov>). These curve files will need to be edited to meet the specifications of TerrSet. The format used in TerrSet is virtually identical to that used by the USGS. It is an ASCII text file with an .isc extension. File structure details may be found in the file formats section of the Help System.

Spectral curve files are stored in a subdirectory of the TerrSet program directory named "waves" (e.g., c:\TerrSet\waves) and a sample of library files has been included.

Spectral curve files do not apply to any specific sensor system. Rather, they are intended as a general reference from which the expected reflectance for any sensor system can be derived in order to create a hyperspectral signature file. This is done using the HYPERSIG module.

For HYPERSIG to create a hyperspectral signature from a library spectral curve file, it will need to access a file that describes the sensor system being used. This is a *sensor band file* with an .sbf extension that is also located in the "waves" subdirectory under the TerrSet program directory. The file must be in ASCII format and file structure details may be found in the file formats section of the Help System.

For comparison, a file for the AVIRIS system, June 1992, has been supplied as a sample (aviris92.sbf) in the "waves" subdirectory of the TerrSet program directory (e.g., c:\TerrSet\waves\aviris92.sbf).

Once the hyperspectral signatures have been created for each of the landcover classes, classification of the imagery can proceed with either the HYPERSAM or HYPERMIN modules.

PROFILE

The PROFILE module offers an additional tool for exploring hyperspectral data. A profile generated over a hyperspectral series will graphically (or numerically) show how the reflectance at a location changes from one band to the next across the whole series. Thus the result is a spectral response pattern for the particular locations identified.

PROFILE requires a raster image of the sample spots to be profiled. Up to 15 profiles can be generated simultaneously,¹⁶ corresponding to sample sites with index values 1-15. A sample site can consist of one or many pixels located in either a contiguous grouping or in several disconnected groups.

The second piece of information that PROFILE will require is the name of a file that contains the names of the TerrSet image files that comprise the hyperspectral series. You have two choices here. You may use either an image group file (.rgf) or a sensor band file (.sbf), since both contain this information. In either case, you will need to have created one of these before running PROFILE. You may wish to choose the sensor band file option, since it can be used with other operations as well (such as HYPERSIG).

Hyperspectral Image Classification

TerrSet offers a range of procedures for the classification of hyperspectral imagery. All but one (HYPERABSORB) work best with signatures developed from training sites, and can be divided into hard and soft classifier types.

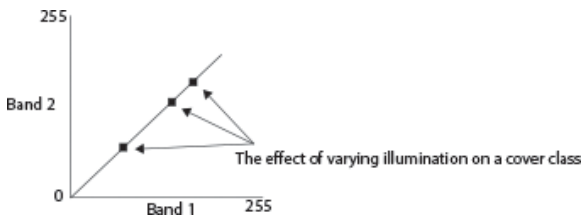
Hard Hyperspectral Classifiers

¹⁶ The display can become quite difficult to read whenever more than just a few profiles are generated at the same time. Under normal use, you may wish to limit the profiling to no more than 5 sites.

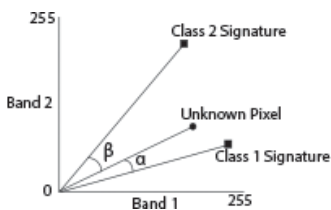
HYPERSAM

HYPERSAM is an implementation of the Spectral Angle Mapper algorithm for hyperspectral image classification (Kruse et al., 1993). The Spectral Angle Mapper algorithm is a minimum-angle procedure that is specifically designed for use with spectral curve library data (although it can also be used with image-based signatures).

The reflectance information recorded in a spectral curve file (.isc) is measured in a laboratory under constant viewing conditions. However, the data in a hyperspectral image contains additional variations that exist because of variations in illumination. For example, solar elevation varies with the time of year and topographic variations lead to variations in aspect relative to the sun. Therefore, there can be significant differences between the spectral response patterns as recorded by the sensor system and those measured in the lab. The Spectral Angle Mapper algorithm is based on the assumption that variations in illumination conditions will lead to a set of signatures that fall along a line connected to the origin of the band space as illustrated in the figure below.



Thus, in the presence of significant illumination variations, it would be anticipated that a traditional distance-based classifier would have some difficulty in identifying the feature in all cases. The Spectral Angle Mapper thus uses a minimum-angle approach. It treats each signature as a vector. Then by comparing the angle formed by an unknown pixel, the origin, and a class mean, and comparing that to all other classes, the class that will be assigned to the unknown pixel is that with the minimum angle, as illustrated in the figure below.



In the above figure, the unknown pixel would be assigned to Class 1 since the angle it subtends with the unknown pixel (α) is smaller than that with Class 2 (β).

HYPERMIN

HYPERMIN is a minimum-distance classifier for hyperspectral data that is specifically intended for use with image-based signatures developed from training sites. It uses a logic that is identical to that of the multispectral hard classifier MINDIST using standardized distances.

Soft Hyperspectral Classifiers

HYPERUNMIX

HYPERUNMIX uses the same approach as the Linear Spectral Unmixing option of UNMIX, except that it uses hyperspectral signature files.

HYPEROSP

HYPEROSP uses a procedure known as Orthogonal Subspace Projection. It is closely related to HYPERUNMIX in that it is based on the logic of Linear Spectral Unmixing. However, it attempts to improve the signal-to-noise ratio for a specific cover type by explicitly removing the contaminating effects of mixture elements. The result is an image that expresses the degree of support for the presence of the signature in question. Note that this measure is not a fraction per se, but simply a measure of support.

HYPERUSP

HYPERUSP is an unsupervised soft classifier based on the logic of HYPERAUTOSIG for the development of signatures (clusters), and HYPERUNMIX for the soft classification.

Hyperspectral Classifiers for use with Library Spectra

TerrSet offers a single classifier specifically designed for use with library spectra (HYPERABSORB). It is similar in basic operation to the TRICORDER (now renamed TETRACORDER) algorithm developed by the USGS.

HYPERABSORB

HYPERABSORB specifically looks for the presence of absorption features associated with specific materials. It follows a logic that has proven to be particularly useful in mineral mapping in arid and extraterrestrial environments. The basic principle is as follows. Particular materials cause distinctive absorption patterns in library spectra. Familiar examples include the massive absorption in the red and blue wavelengths due to the presence of chlorophyll, and the distinctive water absorption features in the middle infrared. However, many minerals exhibit very specific and narrow absorption features related to the movement of electrons in crystal lattices and vibrational effects of molecules. HYPERABSORB measures the degree of absorption evident in pixels as compared to a library spectrum through a process of continuum removal and depth analysis. The Help System gives specific details of the process. However, the basic concept is to co-register the pixel spectrum and the library spectrum by calculating the convex hull over each. This is a polyline drawn over the top of the curve such that no points in the original spectrum lie above the polyline, and in which no concave sections (valleys) occur within the polyline. Using segments of this polyline as a datum, absorption depth is measured for intermediate points. The correlation between absorption depths in the pixel spectrum and the library spectrum then gives a measure of fit, while the volume of the absorption area relative to that in the library spectrum gives a measure of abundance.

The analysis of hyperspectral images is still very much in a developmental stage. We welcome comments on experiences in using any of these procedures and suggestions for improvement.

References and Further Reading

- Ball, G.H., and D.J. Hall, 1965. *A Novel Method of Data Analysis and Pattern Classification*, Stanford Research Institute, Menlo Park, California.
- Clark, R.N., A.J. Gallagher, and G.A. Swayze, 1990, Material absorption band depth mapping of imaging spectrometer data using a complete band shape least-squares fit with library reference spectra, *Proceedings of the Second Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop*, JPL Publication 90-54, 176-186.
- Clark, R.N., G.A. Swayze, A. Gallagher, N. Gorelick, and F. Kruse, 1991, Mapping with imaging spectrometer data using the complete band shape least-squares algorithm simultaneously fit to multiple spectral features from multiple materials, *Proceedings of the Third Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop*, JPL Publication 91-28, 2-3.
- Eastman, J.R., Kyem, P.A.K., Toledano, J., and Jin, W., 1993. *GIS and Decision Making*, UNITAR, Geneva.
- Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., and Goetz, A.F.H., 1993. The Spectral Image Processing System (SIPS)—Interactive Visualization and Analysis of Imaging Spectrometer Data, *Remote Sensing of the Environment*, 44: 145-163.
- Harsanyi, J. C. and Chang, C.-I., 1994. Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach. *IEEE Transactions on Geoscience and Remote Sensing*, 32(4), pp.779-785.
- Richards, J.A., 1993. *Remote Sensing Digital Image Analysis: An Introduction*, Second Edition, Springer-Verlag, Berlin.
- Settle J.J. and N.A. Drake, 1993, Linear mixing and the estimation of ground proportions, *International Journal of Remote Sensing*, 14(6), pp. 1159-1177.
- Shimabukuro Y.E. and J.A. Smith, 1991, The least-squares mixing models to generate fraction images derived from remote sensing multispectral data, *IEEE Transactions on Geoscience and Remote Sensing*, 29(1), pp.16-20.
- Sohn, Y. and R.M. McCoy, 1997, Mapping desert shrub rangeland using spectra unmixing and modeling spectral mixtures with TM data, *Photogrammetric Engineering and Remote Sensing*, 63(6) pp.707-716.
- Wang, F., 1990. Fuzzy Supervised Classification of Remote Sensing Images, *IEEE Transactions on Geoscience and Remote Sensing*, 28(2): 194-201.

Vegetation Indices

by Amadou Thiam and J. Ronald Eastman

Analysis of vegetation and detection of changes in vegetation patterns are keys to natural resource assessment and monitoring. Thus it comes as no surprise that the detection and quantitative assessment of green vegetation is one of the major applications of remote sensing for environmental resource management and decision making.

Healthy canopies of green vegetation have a very distinctive interaction with energy in the visible and near infrared regions of the electromagnetic spectrum. In the visible regions, plant pigments (most notably chlorophyll) cause strong absorption of energy, primarily for the purpose of photosynthesis. This absorption peaks in the red and blue areas of the visible spectrum, thus leading to the characteristic green appearance of most leaves. In the near infrared, however, a very different interaction occurs. Energy in this region is not used in photosynthesis, and it is strongly scattered by the internal structure of most leaves, leading to a very high apparent reflectance in the near infrared. It is this strong contrast, then, most particularly between the amount of reflected energy in

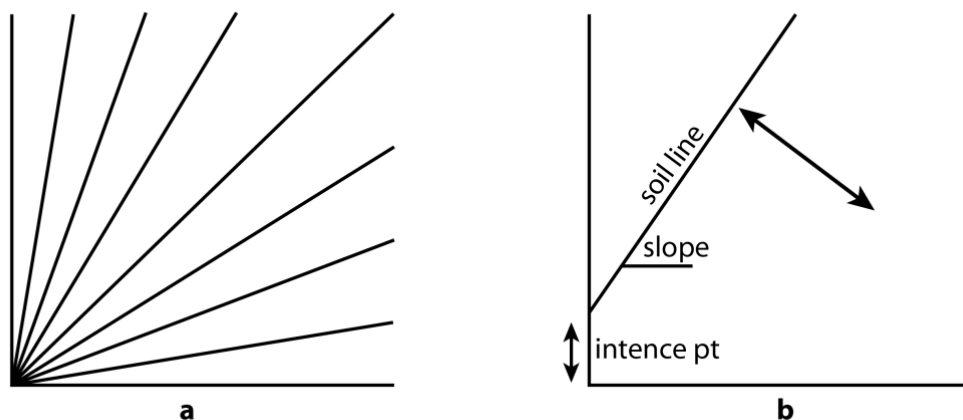
the red and near infrared regions of the electromagnetic spectrum, that has been the focus of a large variety of attempts to develop quantitative indices of vegetation condition using remotely sensed imagery.

The aim of this chapter is to present a set of vegetation index (VI) models designed to provide a quantitative assessment of green vegetation biomass. The proposed VIs are applicable to both low and high spatial resolution satellite images, such as NOAA AVHRR, Landsat TM and MSS, SPOT HRV/XS, and any others similar to these that sense in the red and near infrared regions. They have been used in a variety of contexts to assess green biomass and have also been used as a proxy to overall environmental change, especially in the context of drought (Kogan, 1990; Tripathy et al., 1996; Liu and Kogan, 1996) and land degradation risk assessment. As a consequence, special interest has been focused on the assessment of green biomass in arid environments where soil background becomes a significant component of the signal detected.

This chapter reviews the character of over 20 VIs that are provided by the TASSCAP and VEGINDEX modules in the TerrSet System software. They are provided to facilitate the use of these procedures and to further the debate concerning this very important environmental index. We welcome both your comments on the VIs currently included in TerrSet as well as your suggestions for future additions to the set.

Classification of Vegetation Indices

Jackson and Huete (1991) classify VIs into two groups: *slope-based* and *distance-based* VIs. To appreciate this distinction, it is necessary to consider the position of vegetation pixels in a two-dimensional graph (or *bi-spectral plot*) of red versus infrared reflectance. The slope-based VIs are simple arithmetic combinations that focus on the contrast between the spectral response patterns of vegetation in the red and near infrared portions of the electromagnetic spectrum. They are so named because any particular value of the index can be produced by a set of red/infrared reflectance values that form a line emanating from the origin of a bi-spectral plot. Thus different levels of the index can be envisioned as producing a spectrum of such lines that differ in their slope. In figure a, for example, shows a spectrum of Normalized Difference Vegetation Index (the most commonly used of this group) lines ranging from -0.75 fanning clockwise to +0.75 (assuming infrared as the X axis and red as the Y axis), with NDVI values of 0 forming the diagonal line.



In contrast to the slope-based group, the distance-based group measures the degree of vegetation present by gauging the difference of any pixel's reflectance from the reflectance of bare soil. A key concept here is that a plot of the positions of bare soil pixels of varying moisture levels in a bi-spectral plot will tend to form a line (known as a *soil line*). As vegetation canopy cover increases, this soil background will become progressively obscured, with vegetated pixels showing a tendency towards increasing perpendicular

distance from this soil line (figure b). All the members of this group (such as the Perpendicular Vegetation Index—PVI) thus require that the slope and intercept of the soil line be defined for the image being analyzed.

To these two groups of vegetation indices, a third group can be added called *orthogonal transformation* VIs. Orthogonal indices undertake a transformation of the available spectral bands to form a new set of uncorrelated bands within which a green vegetation index band can be defined. The Tasseled Cap transformation is perhaps the most well-known of this group.

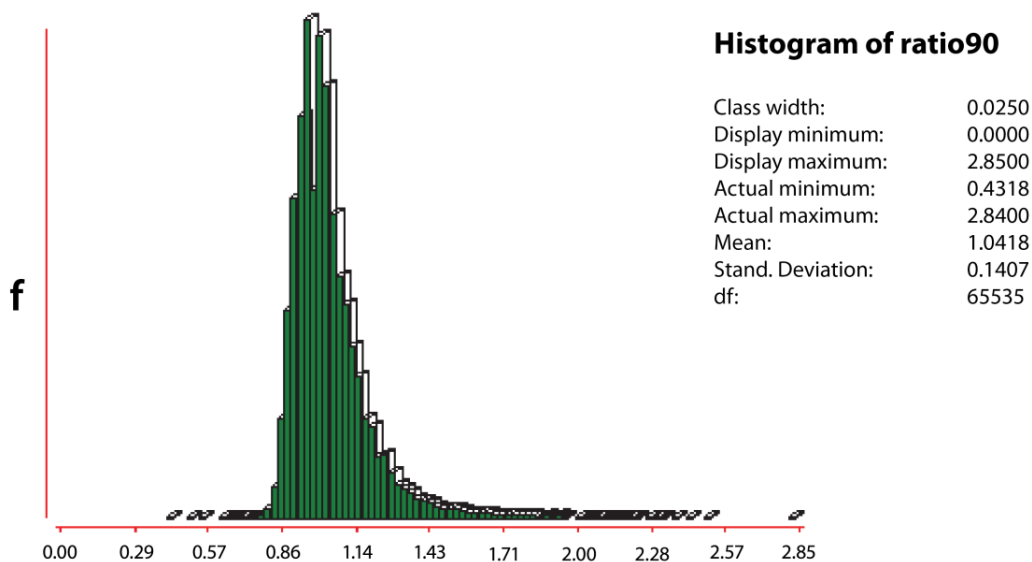
The Slope-Based VIs

Slope-based VIs are combinations of the visible red and the near infrared bands and are widely used to generate vegetation indices. The values indicate both the status and abundance of green vegetation cover and biomass. The slope-based VIs include the RATIO, NDVI, RVI, EVI, TVI, CTVI, and TTVI. The module VEGINDEX in TerrSet may be used to generate an image for each of these VIs.

The Ratio Vegetation Index (RATIO) was proposed by Rouse, et al. (1974) to separate green vegetation from soil background using Landsat MSS imagery. The RATIO VI is produced by simply dividing the reflectance values contained in the near infrared band by those contained in the red band, i.e.:

$$\text{RATIO} = \frac{\text{NIR}}{\text{RED}}$$

The result clearly captures the contrast between the red and infrared bands for vegetated pixels, with high index values being produced by combinations of low red (because of absorption by chlorophyll) and high infrared (as a result of leaf structure) reflectance. In addition, because the index is constructed as a ratio, problems of variable illumination as a result of topography are minimized. However, the index is susceptible to division by zero errors and the resulting measurement scale is not linear. As a result, RATIO VI images do not have normal distributions (figure below), making it difficult to apply some statistical procedures.



The Normalized Difference Vegetation Index (NDVI) was also introduced by Rouse et al. (1974) in order to produce a spectral VI that separates green vegetation from its background soil brightness using Landsat MSS digital data. It is expressed as the difference between the near infrared and red bands normalized by the sum of those bands, i.e.:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

This is the most commonly used VI as it retains the ability to minimize topographic effects while producing a linear measurement scale. In addition, division by zero errors are significantly reduced. Furthermore, the measurement scale has the desirable property of ranging from -1 to 1 with 0 representing the approximate value of no vegetation. Thus negative values represent non-vegetated surfaces.

The Transformed Vegetation Index (TVI) (Deering et al., 1975) modifies the NDVI by adding a constant of 0.50 to all its values and taking the square root of the results. The constant 0.50 is introduced in order to avoid operating with negative NDVI values. The calculation of the square root is intended to correct NDVI values that approximate a Poisson distribution and introduce a normal distribution. With these two elements, the TVI takes the form:

$$TVI = \sqrt{\left(\frac{NIR - RED}{NIR + RED}\right)} + 0.5$$

However, the use of TVI requires that the minimum input NDVI values be greater than -0.5 to avoid aborting the operation. Negative values still will remain if values less than -0.5 are found in the NDVI. Moreover, there is no technical difference between NDVI and TVI in terms of image output or active vegetation detection.

The Corrected Transformed Vegetation Index (CTVI) proposed by Perry and Lautenschlager (1984) aims at correcting the TVI. Clearly adding a constant of 0.50 to all NDVI values does not always eliminate all negative values as NDVI values may have the range -1 to +1. Values that are lower than -0.50 will leave small negative values after the addition operation. Thus, the CTVI is intended to resolve this situation by dividing (NDVI + 0.50) by its absolute value ABS(NDVI + 0.50) and multiplying the result by the square root of the absolute value (SQRT[ABS(NDVI + 0.50)]). This suppresses the negative NDVI. The equation is written:

$$CTVI = \frac{NDVI + 0.5}{ABS(NDVI + 0.5)} \times \sqrt{ABS(NDVI + 0.5)}$$

Given that the correction is applied in a uniform manner, the output image using CTVI should have no difference with the initial NDVI image or the TVI whenever TVI properly carries out the square root operation. The correction is intended to eliminate negative values and generate a VI image that is similar to, if not better than, the NDVI. However, Thiam (1997) indicates that the resulting image of the CTVI can be very "noisy" due to an overestimation of the greenness. He suggests ignoring the first term of the CTVI equation in order to obtain better results. This is done by simply taking the square root of the absolute values of the NDVI in the original TVI expression to have a new VI called **Thiam's Transformed Vegetation Index (TTVI)**.

$$TTVI = \sqrt{ABS(NDVI + 0.5)}$$

The simple **Ratio Vegetation Index (RVI)** was suggested by Richardson and Wiegand (1977) as graphically having the same strengths and weaknesses as the TVI (see above) while being computationally simpler. RVI is clearly the reverse of the standard simple ratio (RATIO) as shown by its expression:

$$RVI = \frac{RED}{NIR}$$

The Enhanced Vegetation Index (EVI) is optimized for non-saturation over high biomass regions (a common issue with the NDVI) as well as improved sensitivity to canopy structural variations, including [leaf area index \(LAI\)](#), canopy type, plant physiognomy, and canopy architecture. This sensitivity to horizontal and vertical noise is considered to complement the NDVI's sensitivity to chlorophyll levels. Adopted by NASA as a standard product in 2000, the EVI applies a canopy background adjustment factor (L) to account for nonlinear, differential NIR and red radiant transfer through a canopy, aerosol resistance coefficients (C1 and C2) along with the visible blue band to account for aerosol influences in the red band, and a gain factor (G) (see Huete et. al., 2002). The equation for the EVI is as follows:

$$EVI = G \times \frac{(NIR-RED)}{(NIR+C1 \times RED-C2 \times BLUE+L)}$$

VEGINDEX's default coefficients for the EVI model are based on NASA recommended values for use with MODIS imagery:

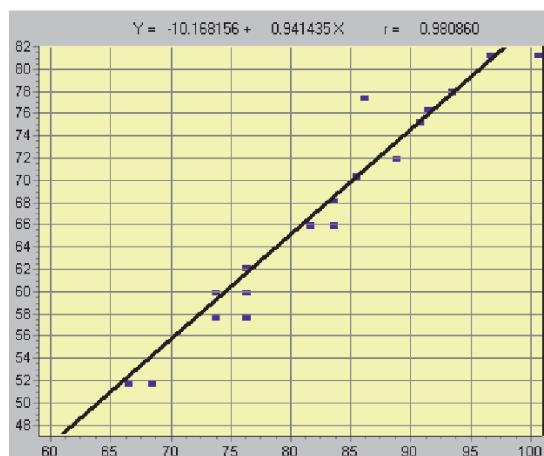
$L = 1$, $C1 = 6$, $C2 = 7.5$, $G = 2.5$

The Distance-Based VIs

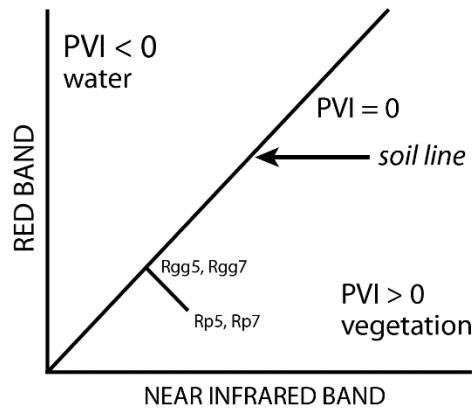
This group of vegetation indices is derived from the Perpendicular Vegetation Index (PVI) discussed in detail below. The main objective of these VIs is to cancel the effect of soil brightness in cases where vegetation is sparse and pixels contain a mixture of green vegetation and soil background. This is particularly important in arid and semi-arid environments.

The procedure is based on the soil line concept as outlined earlier. The soil line represents a description of the typical signatures of soils in a red/near infrared bi-spectral plot. It is obtained through linear regression of the near infrared band against the red band for a sample of bare soil pixels. Pixels falling near the soil line are assumed to be soils while those far away are assumed to be vegetation. Distance-based VIs using the soil line require the slope (b) and intercept (a) of the line as inputs to the calculation. Unfortunately, there has been a remarkable inconsistency in the logic with which this soil line has been developed for specific VIs. One group requires the red band as the independent variable and the other requires the near infrared band as the independent variable for the regression. The Help System for VEGINDEX should be consulted for each VI in the Distance-based group to indicate which of these two approaches should be used.

The figure below shows the soil line and its parameters as calculated for a set of soil pixels using the REGRESS module in TerrSet. The procedure requires that you identify a set of bare soil pixels as a Boolean mask (1=soil / 0=other). REGRESS is then used to regress the red band against the near infrared band (or vice versa, depending upon the index), using this mask to define the pixels from which the slope and intercept should be defined. A worked example of this procedure can be found in the **Tutorial** in the Vegetation Analysis in Arid Environments exercise.



The Perpendicular Vegetation Index (PVI) suggested by Richardson and Wiegand (1977) is the parent index from which this entire group is derived. The PVI uses the perpendicular distance from each pixel coordinate (e.g., Rp5, Rp7) to the soil line as shown in the figure below.



The Perpendicular Vegetation Index (from Richardson and Wiegand, 1977)

To derive this perpendicular distance, four steps are required:

- 1) Determine the equation of the soil line by regressing bare soil reflectance values for red (dependent variable) versus infrared (independent variable).¹⁷ This equation will be in the form:

$$Rg5 = a_0 + a_1 Rg7$$

where $Rg5$ is a Y position on the soil line

$Rg7$ is the corresponding X coordinate

a_1 is the slope of the soil line

a_0 is the Y – intercept of the soil line

- 2) Determine the equation of the line that is perpendicular to the soil line. This equation will have the form:

$$Rp5 = b_0 + b_1 Rp7$$

where $b_0 = Rp5 - b_1 Rp7$

where $Rp5$ = red reflectance

$Rp7$ = infrared reflectance

and

$$b_1 = -1/a_1$$

where

a_1 = the slope of the soil line

¹⁷Check the Help System for each VI to determine which band should be used as the dependent and independent variables. In this example, for the PVI, red is dependent and infrared is independent.

3) Find the intersection of these two lines (i.e., the coordinate Rgg5,Rgg7).

$$Rgg5 = \frac{b_1 a_0 - b_0 a_1}{b_1 - a_1}$$

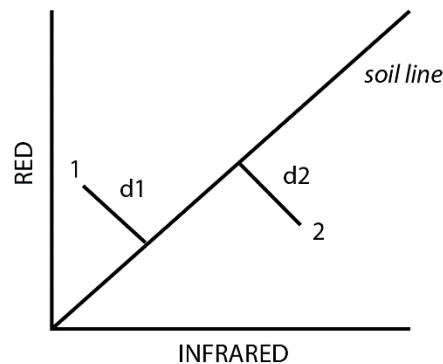
$$Rgg7 = \frac{a_0 - b_0}{b_1 - a_1}$$

4) Find the distance between the intersection (Rgg5,Rgg7) and the pixel coordinate (Rp5,Rp7) using the Pythagorean Theorem.

$$PVI = \sqrt{(Rgg5 - Rp5)^2 + (Rgg7 - Rp7)^2}$$

Attempts to improve the performance of the PVI have yielded three others suggested by Perry and Lautenschlager (1984), Walther and Shabaani (1991), and Qi, et al. (1994). In order to avoid confusion, the derived PVIs are indexed 1 to 3 (PVI₁, PVI₂, PVI₃).

PVI₁ was developed by Perry and Lautenschlager (1984) who argued that the original PVI equation is computationally intensive and does not discriminate between pixels that fall to the right or left side of the soil line (i.e., water from vegetation). Given the spectral response pattern of vegetation in which the infrared reflectance is higher than the red reflectance, all vegetation pixels will fall to the right of the soil line (e.g., pixel 2 in the figure below). In some cases, a pixel representing non-vegetation (e.g., water) may be equally far from the soil line, but lies to the left of that line (e.g., pixel 1 in the figure below). In the case of PVI, that water pixel will be assigned a high vegetation index value. PVI₁ assigns negative values to those pixels lying to the left of the soil line.



Distance from the Soil Line

The equation is written:

$$PVI_1 = \frac{(bNIR - RED + a)}{\sqrt{b^2 + 1}}$$

where

NIR = reflectance in the near infrared band

RED = reflectance in the visible red band

a = intercept of the soil line

b = slope of the soil line

PVI₂ (Walther and Shabaani, 1991; Bannari, et al., 1996) weights the red band with the intercept of the soil line and is written¹⁸:

$$PVI_2 = \frac{NIR - a \times RED + b}{\sqrt{1 + a^2}}$$

where

NIR = reflectance in the near infrared band

RED = reflectance in the visible red band

a = slope of the soil line

b = intercept of the soil line

PVI_3 , presented by Qi, et al (1994), is written:

$$PVI_3 = apNIR - bpRED$$

where

$pNIR$ = reflectance in the near infrared band

$pRED$ = reflectance in the visible red band

a = intercept of the soil line

b = slope of the soil line

Difference Vegetation Index (DVI) is also suggested by Richardson and Wiegand (1977) as an easier vegetation index calculation algorithm. The particularity of the DVI is that it weights the near infrared band by the slope of the soil line. It is written:

$$DVI = g MSS7 - MSS5$$

where

g = the slope of the soil line

$MSS7$ = reflectance in the near infrared 2 band

$MSS5$ = reflectance in the visible red band

Similar to the PVI_1 , with the DVI, a value of zero indicates bare soil, values less than zero indicate water, and those greater than zero indicate vegetation.

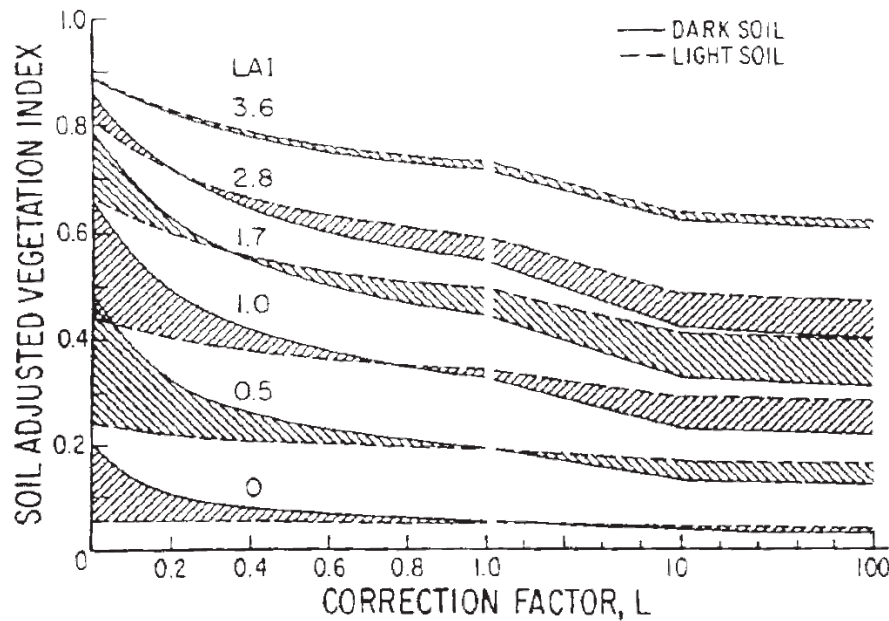
The Ashburn Vegetation Index (AVI) (Ashburn, 1978) is presented as a measure of green growing vegetation. The values in $MSS7$ are multiplied by 2 in order to scale the 6-bit data values of this channel to match with the 8-bit values of $MSS5$. The equation is written:

$$AVI = 2.0MSS7 - MSS5$$

¹⁸ In Bannari, et al. (1996), **a** is used to designate the slope and **b** is used to designate the intercept. More commonly in linear regression, **a** is the intercept and **b** the slope of the fitted line. This has been corrected here.

This scaling factor would not apply wherever both bands are 7-bit or 8-bit and the equation is rewritten as a simple subtraction.

The Soil-Adjusted Vegetation Index (SAVI) is proposed by Huete (1988). It is intended to minimize the effects of soil background on the vegetation signal by incorporating a constant soil adjustment factor L into the denominator of the NDVI equation. L varies with the reflectance characteristics of the soil (e.g., color and brightness). Huete (1988) provides a graph from which the values of L can be extracted (figure below). The L factor chosen depends on the density of the vegetation one wishes to analyze. For very low vegetation, Huete et al., (1988) suggest using an L factor of 1.0, for intermediate 0.5 and for high densities 0.25. Walther and Shabaani (1991) suggest that the best L value to select is where the difference between SAVI values for dark and light soil is minimal. For $L = 0$, SAVI equals NDVI. For $L = 100$, SAVI approximates PVI.



Influence of light and dark soil on the SAVI values of cotton as a function of the shifted correct factor L (from Huete, 1988).

The equation is written:

$$SAVI = \frac{\rho_{nir} - \rho_{red}}{(\rho_{nir} + \rho_{red} + L)} \times (1 + L)$$

where

ρ_{nir} = near infrared band (expressed as reflectances)

ρ_{red} = visible red band (expressed as reflectances)

L = soil adjustment factor

The Transformed Soil-Adjusted Vegetation Index (TSAVI₁) was defined by Baret, et al. (1989) who argued that the SAVI concept is exact **only** if the constants of the soil line are $a=1$ and $b=0$ (note the reversal of these common symbols). Because this is not generally

the case, they transformed SAVI. By taking into consideration the PVI concept, they proposed a first modification of TSAVI designated as TSAVI₁. The transformed expression is written:

$$\text{TSAVI} = \frac{a(\text{NIR} - a\text{RED} - b)}{(\text{RED} + a\text{NIR} - ab)}$$

where

NIR = reflectance in the near infrared band (expressed as reflectances)

RED = reflectance in the visible red band (expressed as reflectances)

a = slope of the soil line

b = intercept of the soil line

With some resistance to high soil moisture, TSAVI₁ could be a very good candidate for use in semi-arid regions. TSAVI₁ was specifically designed for semi-arid regions and does not work well in areas with heavy vegetation.

TSAVI was readjusted a second time by Baret, et al (1991) with an additive correction factor of 0.08 to minimize the effects of the background soil brightness. The new version is named **TSAVI₂** and is given by:

$$\text{TSAVI}_2 = \frac{a(\text{NIR} - a\text{RED} - b)}{\text{RED} + a\text{NIR} - ab + 0.08(1 + a^2)}$$

The Modified Soil-Adjusted Vegetation Indices (MSAVI₁ and MSAVI₂) suggested by Qi, et al. (1994) are based on a modification of the L factor of the SAVI. Both are intended to better correct the soil background brightness in different vegetation cover conditions.

With MSAVI₁, L is selected as an empirical function due to the fact that L decreases with decreasing vegetation cover as is the case in semi-arid lands (Qi, et al., 1994). In order to cancel or minimize the effect of the soil brightness, L is set to be the product of NDVI and WDV (described below). Therefore, it uses the opposite trends of NDVI and WDV. The full expression of MSAVI₁ is written:

$$\text{MSAVI}_1 = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED} + L} \times (1 + L)$$

where

NIR = reflectance in the near infrared band (expressed as reflectances)

RED = reflectance in the visible red band (expressed as reflectances)

$$L = 1 - 2_\gamma \text{NDVI} \times \text{WDVI}$$

where

NDVI = Normalized Difference Vegetation Index

WDVI = Weighted Difference Vegetation Index

γ = slope of the background soil line

2 = used to increase the L dynamic range

range of $L = 0$ to 1

The second modified SAVI, $MSAVI_2$, uses an inductive L factor to:

1. remove the soil "noise" that was not canceled out by the product of NDVI by WDV, and
2. correct values greater than 1 that $MSAVI_1$ may have due to the low negative value of $NDVI \cdot WDV$. Thus, its use is limited for high vegetation density areas.

The general expression of $MSAVI_2$ is:

$$MSAVI_2 = \frac{2\rho_{nir} + 1 - \sqrt{(2\rho_{nir} + 1)^2 - 8(\rho_{nir} - \rho_{red})}}{2}$$

where

ρ_{nir} = reflectance of the near infrared band (expressed as reflectances)

ρ_{red} = reflectance of the red band (expressed as reflectances)

The Weighted Difference Vegetation Index (WDVI) has been attributed to Richardson and Wiegand (1977), and Clevers (1978) by Kerr and Pichon (1996), writing the expression as:

$$WDVI = \rho_n - \gamma \rho_r$$

where

ρ_n = reflectance of near infrared band

ρ_r = reflectance of visible red band

γ = slope of the soil line

Although simple, WDV is as efficient as most of the slope-based VIs. The effect of weighting the red band with the slope of the soil line is the maximization of the vegetation signal in the near infrared band and the minimization of the effect of soil brightness.

The Orthogonal Transformations

The derivation of vegetation indices has also been approached through orthogonal transformation techniques such as the PCA, the GVI of the Kauth-Thomas Tasseled Cap Transformation and the MGVI of the Wheeler-Misra orthogonal transformation. The link between these three techniques is that they all express green vegetation through the development of their second component.

Principal Components Analysis (PCA) is an orthogonal transformation of n -dimensional image data that produces a new set of images (components) that are uncorrelated with one another and ordered with respect to the amount of variation (information) they represent from the original image set. PCA is typically used to uncover the underlying dimensionality of multi-variate data by removing redundancy (evident in inter-correlation of image pixel values), with specific applications in GIS and image processing ranging from data compression to time series analysis. In the context of remotely sensed images, the first component typically represents albedo (in which the soil background is represented) while the second component most often represents variation in vegetative cover. For example, component 2 generally has positive loadings on the near infrared bands and negative loadings on the visible bands. As a result, the green vegetation pattern is highlighted in this component (Singh and Harrison, 1985; Fung and LeDrew, 1987; Thiam, 1997). This is illustrated in Table 1 corresponding to the factor loadings of a 1990 MSS image of southern Mauritania.

Table 1 Factor loadings of the 1990 PCA

	Comp1	Comp2	Comp3	Comp4
MSS1	0.86	-0.46	-0.22	0.01
MSS2	0.91	-0.36	0.19	-0.08
MSS3	0.95	0.25	0.09	0.14
MSS4	0.75	0.65	-0.08	-0.09

The Green Vegetation Index (GVI) of the Tasseled Cap is the second of the four new bands that Kauth and Thomas (1976) extracted from raw MSS images. The GVI provides global coefficients that are used to weight the original MSS digital counts to generate the new transformed bands. The TASSCAP module in TerrSet is specifically provided to calculate the Tasseled Cap bands from Landsat MSS or TM images. The output from TASSCAP corresponding to GVI is xxgreen (xx = the two character prefix entered by the user) by default. The expression of the green vegetation index band, GVI, is written as follows for MSS or TM data:

$$GVI = [(-0.386MSS4) + (-0.562MSS5) + (0.600MSS6) + (0.491MSS7)]$$

$$GVI = [(-0.2848TM1) + (-0.2435TM2) + (-0.5436TM3) + (0.7243TM4) + (0.0840TM5) + (-0.1800TM7)]$$

The negative weights of the GVI on the visible bands tend to minimize the effects of the background soil, while its positive weights on the near infrared bands emphasize the green vegetation signal.

Misra's Green Vegetation Index (MGVI) is the equivalent of the Tasseled Cap GVI and is proposed by Wheeler et al. (1976) and Misra, et al. (1977) as a spectral vegetation index. It is the second of the four new bands produced from an application of the Principal Components Analysis to MSS digital counts. The algebraic expression of the MGVI is:

$$MGVI = -0.386MSS4 - 0.530MSS5 + 0.535MSS6 + 0.532MSS7$$

The principle of the MGVI is to weight the original digital counts by some global coefficients provided by Wheeler and Misra in order to generate a second Principal Component. However, the use of these global coefficients may not yield the same result as a directly calculated second Principal Component, as they may be site specific. The coefficients correspond to the eigenvectors that are produced with a Principal Components Analysis. The eigenvectors indicate the direction of the principal axes (Mather, 1987). They are combined with the original spectral values to regenerate Principal Components. For example PCA1 is produced by combining the original reflectances with the eigenvectors (column values) associated with component 1. Likewise, component 2 (MGVI) is produced by combining the original digital counts with the eigenvectors associated with component 2 as highlighted in Table 2.

Table 2 Eigenvectors of the 1990 PCA

	Comp1	Comp2	Comp3	Comp4
eigvec.1	0.49	-0.51	-0.70	0.08
eigvec.2	0.52	-0.40	0.60	-0.45
eigvec.3	0.55	0.28	0.27	0.74
eigvec.4	0.43	0.71	-0.27	-0.49

The PCA module in TerrSet generates eigenvectors as well as factor loadings with the component images. A site-specific MGVI image can then be produced with Image Calculator by using the appropriate eigenvector values. The following equation would be used to produce the MGVI image for the example shown in Table 18-2:

$$MGVI90 = (-0.507MSS4) + (-0.400MSS5) + (0.275MSS6) + (0.712MSS7)$$

Summary

The use of any of these transformations depends on the objective of the investigation and the general geographic characteristics of the application area. In theory, any of them can be applied to any geographic area, regardless of their sensitivity to various environmental components that might limit their effectiveness. In this respect, one might consider applying the slope-based indices as they are simple to use and yield numerical results that are easy to interpret. However, including the well-known NDVI, they all have the major weakness of not being able to minimize the effects of the soil background. This means that a certain proportion of their values, negative or positive, represents the background soil brightness. The effect of the background soil is a major limiting factor to certain statistical analyses geared towards the quantitative assessment of above-ground green biomass.

Although they produce indices whose extremes may be much lower and greater than those of the more familiar NDVI, the distance-based VIs have the advantage of minimizing the effects of the background soil brightness. This minimization is performed by combining the input bands with the slope and intercept of the soil line obtained through a linear regression between bare soil sample reflectance values extracted from the red and near infrared bands. This represents an important quantitative and qualitative improvement of the significance of the indices for all types of applications, particularly for those dealing with arid and semi-arid environments. To take advantage of these, however, you do need to be able to identify bare soil pixels in the image.

The orthogonal VIs, namely the Tasseled Cap, Principal Components Analysis and the Wheeler-Misra transformation (MGVI), proceed by a decorrelation of the original bands through orthogonalization in order to extract new bands. By this process, they produce a

green band that is somehow free of soil background effects, since almost all soil characteristics are ascribed to another new band called brightness.

Despite the large number of vegetation indices currently in use, it is clear that much needs to be learned about the application of these procedures in different environments. It is in this spirit that the VEGINDEX and TASSCAP modules have been created. However, it has also become clear that remote sensing offers a significant opportunity for studying and monitoring vegetation and vegetation dynamics.

References

- Ashburn, P., 1978. The vegetative index number and crop identification, *The LACIE Symposium Proceedings of the Technical Session*, 843-850.
- Bannari, A., Huete, A. R., Morin, D., and Zagolski, 1996. Effets de la Couleur et de la Brillance du Sol Sur les Indices de Végétation, *International Journal of Remote Sensing*, 17(10): 1885-1906.
- Baret, F., Guyot, G., and Major, D., 1989. TSAVI: A Vegetation Index Which Minimizes Soil Brightness Effects on LAI and APAR Estimation, *12th Canadian Symposium on Remote Sensing and IGARSS'90*, Vancouver, Canada, 4.
- Baret, F., and Guyot, G., 1991. Potentials and Limits of Vegetation Indices for LAI and APAR Assessment, *Remote Sensing and the Environment*, 35: 161-173.
- Deering, D. W., Rouse, J. W., Haas, R. H., and Schell, J. A., 1975. Measuring "Forage Production" of Grazing Units From Landsat MSS Data, *Proceedings of the 10th International Symposium on Remote Sensing of Environment, II*, 1169-1178.
- Fung, T., and LeDrew, E., 1988. The Determination of Optimal Threshold Levels for Change Detection Using Various Accuracy Indices, *Photogrammetric Engineering and Remote Sensing*, 54(10): 1449-1454.
- Huete, A. R., 1988. A Soil-Adjusted Vegetation Index (SAVI), *Remote Sensing and the Environment*, 25: 53-70.
- Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83: 195-213.
- Jackson, R. D., 1983. Spectral Indices in *n*-Space, *Remote Sensing and the Environment*, 13: 409-421.
- Kauth, R. J., and Thomas, G. S., 1976. The Tasseled Cap - A Graphic Description of the Spectral Temporal Development of Agricultural Crops As Seen By Landsat. *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, Purdue University, West Lafayette, Indiana, 41-51.
- Kogan, F. N., 1990. Remote Sensing of Weather Impacts on Vegetation in Nonhomogeneous Areas, *International Journal of Remote Sensing*, 11(8): 1405-1419.
- Liu, W. T., and Kogan, F. N., 1996. Monitoring Regional Drought Using the Vegetation Condition Index, *International Journal of Remote Sensing*, 17(14): 2761-2782.
- Misra, P. N., and Wheeler, S.G., 1977. Landsat Data From Agricultural Sites - Crop Signature Analysis, *Proceedings of the 11th International Symposium on Remote Sensing of the Environment*, ERIM.
- Misra, P. N., Wheeler, S. G., and Oliver, R. E., 1977. Kauth-Thomas Brightness and Greenness Axes, IBM personal communication, Contract NAS-9-14350, RES 23-46.

- Perry, C. Jr., and Lautenschlager, L. F., 1984. Functional Equivalence of Spectral Vegetation Indices, *Remote Sensing and the Environment* 14: 169-182.
- Qi, J., Chehbouni A., Huete, A. R., Kerr, Y. H., and Sorooshian, S., 1994. A Modified Soil Adjusted Vegetation Index. *Remote Sensing and the Environment*, 48: 119-126.
- Richardson, A. J., and Wiegand, C. L., 1977. Distinguishing Vegetation From Soil Background Information, *Photogrammetric Engineering and Remote Sensing*, 43(12): 1541-1552.
- Rouse, J. W. Jr., Haas, R., H., Schell, J. A., and Deering, D.W., 1973. Monitoring Vegetation Systems in the Great Plains with ERTS, *Earth Resources Technology Satellite-1 Symposium*, Goddard Space Flight Center, Washington D.C., 309-317.
- Rouse, J. W. Jr., Haas, R., H., Deering, D. W., Schell, J. A., and Harlan, J. C., 1974. Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. *NASA/GSFC Type III Final Report*, Greenbelt, MD., 371.
- Singh, A., and Harrison, A., 1985. Standardized Principal Components, *International Journal of Remote Sensing*, 6(6): 883-896.
- Thiam, A.K. 1997. Geographic Information Systems and Remote Sensing Methods for Assessing and Monitoring Land Degradation in the Sahel: The Case of Southern Mauritania. Doctoral Dissertation, Clark University, Worcester Massachusetts.
- Tripathy, G. K., Ghosh, T. K., and Shah, S. D., 1996. Monitoring of Desertification Process in Karnataka State of India Using Multi-Temporal Remote Sensing and Ancillary Information Using GIS, *International Journal of Remote Sensing*, 17(12): 2243-2257.

RADAR Imaging and Analysis

The term RADAR is an acronym for Radio Detection And Ranging. RADAR imaging uses radio waves to detect surface characteristics. There are several RADAR Sensing instruments that provide a range of data products to the user community. These include: ERS (European Remote Sensing Satellite), JERS (Japanese Earth Resource Satellite) and the Canadian RADARSAT. RADARSAT is a commercial system (unlike the other systems which are largely scientific missions) intended to provide data for a wide range of uses.

Optical multispectral remote sensing systems like Landsat and SPOT gather data as reflected electromagnetic energy. The source of the energy for these sensors is the sun. Unlike these passive optical systems, RADAR systems such as RADARSAT are active sensing systems, meaning that they provide the energy that is then measured as it returns to the sensor. RADAR systems use energy transmitted at microwave frequencies that are not detectable by the human eye. Most RADAR systems operate at one single frequency. RADARSAT, for example, operates at what is known as C-band frequency¹⁹ (5.3 Ghz frequency or 5.6 cm wavelength) and thus acquires single band data. The sensing instrument used on RADARSAT is known as Synthetic Aperture Radar (SAR). This instrument uses the motion of the satellite and doppler frequency shift to electronically synthesize the large antennae required for the acquisition of high resolution RADAR imagery. The sensor sends a microwave energy pulse to the earth and measures the time it takes for that pulse to return after interaction with the earth's surface.

The Nature of RADAR Data: Advantages and Disadvantages

Since RADARSAT-SAR uses microwave energy, it is able to penetrate atmospheric barriers that often hinder optical imaging. SAR can "see" through clouds, rain, haze and dust and can operate in darkness, making data capture possible in any atmospheric conditions.

¹⁹ Other RADAR frequencies include X - band 12.5-8.0 Ghz (2.4 - 3.75 cm), L - band : 2.0-1.0 Ghz (15-30 cm) and P-band 1.0-0.3 Ghz (30-100 cm). The frequency of the C-band is 8.0-4.0 Ghz (3.75-7.5 cm).

Because microwave energy can penetrate the land surface to appreciable depths, it is useful in arid environments. Additionally, microwave energy is most appropriate for studying subsurface conditions and properties which are relevant to resource prospecting and archaeological explorations. The data gathered by this sensor is of the form known as HH (horizontal transmit, horizontal receive)²⁰ polarized data. The signal sent to the earth surface has a horizontal orientation (or polarization) and is captured using the same polarization. Variations in the return beam, termed *backscatter*, result from varying surface roughness, topography and surface moisture conditions. These characteristics of the RADAR signal can be exploited to infer the structural and textural characteristics of the target objects, unlike the simple reflection signal of optical sensing. RADARSAT-SAR collects data in a variety of beam modes allowing a choice of incidence angles and resolutions (from 20m up to 100m). Since the instrument can be navigated to view the same location from different beam positions, it can acquire data in stereo pairs that can be used for a large number of terrain and topographic analysis needs.

There are, however, some problems that are unique to RADAR imaging. Most, if not all, RADAR images have a speckled or grainy appearance. This is the result of a combination of multiple scattering within a given pixel. In RADAR terms, a large number of surface materials exhibit *diffuse* reflectance patterns. Farmlands, wet soils and forest canopies, for example, show high signal returns, while surfaces like roads, pavements and smooth water surfaces show *specular* reflectance patterns with low signal returns. A mixture of different cover types generates a complex image surface. The data are thus inherently “noisy” and require substantial preprocessing before use in a given analysis task. In mountainous terrain, shadowing or relief displacement occurs because the reflected RADAR pulse from the top of a mountain ridge reaches the antenna before the pulse from the base or lee side of the feature does. This is known as the *layover effect* and is significant in areas of steep slopes. If the slopes facing the antenna are less steep, the pulse reaches the base before the top. This leads to slopes appearing compressed on the imagery—an effect known as *foreshortening*. In urban areas, double reflectance from *corner reflectors* like buildings causes imagery to have a bright, sparkled appearance. Mixed with reflectance from trees and road surfaces, the urban scene makes filtering operations rather challenging. If a bright target dominates a given area, it may skew the distribution of the data values. It must be noted, however, that the RADAR's advantages, including its ability for all-weather imaging, far outweigh its disadvantages.

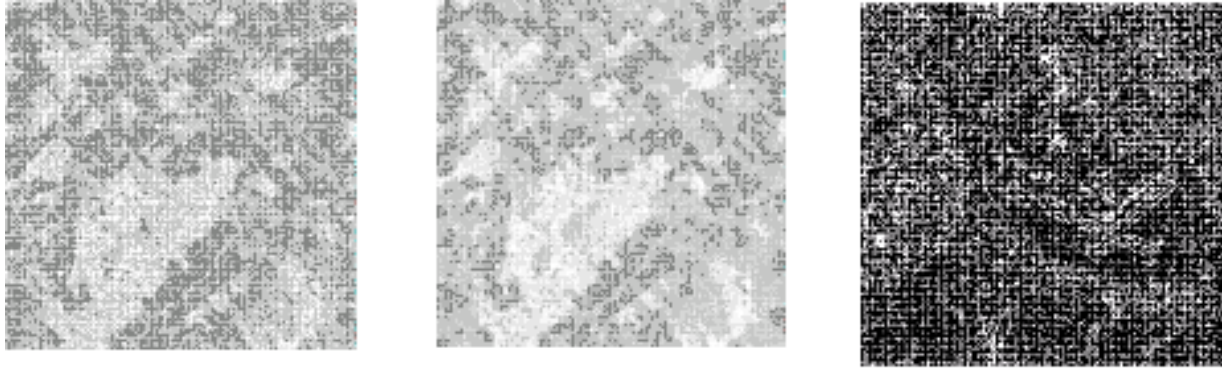
Using RADAR Data in TerrSet

The module named RADARSAT can be used to import RADARSAT data into TerrSet. Furthermore, the generic data import tools of TerrSet can be used to import most other formats. Since each scene is a single band, it can be best viewed using a grey scale palette. The RADAR palette in TerrSet can be used to view RADAR imagery including those scenes that have data distributions that exemplify negatively skewed histograms. An alternative method is to generate a histogram, estimate the range within which a majority of the data values fall, and STRETCH the image to that level. This works very well for visual analysis of the data. (See the Image Exploration exercise in the **Tutorial** manual.)

Currently, the use of RADAR data in environmental analysis and resource management is still in an early stage and is quite limited compared to the widespread use of more traditional image data (e.g., MSS, TM, aerial photography). More sophisticated tools for interpretation of RADAR data are likely to evolve in coming years. Some useful techniques for working with RADAR data (e.g., spatial filter kernels) have been reported in the literature.

In order to facilitate processing and interpretation of RADAR data, TerrSet offers a texture analysis module (TEXTURE) and several options in the FILTER module, including the Adaptive Box Filter. These filtering procedures can be used to minimize some of the speckled “noise” effects in RADAR imagery (although they cannot be completely eliminated) and also provide a quantitative interpretation of RADAR imaged surfaces. The Adaptive Box filter, which is an adaptation of the Lee filter, is highly recommended by Eliason and McEwen (1990) for reducing speckled effects (figure on the left). One must experiment with various filter kernel sizes and threshold options in order to achieve good results. Different filters may work better for different scenes, depending on the mixtures of land surface cover types in the imagery.

²⁰ Other options are vertical transmit, vertical receive (VV), horizontal transmit, vertical receive (HV) and vertical transmit, horizontal receive (VH).



One other method suggested for further reducing speckled effects is *multi-look processing* (Lillesand and Kiefer, 1987). This simply involves averaging (with OVERLAY or Image Calculator) scenes from the same area, acquired at different incidence angles, to produce a smoother image.

Quick-look images can be generated using the MEAN filter in TerrSet (center figure above) to enhance some of the features in image scenes in order to select sites for detailed analysis. The Sobel Edge Detector, the High Pass and the Laplacian Edge Enhancement filters are useful in detecting edges and linear features in imagery. Fault lines, surface drainage patterns, folds, roads and land/water boundaries are useful in various geological, resource management and a variety of environmental and planning applications. Note also that user-defined filters may be tailored to fit specific needs using both the 3 x 3 kernel and the variable size kernel in TerrSet's FILTER module. See the module description of FILTER in the Help System for a detailed explanation of these filtering techniques.

As already mentioned, the RADAR backscatter signal is composed of the various diffuse and specular responses of scene elements to the sent RADAR pulse. The compound pattern of the varying surface responses can be exploited in order to determine the textural characteristics of the land surface. These characteristics can be derived using the TEXTURE module in TerrSet. TEXTURE includes three categories of analysis. The first uses variability in a moving window to assess several different measures, including entropy. The second estimates the fractal dimension of the image surface. The third provides directional edge enhancement filters to enhance edge patterns in different directions. For example, the image on the right above shows a fractal surface derived using the TEXTURE module. Each of the surfaces derived from TEXTURE can be used as input in a classification scheme for RADAR imagery. Thus, instead of using spectral responses, one utilizes scene textural characteristics in classification.

The 24-day repeat cycle of RADARSAT can be exploited in the monitoring of surface phenomena that exhibit temporal variability, as each time step has a different RADAR backscatter signal. Using the COMPOSITE routine in TerrSet, multi-date RADAR imagery can be combined to produce a false color composite (very much like multispectral single scene data) that shows the temporal transitions in given earth surface components like crops or different vegetation types or surface cover types.

It is anticipated that as RADAR data becomes more widely used, there will be accompanying developments in software to exploit the unique character of RADAR imagery.

References

Eliaison, E. M., and McEwen, A. S., 1990. Adaptive Box Filters for Removal of Random Noise from Digital Images, *Photogrammetric Engineering and Remote Sensing*, 56(4): 453-458.

Lillesand, T., Kiefer, R. W., and J.W. Chipman, 2004. *Remote Sensing and Image Interpretation*, John Wiley and Sons, New York.

RADARSAT International, 1995. *RADARSAT Guide to Products and Services*, RADARSAT International, Richmond, B.C., Canada.

CHAPTER SIX

LAND CHANGE MODELER

The Land Change Modeler (LCM) is an integrated software environment within TerrSet oriented to the pressing problem of accelerated land conversion and the very specific analytical needs of biodiversity conservation. Clark Labs worked with Conservation International over a period of several years to develop this suite of tools to be used in a variety of land change scenarios and contexts. In LCM, tools for the assessment and prediction of land cover change and its implications are organized around major task areas: change analysis, change prediction, and planning interventions. Also, there is a facility in LCM to support projects aimed at Reducing Emissions from Deforestation and Forest Degradation (REDD). The REDD facility uses the land change scenarios produced by LCM to evaluate future emissions scenarios.

LCM is a vertical application that sits within TerrSet. However, unlike TerrSet itself, which can be characterized as a horizontal application – a software product meant to fulfill many applications, a vertical application is directed towards a specific application. Earth Trends Modeler is another vertical application within TerrSet.

The Organization of LCM

The Land Change Modeler interface is organized around a set of six major tasks:

- Change Analysis: Analyzing past landcover change.
- Transition Potentials: Modeling the potential for land transitions.
- Change Prediction: Predicting the course of change into the future.
- Planning: Evaluating planning interventions.
- REDD Project: Estimating GHG emissions from REDD projects.
- Harmonize: Formatting land cover maps for use in Land Change Modeler (tab only launches when land cover maps are incorrectly formatted).

The Harmonize tab is not shown by default and will only be invoked if the land cover maps used in the analysis are incorrectly formatted for use with LCM. The land cover maps used by LCM must be formatted to meet the following conditions:

1. The legends in both maps are the same.
2. The categories in both maps are the same and sequential.

3. The backgrounds in both maps are the same and have a value of zero.
4. The spatial dimensions, including resolution and coordinates, are the same.

If any of these conditions are not met, the Harmonize tab will open and allow the user to make changes to the input land cover maps. Any errors must be resolved before continuing with change analysis in LCM.

Within each tab, a series of tasks/analytical stages are presented as a series of drop-down panels. You can have as many drop-down panels open as you wish – they are presented this way simply to accommodate varying screen resolutions.

The first three of the six tabs of LCM are intended for the integrated analysis of land cover change and its projection into the future. As a result, access to almost all panels on these tabs requires the specification of a minimal set of project parameters located on the first panel of the first tab.

Note that the panels on the first three tabs are generally organized around a sequential set of operations that should be followed one after the other.

The Change Prediction Process in LCM

Land change prediction in Land Change Modeler is an empirically driven process that moves in a stepwise fashion from 1) Change Analysis, 2) Transition Potential Modeling, to 3) Change Prediction. It is based on the historical change from time 1 to time 2 land cover maps to project future scenarios.

At the end of the chapter are two flow charts that provide a visualization of the process.

Change Analysis

In Change Analysis, change is assessed between time 1 and time 2 between the two land cover maps. The changes that are identified are transitions from one land cover state to another. It is likely that with many land cover classes the potential combination of transitions can be very large. What is important is to identify the dominant transitions that can be grouped and modeled, termed submodels. Each group, or submodel of transitions can be modeled separately, but ultimately each submodel will be combined with all submodels in the final change prediction. See flow chart at the end of this chapter.

Transition Potential Modeling

The second step in the Change Prediction process is Transition Potential Modeling, where we identify the potential of land to transition. At this stage, we will create transition potential maps that are in essence maps of suitability for each transition. In LCM, a collection of transition potential maps is organized within an empirically evaluated transition sub-model that has the same underlying driver variables. A transition sub-model can consist of a single land cover transition or a group of transitions that are thought to have the same underlying driver variables. These driver variables are used to model the historical change process. For example, if you are modeling deforestation due to agriculture, driver variables to consider would be slopes, proximity to roads, or proximity to previously deforested areas. Transitions are modeled using either a Multi-Layer Perceptron (MLP) neural network, Decision Forest (DF) machine learning, Logistic Regression, Weighted Normalized Likelihood (WNL), Support Vector Machine (SVM), or a similarity-weighted instance-based machine learning tool (SimWeight). Once calibrated, the model is used to predict future scenarios.

Change Prediction

The third and final step in Change Analysis is Change Prediction. Using the historical rates of change and the transition potential model, LCM can predict a future scenario for a specified future date. In its simplest form, the model will determine how the variables influence future change, how much change took place between time 1 and time 2, and then calculate a relative amount of transition to the future date. In order to make the model more robust, Land Change Modeler allows for the incorporation of constraints and incentives, such as zoning maps, and planned infrastructure changes, such as new roads or land cover development. Driver variables can also be dynamic in nature, so that at regular interim intervals they can be recalculated and reenter the model. For example, you may have as one of your driver variables distance from previously deforested areas. This variable could be recalculated at regular intervals as land cover transitions from forest to agriculture. See flow chart at the end of this chapter.

The Change Analysis Tab

The first step in the Change Prediction process is Change Analysis. The Change Analysis tab provides a set of tools for the rapid assessment of change, allowing one to generate one-click evaluations of gains and losses, net change, persistence and specific transitions, both in map and graphical form. The Change Analysis tab consists of four panels, LCM Session Parameters, Change Analysis, Change Maps, and Spatial Trend of Change.

LCM Session Parameters Panel

This panel allows for the specification of the essential files associated with the landcover change analysis of a specific study area.

For the change and prediction analyses, a minimum requirement is the specification of two landcover maps that can be used as the basis of understanding the nature of change in the study region and the means of establishing samples of transitions that should be modeled. Optional inputs include an elevation model and basis roads layer that are used in dynamic road development.

Change Analysis Panel

The Change Analysis panel provides a rapid quantitative assessment of change by graphing gains and losses by landcover category. A second option, net change, shows the result of taking the earlier landcover areas, adding the gains and then subtracting the losses. The third option is to examine the contributions to changes experienced by a single landcover.

Change Maps Panel

This panel provides the ability to create a variety of change maps. There are options to map changes, persistence, gains and losses by land cover category, and map transitions by categories or map exchanges by category.

Spatial Trend of Change Panel

In landscapes dominated by human intervention, patterns of change can be complex, and thus very difficult to decipher. To facilitate interpretation in these contexts, a spatial trend analysis tool has been provided. This is a best fit polynomial trend surface to the

pattern of change. The analytical work done by this option is achieved by a call to the TREND module. Trends up to the 9th order can be calculated.

The Transition Potentials Tab

The Transition Potentials tab groups transitions between two land cover maps into a set of sub-models where each sub-model is identified with a set of driver, or explanatory variables. A transition potential map for each transition is derived which is an expression of time-specific potential for change. There are facilities on this tab to transform driver variables and to explore the potential power of each of the explanatory variables. Variables can be added to the model either as static or dynamic components. Static variables express aspects of basic suitability for the transition under consideration, and are unchanging over time. Dynamic variables are time-dependent drivers such as proximity to existing development or infrastructure and are recalculated over time during the course of a prediction.

Once model variables have been selected, each transition sub-model is modeled using one of six model types: 1) Multi-Layer Perceptron (MLP) neural network, 2) Decision Forest machine learning, 3) Logistic Regression, 4) Weighted Normalized Likelihood (WNL), 5) Support Vector Machine (SVM), or 6) a SimWeight procedure which is based on a modified K-nearest neighbor machine learning algorithm. The SimWeight procedure has been shown to deliver similar results as the MLP procedure, but with minimal parameters (although at the cost of a much longer analysis time). All of these methods are capable of producing good results. However, here are a few perspectives based on experience:

- MLP is extremely capable (especially with complex non-linear relationships) but has the most complex parameters and requires some degree of experience to use it well. However, it does have an automated supervision capability that works quite well. MLP is not recommended for modeling large numbers of transitions simultaneously. It is better to break the problem down into smaller sub-models.
- DecisionForest is the simplest to operate and generally produces good results.
- The output of Logistic Regression is in the form of probabilities rather than transition potentials, except when the option of balanced samples is used. If soft prediction is an important component, then we recommend using balanced samples to produce transition potentials.
- Like MLP, SVM should not be used with large numbers of transitions. It is slower than the others, but it is extremely capable.
- WNL is the fastest and can model very large numbers of transitions simultaneously. However, it is prone to minor over-fitting, which can produce local artifacts.
- SimWeight is very capable but is very slow. We do not recommend it with large training samples.

The result with all the model types is a transition potential map (or probability map for Logistic Regression without a balanced sample) for each transition – an expression of time-specific potential for change.

By default, in LCM each transition is considered to be a separate sub-model. However, transitions can be grouped into larger sub-models and modeled together. This is logical if it can be assumed that the same explanatory variables and driving forces are relevant to account for the transitions experienced. LCM even provides the ability to group all of the transitions into a single sub-model, although this is not commonly logical. Note that SimWeight is the exception in that transitions must each be modeled separately. MLP performs well with small sub-models, but tends to perform poorly with large sub-models. The only technique that is capable of modeling very large sub-models with speed is WNL.

Transition Sub-Models: Status Panel

The Transition Sub-Models: Status panel lists all transitions that exist between the two land cover maps. Transitions can be filtered out using an area threshold and transitions can be assigned to specific sub-models. Transitions should be grouped only if it is assumed that the underlying driving forces of change are the same. The only procedure that can handle very large numbers of transition simultaneously is WNL.

Variable Transformation Utility Panel

The Variable Transformation Utility panel is optional and provides a selection of optional commonly used transformations. These transformations include: evidence likelihood, natural log (ln), exponential (e), logit, square root, and power. These are particularly critical if the Logistic Regression modeling option is chosen since it requires that the variables be linearly related to the potential for transition. The MLP option does not require the variables to be linearly related, but transformation can sometimes make the task easier for it to solve in cases of strong non-linearities, thus yielding a higher accuracy. In general, the transformations are self-evident, but two require special mention.

1. The natural log transformation is effective in linearizing distance decay variables (e.g., proximity to roads).
2. The evidence likelihood transformation is a *very effective* means of incorporating categorical variables into the analysis. For all model types, logistic regression, MLP and SimWeight, variables must either be converted into a set of Boolean (dummy) variables, or transformed using the evidence likelihood transformation option (highly recommended).

The Evidence Likelihood transformation requires two inputs:

- A Boolean map of areas that have gone through the transition being modeled (this can easily be created from the Change Analysis tab).
- A categorical variable or a continuous variable that has been binned into classes (such as with the STRETCH module).

The procedure looks at the relative frequency of pixels belonging to the different categories of that variable within areas of change. In effect, it asks the question of each category of the variable, “How likely is it that you would have a value like this if you were an area that would experience change?”

Transition Sub-Model Structure Panel

The Transition Sub-Model Structure panel provides a table for specifying the explanatory variables to be evaluated. For each sub-model, a list of explanatory variables is specified that will be modeled in the next panel. Each of these variables can be either static or dynamic. Static variables are site variables that do not change over time, such as slope, elevation, etc. Dynamic variables are those that do change over time, such as proximity to development or proximity to roads (assuming dynamic road growth).

If a variable is dynamic, there is the option to identify whether it is dynamic based on a landcover category (such as proximity to urban) or a roads category (such as proximity to secondary roads). There is also the option to specify if the dynamic operation is a distance calculation or whether to use an IDRISI Macro script file. The former is the most common and will calculate distance from the designated landcover or road categories. The MACRO option leaves open an infinite set of possibilities.

When a variable is designated as dynamic and the basis layer type is selected as either roads or landuse, a dialog Box will pop up for specifying the relevant landcover category or road categories from which distance should be calculated in the change prediction stage. If a roads layer was not specified in the LCM Project Parameters tab, you will be able to specify this layer in the dialog and it will update the basis roads layer field in the LCM Project Parameters tab.

Run Transition Sub-Model Panel

The Run Transition Sub-Model panel is where the actual modeling of transition sub-models is implemented. The specific sub-model that will be implemented is that specified in the Sub-Model to be Evaluated drop-down combo box in the Transition Sub-Models: Status panel.

Six methodologies are provided for modeling: a multi-layer perceptron (MLP), DecisionForest, Logistic Regression, WNL, SVM and SimWeight. . In all cases, when the Run Sub-Model button is clicked, samples are extracted from the two landcover maps by the LCM project. These samples are the areas that underwent the transitions being modeled as well as the areas that were eligible to change, but did not. For each modeling option, there are parameters that must be set.

Multi-Layer Perceptron

The MLP option can model multiple transitions at one time. The MLP dialog in LCM has been modified from the existing MLP module found in TerrSet as used for image classification. Initially the dialog for the Multi-Layer Perceptron neural network may seem daunting, but most of the parameters presented do not need to be modified (or in fact understood) to make productive use of this very powerful technique.

As launched by LCM, the Multi-Layer Perceptron starts training on the samples it has been provided of pixels that have and have not experienced the transitions being modeled. At this point, the MLP is operating in automatic mode whereby it makes its own decisions about the parameters to be used and how they should be changed to better model the data. Automatic mode monitors and modifies the start and end learning rate of a dynamic learning procedure. The dynamic learning procedure starts with an initial learning rate and reduces it progressively over the iterations until the end learning rate is reached when the maximum number of iterations is reached. If significant oscillations in the RMS error are detected after the first 100 iterations, the learning rates (start and end) are reduced by half and the process is started again.

All other parameters of the MLP are used by LCM at their normal default values. However, LCM does apply special modifications to the outputs. Since specific transitions are being modeled, LCM masks out of the transition potentials all cases that do not match the from case of any specific transition. For example, if the transition being modeled is from forest to agriculture, values will only exist in pixels that were forest to start with.

In most cases, you can let it run in this mode until it completes its training. However, you are also free to stop its training operation, modify parameters and start it training again. Once the training process has completed, an HTML file will be displayed with information on the training process including information about the relative power of the explanatory variables used. The transition potential maps can then be created. For detailed information on the MLP parameters, see the Help System for MLP in the Transition Potentials tab.

DecisionForest

DecisionForest is a modified version of the free-standing DecisionForest module. In most cases, the default parameters will work well. The number of variables to evaluate at each split defaults to the floor of the square root of the number of independent variables. The number of trees defaults to 100. In the model report, the "Out of Bag" accuracy is the accuracy calculated with the validation data held back (50% of the sample size). The leave one out strategy for variable evaluation looks at the impact on accuracy if one of the variables is left out. Generally, the bigger the drop in accuracy, the more important the variable is. However, this may not be a reliable measure if there is strong inter-correlation between variables.

Logistic Regression

In reality, logistic regression can only model one transition at a time. However, if a sub-model is selected with multiple transitions, LCM models each in turn using the same parameters. Although it will depend on the size of the input data, it is recommended to run logistic regression with the default sampling percentage of 10%, which will reduce the processing time. This model undertakes binomial logistic regression and prediction using the Maximum Likelihood method. See the Help System for a detailed explanation of the logistic regression module parameters. By default, the output is in the form of probabilities of change based on the rate of change in the historic period analyzed. If you wish to have an output more in keeping with the transition potentials of other techniques, you should select the option for balanced samples. If this option is chosen, then equal size samples are chosen for each transition, rather than the default 10% sample. In addition, with the balanced samples option, the output probabilities are based on an assumed prior probability of 50% (similar to MLP). If you are mixing techniques within a single prediction, balanced samples should be used.

WNL

WNL is fast and is designed for cases where a very large number of transitions need to be modeled. The procedure is based on Weighted Normalized Likelihoods and performs a form of Kernel Density Estimation for estimation of the probability distribution associated with change for the independent variables. See Eastman et al. (2019)¹ for a detailed explanation. The speed of WNL, however, comes at the cost of some minor overfitting which can occasionally be seen as some minor spatial artifacts. However, it is unbiased and compares very favorably with the output from MLP.

SVM

Support Vector Machine (SVM) is a very powerful modeling tool that is computationally intensive. Therefore, it is not recommended for large numbers of transitions. It is also difficult to parameterize. Thus, the default in LCM is to set the parameters using grid search. This systematically tests various parameter settings and chooses the combination that yields the highest skill. Also, the default Radial Basis Function (RBF) is the most flexible choice as it can replicate all other kernels. Thus, it suggests to simply use the defaults as set out by LCM.

SimWeight

SimWeight can only model one transition at a time. SimWeight is a Similarity-Weighted Instance-based Machine Learning algorithm. It uses a slightly modified variant of the K-nearest neighbor procedure.² There are two primary parameters that need to be specified – the sample size and K. The sample size only needs to be large enough to be representative. Generally, a sample size from 1000-2000 pixels works fine (half of which will be used for training). The K parameter governs the degree of generality of the solution. Setting K too low will result in over-training. Setting it too high will result in over-generalization. It is recommended to begin with a K about 1/10th of the sample size. Overtraining will be noticeable as spatial artifacts, in which case, increasing K and re-running the model until a smooth result is achieved is recommended. Overtraining in the lowest transition potentials is generally not important. The more significant concern is overtraining in the highest transition potentials.

SimWeight uses half of the sample pixels for validation. The hit rate measures the mean transition potential among the validation pixels that actually went through the transition, while the false alarm rate measures the mean transition potential among pixels that were eligible to change but did not. The difference between these two is the Peirce Skill Score.

The calculation of the relevance weights is an indication of that variable's importance in the model. Relevance is calculated as:

¹ Eastman, J.R., Crema, S., Rush, H.R., and Zhang, K., (2019) "A Weighted Normalized Likelihood Procedure for Empirical Land Change Modeling", *Modeling Earth Systems and Environment*, <https://doi.org/10.1007/s40808-019-00584-0>

² Sangermano, F., Eastman, J.R., Zhu, H., (2010). Similarity weighted instance based learning for the generation of transition potentials in land change modeling. *Transactions in GIS*, 14(5), 569-580.

$$1 - \left(\frac{\sigma \text{ within change area}}{\sigma \text{ of overall image}} \right)$$

The Change Prediction Tab

The Change Prediction tab provides the controls for a dynamic landcover change prediction process. After specifying the end date, the quantity of change in each transition can either be modeled through a Markov Chain analysis or by specifying the transition probability matrix from an external (e.g., econometric) model. Two basic models of change are provided: a hard prediction model and a soft prediction model. The hard prediction model is based on a competitive land allocation model similar to a multi-objective decision process. The soft prediction yields a map of vulnerability to change for the selected set of transitions. In general, we prefer the results of the soft prediction for habitat and biodiversity assessment. The hard prediction yields only a single realization while the soft prediction is a comprehensive assessment of change potential.

In setting up the change prediction analysis, the user can specify the number of dynamic reassessment stages during which dynamic variables are updated. This also includes the optional dynamic growth (intensification) of the road network. At each stage, the system also checks for the presence of planning interventions (see below), including incentives and constraints and major infrastructure improvements.

The Change Prediction tab uses information from several other tabs. Of critical importance, all included transitions in the Transition Potentials tab must have been already modeled using MLP, DecisionForest, WNL, SVM, SimWeight, or logistic regression, i.e., a transition potential image must exist for each included transition. There are options to include infrastructural changes or incentives and constraints as specified on the Planning tab.

Change Demand Modeling Panel

The default procedure for determining the amount of change that will occur to some point in the future is by means of a Markov Chain. A Markovian process is one in which the state of a system can be determined by knowing its previous state and the probability of transitioning from each state to each other state. To determine this, LCM makes a call to TerrSet's MARKOV module at the time a prediction is run. Using the earlier and later landcover maps along with the date specified, MARKOV figures out exactly how much land would be expected to transition from the later date to the prediction date based on a projection of the transition potentials into the future. Note that this is not a simple linear extrapolation, since the transition potentials change over time as the various transitions in effect reach an equilibrium state.

The default Markov transition probabilities are calculated by entering the end prediction date. The resulting Markov matrix can be viewed and edited. Alternatively, one can specify a transition probability file from some other projection tool, such as an econometric model.

Dynamic Road Development Panel

This panel sets the parameters for dynamic road development. Dynamic road development is a procedure that tries to predict how roads will develop in the future.³ Three levels of roads are recognized: primary, secondary and tertiary, which must be coded with

³ Dynamic road development in TerrSet was inspired by the pioneering work of the DINAMICA team. See Soares-Filho, B.S.; Assunção, R.M.; Pantuzzo, A. Modeling the spatial transition probabilities of landscape dynamics in an Amazonian colonization frontier. *BioScience*, v. 51, p.1039-1046, 2001.

integer values 1, 2 and 3, respectively. Primary roads can only grow by extending their endpoints (if endpoints exist within the map). Secondary roads can grow as new branches off of primary roads and they can extend themselves. In a similar manner, tertiary roads can grow as new branches off of secondary roads, and they can extend, themselves. The following options control the growth patterns of new road end-points and new road routes:

Road Growth Parameters

The critical control parameters for dynamic road development are *road spacing* and *road length*. The former dictates the frequency with which roads are generated along a route of superior class. Specifically it is the minimum distance that must separate roads along a route of superior class. The latter dictates the maximum length a road class will grow in each dynamic stage. The actual length of any new segment will fall randomly within that range.

Mode of End-Point Generation

Within the limits of the controlling parameters, new road end-points can be generated either randomly or by means of a procedure that looks for the location of highest transition potential, but with a stochastic perturbation. Rather than picking the location with the absolute highest transition potential within the growth length parameter, a small random perturbation is added to the transition potentials such that there is a large chance it will pick a location very similar to the highest transition potential and an increasingly less likely chance of taking one that is quite different.

Using the stochastic highest transition potential option allows for the selection of only those transitions that are relevant for road growth. For example, a model might include transitions related to declines in agriculture as well as urbanization. Declines in agriculture may not be a basis for road growth.

Mode of Route Generation

Once a new endpoint for a road has been generated, two options are provided for how the route is selected in joining up that location to the existing road network. The default option is the minimum gradient route. This route is a balance between trying to achieve a short route and the need to avoid steep slopes as much as possible. Alternatively, the highest transition potential route balances the need for a short route with the desire to link up as many areas of high transition potential as possible (on the assumption that these are areas that will have a high likelihood of needing a road connection in the future).

Using the highest transition potential route option allows for the selection of only those transitions that are relevant for road growth. These are the same transitions that would be relevant for end point generation (see above).

Skip Factor

In our experience, we have found that it is sometimes more efficient to not build roads at every stage, but build them only after several stages have passed. Thus, there is a skip factor parameter to set stages for road building.

Change Allocation Panel

See also: Jiang, Ziyang. "The Road Extension Model in the Land Change Modeler for Ecological Sustainability of IDRISI." 15th ACM International Symposium on Advances in Geographic Information Systems, Nov. 7-9, 2007, Seattle, WA.

The Change Allocation Panel parameterizes and initiates the actual prediction process. The process is based on the transition potential maps developed from the Transition Potentials panel and the prediction date set in the Change Demand Modeling panel. The following parameters can be set for predicting landcover change:

Recalculation Stages

This option controls how much of the total quantity of change to be allocated throughout the prediction process from the later landcover date to the specified prediction date. By default, a recalculation stage of 1 will allocate the entire quantity of change to the prediction date. By increasing the recalculation stages, that total quantity of change will be divided linearly to each stage in the allocation process. For example, if you are allocating over a 10 year period 1000 ha. and predicting from a later year of 2011 to 2020, setting recalculation stages to 2 will allocate half the total quantity of change to year five (2015) and the other half to year 10 (2020).

The recalculation stage option also dictates the frequency with which dynamic elements are recalculated. Dynamic elements include dynamic variables, dynamic road building, infrastructure changes, and incentives and constraints. For each of these elements, the total quantity of change is allocated as to the number of recalculation stages in a linear fashion state above, but the dynamic elements are used to alter the location of this quantity at each stage.

The use of dynamic variables is set in the Transition Sub-Model Structure panel on the Transition Potentials tab. A dynamic variable is one which varies in character with time. For example, one of the variables associated with a specific transition might be distance from deforested areas. As time progresses, the extent of this deforested area will increase, thereby changing this distance variable. All explanatory variables that are indicated as being dynamic are recalculated at each stage.

Dynamic road building is undertaken (unless a skip factor has been specified) through the Dynamic Road Development panel. Dynamic road building is a predictive modeling of the development of roads over time.

Infrastructural changes are reviewed and incorporated as necessary. These are specified on the Planning tab. For example, this could include known new roads to be developed at specific time periods in the future.

Incentives and constraints are applied to each transition potential map in the model. At each recalculation stage, the transition potential map associated with each transition is essentially multiplied by the constraint and incentive map during the change prediction process. These incentive and constraint maps are also specified on the Planning tab

Hard Versus Soft Prediction

LCM offers two modes of change prediction: hard and soft. A hard prediction is a commitment to a specific scenario. The result is a landcover map with the same categories as the inputs. The hard prediction procedure used by LCM is based on TerrSet's multi-objective land allocation (MOLA) module. LCM looks through all transitions and creates a list of host classes (classes that will lose some amount of land) and a list of claimant classes (classes that will acquire land) for each host. The quantities are determined from a run of the MARKOV module. A multi-objective allocation is then run to allocate land for all claimants of a host class. The results of the reallocation of each host class are then overlaid to produce the result. The module that performs this work is CHGALLOC, which is an internal module that does not exist in the menu system. When running, it will report the number of passes, which is identical to the number of host classes.

In contrast, the soft output is a continuous mapping of vulnerability to change. It doesn't say what *will* change, but rather, the degree to which the areas have the right conditions to precipitate change. It is simply an aggregation of the transition potentials of all selected transitions.

When creating the soft prediction, there is the option to select which transitions to include in the portrait of vulnerability. By default, all transitions used to develop the prediction model are selected, in which case the vulnerability to any kind of change is modeled. More typically, you would select only certain transitions to include. For example, if your interest is in forests, you might include all transitions that relate to the loss of forest cover.

There is also the option to set aggregation type for the soft prediction. The soft prediction is based on the current state (during the prediction) of transition potentials for each of the selected transitions. These are then aggregated to produce the soft output for each stage. Two aggregation options are provided: *maximum* and *logical OR*. The former characterizes a pixel by the maximum transition probability that exists at that location for the included transitions. The second calculates the logical OR of these transition potentials. This latter option treats a location as being more vulnerable if it is wanted by several transitions at the same time. For example, if a certain pixel is evaluated as 0.6 as its potential to transition to one cover type and 0.7 to another cover type, the former option would calculate its vulnerability to change as 0.7 while the latter would evaluate it at 0.88 $((A+B)-(A*B))$. It is left to the user to decide which is the more appropriate in the context of the study being undertaken.

Display Options

LCM provides several options for display of the prediction. One is to display the intermediate stage images (as opposed to only the final prediction). This option should be used with care, as Windows display memory be rapidly exhausted, putting the entire Windows system into an unstable state. The limits here will depend, in part, on how much RAM is installed on your system.

A second option for display is to create an AVI video file. In TerrSet, this file can be played in TerrSet's Media Viewer – a utility provided under the File Display menu. It can also be played with programs such as Microsoft Media Player and can be inserted into a Microsoft PowerPoint presentation. For long sequences, a frame rate of 0.25 generally works well, but slower rates may be more appropriate for slower sequences.

Validation Panel

The Validation panel allows you to determine the quality of the predicted land use map in relation to a map of reality. It does this by running a 3-way crosstabulation between the later landcover map, the prediction map, and a map of reality. The output will illustrate the accuracy of the model results where:

A | B | B = Hits (green) – Model predicted change and it changed

A | A | B = Misses (red) – Model predicted persistence and it changed

A | B | A = False Alarms (yellow) – Model predicted change and it persisted

The REDD Project Tab

The REDD Project tab is a modeling tool for calculating the estimated greenhouse gas (GHG) emission reduction that would result from the implementation of a Reducing Emissions from Deforestation and Forest Degradation (REDD) project. REDD is a climate change mitigation strategy intent on the conservation of forests for the sequestration of carbon. An important component of the REDD analysis is the assessment of historical land cover change rates and patterns, in particular deforestation, along with the driving forces of these changes. REDD analysis utilizes the results from the change prediction modeling to assess a project's potential for carbon sequestration.

The result of the REDD analysis is a REDD Project Design Document (PDD) following the methodology developed by the World Bank's BioCarbon Fund and submitted for approval to the Verified Carbon Standards (VCS). The PDD is a series of spreadsheets that assess the carbon impact of the project, leakage and reference areas for specified intervals throughout the life of the project.

The REDD analysis follows a sequence of three panels on the REDD Project tab.

Calculate CO₂ Emissions Panel

This panel allows for the identification of the types of carbon pools included in the analysis as well as their respective carbon densities in terms of tons of carbon per hectare (tC ha⁻¹). Six carbon pools can be included in the analysis: above-ground, below-ground, dead wood, harvest wood products, litter, and soil organic carbon. For all carbon pools other than below-ground, two options for input type are provided, either a constant value of the carbon density of the pool or a raster image of the carbon density. When the below-ground carbon pool is included, there are also the options to specify carbon density as a percentage of the above-ground pool or to use the Cairns calculation.⁴

Calculate Non-CO₂ Emissions Panel

This panel allows for the inclusion of non-CO₂ emissions when deforestation is due to fire. Specifically, methane (CH₄) and nitrous oxide (N₂O) emissions can be estimated per carbon pool, excluding harvested wood products and soil organic carbon, when deforestation is due to fire. For each land cover class being modeled in the change prediction process, the proportion of the forest burned is specified, along with the percentage of it being burned for each of the above-ground, deadwood and litter carbon pools. Also, the combustion efficiency of each of these three carbon pools must be specified.

Calculate Net GHG Emissions Panel

The final panel, Calculate Net GHG Emissions, uses the information in the previous REDD tables to estimate the net GHG emissions based on a REDD project's estimated effectiveness. This effectiveness rate is calculated by specifying, for each REDD reporting interval, a leakage rate and success rate. The success rate reflects how much deforestation inside of the project area will be prevented at each stage. The Leakage rate represents the amount of forest that will be deforested in the leakage area as a result of displacement from protecting the project area. The difference of these two rates is the effectiveness of the REDD project for each reporting interval.

The following is the list of tables produced as a result of the REDD analysis:

1. List of carbon pools included or excluded in the proposed REDD project activity.
2. List of sources and GHG in the proposed REDD project activity.
3. List of land cover classes with their respective average carbon density per hectare (tC ha⁻¹) in different carbon pools.
4. Baseline deforestation activity data per land cover class in project area, leakage area and reference area.
5. Baseline carbon stock changes per land cover class in project area, leakage area and reference area.

⁴ Cairns, M.A., Brown, S., Helmer, E.H., Baumgardner, G.A. 1997. Root biomass allocation in the world's upland forests. *Oecologia*, 111(1), 1-11.

6. List of LULC classes with their respective average emission per hectare ($\text{tCO}_2\text{e ha}^{-1}$) in different sources.
7. Baseline non- CO_2 emission per LULC class in project area, leakage area and reference area.
8. For each stage, the output represents the increase in GHG emissions due to leakage from the project area. This will reduce the overall effectiveness of the project by decreasing the baseline carbon stocks. This is calculated for both CO_2 and non- CO_2 emissions.
9. For each stage, the output represents the ex-ante net anthropogenic GHG emission reductions (C-REDD), accounting for reductions in the carbon baseline (C-Baseline) due to leakage (C-Leakage) and the project's actual success rate (C-Actual). The final calculation is:

$$\text{C} - \text{REDD} = (\text{C} - \text{Baseline}) - (\text{C} - \text{Actual}) - (\text{C} - \text{Leakage})$$

Additional LCM References

Fuller, D. O., Hardiono, M., and Meijaard, E. 2011. Deforestation Projections for Carbon-Rich Peat Swamp Forests of Central Kalimantan, Indonesia. Environmental Management.

Retrieved from www.ncbi.nlm.nih.gov/pubmed/21359865.

Sangermano, F., Toledano, J., Eastman, J.R. 2012. Land cover change in the Bolivian Amazon and its implications for REDD+ and endemic biodiversity. Landscape Ecology 27(4), 571-584. <http://dx.doi.org/10.1007/s10980-012-9710-y>

The Planning Tab

The Planning tab offers tools for incorporating planning interventions into the change prediction process.

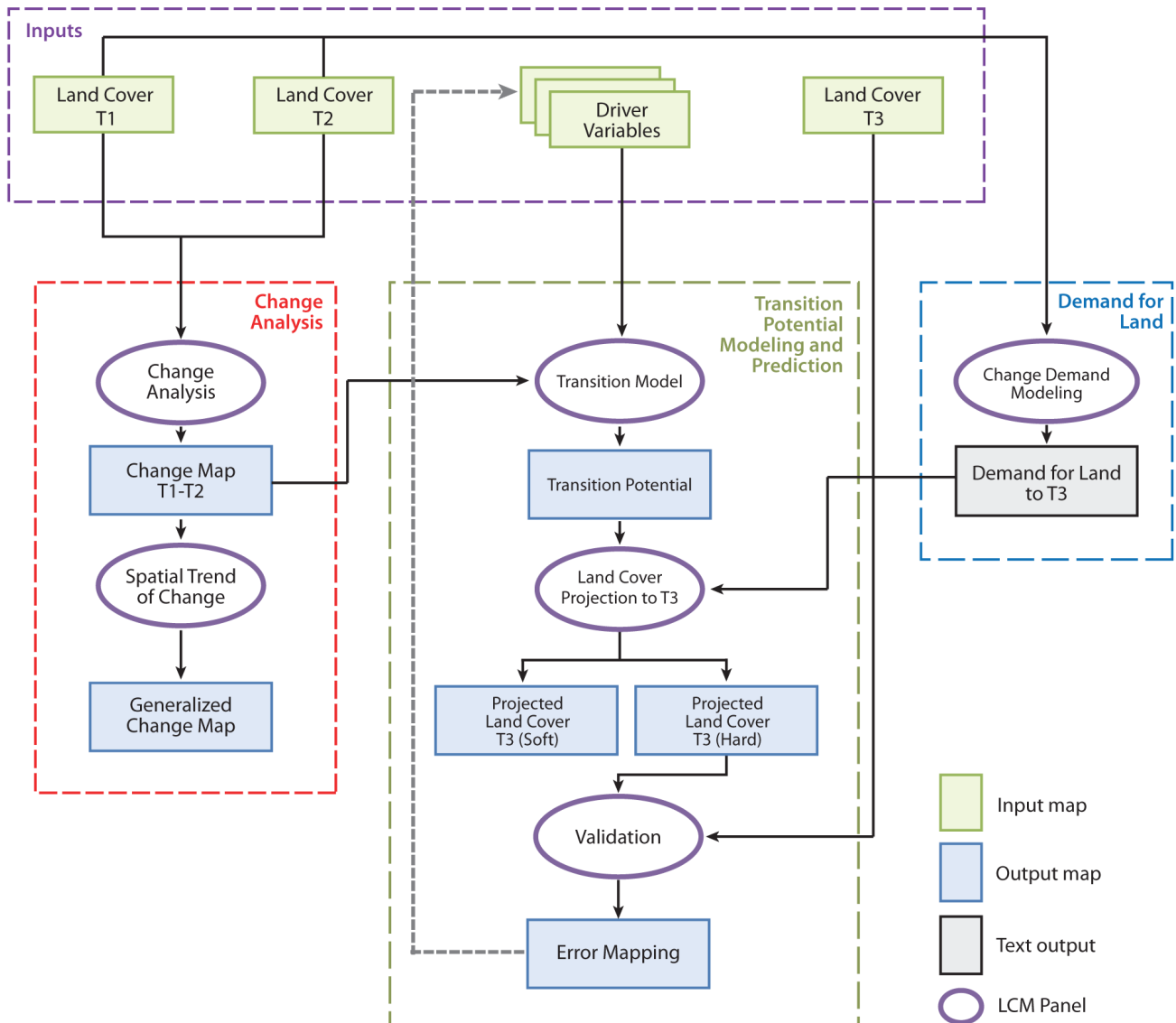
Constraints and Incentives Panel

The Constraints and Incentives panel provides the ability to assess the impacts of existing and proposed reserved areas, such as tax incentives, for redirecting the course of change. These interventions are integrated with the change prediction process. Constraints and incentives are incorporated by specifying an incentive/constraint map for each of the transitions in the model. Constraints and incentives are handled in a unified fashion. Values of 0 on the map are treated as absolute constraints while values of 1 are unconstrained and consequently have no impact. Values less than 1 but above 0 act as disincentives while values greater than 1 act as incentives. The way the constraints and incentives feature works is that the transition potentials associated with each transition are *multiplied* by the incentives/constraints map.

Planned Infrastructure Changes Panel

This panel allows for infrastructure modifications, e.g., major road developments, during the change prediction process. This planned intervention is incorporated through the specification infrastructure maps and the year they become effective. The Change Allocation panel of the Change Prediction tab checks this list with each stage of the prediction and adds each new road when the date of the active stage is equal to or greater than the infrastructure date. In addition, new infrastructural components can be developed by specifying the end-points and allowing the system to develop the least-cost engineering routes.

Change Analysis, Model Training, and Validation



CHAPTER SEVEN

HABITAT AND BIODIVERSITY MODELER

The Habitat and Biodiversity Modeler is intended to model species and habitat impacts, including biodiversity. Those familiar with Land Change Modeler will note that the tools in this application are those previously available in Land Change Modeler. Because of the extensiveness of these tools, we have now included them in a single vertical application. The tools include:

1. Habitat Assessment
2. Habitat Change and Gap Analysis
3. Species Modeling including Maxent
4. Biodiversity Analysis and the ability to import IUCN species range data
5. Landscape Pattern Analysis and Change Process
6. Corridor Planning
7. Reserve Planning using Marxan

Species Tab

The species tab models habitat, species gap analysis, refines existing species polygons, and habitat suitability and distribution.

Habitat Assessment Panel

The Habitat Assessment panel allows one to assess the status of habitat on an animal species-specific basis. Based on any existing or predicted land cover maps and an optional map of species-specific habitat suitability, the habitat assessment tool develops a map with five categories: primary habitat, secondary habitat, primary potential corridor, secondary potential corridor and unsuitable. Important parameters that control this process include home range sizes, buffers based on sensitivity to humans and the ability to cross gaps within home ranges and during dispersal. The resulting map can be used to estimate maximum populations and serves as a primary resource in the planning for corridors. The habitat suitability map can be created with the Habitat Suitability / Species Distribution panel.

Important terms and parameters that need to be specified include:

- ***Habitat and Potential Corridor***

The habitat assessment map produced by this analysis includes five categories of habitat status. Below, they are indicated with a possible interpretation. However, they can be interpreted in any way that seems appropriate to the study under consideration.

1. **Primary Habitat.** This is habitat that meets all the necessary life needs in terms of home range size, access to summer and winter forage, etc. Issues other than minimum area and required buffer size are specified by a minimum suitability on a habitat suitability map.
2. **Secondary Habitat.** This includes areas which have the designated habitat cover types, but which are missing one or more requirements (such as area or minimum suitability level) to serve as primary habitat. Secondary habitat areas provide areas of forage and safe haven for dispersing animals as they move to new areas of primary habitat.
3. **Primary Potential Corridor.** Areas of primary potential corridor are non-habitat areas that are reasonably safe to traverse, such as at night.
4. **Secondary Potential Corridor.** There are areas that are known to be traversed by the species in question, but which constitute much riskier cover types.
5. **Unsuitable.** These are areas that are not suited for habitat or corridors.

▪ ***Include as Potential Habitat***

Any land cover type from the input land cover maps can be included in the study and can be selected for those that are associated with habitat for the species in question.

▪ ***Gap Distance Within Range***

This parameter is concerned with gaps within the home range of the species of concern. Gap distances do not need to be specified by cover types included as potential habitat components.

▪ ***Gap Distance Outside Range***

This parameter is concerned with gaps that the animal is capable of crossing when dispersing. This parameter is important in determining which areas can serve as potential corridors. In addition, this parameter effectively establishes the maximum length of the corridor.

▪ ***Minimum Core Area***

This constitutes, in the case of primary habitat, the minimum home range area of the species involved, exclusive of any buffers (hence the use of the term core). For secondary habitat areas, the core area is more likely related to forage abundance.

▪ ***Minimum Edge Buffer***

This is the size of buffer needed as distance from human activity. For potential corridor areas, this therefore constitutes half the necessary corridor width.

▪ ***Minimum Habitat Suitability***

The inclusion of a habitat suitability model is optional but strongly recommended. For each of the main habitat/corridor categories, a minimum suitability can be specified for inclusion in that category. A general strategy for development of this layer is as follows:

1. Develop separate suitability maps for each of the primary and secondary habitat and potential corridor categories. The Habitat Suitability / Species Distribution panel provides a variety of tools for empirically developing this. However, the multi-criteria evaluation (MCE) option will most often be the tool of choice since the suitability mapping will be based on published reports of species/landscape associations.

2. Rescale the range of the primary habitat suitability map to a range of 0.75-1.0 using the STRETCH module. Then rescale the secondary habitat map to a 0.5-0.75 range; the primary potential corridor map to a 0.25 - 0.5 range and the secondary potential corridor map to a 0 - 0.25 range. Combine these four maps using the cover option in OVERLAY. The result will be a single map layer that ranges in value from 0.0-1.0. The default thresholds in HBM are set for 0.75, 0.5 and 0.25 in the decision for allocating land to the basic categories (before consideration of minimum area, gap crossing and buffer considerations). All areas with a value of 0 are by definition unsuitable.

In practice, the user is free to establish whatever thresholds are meaningful and logical in the context of their study.

Habitat Change / Gap Analysis Panel

This panel is used for two kinds of analyses: an analysis of change in habitat status (created by means of two runs of the Habitat Assessment panel) and Gap Analysis by comparing the results of one run of the Habitat Assessment panel and a protection layer map. In the case of habitat change, a graph is produced of gains and losses that can be altered with one of net change.

With gap analysis, the protection map can be either a simple Boolean image showing areas that are protected or not, or a multi-level integer map showing various protection levels. The result is simply a crosstabulation of habitat categories and protection levels.

Species Range Polygon Refinement Panel

This panel allows for the refinement of range polygon maps of species distributions developed by experts who draw the ranges onto map bases. This information is exceptionally valuable, but subject to error as a result of imprecision in the base maps, projection and geodetic datum errors, and limited geographical extent of expertise (i.e., the expert delineates only in the areas where she or he has expertise).

The underlying principle of the refinement process is to uncover the common environmental logic of the areas delineated by the range polygon. It does this by creating clusters of environmental conditions according to a set of environmental variables that the user believes can characterize the niche of the species. It then compares these clusters with the range polygon to determine the proportional inclusion of clusters within the range polygon. Clusters that fall wholly or largely within the polygon are assumed to describe essential components of that niche. Those that fall mostly or wholly outside are assumed to be unlikely components. The polygon is thus refined by removing areas that fall below a designated confidence. In addition, another option is provided to simply output a confidence map that can be used in conjunction with the original range polygon by the Weighted Mahalanobis Typicality procedure in the Habitat Suitability / Species Distribution panel. This is the default option and the one we generally recommend.

Environmental Variables and Cluster Development

The critical component of this analysis is the production of environmental clusters. For this you will need to supply a set of environmental variables that can describe basic environmental conditions. Because of the clustering technique used, this is limited to a maximum of seven variables. To stay within this limit, we strongly recommend the use of Principal Components Analysis as a way of reducing a larger set of variables to a smaller set of highly informative components. That said, you should avoid the inclusion of components with very low explanatory power.

What variables should be used? This should be decided in the context of the species being modeled. Variables can be divided into direct, resource or indirect gradients. Direct gradients are those variables that affect directly the physiology of the species (e.g. temperature), while resource gradients are those that are consumed (e.g. water, nutrients), and indirect gradients are those that do not have a direct effect on the species, but affect the distribution of a direct or resource gradient (e.g. elevation, aspect). Common variables used are elevation, slope and aspect (because of their relationship to temperature and soil moisture), the standard deviation and mean annual NDVI (as a statement of realized long term and seasonal habitat conditions), the standard deviation and

mean annual temperature (as a statement of variations in energy availability), the standard deviation and mean annual precipitation (as a statement of variations in water availability), and minimum and maximum temperature and precipitation (which represent extreme environmental conditions).

Output Options

Four output options are provided:

1. **Presence.** This is a refined range polygon where areas that are poorly associated with the core environmental characteristics of the original range polygon are removed.
2. **Presence/Pseudo-Absence.** The output is the same as the above except that areas that are extremely unlikely to be associated with the core environmental characteristics of the original range polygon are treated as absence while only those that have a close association are considered as presence. This option is provided to allow the use of modeling procedures that require absence data (such as logistic regression). However, bear in mind that the absences are really pseudo-absences. To account for sampling issues, the absence pixels are chosen as a random sample of those that meet the lower threshold criterion such that the number matches (given some variance associated with the random selection process) the number of presence pixels.
3. **Confidence.** This is the default option and the one we generally recommend. Each pixel within the original polygon is assigned a confidence value from 0-1 based on how well it fits the general nature of a coherent pattern of environmental conditions (as will be explained further below).
4. **Thresholded Confidence.** This option is the same as the above, except that areas that fall below a minimum specified confidence are forced to have a confidence of zero.

For all options except the Confidence output, an upper and/or lower threshold must be specified, to establish areas of presence or absence. The default thresholds will serve as a general guideline of the values that would be used. In general, for presence, you are looking for a value that separates a clear group of clusters that strongly overlap the range polygon, while for absence you want to isolate clusters that have very little or no presence in the polygon. In many instances, this is very hard to do, which is why we recommend the use of the Confidence option coupled with the Weighted Mahalanobis Typicality procedure in the Habitat Suitability / Species Distribution panel. Using this option, no decision needs to be made.

There is also a background mask option which is quite important to the use of this procedure. If you are modeling a land species and are working in an area with significant ocean areas, you should provide a mask image to remove these from the calculations of proportional areas. For marine species, clearly the opposite applies.

Habitat Suitability / Species Distribution Panel

This panel provides a set of tools for developing habitat suitability and species distribution maps. The specific options available depend upon the nature of the training data, if any, that will be used: presence only, presence/absence, abundance or none. In all cases, you will need to specify a set of environmental variables that define the species habitat or niche.

Environmental Variables: Habitat Suitability Mapping

For habitat suitability mapping, the variables used likely relate to habitat land cover types, proximity to summer and winter foraging areas, proximity to human disturbance and so on. All variables specified must be continuous variables unless the multi-criteria

evaluation (MCE) option is used. For all but the MCE option, categorical variables should be converted to a series of Boolean layers (also known as dummy variables). For the instance where MCE is used, an assignment procedure is provided that assigns suitabilities to categorical variable classes. Also with the MCE option, Boolean constraints can be added.

Environmental Variables: Species Distribution Modeling

The variables that should be used for species distribution modeling should be decided in the context of the species being modeled. Generally you would include variables that relate to the seasonal and interannual availability of energy and water. Commonly used factors include elevation and slope (because of their relationship to temperature and soil moisture), the first and second principal components of mean monthly Normalized Difference Vegetation Index (NDVI) imagery (as a statement of realized long term and seasonal growing conditions), the long term coefficient of variability in NDVI (as a statement of interannual variability), and the first two components of mean monthly precipitation and temperature.

No Training Data - MCE

The Multi-Criteria Evaluation option is designed for cases where training data are not available but where studies are available to guide the development of a suitability or distribution map by means of a multi-criteria evaluation. This option utilizes the special techniques available in the TerrSet modules FUZZY and MCE.

When using the MCE option, the first and very important stage in the analysis is to convert each of the environmental variables to factors. The difference between the two is that a variable is unscaled with respect to the model while a factor is scaled to a specific numeric range using a scaling procedure that is directly related to the expression of suitability. For example, if one were modeling a species that is sensitive to humans, a distance from human settlement layer might be used. Suitability would clearly be worst within and immediately next to areas of human occupation. As you move farther away, the land is becoming increasingly better, up to a limit. It might be that, once one reaches a distance of 2 kilometers, being further is now irrelevant - it is far enough away. In this case, maximum suitability (on the basis of this variable alone) will have been reached. Thus, we should rescale the variable such that suitability is 0 at the edge of human occupation and increases in value until it reaches its maximum at 2 km, and remains at that value for all greater distances. In the transition of multi-criteria evaluation, this process is known as standardization, but in reality one is recasting the data into an expression of membership in the fuzzy set of suitable lands.

Two options are provided for standardization: a call to the FUZZY module in TerrSet or a call to the ASSIGN module. The former is designed for the standardization of continuous variables such as in the example above while the latter is intended for the standardization of categorical variables. Note that in contrast to the standardization used in TerrSet's multi-objective decision making procedure, standardization here uses a 0.0-1.0 scaling range.

As with the MCE module itself, there are several aggregation options that dictate how the factors will be combined to create a single suitability map. The default is weighted linear combination (WLC), which is appropriate when you wish the factors to trade off (i.e., to allow poor qualities to be compensated by good qualities). The Minimum operator allows no trade-off and characterizes each location by its worst quality. This is clearly the most conservative operator. The Maximum operator also allows no trade-off, but characterizes locations by their best quality.

Presence Data

Presence data is probably the most common form of training data for species modeling - it records where the species has been observed, but not where it has been observed to be absent. Three procedures are available for dealing with these data: MaxEnt, Mahalanobis Typicality, and a Weighted Mahalanobis Typicality procedure.

- **MaxEnt**

Habitat and Biodiversity Modeler provides an interface to the MaxEnt program for modeling presence-only species data. MaxEnt is a widely used species distribution model that employs a maximum entropy approach to estimate the probability distribution of a particular species. Maximum entropy is a machine learning algorithm that estimates the probability distribution of a species by finding the maximum entropy distribution subject to the constraint that the expected mean equals the empirical mean of the distribution. This method falls into the category of a use-availability model, as it uses a sample of the environmental conditions present in the study region (called background) to define the species probability distribution. MaxEnt has been widely used as a presence-only species distribution model, as it has been described to have higher predictive accuracy than other presence-only methods, although careful attention should be placed when sample bias exists. One advantage of MaxEnt is that it is able to represent complex species-gradients relationships, as it uses transformation of the variables, called features (e.g., if product features is selected, the product of all possible pair-wise combinations of variables is used, allowing for the fitting of simple interactions).

The MaxEnt software, developed by Steven Phillips, Miroslav Dudik and Robert Shapire, can be downloaded for free from: https://biodiversityinformatics.amnh.org/open_source/maxent/, and must be installed before using this option in HBM. All MaxEnt functionality has been implemented in this version.

▪ ***Mahalanobis Typicality***

The Mahalanobis Typicality option assumes that the underlying species distribution is normal with respect to environmental gradients. However, our tests have shown that it performs reasonably even with mildly skewed data. The output is in the form of typicality probabilities - an expression of how typical the pixel is of examples it was trained on. Thus a value of 1.0 would indicate a location that is identical to the mean of environmental conditions that were evident in the training data. However, be careful about the interpretation of low typicalities. Since typicalities express the full range of variability, a low typicality may be unusual, but still legitimately a location that is part of the species' range. If you are looking for a threshold for when to consider an area as being unlikely to be part of its range, it is likely to be a very low value (e.g., 0.001). As an illustration of this concept, consider the case of a blue lobster. Blue lobsters are very rare, but they are still lobsters! See the Help System for MAHALCLASS for further information.

▪ ***Weighted Mahalanobis Typicality***

This option requires both a training site file and a confidence (weight) file. It was intended that this option would be used with the confidence output of the Species Range Polygon Refinement panel. A confidence/weight image contains values from 0.0-1.0 that express the degree of confidence that the pixel is truly a member of the species' range. Habitat and Biodiversity Modeler uses this file along with a corresponding training file and submits them to the FUZSIG module for developing the signature statistics that are needed by MAHALCLASS. FUZSIG creates a weighted multivariate mean and variance/covariance matrix based on the confidence weights.

Presence / Absence Data

Two modeling approaches are available for presence/absence data: multi-layer perceptron neural network and logistic regression.

▪ ***Multi-Layer Perceptron***

This option uses a subset of the options available in the MLP module utilizing the loaded environmental variables. Please refer to the Help System for MLP regarding this option.

▪ ***Logistic Regression***

This option uses a subset of the options available in the LOGISTICREG module utilizing the loaded environmental variables. Please refer to the Help System for LOGISTICREG regarding this option.

Abundance Data

Using abundance data, the sole modeling approach is multiple regression. This option calls the MULTIREG module in TerrSet using the loaded environmental variables. Please refer to the Help System for MULTIREG regarding this option.

Species Data Formatting

The input to these procedures can be vector, raster, XY-Text or XY-CSV. XY-Text is a text file format suitable for presence point data, where each location is referenced by an X and Y coordinate separated by one or more spaces or tabs. XY-CSV (comma separated values) is similar except that the X and Y pair are separated by a comma. For all other tabular formats, we recommend that you load the data into Database Workshop and output the data as a vector file. Database Workshop can accept a wide range of formats (including DBF, ACCDB, XLS and CSV) and allows you to sort and subset before outputting to a vector or raster layer.

Biodiversity Tab

The Biodiversity tab allows for the import of IUCN species range polygons and the assessment of species biodiversity.

Subset IUCN Species Ranges Panel

This tool allows you to import IUCN species data and to subset the data according to the IUCN Red List status: least concerned, near threatened, vulnerable, endangered, critically endangered, extinct in the wild, extinct, and data deficient. Resulting data can be used to create biodiversity maps using the Biodiversity Analysis Panel.

Biodiversity Analysis Panel

The Biodiversity Analysis panel provides the ability to produce a spatially explicit mapping of:

1. **Alpha Diversity:** the total number of considered species at each location. Alpha Diversity is computed simply as the richness of species at each location - i.e., it is the total number of species found at each location.
2. **Gamma Diversity:** the total number of considered species over a large region. Gamma Diversity is calculated as the richness of species over a region. Thus the value recorded at any pixel represents the richness within the region to which it belongs and not the richness at that particular spot.
3. **Beta Diversity:** the ratio of Gamma to average Alpha Diversity over a large region, and thus a measure of the turnover of species. There are many measures of beta diversity that have been proposed. The measure used here is the original Whittaker's beta diversity.
4. **Sorensen Dissimilarity:** a measure of species compositional dissimilarity. Sorensen's Dissimilarity is measured as 1 minus Sorensen's Index, where Sorensen's Index is computed as the number of species that are common between the pixel and the region to which it belongs divided by the average alpha within the region.
5. **Range Restriction:** a continuous measure of vulnerability that can also be interpreted as a measure of endemism. The Range Restriction Index is based on a comparison of the area over which the species is found relative to the entire study region. It is intended for continental or global scale analyses and should include a mask file to mask out water areas for land species or vice versa for marine species. The index ranges from 0-1 with high values indicating that the majority of species present

at that location have restricted ranges. Note that the index is continuous and does not rely on a threshold area to define range restriction. The formula for the Range Restriction Index is as follows:

$$RRI = \frac{\sum_{i=1}^n \left(1 - \left(\frac{range_{area}}{total_{area}} \right) \right)^2}{Alpha_Diversity}$$

where alpha diversity is expressed as richness (the number of species) and the total area is the total area of the image minus any masked areas.

In all cases, the input for this analysis is in the form of species range polygons. Three input formats are supported. The first is a vector composite polygon where all species polygons are contained within the same vector file. The second is a vector group file that lists the names of a set of vector files that contain the range polygons for a single species. The third is a raster group file that lists a set of raster files that contain the rasterized range polygons of a single species.

With the exception of Alpha Diversity, the data must ultimately be converted to a raster form for analysis. Thus, if a vector group file is supplied, each file is rasterized (using the spatial characteristics of the reference file) and a raster group file is created with the same name as the vector group file. If a vector composite file is used, it is first broken out into a set of separate vector files, along with a vector group file of the same name as the vector composite file. These vector files are then in turn rasterized.

▪ **Regional Definition**

All measures except Alpha Diversity and the Range Restriction Index require the definition of a region over which the index is calculated. Three options are provided. The vector and raster region polygon options will yield a mapping where all pixels within a region (such as an ecoregion) will have the same index value. The focal zone option, however, is quite different and can produce a different value at each pixel location.

The focal zone option calculates values by comparing the species composition in each pixel to those in a circular zone surrounding it. To use the focal zone option, you must set the focal zone diameter (e.g., 50 km). This focal zone is moved successively over every pixel in the image. As a consequence, the analysis does take considerable time to complete. Continental scale analyses at a moderate resolution (e.g., 1 km) are probably best set up at the end of the day so that they can run overnight.

Landscape Analysis Tab

The Landscape Analysis tab performs landscape pattern and change process analysis.

Landscape Pattern Analysis Panel

This panel permits analyses of landscape pattern of any landcover maps, or landcover prediction. Options include:

Normalized Entropy

This measure is Shannon's Entropy measure normalized by the maximum entropy for the number of landcover classes involved. Another common term for this measure is Diversity. It is calculated over the local neighborhood of each pixel, defined as a 3x3, 5x5 or 7x7 neighborhood. The formula is as follows:

$$E = - \sum \left(\frac{p \times \ln(p)}{\ln(n)} \right)$$

where p is the proportion of each class within the neighborhood, ln is the natural logarithm and n is the number of classes. The result is an index that ranges from 0-1 where 0 indicates a case where the landcover is uniform within the neighborhood and 1 indicates maximum diversity possible of landcovers within the neighborhood.

Relative Richness

This is another measure of diversity of cover classes, measured as:

$$R = n/n_{max} \times 100$$

where n is the number of different classes present in the neighborhood and nmax is maximum number of classes possible.

Edge Density

Edge Density is a simple measure of fragmentation. Edge density is tabulated as the number of adjacent pairs of pixels within the neighborhood that are different from each other relative to the maximum number of different pairs possible.

Patch Area

Patch Area groups adjacent pixels of similar landcover category into patches, calculates their areas, and outputs an image where each pixel expresses the area of the patch to which it belongs.

Patch Compactness

Patch Compactness groups adjacent pixels of similar landcover category into patches, calculates their compactness, and outputs an image where each pixel expresses the compactness of the patch to which it belongs. Compactness is calculated as:

$$C = SQRT(A_p/A_c)$$

where SQRT is the square root function, A_p is the area of the patch being calculated, and A_c is the area of a circle having the same perimeter as that of the patch being calculated.

Landscape Change Process Analysis Panel

The Landscape Change Process Analysis panel compares two landcover maps and measures the nature of the change underway within each landcover class. It does this by using a decision tree procedure that compares the number of landcover patches present within each class between the two time periods to changes in their areas and perimeters. The output is in the form of a map where each landcover class is assigned the category of change that it is experiencing. The interpretation of the categories is as follows:

Deformation: the shape is changing.

Shift: the position is changing.

Perforation: the number of patches is constant but the area is decreasing.

Shrinkage: the area and perimeter are decreasing but the number of patches is constant.

Enlargement: the number of patches is constant but the area is increasing.

Attrition: the number of patches and the area are decreasing.

Aggregation: the number of patches is decreasing but area is constant or increasing.

Creation: the number of patches and area are increasing.

Dissection: the number of patches is increasing and the area is decreasing.

Fragmentation: the number of patches is increasing and area is strongly decreasing.

Note, however, that while the output is in the form of a map, it is not spatially explicit - i.e., the process attributed to a landcover category is uniform over the entire study area.

The Planning Tab

The Planning tab offers a corridor planning tool to develop biological corridors as well as an interface to the Marxan software for reserve planning.

Corridor Planning Panel

This panel is used to build biological corridors. The primary inputs are Boolean maps of the two terminal regions and a habitat suitability map. Optional inputs include a development suitability map, a conservation value map and a protected lands map. There are options to specify the ideal corridor width and the number of branches. The first branch is by definition the best route. Successive branches are of lower quality.

HBM builds corridors using a cost distance procedure. The first step involves an aggregation of the various suitability/value maps. In general, the effect is such that conservation value increases suitability for the corridor while development value decreases it (although only in unprotected lands). Once an aggregate suitability map has been created, the suitabilities are converted to frictions and a cost distance is calculated from one of the terminal regions. A least-cost path is then run from the other terminal region back to the first. After this, a second cost distance is run from the least-cost path, after which it determines the mean relationship between cost distance and spatial distance to determine a cost threshold to use in constructing the corridor. If additional branches need to be built, the suitability of already selected corridor areas is reduced to zero and the process is repeated.

Marxan Panels

Two Marxan panels within the Habitat and Biodiversity Modeler Planning tab are meant to interface with the Marxan software. Marxan is freeware developed at The University of Queensland intended for conservation planning. Marxan provides techniques for reserve system design and performance as well as tools for developing multi-use zoning plans for natural resource management. Marxan can be used to identify areas that meet biodiversity targets, taking into account minimum costs.

Marxan is not provided with IDIRISI but can be downloaded for free from The University of Queensland website at: <http://www.uq.edu.au/marxan>.

CHAPTER EIGHT

GEOSIRIS

Carbon emissions due to deforestation and land use change represent nearly one-fifth of global greenhouse gas emissions, second only to the energy sectorⁱ. This sector has been cited as a potentially cost-efficient means of reducing emissions when compared to other sectors.ⁱⁱ One approach is the implementation of individual REDD (Reducing Emissions from Deforestation and Forest Degradation) projects, where a specific parcel of land is protected, and the change in deforestation and carbon emissions due to this protection is quantified. This approach is accessible in TerrSet on the REDD Project tab of the Land Change Modeler (LCM).

The GeOSIRIS model takes a different approach. Instead of protecting a specific parcel of land from deforestation, GeOSIRIS projects work on the regional or national scale, where a specific price is set to each ton of carbon dioxide emitted (\$/tCO₂e). The GeOSIRIS model assumes forest users face a trade-off between the agricultural revenue obtained from deforesting land, and the carbon revenue obtained by protecting it. Given a proposed carbon price, and maps of previous deforestation and other variables, the model predicts how deforestation, carbon emissions, and agricultural and carbon revenue would change if such a carbon-trading policy were implemented.

The GeOSIRIS model was originally developed as OSIRIS by Jonah Busch at Conservation International, as a transparent decision support tool for REDD+ policy makers.¹ OSIRIS stands for Open Source Impacts of REDD+ Incentives Spreadsheet, and the original versions of these tools are Excel spreadsheets available at Conservation International's website. As the OSIRIS tool was designed to be transparent and open-source, all calculations can be seen within this spreadsheet. The implementation of OSIRIS in TerrSet allows you to see the OSIRIS results in a spatial context, thus the name: GeOSIRIS.

This chapter goes into detail of how to use the GeOSIRIS module in TerrSet, as well as, the mathematical underpinnings of different aspects of the model. You may find it useful to review the REDD project tab in the Land Change Modeler (LCM), to see a different approach to reducing carbon emissions related to deforestation.

Outline of GeOSIRIS Panels

The GeOSIRIS model follows a sequence of seven panels.

¹ Busch, Jonah, et al. "Structuring economic incentives to reduce emissions from deforestation within Indonesia." *Proceedings of the National Academy of Sciences* 109.4 (2012): 1062-1067.

Busch, Jonah, et al. "Comparing climate and cost impacts of reference levels for reducing emissions from deforestation." *Environmental Research Letters* 4.4 (2009): 044006.

GeOSIRIS Project Parameters

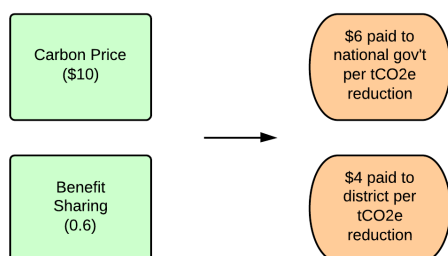
In this panel you can create, load, and save your GeOSIRIS project. You can also choose to overlay a vector file on the result images or use a custom palette for images.

External Factors

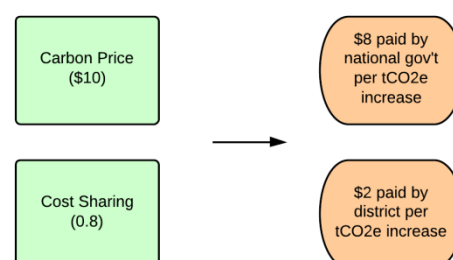
In this panel you specify (i) the carbon price for emission reductions or increases, and (ii) the country national reference level of emissions. The country national reference level of emissions specifies the baseline against which emission increases or decreases are measured.

Decision on National REDD+ Rules and Incentives

This panel determines how carbon revenue and penalties are shared between the national government and other administrative levels, such as districts or provinces that are specified in the Input Image Files panel. The GeOSIRIS model allows for a portion of carbon revenues and penalties to be shared, which can increase participation and emission reductions in the GeOSIRIS projectⁱⁱⁱ. The diagrams below show examples of how carbon revenue and penalties would be shared between the national government and the district level.



**Carbon revenue sharing with districts,
given a world carbon price of \$10 and a
benefit sharing proportion of 0.6.**



**Carbon penalty sharing with districts,
given a world carbon price of \$10 and a
cost sharing proportion of 0.8.**

Model Parameters

The model parameters are economic and affect the price of agriculture. The price elasticity is a measure of how sensitive the domestic production price of agriculture is to the change in deforestation. The exogenous increase in agricultural price is one portion of the final change in agricultural price (see the Equilibrium Step section below).

Input Image Files

This panel contains the input maps for the project area, the forest area, carbon (above/below ground, soil, and peat) and the administrative levels, like province or district on which GeOSIRIS is administered (see the Geographic Approach section). Maps can be Boolean or Fractional, which determines the regression type that can be run in the Effective Opportunity Cost panel (see the section Boolean vs. Fractional Maps and Logistic vs. Poisson Regression)

Effective Opportunity Cost

This panel calculates a regression between deforestation and agricultural revenue and other independent variables. There are two main outputs of this panel: (1) the regression coefficient relating agricultural revenue to deforestation, and (2) the effective opportunity cost image, which displays the relative probability of a pixel being deforested, based on a linear combination of the regression coefficients for all independent variables. (See the sections *Boolean vs. Fractional Maps* and *Logistic vs. Poisson Regression*, *Regression Coefficient Stratification*, and *Effective Opportunity Cost* sections for more information).

Output Parameters

The output parameters panel calculates the final change in the domestic agriculture price, based on an iterative process. The process ends when the proportional change in the agriculture price for two successive iterations is less than the model precision input.

Important GeOSIRIS Calculations

Emission Factor

The emission factor map is used to calculate how many tons CO₂ will be emitted per hectare of deforestation. There are three components for the emission factor in the GeOSIRIS model: above and below-ground carbon, soil carbon, and peat.

The equations for the emission factor at each pixel are:

$$E = (AB + SC * f_s) * 3.67 \text{ where } Peat = 0$$

$$E = AB * 3.67 + f_p \text{ where } Peat > 0$$

$$E = \text{emission factor} \left(\frac{tCO_2e}{ha} \right) \text{ (pixel level)}$$

$$AB = \text{above and below ground carbon (pixel level)}$$

$$SC = \text{soil carbon (pixel level)}$$

$$f_s = \text{soil carbon factor (input parameter)}$$

$$f_p = \text{emission factor for peat soil (input parameter)}$$

Note that when the peat value of pixel is non-zero, the emission factor equation does not depend on the depth of the peat soil, only the peat emission factor f_p .

Boolean vs. Fractional Land Ccover Type

You may have noticed in the Input Image Files panel that you have the option of selecting Boolean or Fractional land cover type. This refers to the data type of the (1) reference area, (2) forest before REDD, and (3) deforested area maps. When the Boolean option is selected, these maps must be Boolean images where a value of 1 means a pixel is either (1) completely contained in the reference area, (2) has 100% forest cover, and (3) 100% deforestation, respective of the three maps listed above.

If the data type is fractional, then these images change in the following ways: (1) the Reference Area map is renamed Land Area, with values in hectares. Pixels can have an area value representing only a fraction of the pixel, which is often the case for pixels along the coastline. (2) the Forest Before REDD image has values in hectares. (3) The Deforested Area map has value in hectares.

The data type also affects whether the GeOSIRIS Modeler uses a logistic or Poisson regression in the Effective Opportunity cost panel. See the Logistic vs. Poisson Regression section for more details.

The following table lists the discrepancies between Boolean and Fractional Landcover Types.

	Boolean	Fractional
Input maps that change (with units)	Reference Area (0 or 1)	Land Area (ha)
	Forest Before REDD+ (0 or 1)	Forest Before REDD+ (ha)
	Deforested Area (0 or 1)	Deforested Area (ha)
Input maps that do not change	Biomass Carbon (tC/ha)	Biomass Carbon (tC/ha)
	Soil Carbon (tC/ha)	Soil Carbon (tC/ha)
	Peat (tC/ha)	Peat (tC/ha)
	Independent Variables	Independent Variables
Regression Type	Logistic or External Model	Poisson or External Model

Logistic vs. Poisson Regression

In the Effective Opportunity Cost panel, a regression is used to calculate the regression coefficient relating agricultural revenue (and other variables) to deforestation. The regression type can be logistic (for Boolean data) or Poisson (for fractional data). This section briefly describes the difference between these two regression types. Please note that while the regression coefficient displayed in GeOSIRIS modeler is for the agricultural revenue variable, all coefficients are used in calculating the Effective Opportunity Cost image (described in the *Effective Opportunity Cost* section).

Logistic Regression

A logistic regression analysis in TerrSet uses the LOGISTICREG module to perform a binomial logistic regression. In this regression, the dependent variable (deforestation) must only have values of 0 and 1. The logistic regression equation is.

$$P(y = 1 | X) = \frac{e^{\sum_{i=1}^N B_i X_i}}{1 + e^{\sum_{i=1}^N B_i X_i}}$$

where:

P = the probability of the dependent variable being 1
 X_i = independent variables ($X_0 = 1$ for the constant term)
 B_i = variable coefficients (or parameters)

For more information, please see the related TerrSet Help system.

Poisson Regression

A Poisson regression is generally used to model count data, such as the number of cases of diabetes occurring in a population within a given period of time. For the GeOSIRIS modeler, deforestation can be interpreted as count data by thinking of each pixel as composed of smaller subsections which may be individually deforested. For example, a 3 km by 3 km pixel will have an area of 900 ha, which can be thought of as 36 individual 25 ha sections, each of which can be individually deforested.

The Poisson regression equation is

$$E(Y | X) = e^{\sum_{i=0}^N B_i X_i}$$

where:

$E(Y | X)$ = the expected "count" of deforestation (Y) given certain input conditions (X)
 X_i = independent variables ($X_0 = 1$ for the constant term)
 B_i = variable coefficients (or parameters)

There is technically no upper bound for the response variable Y in a Poisson regression. If any pixels are predicted to have a greater area of deforestation than their currently forested area, then the predicted deforestation area is scaled down to equal the amount of currently forested area.

Regression Coefficient Stratification

The regression step of the GeOSIRIS modeler calculates the relationship between deforestation and several independent variables, including agricultural revenue. There are several options to *stratify* this regression, where GeOSIRIS will run a separate regression for several different classes. These classes can be based on either the amount of preexisting forest cover (for quantile or user-defined classification), or geographic regions, such as provinces or districts (for geographic stratification). What follows is a brief overview of each stratification option.

Single Forest Cover Class – No stratification

All pixels with preexisting forest cover are included (i.e., where Forest Before REDD+ is non-zero.)

Multiple Forest Cover Classes

Quantile Classification – The study area is divided into (approximately) equally sized areas based on forest cover. As an example, a quantile classification with 4 classes may be divided into regions of 0-40% forest cover, 40-80% forest cover, 80-95% forest cover, and 95-100% forest cover. (Note: these regions do not overlap, and the first region does not contain areas with exactly 0% forest cover. More precise boundary values are 0.001%-40% for the first class, 40.001%-80% for the second class, etc.)

User-Defined Classification – The study area is divided into forest cover classes defined by the user. The user specifies the upper bounds for the classes. For example, if there are 3 classes, and the user defines upper bounds as 35%, 90%, and 100% forest cover, then the forest cover classes will be regions of 0-35% forest cover, 35.001% - 90% forest cover, and 90.001% - 100% forest cover.

Geographic Stratification – The study area is divided by geographic region, such as districts or provinces.

	Boolean	Fractional
Logistic – Single Coefficient	Yes	No
Logistic – Geographic Stratification	Yes	No
Poisson – Single Coefficient	No	Yes
Poisson – Quantile Classification	No	Yes
Poisson – User-Defined Classification	No	Yes
Poisson – Geographic Stratification	No	Yes
External Model – Single Coefficient	Yes	Yes
External Model – Quantile Classification	No	No
External Model – User-Defined Classification	No	Yes
External Model – Geographic Stratification	Yes	Yes

Effective Opportunity Cost Score

The Effective Opportunity Cost Score image is produced during the regression step of the Effective Opportunity Cost panel. There are no units for the image – it displays the relative probability of deforestation at each pixel. The score is a combination of the independent variable coefficients. The equation for the score is:

$$OC = \frac{B_0 + B_1X_1 + B_2X_2 + \dots + B_NX_N}{B_1}$$

where:

OC = Effective Opportunity Cost Score

B_0 = Constant term

B_1 = Agricultural revenue coefficient

B_i = Variable coefficients

X_1 = Agricultural revenue variable

X_i = Other independent variables (for $i = 2 \dots N$)

In other words, the effective opportunity cost score at a pixel is a linear combination of the constant term and the values of the independent variables at that pixel multiplied by their respective regression coefficients, all divided by the regression coefficient for agricultural revenue.

The score is used to calculate the fraction of a pixel which will be deforested.

Geographic Approach

The GeGeOSIRIS model can be administered at different sub-national levels, such as the district or provincial level. You can enter image files In the Administrative Levels table of the Input Image Files panel.

This section outlines how the geographic approach works, including the Province and District-Level decision maps in GeOSIRIS.

Calculating the Proportional Change in Agricultural Price

This section details how the final proportional change in agricultural price is calculated in the Output Parameters panel. This calculation uses an iterative loop and two input parameters, the Model Precision and the Maximum Number of Iterations.

The proportional change is calculated as the sum of endogenous change and exogenous change.

$$\text{Change in Agricultural Price} = \text{Endogenous Change} + \text{Exogenous Change}$$

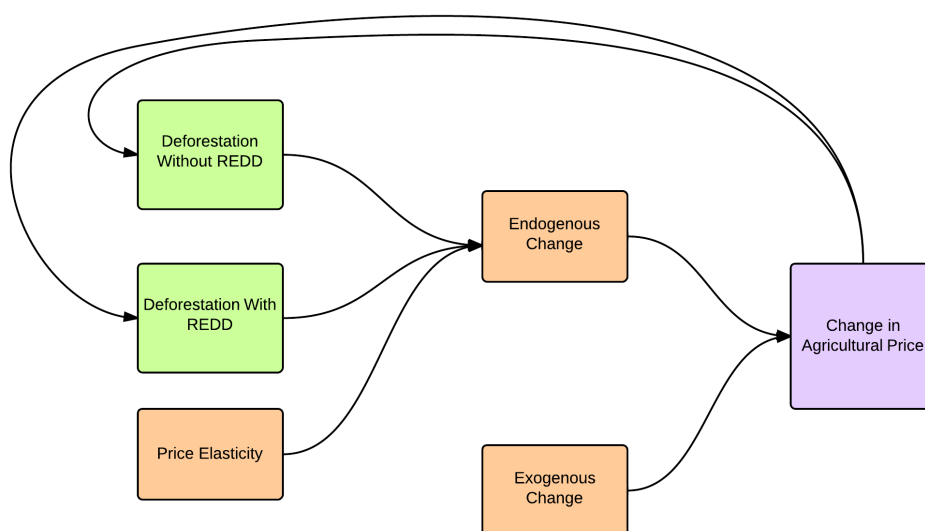
The exogenous change is entered as an input parameter in the Model Parameters panel. For each iteration of the model, the endogenous change is calculated based on the amount of deforestation with and without REDD, and the price elasticity.

$$\text{Endogenous Change} = \left(\frac{\text{Deforestation without REDD}}{\text{Deforestation with REDD}} \right)^e$$

where the exponent $e = \text{price elasticity}$.

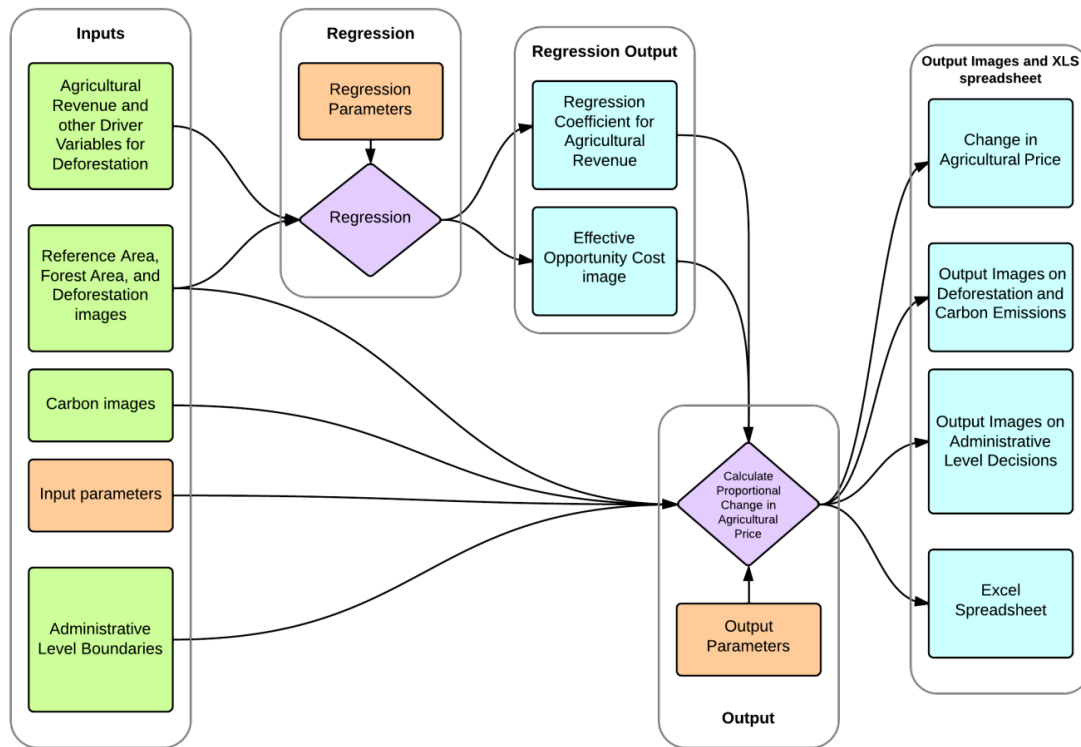
The model compares two successive values in the change in agricultural price to see if they are within the model precision value. If they are, then the most recent iteration's value is used for the final calculations. The model will continue to run until two successive values meet the Model Precision criteria, or the maximum number of Iterations is exceeded, in which case the model terminates without performing any final calculations.

The chart below shows the steps in this calculation.



Overall process

The overall GeGeOSIRIS model has two main steps, a regression step, where the regression coefficient(s) and Effective Opportunity Cost image are calculated, and the final output step, where the proportional national change in agricultural price and output images are calculated and the summary Excel spreadsheet is produced. You can see a flowchart of these steps in the diagram below. It is possible to revise model parameters after the regression or final output steps if you would like to vary the model.



List of GeGeOSIRIS Parameters and Images

The following is a list of all GeOSIRIS parameters and images.

Input Images

Reference Area: a Boolean image of the REDD+ project area, for Boolean image type. (units: Boolean)

Land Area: a continuous image of the land area per pixel, for Fractional image type. (units: hectares)

Forest Area Before REDD+: depicts where forest area existed at most recent time point, before the proposed REDD+ project. This is a Boolean map for the Boolean image type, and a continuous map for the Fractional image type. (units: Boolean for Boolean image type, hectares for Fractional image type)

Biomass Carbon: above and below ground biomass. (units: tons carbon per hectare)

Soil Carbon: carbon content in soil. (units: tons carbon per hectare)

Peat: carbon content in peat soil. (units: Boolean, 1 = presence and 0 = absence.)

Deforested Area: observed amount of deforestation within the previous 5 years. This is a Boolean map for the Boolean image type, and a continuous map for the Fractional image type. (units: Boolean for Boolean image type, hectares for Fractional image type)

Administrative Levels: Maps of sub-national administrative jurisdictions, such as districts or provinces. (units: jurisdictions should be numbered with positive integers. They do not need to be consecutive).

Agricultural Revenue: The agricultural revenue variable should represent the Net Present Value (NPV) of the agricultural revenue of land at that pixel for a specified time frame (e.g., 5 years), usually with a discount factor applied to represent reduced revenue in future years. (units: U.S. Dollars per hectare)

Other Independent Variables: Other potential driver variables for deforestation. Examples include elevation, slope, logarithmic distance to the nearest road or provincial capital, or the amount of area per pixel included in a national park, or a timber plantation. (units: varies based on driver variable)

Effective Opportunity Cost (External Model Only): displays the relative probability of deforestation at each pixel. This is a composite index of all driver variables affecting deforestation. (units: index units)

Input Parameters

World Carbon Price: Global carbon price for REDD+ project. (units: U.S. dollars)

Country national reference level as proportion of business-as-usual (BAU) emissions: Proportion of recent estimated emissions (based on observed deforestation). This parameter is used to create a baseline level of emissions which will determine carbon revenue and penalties.

Benefit Sharing: Proportion of carbon revenue shared by subnational entities with the national government.

Cost Sharing: Proportion of carbon penalties shared by subnational entities with the national government.

Price elasticity: This parameter represents how sensitive the domestic production price is to change in deforested area.

Exogenous increase in price of agriculture: This parameter represents how much the agricultural price on newly deforested land will increase due to factors outside of this particular project, such as the implementation of a global REDD+ mechanism.

Soil-carbon percent: percent of carbon contained in soil which will be emitted as carbon dioxide in the case of deforestation.

Emission value (peat soil): amount of carbon dioxide equivalent emitted from peat soil during deforestation. (units: tons CO₂ equivalent per hectare).

% of BAU: This parameter is used to calculate the reference level of emissions for each subnational jurisdiction, which is used to determine carbon payments/ penalties.

% of Emission Rate: the percent of the national emission rate to be used as the reference-level floor.

Proportion of potential agricultural revenue retained after production costs: the proportion of agricultural revenue retained after production costs are considered.

Regression sampling proportion: Proportion of pixels used in regression calculation.

Output Images and Parameters (Regression)

Effective Opportunity Cost: the relative probability of deforestation at each pixel. This is a composite index of all driver variables affecting deforestation. (units: index units)

The equation for the score is:

$$OC = \frac{B_0 + B_1X_1 + B_2X_2 + \dots + B_NX_N}{B_1}$$

where:

OC = Effective Opportunity Cost Score

B_0 = Constant term

B_1 = Agricultural revenue coefficient

B_i = Variable coefficients

X_1 = Agricultural revenue variable

X_i = Other independent variables (for $i = 2 \dots N$)

Regression Coefficients: the regression coefficients relate the net present value (NPV) of agricultural revenue with deforestation.

Output Images and Parameters (Output Step)

Carbon emissions with REDD: carbon emissions based on modeled deforestation over given period with REDD+ project in place. (units: tCO₂ / 5 years)

Carbon emissions without REDD estimated: estimated carbon emissions based on observed deforestation during a given time period. (units: tCO₂ / 5 years).

Carbon emissions without REDD modeled: carbon emissions based on modeled deforestation, without a REDD+ project in place. (units: tCO₂ / 5 years) (units: tCO₂ / 5 years).

Change in emissions due to REDD: change in emissions from modeled carbon emissions without REDD to modeled carbon emissions with REDD. (units: tCO₂ / 5 years).

Deforestation with REDD: modeled deforestation over give time period with REDD+ project in place. (units: ha / 5 years).

Deforestation without REDD (modeled): modeled deforestation over given time period without a REDD+ project in place. (units: ha / 5 years).

Effective opportunity cost score scaled for production costs: this image represents the effective opportunity cost score created during the Regression step, scaled for agriculture production costs. (units: index units)

Emission Factor: The total potential carbon dioxide that can be emitted from all sources: above / below ground carbon, soil carbon, and/or peat. (units: tCO₂ emitted / ha)

The equations for the emission factor at each pixel are:

$$E = (AB + SC * f_s) * 3.67 \text{ where } Peat = 0$$

$$E = AB * 3.67 + f_p \text{ where } Peat > 0$$

$$E = \text{emission factor} \left(\frac{tCO_2e}{ha} \right) \text{ (pixel level)}$$

AB = above and below ground carbon (pixel level)

SC = soil carbon (pixel level)

f_s = soil carbon factor (input parameter)

f_p = emission factor for peat soil (input parameter)

Emittable forest carbon in parcel before REDD: Atmospheric CO₂ equivalent of total above-ground, below-ground, and soil biomass, which is susceptible to release during deforestation. (units: tCO₂)

The equation for the emittable forest carbon is:

$$FC = FA * EF$$

FC = emittable forest carbon

FA = Forest Area before REDD

EF = Emission Factor

Fraction deforested without REDD: fraction of forest area modeled to be deforested without a REDD+ project in place. (units: fractional units)

Raw agricultural revenue scaled for production costs: Net-present value (NPV) of agricultural revenue at a pixel, scaled to include production costs. (units: \$/ha)

Site level baseline: The site-level (pixel-level) baseline of carbon emissions against which carbon increases or decreases are measured. Increases will result in carbon penalties while decreases will result in carbon revenue for the participating nation. (units: tCO₂ / 5 years).

The site-level baseline is calculated based on parameters for the particular administrative level applied (e.g., district). The baseline is the maximum of two values: the modeled business-as-usual emissions, (potentially multiplied by a scaling factor) and a baseline floor, which is calculated as a proportion of the total emittable forest carbon stock.

The equation for the site level baseline is:

$$B = \max(CE_m * BAU_a, F * E * f_a)$$

B = site – level baseline (units: tCO₂ / 5 years)

CE_m = carbon emissions without REDD (modeled) (units: tCO₂ / 5 years)

BAU_a = % of BAU for the administrative level

F = forest before REDD (units: ha)

E = emission factor (units: tCO₂ / ha)

f_a = administrative – level floor as proportion of total forest carbon stock

Would be parcel deforestation if jurisdiction opts into REDD: amount of deforestation if jurisdiction opts into REDD. (units: ha / 5 years)

Would be parcel deforestation if jurisdiction opts out of REDD: amount of deforestation if jurisdiction opts out of REDD. (units: ha / 5 years)

Would be parcel emissions if jurisdiction opts into REDD: amount of carbon emissions if jurisdiction opts into REDD. (units: tCO₂ emitted / 5 years)

Would be parcel emissions if jurisdiction opts out of REDD: amount of carbon emissions if jurisdiction opts out of REDD. (units: tCO₂ emitted / 5 years)

Would be relative loss from foregone agriculture at parcel: opportunity cost of forest land that is not deforested and converted to agriculture due to a REDD+ project (units: \$, net present value)

District/Province Level Decisions: Decision of a district or province to participate in a REDD+ project. (1 = opt-in, 0 = opt-out)

Proportional Change in Agricultural Price: proportional change in agricultural price if a REDD+ project is in place

<http://www.un-redd.org/AboutREDD/tabid/102614/Default.aspx>

<http://www.un-redd.org/AboutREDD/tabid/102614/Default.aspx> (question 6)

Busch, Jonah, et al. "Structuring economic incentives to reduce emissions from deforestation within Indonesia." Proceedings of the National Academy of Sciences 109.4 (2012): 1062-1067.

CHAPTER NINE

ECOSYSTEM SERVICES MODELER

The Ecosystem Services Modeler (ESM) contains 15 ecosystem service models that are closely based on the InVEST toolkit developed by the Natural Capital Project - a partnership between the Wood's Institute for the Environment at Stanford University, The Nature Conservancy, the World Wildlife Fund and the Institute on the Environment at the University of Minnesota. In a few instances, we have modified the InVEST models when the underlying procedures in the Clark Labs' software offer added value. In all cases, however, the fundamental spirit and algorithmic procedures developed by the Natural Capital Project have been maintained.

While the InVEST documentation available on-line from the Natural Capital Project is equally relevant to the implementation in ESM, it should be noted that there may be version differences between the two implementations. As well, the version of the InVEST model implemented in TerrSet may not be the latest version documented by InVEST. The 15 models include:

1. Water Yield
2. Hydropower
3. Water Purification
4. Sediment Retention
5. Carbon Storage and Sequestration
6. Timber Harvest
7. Habitat Quality and Rarity
8. Crop Pollination
9. Habitat Risk Assessment
10. Offshore Wind Energy
11. Aesthetic Quality
12. Overlapping Use
13. Coastal Vulnerability
14. Marine Aquaculture
15. Wave Energy

Water Yield

Water Yield model measures the average annual runoff, or water yield, at the watershed, subwatershed and the pixel level. The water yield output images are required to run the Hydropower and the Water Purification models (also found within ESM).

Average annual surface runoff is calculated as:

$$Y_{xj} = \left(1 - \frac{AET_{xj}}{P_x}\right) \times P_x$$

where AET_{xj} is the annual actual evapotranspiration on pixel x within land cover type j , and P_x is the average annual precipitation on pixel x . The fraction AET_{xj}/P_x is an approximation of the Budyko curve. It represents the evapotranspiration partition of the water balance. This fraction is determined by a dryness index R_{xj} and a dimensionless parameter, ω_x , that denotes the ratio of plant available water content to expected precipitation during the year. R_{xj} is determined by the ratio of the product of vegetation evapotranspiration and reference evapotranspiration to the annual precipitation.

Global maps of yearly evapotranspiration rates are available from the FAO-UN GeoNetwork:

(www.fao.org/geonetwork/srv/en/main.home) website. Otherwise, potential evapotranspiration can be calculated using precipitation and temperature data and the modified Hargreaves equation.¹ Additionally, there is often regional or local data available for potential evapotranspiration, particularly in the United States.

Average annual precipitation for a study area may be interpolated using data from local climate stations. Otherwise, global data is freely available through the Global Precipitation Climatology Project (<http://cics.umd.edu/~yin/GPCP/main.html>) and the Tropical Rainfall Measuring Mission (<http://trmm.gsfc.nasa.gov/>). Note that snow precipitation should be included in this measure.

Other data necessary to calculate water yield include soil depth (mm) and plant available water fraction ($PAWF$). $PAWF$ is defined as the proportion of water holding capacity (the amount of water held between soil field capacity and wilting point) that can be absorbed by a plant – this is also known as the Available Water Content. Generally, plant available water is considered to be 50% of the soil water holding capacity. Global soil depth data is available through NOAA and the FAO. Additionally, free soil data is available for the United States from the USDA NRCS Soil Data Mart – this is an extensive resource containing geospatial information for many soil properties. Alternatively, NASA's Global Change Master Directory website contains a host of datasets, including regional and global soil water holding capacity.

Outputs

The Water Yield model produces three outputs. The three images produced all measure average annual water yield (mm), but at different scales: pixel-level, subwatershed, and watershed. Water yield is calculated at these three different scales as follows: if the scale is at the pixel level, divide each pixel value by the area of the pixel to get the standard mm/m² units of water yield. If the scale is at the subwatershed or watershed level, divide by the area of each subwatershed or watershed.

Hydropower

The Hydropower Model is used to estimate the potential energy production from a reservoir based on its water yield and dam practices. There is the option of assessing the economic value of this energy based on consumptive demand, energy prices and the cost of dam maintenance. The Hydropower model works in three stages. First, the model calculates water scarcity based on water

¹ Droogers, P. and R. Allen. (2002). Estimating reference evapotranspiration under inaccurate data conditions. *Irrigation and Drainage Systems* 16: 33-45.

yield and consumptive use. Next, the model estimates the amount of energy produced by a supply of water. Finally, the value of that energy is calculated.

The Hydropower model first calculates water scarcity; this accounts for water yield and water consumption in each watershed. Note that this value only accounts for consumptive demand for water (water not returned to the water resource system). The Hydropower model requires a water yield image. This can be produced within ESM's Water Yield model. For more information about how this image is produced and what it represents see the Water Yield section. Each land cover type is assumed to consume a certain amount of water (e.g., "residential" water use = population \times per capita water use). These values may be obtained from agricultural, forestry, or hydrology journals or from local experts. The demand or use of water is then summed along flow paths throughout the watershed, and the amount of water that reaches the reservoir (or dam) is defined as the difference between the water yield previously calculated and the total volume of water consumed in the watershed upstream of the reservoir.

Lastly, the model estimates the amount and value of energy produced given the net supply of water from the watershed. The energy produced is then redistributed to indicate the proportion contributed by each subwatershed. Power produced by a dam or reservoir is calculated as the product of water density, flow rate, water height behind the dam at the turbine, and the gravity constant. This model relies on the assumption that water flow is released at a steady rate continuously throughout the year. The annual inflow calculated in the previous step is converted to a per second rate, and the resulting power (in watts) is multiplied by the number of hours in a year to produce the kilowatt-hour value for electric energy.

The annual average energy production from a hydropower facility is calculated as the product of turbine efficiency, the proportion of inflow water designated for hydropower production, the annual average inflow of water and the height of water behind the dam at the turbine. The annual power produced by a dam is then converted to the net present value of hydroelectricity, accounting for market value of electricity, annual average energy production, total annual operating costs and the annual market discount rate for the remainder of the lifetime of the dam. The net value is then redistributed so the user may see how much energy and revenue is generated by each subwatershed, according to the proportion of water supply generated from each one.

Outputs

The Hydropower model yields six possible outputs. The two parameters measured, water scarcity (m^3) (total water consumptive use) and water value (USD or other currency) (total environmental service provided, in economic terms), are calculated at three different scales: pixel level, subwatershed level, and watershed level.

Water Purification: Nutrient Retention

Water purification nutrient retention refers to the ability of soil and vegetation to retain nutrients imported from upstream watersheds. The Water Purification model calculates the natural retention for nitrogen (N) and phosphorous (P) pollutants. The model works in three steps: (1) it calculates the average annual runoff (or water yield) from each pixel in the landscape, (2) it estimates the amount of nutrient pollutants retained by each vegetated surface in the landscape and then aggregates the results to subwatershed and watershed scales, and (3) it determines the value of the service provided by each subwatershed based on the avoided cost of water treatment.

Water purification begins by estimating the amount of pollutant that is exported from each pixel. The amount of the pollutant export is based on export coefficients, which represent the quantity of nutrient or sediment generated per unit area per unit time ($\text{g ha}^{-1} \text{yr}^{-1}$). Use of the export coefficients assumes that a given land use activity will yield a specific quantity of nutrients or sediment to a

downstream reservoir. Comprehensive lists of export coefficients have been compiled by the US EPA and in Reckhow et al. (1980)² and Lin (2004)³.

The export coefficient for a given land use type is multiplied by an export adjustment factor, or the ratio of runoff (water yield) for a given land use type to the average runoff in the watershed of interest. The adjustment factor accounts for land use types that display above or below average runoff values. For example, pasture and agricultural fields tend to have larger than average runoff values in comparison with surrounding forested land cover types. Multiplying the export coefficient by the export adjustment factor results in an Adjusted Loading Value (ALV) for every pixel.

Once the export load or ALV is calculated, it is possible to estimate the amount of load retained by the downstream parcels. Water flows through the landscape according to the slope and aspect of land, and a parcel of land (pixel) is able to retain some of the nutrients in the water based on the presence of vegetation and its ability to hold pollutants. The direction of flow is calculated with the FLOW module in TerrSet. From there, the amount of cumulative pollutant input and export is calculated using the ALV, and nutrient retention is calculated as a function of root depth and retention capacity by vegetation. This model does not account for the saturation of nutrient uptake.

The valuation of services provided by the natural landscape must account for an amount of nutrient loading that is permissible in a reservoir. Often there is some threshold or water quality standard that, once passed, specifies a need for water treatment. Only if this critical threshold is reached does the model calculate the economic value of nutrient retention services based on the avoided cost of treatment. The value of retention services for subwatershed w is:

$$\text{Value}_w = \text{cost}(p) \times \text{retained}_w \times \sum_{t=0}^{T-1} \frac{1}{(1+r)^t}$$

Where $\text{cost}(p)$ is the annual treatment cost in \$/kg for the pollutant of interest p , retained_w is the total pollutant retained by subwatershed w , T is the time span of the model (typically the remaining lifetime of a reservoir), and r is the annual market discount rate used to calculate net present value.

Outputs

There are nine possible outputs from the Water Purification model. Three of those outputs depict water yield at multiple scales, four express nutrient retention and export, and two contain economic valuation information.

▪ **Water Yield (mm)**

Maps of average annual water yield at the pixel, subwatershed and watershed scales. If the scale is on a pixel level, divide each pixel value by the area of the pixel to get the standard mm/m² units of water yield. If the scale is at the subwatershed/watershed level, divide by the area of each subwatershed/watershed.

▪ **Nutrient Retention (kg)**

The total amount of nutrient (nitrogen, phosphorus, or both) retained (kg).

▪ **Nutrient Export (kg)**

The total amount of nutrient (nitrogen, phosphorus, or both) exported (kg) to streams.

² Reckhow, K.H., M.N. Beaulac, and J.T. Simpson. 1980. Modeling Phosphorus Loading and Lake Response under Uncertainty: A Manual and Compilation of Export Coefficients, East Lansing, Michigan, Michigan State University.

³ Lin, J. 2004. Review of Published Export Coefficient and Event Mean Concentration (EMC) Data. Westlands Regulatory Assistance Program.

▪ Valuation

The value of avoided cost of water treatment due to nutrient retention (nitrogen, phosphorus, or both) at a subwatershed scale.

Sediment Retention

The Sediment Retention model estimates a watershed's ability to retain sediment based on geophysical, climatological and ecological factors. The economic value is calculated as the avoided cost of sediment removal. The basic sequence is as follows: the model calculates the mean annual soil loss from each pixel in a watershed, determines the amount of that soil that may arrive at a particular cell, calculates the retention of sediment, and finally assesses the annual cost of removing the accumulated sediment. Erosion and sedimentation rates are influenced primarily by four factors: precipitation rates, soil structure, vegetation cover, and topography. This model accounts for two major relationships:

High rainfall-intensity results in a high risk of erosion and transportation, or high erosivity.

A high proportion of sand in soils results in a high risk of erosion and transportation, or high erodibility.

The long-term, mean annual rate of erosion (A) is calculated using the Universal Soil Loss Equation (USLE):

$$A = R \cdot K \cdot LS \cdot C \cdot P$$

The USLE is the product of R , the rainfall erosivity factor, K , the soil erodibility factor, LS , the slope-length factor, C , the crop/vegetation and management factor, and P , the support practice factor. A is the soil loss in tons per year, and is compared to a reference soil loss value that is considered to be "tolerable" soil loss.

The rainfall erosivity factor, R , represents a measure of the erosive force of rain in a normal precipitation year, and is based on the kinetic energy of falling rain. Two components make up the R -factor: the total storm energy and the maximum 30 minute storm intensity. The R -factor is the sum of the product of two components for all major storms in an area during an average year.⁴ The maximum 30 minute storm intensity is known as the annual Erosion Index (EI_{30}). Both components require a continuous record of rainfall intensity for calculation. While localized R values are difficult to find, regional R values have been determined through quantitative field measurements for many areas of the world and may be found in scientific literature. Depending on the size of the study region, it is likely that the landscape will have only one R value.

The soil erodibility factor, K , is a measure of the susceptibility of soil particles to detach and be transported by rainfall or runoff. The magnitude of K depends on the texture, structure, permeability, and organic matter content of soil. K ranges from 0.7 (for the most fragile, erodible soil) to 0.1 (for highly stable soil), and reflects the inherent differences in soil erosion rates when all external forces (infiltration rate, permeability, water capacity, rain splash, abrasion, etc.) are held constant. This factor may be calculated according to the equation:

$$K = 1.292 \times [2.1E^{-6}(12 - P_{om}) f_p + 0.0325(S_{struc} - 2) + 0.025(f_{perm} - 3)]$$

$$\text{and} \quad f_p = P_{silt}(100 - P_{clay})^5$$

⁴ For more information about R , see <http://www.fao.org/docrep/t1765e/t1765e0e.htm>

⁵ For information about K , see http://mepas.pnnl.gov/mepas/formulations/source_term/5_0/5_32/5_32.html

where P_{om} is the percent organic matter, f_p is the particle size parameter, S_{struc} is the soil structure index, and f_{perm} is the profile-permeability factor. An alternative equation is offered in Wischmeier and Smith (1978)⁶. If K cannot be directly measured for a study area, several K -factor tables have been developed and published^{7,8} and may be used for general analysis.

Slope Length (LS) is the distance that a drop of rainfall or sediment will flow before it loses energy. In general, the steeper the slope of a field, the greater the amount of soil loss from erosion by runoff and rainfall. Therefore, larger values of LS indicate more erosion and sediment accumulation. This is important for development planning. For example, consolidation of many small fields into one large field will increase the length of LS , permitting greater erosion potential. In this model, a slope threshold value (%) is defined such that LS is calculated separately for low slopes (slope % < threshold %) and high slopes (slope % > threshold %). The slope threshold represents the slope at which farmers begin using specialized practices, such as terracing, to account for the steepness of the slope. The default slope threshold is 75%, but the user may adjust this according to the study region.

For low slopes:

$$LS = \left(\frac{flowacc \cdot pixel\ size}{22.13} \right)^{nn} \left(\frac{\sin(slope \cdot 0.01745)}{0.09} \right)^{1.4} \times 1.16$$

where $flowacc$ is the accumulated water flow to each pixel, determined by the FLOW module, $pixel\ size$ is the spatial resolution, and nn ranges from 0.2 to 0.5, depending on the slope.

$$LS = 0.08\lambda^{0.35} \cdot slope^{0.6} \quad \text{and} \quad \lambda =$$

For high slopes:

$$pixel\ size \quad \text{if} \quad flowdir = 1, 4, 16 \text{ or } 64$$

$$1.4 \cdot pixel\ size \quad \text{if} \quad flowdir = \text{other}$$

where $flowdir$ is the flow direction within the pixel, determined by the FLOW module.

The crop/vegetation and management factor, C , represents the relative effectiveness of soil and crop management systems in terms of preventing soil loss. It includes the effects of vegetation cover, crop sequence, productivity level, length of growing season, tillage practices, residue management, and the expected time distribution of erosive rainstorms. Values for C may be found in scientific literature (see references 11 and 12 on previous page), and online at: http://topsoil.nserl.purdue.edu/usle/AH_537.pdf (U.S. DOA).

Finally, the support practice factor P (also called the conservation practice factor) is the ratio of soil loss with a specific support practice to soil loss with straight -up-and-down slope farming. For example, contouring is a practice which involves field operations such as plowing, planting, cultivating, and harvesting, approximately on the contour following the natural curves of a landscape. Terracing is another method that slows the runoff of water and thus reduces the amount of eroded topsoil. In general, P is higher on steeper slopes.

The ability of vegetation to retain sediment in a given pixel is calculated by subtracting the USLE estimated annual rate of erosion (A) from the erosion that would occur if the pixel were classified as bare soil, with no vegetation. The bare soil estimate (BSE) is calculated as the product of R , K , and LS above, removing crop and practice support factors from the USLE.

Note that the USLE itself does not account for sediment that has eroded from areas upstream, which may be trapped by vegetation in downstream pixels. In the Sediment Retention model, the total sediment retained by pixel x ($sret_x$) is the sum of the sediment

⁶ Wischmeier, W.H and D.D. Smith. 1978. Predicting rainfall erosion losses – a guide to conservation planning. U.S. Department of Agriculture, Agriculture Handbook No. 537.

⁷ Stone, R.P. and D. Hilborn. 2000. FACTSHEET: Universal Soil Loss Equation (USLE). Ministry of Agriculture, Food and Rural Affairs.

⁸ Goldman, S, K. Jackson, R. Bursztynsky. Erosion and Sediment Control Handbook. March 1986. McGraw-Hill, Texas.

removed from flow by the pixel itself and the sediment removed along the hydrologic flowpath. The cumulative sediment retention is determined based on flow direction within the watershed, which is calculated using the FLOW module in TerrSet with a DEM input.

There are two services associated with sediment retention in a watershed: water quality improvement and avoided sedimentation of reservoirs (dredging). In ESM, the user can calculate both of these outputs, and estimate the economic value of each service. The services provided to water quality depend on the allowed annual load of sediment to a water body, which is typically specified by national or local government. A water body is not considered to receive a benefit if the annual load is less than the allowed annual load. Assuming each pixel in the scene received an equal proportion of the allowable load,

$$\text{sed_ret_wq}_x = \text{sret}_x - \frac{\text{wq_annload}}{\text{num_pixels}}$$

where sret_x is the total sediment retained by pixel x , wq_annload is the annual allowed sediment load and num_pixels is the total number of pixels in the watershed (all of them contribute to either sediment retention or soil loss).

The avoided cost of sediment removal in reservoirs depends on the *dead storage capacity*, which refers to the amount of space at the bottom of the reservoir that is engineered to allow sediments to settle. This space is constructed specifically for water quality control and to avoid the costs of dredging, so a reservoir does not receive benefits from a watershed's ability to retain sediment until the dead storage is filled. The service associated with dredging is calculated as:

$$\text{sed_ret_dr}_x = \text{sret}_x - \frac{\text{dr_deadvol} \times 1.26}{\text{dr_time} \times \text{num_pixels}}$$

where dr_deadvol is the engineered dead storage volume of the reservoir, dr_time is the remaining lifetime of the reservoir and 1.26 (tons m^{-3}) is a constant that represents sediment density.

The model sums and averages the sediment export and retention per pixel to the subwatershed level, and provides separate outputs for retention services for water quality, retention services for dredging, and total sediment export. The final step is the economic valuation for avoided costs of water treatment and dredging. The following equation is used to calculate the value of each subwatershed based on services provided to the reservoir:

$$\text{sed_Value}_1 = \text{cost}(s) \times \text{sret_sum} \times \sum_{t=0}^{T-1} \frac{1}{(1+r)^t}$$

where sed_Value_s is the value of sediment retention for subwatershed s over T years, T is either the period of time for which the land cover is constant (when valuing water quality services) or the length of the reservoir life (when valuing dredging services), sret_sum is the total sediment retention for either water quality (sret_sum_wq) or dredging (sret_sum_dr), $\text{cost}(s)$ is the marginal cost per m^3 of sediment removal for water quality treatment or dredging, and r is the annual market discount rate. The $\text{cost}(s)$ ($\$ \text{m}^{-3}$) value depends on the methods used for sediment removal, which are likely to vary across reservoirs. Additionally, reservoir lifetime (dr_time), constant land cover duration (wq_time), dead storage capacity (dr_deadvol), and the annual sediment load (wq_annload) are all variables that are specific to the reservoir or watershed, so the user must input each variable separately for each watershed.

There are some limitations to the USLE approach. The USLE is only capable of predicting erosion from sheet wash, meaning erosion from plains with gentle slopes. Therefore, it should only be applied in areas with slopes between 1 and 20%. Additionally, the relationship between precipitation and kinetic energy is not straightforward in mountainous regions. The model is also sensitive to the way in which land cover and land use types are grouped, so it is important to consider carefully the focus of the project and the land cover types that are being modeled.

Outputs

There are nine outputs from the Sediment Retention model. The first five estimate physical values of sediment retention and export and include the following:

- ***Sediment Export***

Total sediment export, in metric tons, for every pixel in the landscape

- ***Sediment Retention for Water Quality***

Sediment retention, in tons, adjusted for the maximum allowable sediment threshold for water quality, at both the pixel and subwatershed scales. This is a biophysical expression of sediment retention services provided by the landscape.

- ***Sediment Retention for Dredging***

Sediment retention, in tons, adjusted for the designated reservoir dead volume, at both the pixel and subwatershed scales. This is a biophysical expression of sediment retention services provided by the landscape.

The remaining outputs are calculated in the Valuation Panel:

- ***Water Quality Regulation Valuation***

The economic benefit of sediment retention, in USD or other currency, in terms of avoided cost of water treatment. This is summarized on a subwatershed and a watershed level.

- ***Avoided Dredging Valuation***

The economic benefit of sediment retention, in USD or other currency, in terms of avoided cost of dredging. This is summarized on a subwatershed and a watershed level.

Carbon Storage and Sequestration

The Carbon Storage and Sequestration model uses current and future land cover maps and current data on wood harvest rates, harvest product degradation rates, and stocks in four carbon pools to estimate the amount of carbon currently being stored by a particular landscape or the amount sequestered over time.

Terrestrial ecosystems currently store four times more carbon than the atmosphere and are extremely important in moderating carbon-driven anthropogenic climate change. However, land cover change due to timber harvesting, deforestation or fire may alter the amount of carbon a landscape is capable of storing. Therefore, land management practices are essential in the maintenance and sustainability of terrestrial carbon pools.

The amount of carbon stored in a land parcel depends on the size of four primary carbon sinks: aboveground biomass, belowground biomass, soil organic matter, and dead organic matter. Aboveground biomass is all the living biomass above the soil including the stem, stump, branches, bark, seeds and foliage⁹. Belowground biomass, as you might guess, is the living biomass of roots underground. Soil organic matter is the largest carbon pool on land, and is made up of the living biomass of micro-organisms, fresh and partially decomposed residues, and humus. Dead organic matter, or detritus, is typically made up of leaf litter and other organic matter mixed with soil. A fifth, optional carbon pool is the harvested wood products (HWP) pool, specified by parcel, which includes the amount of carbon stored in the wood that is harvested from a forested land parcel, the frequency of harvest and the decay rate of these wood products. It is assumed that the carbon stored in HWP is released as the product decays.

⁹ 2003 Good Practice Guidance for Land Use, Land-Use Change and Forestry. Glossary. IPCC. J. Penman et. al. eds.

This model uses a simplified carbon cycle to estimate the net amount of carbon stored in a region, the total biomass removed from a forested and harvested area, and the economic value of the carbon sequestered in the remaining carbon pools. Note the difference between the amount of carbon *stored* in a land parcel, and the amount *sequestered*. Sequestration indicates long term storage of carbon in a carbon pool. Market prices currently relate only to sequestration over time based on the price per metric ton of carbon traded in the carbon market. Outputs of the model are in Mg of carbon per pixel and the value of carbon sequestered in dollars. Due to the temporal aspect of carbon sequestration, economic valuation can only be estimated if a future land cover image is available.

Land cover maps and the amount of carbon stored in each carbon pool are used in the analysis. An estimate of the amount of carbon stored in the four fundamental carbon pools is required for each land cover type. This estimate is then applied across the entire landscape to produce a map of carbon stored in the study region. If possible, include elevation classes, multiple disturbance areas, or landscape dynamics. For example, if the perimeters of agricultural fields in the study region are regularly burned for maintenance, the amount of carbon stored in the aboveground biomass and dead organic matter pools should reflect this portion of the fields not storing carbon.

It is not necessary to have information for all carbon pools. If data for only one or two of the four pools is provided, the model will run but the amount of carbon stored in the landscape will be underestimated. The HWP pool is optional because not all landscapes include harvested timber land. In addition, studies have shown that wood products and landfills account for less than 7% of forest-related carbon stocks,¹⁰ so HWP may be considered negligible in its influence on carbon sequestration. Of course, the proportion of carbon stored in harvested wood products varies across landscapes and land use practices.

The unit for all carbon pools in each land cover type is MgC ha⁻¹. Often, users are able to obtain CO₂ data and must convert this information to C by multiplying MgCO₂ha⁻¹ by 0.2727 (or $\frac{3}{11}$). Carbon pool data may be obtained from sample plot field measurements or from existing forest inventory data that is not more than five years old.¹¹ If field data is not available, additional resources and suggested methods of carbon pool calculations can be found in the 2006 IPCC Guidelines report.¹²

If the HWP option is used, the user must specify a raster group file containing information about the amount of carbon removed from the parcel over the course of a harvest period, the frequency of harvest periods per year, the average decay rate of the products made from the harvested wood, removed biomass conversion expansion factor (allows for the conversion of values of mass of harvested wood into volume), the start year of harvesting, and the average carbon density of the wood removed. Each of these variables must be input in the form of raster images with a specific naming convention (see Help for the model) so that the information is spatially explicit.

The amount of carbon removed (Mg ha⁻¹) from a harvested plot of land is averaged over a typical harvest period, and can be estimated from surveys, market analyses, or expert opinion. These values apply only to the wood that is physically removed from the parcel, and does not include the slash and vegetation waste left behind. For a more accurate model, the user might estimate the amount of carbon released by the slash that is left to decay in the forest and use it to calculate the *net* carbon removed during timber harvest. However, these values are difficult to calculate, and a more general model output will suffice if the user is aware of its limits. Research in this field continues to progress; to find estimates of carbon per unit roundwood or other harvested wood simply search peer-reviewed journals or IPCC publications. The USDA Forest Service also releases technical reports with estimates of carbon per unit roundwood by region in the U.S. and other parameters required for carbon modeling.

The average decay rate of a wood product is calculated using the half-life of the product, in years. Half-life estimates of wood products must be obtained from previous literature or data collection. The *2006 IPCC Guidelines for National Greenhouse Gas Inventories*¹² report includes a chapter on HWP including methods for calculating carbon stored in hard wood and paper products, as

¹⁰ Birdsey, R. and M. Cannell. 1992. Carbon in U.S. Forests and Wood Products, 1987-1997: State-by-State Estimates. USDA Forest Service, Northeast Research Station and U.S. Environmental Protection Agency, General Technical Report.

¹¹ REDD Methodological Module. Estimation of carbon stocks and changes in carbon stocks in the above-ground biomass carbon pool. Avoided Deforestation Partners. April 2009.

¹² 2006 IPCC Guidelines for National Greenhouse Gas Inventories: Agriculture, Forestry and Other Land Use. Volume 4. S. Eggleston, et al. eds.

well as half-life and decay rate estimates. If the harvested wood is used for multiple products, the user must use an average half-life and decay rate or identify on the slowest decay rate of all the products.

The average tree volume per ton of wood removed is calculated using the logic applied in the Timber Harvest model. The user must include a Biomass Conversion Expansion Factor (*BCEF*), which converts the biomass of harvested wood into a volume amount, or vice versa. See the Timber Harvest model section for description and data sources. The biomass of harvested wood is calculated using the average carbon density or carbon fraction of wood removed from the harvested parcel. This represents the average amount of carbon, in Mg, in one Mg of dry wood. Estimated carbon density values may be found in Table 4.3 of the *2006 IPCC Guidelines for National Greenhouse Gas Inventories*¹². If no values seem suitable, use a carbon density of 0.5.

If the user includes both a current and a future land use map, the model will estimate the net change in carbon stock over time, which may result in either sequestration or loss (release) to the atmosphere. The output will be the total amount of carbon sequestered over the model run time and the economic value of the carbon sequestered. Economic valuation depends on the current market price of carbon per metric ton. Here the user may use the actual market price according to the carbon market, but these values change frequently due to policies, subsidies, and other legislative influences. For this reason, the Social Cost of Carbon (SCC) may be a better estimate of carbon price. The SCC estimates the benefit society will gain by avoiding the damage caused by each additional metric ton of CO₂ released into the atmosphere.¹³ Estimates for SCC may be found in Nordhaus 2011¹⁴, Tol 2008¹⁵, and Greenstone et al. 2011¹⁶, under different market scenarios. For the purposes of this model, any SCC values extracted from literature must be converted from cost of CO₂ to cost of carbon.

One limitation of this model lies in the assumption that any change in carbon storage is due to abrupt changes in land cover type. In reality, many landscapes are slowly recovering from disturbances or are experiencing natural progression, which is not strictly accounted for in the model. This means that if no land use change is captured in the current and future land use scenarios, the net sequestration will be zero.

Outputs

There are five potential raster outputs from this model estimating carbon storage, carbon sequestration, and economic value of carbon.

▪ **Current/Future Stored Carbon (Mg)**

The amount of carbon, in Mg, stored in each pixel across the landscape. This value is the sum of carbon stored in each pool included in the model.

▪ **Sequestered Carbon (Mg)**

The total sequestered carbon for each pixel based on the difference in stored carbon from current to future land use scenarios. Positive values indicate sequestered carbon, while negative values indicate carbon lost to the atmosphere. Areas with large negative or positive values will have the largest changes in land cover or land use, or drastic changes in harvest characteristics.

▪ **Value of Currently Stored Carbon**

¹³ Bell, Ruth G. 2011. The Social Cost of Carbon and Climate Change Policy. World Resources Institute.

¹⁴ Nordhaus, W. 2011. Estimates of the Social Cost of Carbon: Background and Results from the RICE-2011 Model. Cowles Foundation Discussion Paper No. 1826. Yale University.

¹⁵ Richard S. J. Tol (2008). The Social Cost of Carbon: Trends, Outliers and Catastrophes. *Economics: The Open-Access, Open-Assessment E-Journal*, Vol. 2, 2008-25. <http://dx.doi.org/10.5018/economics-ejournal.ja.2008-25>.

¹⁶ Greenstone, M, E. Kopits, and A. Wolverton. 2011. Estimating the Social Cost of Carbon for Use in U.S. Federal Rulemakings: A Summary and Interpretation. MIT CEEPR WP 2011-006.

The market price of carbon, in US dollars or other currency, of all the carbon stored in each pixel across the landscape for the current land use.

▪ **Value of Sequestered Carbon**

The economic value, in U.S. dollars or other currency, of carbon sequestered from the current to the future land use scenarios. Each pixel has a dollar value. Negative values indicate the cost of carbon emissions, while positive values correspond to the benefit gained from sequestration.

If the Harvest Rate option is used, the module also produces the following outputs:

Bio_HWP_cur and Bio_HWP_fut

Represent the biomass of wood removed since the start date of harvest, for current and future land cover.

C_HWP_cur and C_HWP_fut

Represent the amount of carbon stored in the harvested wood products for current and future land cover.

Vol_HWO_cur and Vol_HWP_fut

Represent the volume of wood removed since the start date of harvest for the current and future land cover.

Timber Harvest

The Timber Harvest model maps the ecosystem services provided by the legally recognized harvest of roundwood¹⁷ timber from forest parcels. Parcels may include a whole forest or a part of a forest, but they should only include managed areas. The model produces three outputs in vector format: the total biomass, total volume, and the net present value (NPV) of roundwood timber harvested from each parcel over the user-defined time period.

The total biomass ($T_{biomass}$) of timber for forest parcel x is measured in tons and estimated as:

$$T_{biomass} = Parcel_{area_x} \cdot \frac{Perc_{harv_x}}{100} \cdot Harv_{mass_x} \cdot \frac{T_x}{Freq_{harv_x}}$$

where $Parcel_{area_x}$ is the area of forest parcel x in hectares, $Perc_{harv_x}$ is the percent of forest parcel area that is harvested, and $Harv_{mass_x}$ the mass of timber harvested per hectare. T_x is the number of years for which the model will run, and $Freq_{harv_x}$ is the frequency, in years, of harvest periods.

The total volume (T_{volume}) of timber for forest parcel x is based on the Biomass Conversion and Expansion Factor (BCEF), which converts the mass of dry harvested timber into a volume amount:

$$T_{volume} = T_{biomass} \cdot \frac{1}{BCEF_x}$$

BCEF (Mg dry wood per m³ wood) is a function of tree species and stand age, and is equal to the ratio of above-ground biomass to growing stock, which is the volume of all living trees in a given area of forest or wooded land that have more than a certain diameter

¹⁷ Roundwood is defined in the FAO Forest Harvesting Glossary as: "Wood in its natural state as felled, with or without bark. It may be round, split, roughly squared or in other forms."

at breast height.¹⁸ If BCEF data is not readily available, the user may find suitable values in Appendix 3a.1 in the IPCC in 2003⁹ report. BCEF values are also listed in Table 4.5 in the *IPCC 2006*¹² report. The USDA report on U.S. timber production also provides useful information.¹⁹ Otherwise, use a BCEF value of 1 for each parcel.

The economic valuation is done in two steps. First, the monetary value per hectare, V_x , for parcel x is calculated:

$$V_x = \frac{\text{Perc_harv}_x}{100} \cdot (\text{Price}_x \cdot \text{Harv_mass}_x - \text{Harv_cost}_x)$$

where Price_x is the marketplace value of harvested roundwood from each parcel, and Harv_cost_x is the cost per hectare of harvesting Harv_mass . Price reflects the price paid to harvesters at mills or other timber processing sites, and does not account for harvest or maintenance cost.

The second step is the calculation of the net present value (NPV) which includes the sum of V_x over all the harvest periods, the parcel maintenance cost and the market discount rate r . Parcel maintenance cost is the annual cost per hectare of maintaining the timber parcel, which may include the cost to replant or care for the stand, the cost of harvest, delivering wood, taxes, pesticides, etc. If the timber is harvested from a natural forest with no treatment or thinning necessary, the maintenance cost may be zero.

This model does not account for entities without legal rights to remove wood; only legal harvest is accounted for. Additionally, harvest parameters and market values are considered to be constant throughout the user-defined time period. In reality, these values change from year to year. However, the model is meant to provide a general estimate of ecosystem services provided by a forest parcel. The user may enter different parameter values for every year by running several models, one for each year in question.

Outputs

There are three vector file outputs from the Timber Harvest model.

- **Total Biomass Harvested**

The total biomass of timber, in Mg, harvested from each parcel.

- **Total Volume Harvested**

The total volume of timber, in m³, harvested from each parcel.

- **TNPV**

Total net present value, measured in the currency provided by the user, of the timber harvested from each parcel.

Habitat Quality and Rarity

The Habitat Quality and Rarity model estimates the extent and degradation of species habitat in response to threats and land cover change in order to generate maps of habitat quality and rarity. The results allow for a rapid assessment of regional habitat conditions that can be used as a proxy for more in-depth investigation regarding the status of species fitness. The model may be performed on a single species of interest or a group of species sharing similar needs for habitat (e.g., freshwater-dwelling reptiles). Using land cover

¹⁸ Marklund, L.G. and D. Schoene. 2006. Global Forest Resources Assessment 2005: Global Assessment of Growing Stock, Biomass and Carbon Stock. FAO: Forest Resources Assessment Programme. 55 p.

¹⁹ Howard, James L. 2007. U.S. timber production, trade, consumption, and price statistics 1965 to 2005. Research Paper FPL-RP-637. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 91 p.

data and a series of threat images, the user can produce maps displaying the extent and degradation of species habitats and the changes experienced by those habitats over time. This regional-scale vulnerability assessment describes the response of habitat types to different threats and land cover change. The model compares current and future land cover scenarios to a baseline (historic) land cover image depicting the land as it would be without human influence. As not all threats are equally damaging to all habitats, the model estimates the impact of one threat over another through a series of user-defined weights. Additionally, the influence of a threat is considered to degrade across space over user-defined distances by means of a linear or exponential function.

Habitat Quality

Hall et al. (1997)²⁰ define habitat quality as “the ability of the environment to provide conditions appropriate for individual and population persistence.” In this ESM model, the output is a continuous raster image where lower pixel values indicate lower habitat quality for a species based on the availability of resources for the survival, reproduction, and population persistence of the species. Locations that have the highest quality values are habitats that exhibit relatively intact structure, are far from developed areas and function in conjunction with the baseline historic image, i.e., a landscape that was free of human influence. An alternative definition of habitat quality²¹ deals with persistence at the individual level, and is the per capita contribution to population growth expected from a given habitat. The Habitat Quality and Rarity model only provides information for population-level habitat fitness.

At its most basic level, the model requires a current land cover map, a series of threat images, and associated databases with information on the intensity of threats and the sensitivity of habitats to those threats. For the purpose of this model, human-modified land cover types are considered threats and are defined as the alteration of habitats through the process of fragmentation, edge effects, and land conversion. The impacts of individual threats are determined based on the following four factors:

w : The relative weight of each threat impact variable to the habitat area is a user-defined weight ranging from 0 to 1 that is stored within the Threat Table and represents the potential threat posed to the species habitat. For example, if industrial land cover is given a threat weight of 1 and agricultural lands are set to have a threat weight of 0.5, the model considers the threat of urban land to be twice as great as that of agricultural lands. This model normalizes all the threat weights for a given habitat to sum to 1.

Typically, habitats closer to a threat source will experience higher impacts than habitats farther away. Therefore, the distance between the threat source and the species habitat, and how that threat translates across space, is an important factor. A linear or exponential distance decay function can describe the rate of change in impact, i :

$$i = 1 - \left(\frac{d_{xy}}{d_{r\max}} \right) \quad \text{if linear}$$

$$\exp \left(-2.99 \left(\frac{d_{xy}}{d_{r\max}} \right) \right) \quad \text{if exponential}$$

where d_{xy} is the linear distance between pixels x and y and $d_{r\max}$ is the maximum effective distance of the threat r across space. If impact $i_{rxy} > 0$ then the pixel value in x is located within the threat disturbance zone, which defines the furthest extent of the area impacted by threat.

a : The risk or accessibility of threat encroachment into a given habitat is a result of the amount of legal, institutional, social, or physical protection from a given disturbance. The model assumes that stronger protection will reduce the severity of a nearby threat, regardless of threat type. Accessibility may be added to the model as an image ranging from 0 to 1, where 1 indicates that threats have complete access to habitats and 0 represents a habitat that is perfectly protected.

²⁰ Hall, L.S., Krausman, P.R. and Morrison, M.L. 1997. The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin* 25(1): 173-182.

²¹ Johnson, M., 2007. Measuring Habitat Quality: A Review. *The Condor* (109): 489-504.

s : The relative sensitivity of habitats to degradation by imposing threats is different for each habitat and each threat, and this model assumes that highly sensitive habitats are more likely to degrade over time. Sensitivity values are on a continuous range from 0 to 1, where values closer to 1 indicate that a habitat is highly sensitive to a threat. This definition of habitat sensitivity differs from a broader definition which accounts for the rarity of the habitat type, but sensitivity scores may be extracted from existing ecology and conservation literature.

The total threat level T in a given pixel within a particular habitat type is the product of the four factors above for every pixel in a threat image, summed across all threat images. This means that threat level is equal to 0 if either accessibility a is 0 (or the area is perfectly protected), or if the habitat sensitivity s is 0 (or the habitat is not influenced by the threat), then the threat level is equal to 0. Additionally, because the threat weights are normalized within the model, we can consider the total threat score to be a weighted average of all threat levels, and as the threat level increases the overall habitat quality decreases.

This model uses a half-saturation function that accounts for habitat suitability to convert the threat score into a measure of habitat quality for pixel x in habitat type j :

$$\text{Quality}_{xj} = HS_j \left(1 - \left(\frac{T_{xj}^z}{T_{xj}^z + k^z} \right) \right)$$

where z is a constant equal to 2.5, HS_j is a habitat suitability score, and k is a *half-saturation* constant that normally needs to be adjusted for each individual case. However, in ESM, k is determined automatically such that the quality values will range from 0 to HS_j . The habitat suitability score HS_j is a number provided by the user indicating the suitability of land cover type j for a particular species habitat. A value of 0 indicates that it is entirely unsuitable (e.g., roads tend to be 100% unsuitable for most species), and a value of 1 indicates that the land cover type is entirely suitable (e.g., forest). The scaling factor in the interface controls the scaling of the quality scores. This should normally be left at 0.5. However, it can be changed to give greater emphasis to higher or lower quality values.

Habitat Rarity

This section of the model deals with habitat rarity, or the relative abundance of a habitat compared to its abundance in a baseline or historic land cover map. A baseline map is essentially a map of the natural environments devoid of man-made land cover classes. Here, habitats that were abundant in the baseline land cover map and have been dramatically reduced in the current or future scenario are considered rare because they are being replaced by other land cover classes.

Habitat rarity for habitat type j is defined as:

$$\text{Rarity}_j = 1 - \frac{N_j}{N_{j \text{ baseline}}}$$

where $N_j / N_{j \text{ baseline}}$ is the ratio of the number of pixels in land cover type j in the current or future map to the number of pixels in land cover type j in the baseline map. Based on this equation, a habitat with a Rarity_j value close to 1 indicates a major loss in a natural land cover type, which suggests that the habitat is rare and should be considered for protection. Once the Rarity_j measure is determined for each habitat type j , the quantification of the overall rarity for pixel x is:

$$\text{Rarity}_x = \sum_{j=1}^x \sigma_{xj} \text{Rarity}_j$$

where σ_{xj} is a Boolean indicator that equals 1 if pixel x belonged to habitat or land cover type j in the past and still belongs to land cover type j in the current or future scenarios, and $\sigma_{xj} = 0$ otherwise.

Outputs

The Habitat Quality and Rarity model produces three outputs:

▪ *Bio_Degradation*

The spatial distribution of habitat degradation across the landscape. A value of 0 indicates a non-habitat location, and any value greater than 0 indicates a degree of habitat degradation relative to other pixels.

▪ *Bio_Quality*

The habitat quality assessment for the study region, where values close to 1 suggest an area with high habitat quality. Pixels with a value of 0 are non-habitat locations.

▪ *Bio_Rarity*

The relative magnitude and extent of habitat rarity in the landscape. This output will only be generated if a baseline land cover image was supplied as an input. Values range from 0 to 1, where 1 indicates high habitat rarity and 0 indicates either abundance, or that the rarity could not be determined because the land cover category in question was not present in the baseline land cover map.

Crop Pollination

The Crop Pollination model evaluates the abundance and the economic value of wild bees as pollinators for agricultural crops. Animal pollinators are responsible for 35% of the global crop production²². As the quantity and diversity of bees decline in many parts of the world, it is important to assess the potential loss of pollination services provided by wild bees (native or feral) and to address habitat protection and agricultural vulnerability.

The abundance of wild bees depends on nesting site availability, food (flower) abundance in nearby areas, the number and type of species of bees present, and the flight range of each bee species. The model works in two steps: first, a bee abundance index is calculated for every pixel in an image based on nesting sites, flower resources and flight ranges. In the second step, the index and flight ranges are used to estimate the number of bees visiting agricultural sites. An economic value is calculated based on the market value of the crop benefiting from the pollination service.

The model is based on a land-use or land-cover map that includes natural and managed land cover types. Each land cover type must have an associated index value (ranging from 0 to 1), indicating the relative abundance of nesting sites and flower resources. Nesting site abundance may be broken up into the different kinds of nesting sites preferable to specific species of bees. According to the USDA, about 30% of the 4,000 native bee species in the U.S. are wood-nesters that build their nest inside hollow tunnels (in twigs, logs, etc.). The remaining 70% prefer underground nests, where the bees burrow tunnels and chambers in the soil. Bumble bees, often the most productive pollinators, construct nests in small cavities, either underground or underneath fallen vegetation.

Flower abundance information may also be broken up according to the season, if seasonality is important in the study region. The model also includes the typical foraging distance or flight range for each species of wild bee. Foraging distance determines the spatial scale at which wild bees can provide pollination services to crops. Values for foraging distance may be obtained from quantitative field estimates, existing literature, or from proxies such as body size. For example, Greenleaf et al. (2007)²³ found that foraging distance increased with body size non-linearly according to the equation:

$$\log Y = \log(-1.643) + 3.242 \cdot \log X$$

²² Pollinator Biology and Habitat. *New England Biology Technical Note, April 2009*. Natural Resources Conservation Service and the USDA.

²³ Greenleaf, Sarah, Neal Williams, Rachael Winfree and Claire Kremen. 2007. Bee foraging ranges and their relationship to body size. *Oecologia*, 153: 589-596.

where Y is the foraging distance in km and X is the intertegular span (IT span), or the distance between wing bases on a wild bee. Note that the IT span is related to the drop body mass:

$$\text{IT span} = 0.77(\text{mass})^{0.405} \quad (R^2 = 0.96; \text{ mass in mg and IT in mm}) \quad (\text{Cane } 1987^{24}).$$

Bee Abundance Indices

First, a pollinator abundance index is calculated for every pixel in the scene, based on the availability of nesting sites (h_n) and flower abundance in surrounding pixels (h_f). Flower abundance in close pixels will be weighted more heavily than flower abundance in very far pixels, and the maximum distance depends on the foraging distance of the species (α).

The pollinator abundance index P for pixel x is therefore:

$$P_x = N_j \cdot \frac{\sum_{m=1}^M F_j e^{-\frac{D_{mx}}{\alpha}}}{\sum_{m=1}^M e^{-\frac{D_{mx}}{\alpha}}}$$

where N_j is the nesting suitability of land cover type j , F_j is the relative abundance of flower resources in land cover type j , D_{mx} is the Euclidean distance between all other pixels (m) and x , and α is the average foraging distance for a wild bee species.

Foraging frequency in pixel x decreases exponentially with distance, such that pixels farther away contribute less to the total floral resource availability. It is assumed that pollinators forage in all directions equally. Values for N_j and F_j may be obtained from experts or from quantitative field data. If field data is not available, some values may be found in existing literature. For example, Lonsdorf et al. (2009)²⁵ provide supplementary data tables including values for nesting suitability and floral resources over a range of nesting types and land cover types in California, Costa Rica, New Jersey, and Pennsylvania.

Each pixel is given an abundance index value, ranging from 0 to 1, indicating the presence of pollinators in that pixel. In this sense, the output is a potential pollinator service map, accounting for the supply of the service but not the demand.

To account for demand, the model calculates the abundance index of pollinators on each agricultural site according to the land cover map. The model produces a “farm abundance” map, which determines the relative abundance of bees that travel from the source pixel x to forage on a crop in agricultural pixel o according to the equation:

$$P_{ox} = \frac{P_x e^{-\frac{D_{ox}}{\alpha}}}{\sum_{m=1}^M e^{-\frac{D_{om}}{\alpha}}}$$

where D_{ox} is the Euclidean distance between source pixel x and the agricultural pixel o , and P_x is the pollinator abundance of pixel x derived above. In this equation, the numerator is the distance-weighted proportion of pollinators supplied by pixel x that visit agricultural pixel o , and the denominator normalizes the result by the total spatial range of the pollinators.

In ESM the above formulas are implemented as follows:

Calculation of the proportion of nesting bees in each land cover (HN_<BeeSpecies>)

²⁴ Cane, J. 1987. Estimation of Bee Size Using Intertegular Span (Apoidea). Journal of the Kansas Entomological Society. 60 (1): 145-147.

²⁵ Lonsdorf, E., C. Kremen, T. Ricketts, R. Winfree, N. Williams, and S. Greenleaf. 2009. Modelling pollination services across agricultural landscapes. Annals of Botany, 103 (9): 1589-1600.

This is done by assigning to each land cover a value corresponding to the product between the use of the nesting guide from the pollinator species table (field NS_) and the availability of that nesting site in each land cover, from the Land Cover attribute table (field N_).

▪ **Calculation of flower availability**

The calculation of flower availability is done in two steps. The first step assigns to each land cover a value corresponding to the product between the seasonal activity of the pollinator, from the pollinator species table (field FS_) and the relative abundance of flowers in each land cover for that season, from the land cover attribute table (field F_).

In the second step, the resulting map is processed to accommodate for the foraging distance of the pollinator (Alpha value extracted from the pollinator species table), using a FILTER contextual operation. The FILTER has a radius equal to Alpha and is down weighted from the central kernel so that cells further away are less available. The weights in the filter are calculated as: $e^{(-D/\alpha)}$, where D is the distance from the center of the filter kernel to the center of all other kernels and alpha is the foraging distance specified in the pollinator table.

▪ **Calculation of pollinator source abundance index (SUP_<BeeSpecies>)**

The pollinator source abundance index is the relative abundance of each pollinator nesting in each landscape and is calculated by multiplying the availability of nesting sites (HN_<BeeSpecies>) times the availability of flower resources (HF_<BeeSpecies>).

▪ **Calculation of pollinator abundance in each land cover (FRM_<BeeSpecies>)**

The abundance of each pollinator species in each land cover is calculated based on the foraging distance of each pollinator. It is calculated through a FILTER operation on the pollinator abundance index calculated above. As with the flower availability, the FILTER is based on the Alpha value for each bee, and down weights locations that are further away.

The above steps are done for each pollinator and then combined by calculating the average across pollinator abundance in each land cover (FRM_AVG) and average pollination source abundance (SUP_TOT).

If future land cover conditions are input to the model, the process is repeated, replacing the current land cover for future land cover.

Economic Value

The economic value of pollinator services is based on the level of crop dependency on wild bee pollinators. Some crops (e.g., blueberries) depend more heavily on wild bees for fertilization than others. Additionally, the model assumes that the dependency of crop yield on wild bee pollination saturates; the yield increases with pollinator abundance with diminishing returns. The expected yield Y (FRM_val) of a crop c on farm pixel o is calculated as:

$$Y_o = 1 - v_c + v_c \cdot \frac{P_\alpha}{P_\alpha + \kappa_c}$$

where v_c is the proportion of the total crop yield attributed only to wild pollination such that a value of 1 indicates that the crop c 's yield depends entirely on wild pollination. The κ_c term is a half-saturation constant that represents the abundance of pollinators required to reach 50% of the pollinator-dependent yield.

Next, the model calculates the economic benefits supplied by the source pixels according to the expected yield Y_o . To do this, the partial contribution of each bee to the total farm abundance index is calculated as the ratio between pollinator abundance in each land cover (Step 4) and the total pollination abundance (FRM_Total), the sum of the pollinator abundance for all species of bees. The partial contribution is multiplied by the yield and added across species to obtain the total services provided by pollination (SUP_VAL).

The same distance-weighting procedure described above is used to calculate the services provided by each pollinator according to their distance to farm α . Finally, the total services provided (PS) by a pollinator in pixel m to all farms O is calculated as:

$$PS_m = v_c \cdot \sum_{o=1}^O V_o \frac{P_{ox}}{P_o}$$

where V_o is the crop value of crop c in farm α .

This results in a map where each pixel has an associated relative economic benefit according to the crops being pollinated and the abundance of different pollinators visiting farms from each pixel.

There are some limitations to the crop pollination model. Pollinator preference for nesting sites is represented by a Boolean indicator, so nesting suitability is not weighted on a per species basis. The amount of floral resources is described based on land cover type, and it not assigned for species genera of bees. Therefore, it is assumed that each species or genus forages for the same resources and prefers the same land cover types. Additionally, many land use and land cover maps do not specifically delineate transition zones that may be areas of high suitability for nesting. For example, many wild bee species prefer to nest in the border region between forest and agricultural fields, which is not considered. Finally, much of the data necessary to run the model is unavailable or difficult to find for many study areas.

Outputs

There are six possible outputs for this model: three for a current year scenario, three for a future year scenario (optional). The outputs describe the pollinator abundance index for all pixels in the study region, the pollinator abundance index in agricultural areas, and the pollination service value in relative dollar amounts (optional).

- ***Total nesting map for all species/guilds (prefix_SUP_TOT)***

The Abundance Index, ranging from 0-1, represents all pollinators in every pixel in the study region. This represents the likely abundance of pollinator species nesting on each pixel across the landscape.

- ***Visiting pollinators on agricultural sites (prefix_FRM_AVG)***

The Abundance Index, ranging from 0-1, represents the likely average abundance of pollinators visiting each agricultural pixel.

- ***Relative value of pollination services (prefix_SUP_VAL)***

This is a relative valuation of the crop yield due to wild pollination on each agricultural pixel. It is based on the Farm Abundance Index, and uses the saturating yield function to convert abundance into value units. It represents the contribution of wild pollinators to crop yield, and although the units are not dollars, the output does represent an economic value.

Habitat Risk Assessment

The Habitat Risk Assessment model uses information on habitat distribution and human activities to produce maps of habitat risk. The level of risk experienced by a habitat depends on the degree of exposure of the habitat to human activities and the impacts that result from such exposure. Exposure is a function of the overlap between habitat and human activities, the time period over which the overlap occurs, the intensity of the human activity, and the level of impact mitigation. Impact is a function of the amount of habitat loss, the resulting change in habitat structure, and the ability of a habitat to adapt or recover from disturbances. The model works in three steps: 1) first it calculates exposure and impact scores for each combination of habitat and stressor activity, 2) next it combines these scores into a risk factor for each combination, and 3) finally it aggregates the risk factors to obtain the total risk for

every habitat. Note that in this model, aggregated risk in natural environments is additive rather than synergistic (acting together) or antagonistic (acting against each other), due to the limited knowledge of how stressors interact with one another in a given habitat.

Exposure and impact scores are calculated based on variables provided by the user that are divided into categories of high, medium, and low. Exposure of habitats depends on the spatial overlap (evaluated by the model using the input maps) and the temporal overlap (based on information provided by the user) of habitats and stressors, as well as the intensity of each stressor on a habitat and the consequent management strategy. The impact of exposure depends on the change in habitat size, the loss of structure, and the frequency of natural disturbance (since habitats that experience high frequency of natural disturbance may be more capable of recovering from anthropogenic disturbance).

Additional considerations include the natural mortality rate of habitats, the recruitment rate, the age at maturity, and the connectivity rate. Natural mortality is rated from low (e.g. 0-20%) to high (e.g. > 80%) according to the best knowledge of the user, and is included because habitats with high overall natural mortality tend to be more reproductively active and are better able to adapt to disturbance. The recruitment rate is the length of time for a population to be replaced through reproduction and survival within a habitat; habitats with fast recruitment rates have a higher potential for recovery because there is a higher likelihood that the species will re-establish a population in a stressed environment. The age at maturity of a habitat is the time it takes for an entire habitat to reach maturity, and it is rated from low (e.g. < 1 year) to high (e.g. > 10 years). Habitats with a low age at maturity are more likely to recover quickly from a stressor. Finally, the connectivity rating ranges from low (high dispersal) to high (low dispersal) and describes the distribution of habitat patches. Habitat patches that are highly connected or concentrated in one area facilitate the flow of energy, matter and organisms, enhancing the potential for habitats and populations to recover from disturbance.

Although the rating systems presented in this model appear to be data intensive and largely subjective, it is possible to get reasonable estimates from existing literature or field research, and the user may define the ranges and legend for the ranking system however he or she chooses. Additionally, for each input parameter, the user may enter their importance weight (more important, equal importance, less important or unrated) that will condition the impact of that parameter in the final risk map.

Because this model uses a ranking or rating system for attribute information, results should be interpreted on a relative scale within the study area, and should not be compared across study areas.

Outputs

▪ ***Ecosystem Risk - <prefix>+EcoRisk***

Ranked cumulative ecosystem risk that is the sum of all risk scores for all habitats included in the analysis

▪ ***Recovery Potential- <prefix>+Recov***

Value that represents the potential for recovery of each cell based on the underlying habitat. It gives an indication of the resilience to human stressors.

▪ ***Cumulative Habitat Risk- <prefix>+CumRisk_+<habitat>***

Represents the cumulative risk of all stressors for each habitat

▪ ***HTML output***

HTML file showing the cumulative exposure and consequence value for each habitat to all stressors, and for each habitat the exposure to each stressor and its consequences.

Offshore Wind Energy

The Offshore Wind Energy model assesses the potential for establishing offshore structures that would capture wind energy and their benefits. Increased attention is being given to wind because of its potential to be a clean source of energy production. Offshore wind energy is becoming more and more important because of the trend in stronger offshore winds and the availability of energy loading centers along the coast.

The Offshore Wind Energy model produces five outputs: (1) power density, (2) harvested energy, (3) annual offset of carbon emissions, (4) wind farm value and (5) levelized cost of energy. The model works in two stages, first it creates an energy calculation and then it assesses the value of the wind energy.

Wind Power Density

Wind power density and potential is used to pinpoint offshore areas which have the potential to produce large amounts of energy and are in this sense suitable for wind farms. Wind power density is measured in W/m² and estimated as:

$$PD = \frac{1}{2} \rho \sum_{j=1}^c (f(V_j) \cdot V_j^3)$$

Where ρ is the average air density in kg/m³, V is the wind speed measured in m/s, and $f(V)$ is the probability density function of V . Mean air density, ρ is estimated using the equation $\rho = 1.225 - (1.194 \cdot 10^{-4})Z$ (NOAA 1976)²⁶. Z is the height of the wind turbine, in meters, at which wind power density is calculated.

Wind speed (V), which gets used both in the power density and probability density function equations, can be calculated based on the hub height (Z) of the wind turbine (Elliott et al., 1986)²⁷:

$$\frac{V}{V_r} = \left(\frac{Z}{Z_r} \right)^\alpha$$

Where V is wind speed, in meters/second, at hub height Z , in meters. V_r is the wind speed, in meters/second, at the reference height Z_r , in meters, in which wind speed data is known. The power law exponent, α , is given a default value of 0.11, based off of the equation $\alpha = 0.11 \pm 0.03$ for ocean surfaces within average atmospheric conditions.

Probability density function of wind data, $f(V)$, is generally measured using either a Rayleigh or a Weibull distribution (Manwell et al. 2009)²⁸. This model relies on the Weibull distribution because of its ability to calculate a larger range of wind patterns (Celik 2003²⁹; Manwell et al. 2009²⁸). It is calculated as such:

$$f(V) = \frac{k}{\lambda} \left(\frac{V}{\lambda} \right)^{(k-1)} e^{-\left(\frac{V}{\lambda} \right)^k}$$

²⁶ National Oceanic and Atmospheric Administration. 1976. U. S. Standard Atmosphere. NOAA- S/T76-1562, Washington, DC.

²⁷ Elliot D. L., C. G. Holladay, W. R. Barchet, H. P. Foote, and W. F. Sandusky. 1986. Wind energy resource atlas of the United States. DOE/CH 10093-4. Solar Technical Information Program. Richland, Washington.

²⁸ Manwell, J. F., J. G. McGowan, and A. L. Rogers. 2009. Wind energy explained: Theory, design and application. John Wiley & Sons Ltd., West Sussex, United Kingdom.

²⁹ Celik, A. N. 2003. A statistical analysis of wind power density based on the Weibull and Rayleigh models at the southern of Turkey. Renewable Energy 29: 509-604.

The user must enter values for k , the shape factor, λ , the scale factor, and V , the wind speed as explained in a previous equation. While the shape factor describes the profile of the curve, the scale factor describes the curve's range and maximum. The shape factor varies along a gradient, where, a larger k value is associated with consistent wind speeds and a smaller k indicates a greater variety of wind speeds and a greater likelihood of having very low or very high wind speeds. Inputs k and λ were calculated using the MATLAB function, *wblfit*. This prediction relies on wind time series data to estimate the Weibull distribution's maximum likelihood estimates for the two parameters. Additional information on this function can be found at: <http://www.mathworks.co.kr/kr/help/stats/wblfit.html>.

Harvested Energy

The amount of harvested energy from any given wind turbine is contingent on its physical characteristics as well as the wind conditions surrounding its specific location (Pallabazzer 2003³⁰; Jafarian & Ranjbar 2010³¹). This model estimates harvested energy using output power of a wind turbine and wind speed statistics. Although output power can be calculated using wind turbine output power curves, using a polynomial approach allows for less user input and a certain level of flexibility when it comes to assessing different turbines. The estimation of output power is explained by the following polynomial model approach (Jafarian & Ranjbar 2010)³¹:

$$OP = \begin{cases} 0 & V < V_{cin} \text{ or } V > V_{cout} \\ \frac{P_{rate}}{(V^m - V_{in}^m)/(V_{rate}^m - V_{in}^m)} & V_{rate} < V < V_{cout} \\ & V_{cin} \leq V \leq V_{rate} \end{cases}$$

Harvested energy (HE) is measured in MWh and is estimated as:

$$HE = nday \frac{\rho}{\rho_o} prate \left(\int_{V_{cin}}^{V_{rate}} \frac{V^m - V_{cin}^m}{V_{rate}^m - V_{cin}^m} f(V) dV + \int_{V_{rate}}^{V_{cout}} f(V) dV \right) (1 - lossrate)$$

In addition to the polynomial components, this equation also factors in the number of days of energy output (365 for an annual assessment), $nday$, standard atmospheric air density (1.225 kgm^{-3} for sea level conditions in the US), P_o , and energy losses as a result of downtime, conversion inefficiencies and electrical deficits (default set to 0.05), $lossrate$. The harvested energy equation calculates harvested energy from a single wind turbine, but it can be multiplied with the total number of turbines on a wind farm to estimate the total farm's output harvested energy.

Yearly Carbon Emission avoided

The carbon offset, due to the avoidance of carbon dioxide emissions, is estimated with the conversion factor:

$$6.8956 \cdot 10^{-4} \text{ metric tons CO}_2/\text{kWh}$$

This equation converts carbon-free wind power to the amount of CO_2 emissions avoided annually.

³⁰ Pallabazzer, R. 2003. Provisional estimation of the energy output of wind generators. *Renewable Energy* 29: 413-420.

³¹ Jafarian, M., and A. M. Ranjbar. 2010. Fuzzy modeling techniques and artificial neural networks to estimate annual energy output of a wind turbine. *Renewable Energy* 35: 2008-2014.

Wind Farm Value

The power generated, throughout the expected lifetime of a wind farm, contains a monetary value, quantified based on the net revenue a farm operator would accumulate. This value is considered the Net Present Value (NPV) of energy and for each wind farm is calculated as such:

$$NPV = \sum_{t=1}^T (R_t - C_t)(1 + i)^{-t}$$

The components of this equation include R_t , the total revenue generated in year t , C_t , the total costs in year t , T , the expected lifetime of a wind farm, and i , the weighted average cost of capital (WACC). Total yearly revenue is calculated as $R_t = sE_t$, where s is the price per kWh and E_t is the amount of kWh the wind farm generates. It should be noted, that $E_t = 0$ during the first year of construction. T and i can be modified as necessary, the former within the global parameters .csv file and the latter within the valuation section of the model's interface. WACC may be a suitable approach for evaluating the monetary value over time if you are working with a project that has a significant level of risk associated with initial and ongoing costs, and is supported financially by debt and equity. WACC is included into calculations in the same way as a discount rate would; if appropriate, input a discount rate instead. Although not included as a default value, Levitt et al. (2011)³² recommend a WACC value of 0.116; this is higher than typical discount rates and so your analysis may require a different value.

Revenue (R_t) gets entered into the model as a set rate for the whole lifetime of a wind farm, measured in kilowatt hours (kWh) of energy. Income comes in the form of either a definite energy price over a period of time or a result of a negotiation between the farm operator and consumers. The equation for revenue is: $R_t = sE_t$, where E_t is the amount of kWh produced by a wind farm and s is the price paid per kWh. No default value is available for this input because local procedures and location play important roles, and so research should be conducted to produce an accurate value.

Costs (C_t) include both one-time and ongoing costs. One-time costs are generally seen during the construction and deconstruction phases, including payments for turbines, foundations, electrical transmission equipment, and removal of equipment among other miscellaneous development expenses.

Turbine and foundation costs are based on unit costs. User-defined information can be specified as necessary, although cost data for 3.6 MW and 5.0 MW class turbines and monopole and jacketed foundations are included. The model assumes one foundation per turbine, with monopole foundations generally associated with 3.6 MW turbines and jacketed foundations with 5.0 MW or deep water. Foundation-based turbines are oftentimes constrained to depths of roughly 60 meters, but the model gives the user the opportunity to enter floating turbines by simply defining the appropriate unit cost for this farm design. The table below summarizes existing turbine costs, based on 2012 US dollars, and can be used as default costs.

Farm	Location	Capacity	# of Turbines	Total MW	Unit Cost (\$mil)
Riffgat	UK	3.6	30	108	6.3
Sheringham Shoal	UK	3.6	88	317	5.65
Greater Gabbard	UK	3.6	140	504	6.03
Butendiek	Germany	3.6	80	288	5.54
London Array	UK	3.6	175	630	6.29

³² Levitt, A., W. Kempton, A. Smith, W. Musial, and J. Firestone. 2011. Pricing offshore wind energy. Energy Policy 39: 6408-6421.

Amrumbank	Germany	3.6	80	288	6.41
Global Tech 1	Germany	5	80	400	10.4
Borkum 2	Germany	5	40	200	10.6

Foundation unit costs are slightly more difficult to estimate, but data from major foundation manufacturers is included below. Ramboll estimates that monopile foundations with 3.6 MW turbines are \$2 million each. Burbo and Rhyl Flats estimate \$1.9 million to \$2.2 million respectively for the same foundation. Nordsee Ost estimates jacketed foundations with 5.0 MW turbines to be \$2.74 million, and Ormonde estimates \$2.43 million for the same structure. Deepwater (40 meter) gravity foundations at Global Tech 1 are estimated at \$6.65 million each in 2012 by the European Energy Programme for Recovery. The total cost of turbines and foundations is calculated by multiplying the number of wind turbines by the unit cost of a turbine and its foundation.

Electricity transmission equipment varies based on wind farm design, size, distance from shore and use of either alternating current power or direct current power. This variation makes it difficult to estimate the cost of electricity transmission equipment. In order for the model to be used in several contexts and without extensive modeling, data from 20 wind farms was gathered from the U.K. Ofgem tender process and used to create a least squares regression equation that would relate total transmission costs to farm capacity and cable distance. The least squares regression equation is represented as:

$$TransCost = \beta_0 MW + \beta_1 TotCable + \epsilon$$

The model includes the impact of resistance which tends to cause losses within electrical transmissions. These losses depend on the current type, alternating current or direct current. Research points to a transition point at 60km, where alternating current is most cost efficient with distances less than or equal to 60km and direct current is most efficient above 60km (Carbon Trust, 2008³³; UMaine, 2011³⁴). The length of transmission cable is important for this calculation and it can be measured by either creating a .csv file that includes latitude and longitude values for all grid connection points or by modeling grid locations with fixed parameters. When lacking a sufficient grid connection map the fixed parameter is a good option. This choice identifies the expected average distance over land along the coast in which cables are expected to travel to reach their grid connections. The default parameter value is 5.5 km, the average overland cable distance within the UK.

In addition to cables connecting wind turbines to coastlines, wind farms also consist of cables which connect turbines to one another, considered array cables. Based on data, there appears to be a simple linear relationship between array cable length and turbine number. This data points to a length of 0.91km of cable per turbine, at an estimated cable cost of \$260,000/km.

Wind farm operators are also responsible for procuring finances for manufacturing and payments associated with construction phases. The model estimates this value at 5% of total costs, based on research that suggests 2%-8% of capital costs (AWS Truewind, 2010³⁵; Blanco, 2009³⁶). Installation costs are an additional one-time expense; this comprises roughly 20% of total expenses (Kaiser and Snyder, 2012)³⁷. In addition to installation costs, there are also costs for retiring and destroying a facility ($t=T$), a one-time cost that typically amounts to 2.6%-3.7% (Kaiser and Snyder, 2012)¹¹. The model's default value is set to 3.7%.

³³ Carbon Trust. 2008. Offshore wind power: Big challenge, big opportunity. Report on behalf of the Government of the United Kingdom.

³⁴ UMaine. 2011. Maine deepwater offshore wind report. <http://www.deepcwind.org/docs/OfficialOffshoreWindReport-22311.pdf>

³⁵ AWS Truewind. 2010. New York's offshore wind energy development potential in the Great Lakes. Feasibility Study for New York's State Energy Research and Development Authority.

³⁶ Blanco, M. 2009. The economics of wind energy. Renewable and Sustainable Energy Reviews 13: 1372-1382.

³⁷ Kaiser, M. and B. Snyder. 2012. Offshore wind capital cost estimation in the U. S. Outer Continental Shelf: A reference class approach. Marine Policy 36: 1112-1122.

While all of the costs up to this point have been one-time costs, there are also the on-going costs of maintaining and operating (O&M) a wind farm facility. The model's annual 3.5% default value for these expenses is based on research that points to values between 3% and 3.5% (Boccard, 2010)³⁸. Although default values are set for the values costs within this section, each can be user-defined by navigating to the .csv file.

Compiling the above information, costs can be estimated through the following equation:

$$C_t = \text{Turbine} + \text{Foundation} + \text{TransCost} + \text{Decommission} + \text{O\&M}$$

Levelized Cost of Energy

Levelized cost of energy (LCOE) measures the minimum price that a wind farm operator would need to obtain for energy in order to pay for all outward costs. Obtaining a price lower than the LCOE would be a losing proposition, financially. The output value is represented as \$/kWh and is calculated with the following equation:

$$LCOE = \frac{\sum_{t=1}^T \frac{O\&M \cdot CAPEX}{(1+i)^t} + \frac{D \cdot CAPEX}{(1+i)^T} + CAPEX}{\sum_{t=1}^T \frac{E_t}{(1+i)^t}}$$

Where *CAPEX* stands for the initial capital expenditures and includes all transmission, turbine and foundation costs, *O&M* stands for operations and management, *D* represents decommissioning facility destruction expenses, *E_t* stands for the energy produced annual in kWh, *i* stands for a discount WACC rate, and *t* stands for the annual time step, where $t = \{1 \dots T\}$.

Limitations

There are certain limitations to the predictability of this model. Default wind data is available but should be used for modeling on larger scales, either global or regional at 4 or 60 minutes spatial resolution. Rather than using power curves and polynomial equations, the calculation of harvested energy can sometimes be done more accurately through the use of various technological devices. If these data are available then adjustments can be made to the harvested energy section.

Although inflation can be included in the valuation of wind energy, by altering the discount rate, this assessment is not time-sensitive. Prices cannot be altered as energy prices fluctuate over time. To modify the model for your cost projection needs it is possible to establish an equation for foundation costs which would increase in deeper waters. Installation costs can also be modified based on the condition of the seafloor.

Using floating turbines is slightly more difficult than foundation-based assessments, but it can be done if accurate depth, distance constraints, and costs (entered in as "foundation costs") are used. Electrical transmission calculations are reliant on areas that are no more than 60 meters deep or 200 kilometers from shore; the model is less appropriate for predicting areas that are outside these limits.

Outputs

There are five raster file outputs from the Offshore Wind Energy model.

▪ Power Density

³⁸ Boccard, N. 2010. Economic properties of wind power: A European assessment. Energy Policy 38: 3232-3244.

Power density, measured in W/m^2 , for a wind farm centered on a given pixel.

- ***Harvested Energy***

Annual harvested energy for a wind farm centered at a given pixel.

- ***Yearly Carbon Emission avoided***

The yearly carbon emissions avoided, measured in tons, for a wind farm whose center resides on a given pixel.

- ***Wind Farm Value***

The total value, at present, of a wind farm centered at a given pixel.

- ***LCOE***

Levelized cost of energy, measured in $\$/\text{kWh}$, is the lowest price a wind farm developer can obtain from wind energy in order to negate expenses and have the project's costs equal zero.

Aesthetic Quality

The Aesthetic Quality model performs a viewshed analysis to determine whether new development or management areas will impact visibility from specific vantage points. In essence, it produces the “visual footprint” of planned land use. For example, one may calculate the visual footprint of a proposed wind turbine or wave conversion facility from the vantage point of a coastal home, a local scenic point of interest, or the town center. Likewise, the model may calculate the visual footprint of a new city park, so that planners can estimate increases in property values as a result of additional scenic views. Note that while the model accounts for topography, it does not require information about structures or vegetation that may constrain the viewshed. For example, the height and dimensions of a building may not be included in a digital elevation model, but it might block the view at a certain location. However, height information for these obstructions may be added to the model optionally if data for trees, buildings, etc., are available.

Using a DEM and a rasterized point or polygon file with the locations and heights of the new or proposed development, the VIEWSHED module in TerrSet conducts a viewshed analysis over a defined region of interest. Viewer height may be added in order to determine whether people will be able to see the new development (using an average human height in meters), or to evaluate the visual impact for skyscrapers, housing developments, or one particular building. Additionally, users may include information concerning the haziness or air condition of the region by estimating the visual contrast of the proposed development. Visual contrast is a measure of how visible an object is compared to the background. In very hazy or foggy conditions, visual contrast is virtually zero and the object cannot be seen. In clear conditions, visual contrast may be close to 1 (or 100%), and the visibility of the object will depend entirely on distance from the viewer and the curvature of the earth.

Outputs

Without considering atmospheric conditions, the Aesthetic Quality model output is a continuous proportional map where in-site pixels are given a value between 0 and 1 indicating the proportion of facilities that are visible from the pixel. For example, if 25 of the 50 proposed wind turbines (to be built off the coast of an important tourist center) are visible from a golf course, it will be given a value of 0.5.

When haziness or other air conditions are included in the analysis, the output is a visual contrast map for the nearest object or facility, with values ranging from 0 to 1 where 1 indicates that the object or facility is perfectly visible to the viewer. An additional output that considers atmospheric conditions is a proportional visual contrast map. Here, the visual contrast is calculated for every facility and the sum of those values is divided by the total number of facilities, for every pixel in the study area.

In addition to the viewshed outputs, summarized statistics may be produced for particular regions of interest (e.g., parks, conservation areas, residential zones) or populations. Here, areas or populations impacted are indicated by a visual contrast value or a proportional value greater than 0. Regions of interest are quantified in terms of the percent of the specified area from which at least one facility is visible. This involves a simple calculation of the area of the special interest region, and a summation of the areas of pixels that are visually impacted. The population impacted within the viewshed is based on the number of people living close to the shore; the viewshed analysis estimates the number of people impacted by visible sites.

Overlapping Use

The Overlapping Use model calculates the frequency and importance of human activity within a management zone. Knowing how an area is used by humans is a key aspect in conservation planning and management. The model identifies the areas that are important for land management, such as recreational use. Inputs include the study region, layers for different land management activities and their importance.

The model is able to estimate overlapping use potential either from a raster landscape or seascape image (a grassy field, coastal waters, a lake, etc.), from a rasterized image representing management areas. Maps of the locations of human activities are also required.

In the simplest model, each human activity is given an equal weight and the importance is calculated as the frequency of activities occurring in each pixel across the landscape or within each management area. A more complex model includes weights of importance for each activity. *Inter-activity* weights measure the importance of each activity relative to other activities in the model. For example, a user may wish to assess the importance of a national forest based on its use for camping, hiking, and wildlife conservation. While camping and hiking generate revenue for the forest, wildlife conservation has been a top priority in this region for years and is one of the main reasons people visit the forest. Conservation may be considered three times as important as hiking and camping, so their respective weights will be entered in an input table as 3, 1, and 1. The Inter-activity weights are assigned within the Recreation Table (.csv). This table consists of three columns. The first column lists the names of the recreational use images to be included in the analysis. The second and third columns are optional and list the inter-activity weights and the buffer distance. The inter-activity weights were described above, and if not set the module will assume that all activities have the same importance. The buffer distance specifies an area of influence of the activity in meters.

Intra-activity weights provide spatially explicit information that indicates the relative importance of various locations for a particular activity. For example, bass fishing may be a significant activity for East Beach in Martha's Vineyard, but the northern end of the beach draws the largest number of fishing tourists every year. Therefore, in a map provided by the user, the northern part of East Beach will have a higher importance index than other areas of the beach. The intra-activity weights are spatially explicit. If no weights given then all areas are considered equally important. The *intra-activity* weight is a Boolean parameter representing presence (=1) or absence (=0) of the activity in each pixel or management zone, and the *inter-activity* weight is set to 1 for all activities (equal importance for all activities).

If weights are included in the model, the importance of each pixel or management zone in the landscape is calculated as the sum of the product of the weights (inter * intra) divided by the total number of uses.

The user is also able to include a buffer for analysis around each activity, which may be important if the user has very small polygons for the location of an activity and wants to include a larger area. If a buffer area is entered, the model will automatically include the larger area in the analysis.

The weights and buffer parameters are optional inputs in the overlapping use model. Additional optional inputs include points that indicate hotspots of human use or access points, and a distance decay rate that dictates the way importance of a location changes as one moves towards or away from those hotspots. For example, sailing, swimming, and fishing may all be used in a recreational use analysis for a large lake, but the importance of those activities, or the amount of human use, is strongest near the public dock. Therefore, it is important that planning account for the importance of the zone surrounding the public dock, even more so than other

areas around the lake. In this model, the optional distance decay factor will be applied to the activity importance score calculated above. If the initial importance score for a pixel or management zone is Imp_0 , then the importance at a distance d with a decay rate of λ is:

$$Imp_d = Imp_0 e^{-\lambda d}$$

which simply states that the importance of a pixel decreases by λ amount every distance unit away from the hotspots. If ground reference data is unavailable to calculate the value of λ , we recommend that the user test multiple values for λ until a reasonable statement of decay is reached.

Outputs

The Overlapping Use model produces two outputs.

- ***Overlapping Use Importance***

A raster image containing the importance score for each pixel or management zone in the study area.

- ***Frequency of Overlapping Use Activity***

A raster image depicting the number of overlapping use activities that occur within each pixel or management zone.

Coastal Vulnerability

The Coastal Vulnerability model evaluates the exposure of coastal communities to storm-induced erosion and inundation. Since coastal habitats play an important role in protecting coastlines, scenarios of habitat modifications can also be performed to evaluate the effect of those changes on coastal vulnerability. For each coastline segment in the study area, vulnerability ranks from seven environmental variables are combined to produce a qualitative Vulnerability Index that represents the level of physical exposure of that coastal area. Geophysical characteristics, as well as the erosion buffer capacity provided by natural habitats, are included in the calculation. This model also produces a population density map to highlight the coastal communities that are vulnerable to damage from storms.

The following inputs can be used to calculate the Vulnerability Index: 1) coastline geomorphology, 2) elevation, 3) coastal habitats, 4) sea level rise, 5) wind exposure, 6) wave exposure, and 7) surge potential. Wind and wave exposure are calculated from fetch distance and wind speeds analysis within the model. Each of the optional input variables (except for storm surge) is ranked on a linear scale from 1-5 in order of increasing vulnerability to erosion and inundation from storms. According to this ranking system, a value of 1 represents low exposure and a value of 5 represents very high exposure (Gornitz, 1990)³⁹. The table below shows suggested rank values for each input variable, but the user may use their discretion in ranking geomorphology and habitat types. Rankings for variables marked (*) are calculated automatically by the model.

³⁹ Gornitz, V. 1990. Vulnerability of the East Coast, U.S.A. to future sea-level rise. Proceedings of the Skagen Symposium, J. of Coastal Res. Special Issue No. 9.

	Very Low	Low	Moderate	High	Very High
Rank	1	2	3	4	5
Wind Exposure*	<=20th Percentile	<=40th Percentile	<=60th Percentile	<=80th Percentile	>80th Percentile
Wave Exposure*	<=20th Percentile	<=40th Percentile	<=60th Percentile	<=80th Percentile	>80th Percentile
Storm Surge*	<=20th Percentile	<=40th Percentile	<=60th Percentile	<=80th Percentile	>80th Percentile
Coastline Geomorphology	Rocky; high cliffs; fjord; fiard, high seawalls	Medium cliff; indented coast, bulkheads and small seawalls	Low cliff; glacial drift; alluvial plain, revetments	Cobble beach; estuary; lagoon; bluff	Barrier beach; sand beach; mud flat; delta
Natural Habitats	Coral reef; mangrove; coastal forest	High dune; marsh	Low dune	Seagrass; kelp	No habitat
Relief	<=20th Percentile	<=40th Percentile	<=60th Percentile	<=80th Percentile	>80th Percentile
Current or Future Sea Level Change	Current rate < 1.8 mm/yr or Future net decrease	Current rate 1.8 – 2.5 mm/yr	Current rate 2.5-3 mm/yr or future change -1 to +1m	Current rate 3 – 3.4 mm/yr	Current rate > 3.4 mm/yr or Future net rise

Wind and wave exposure are important vulnerability variables, but difficult to calculate. The Coastal Vulnerability model uses a fetch procedure in TerrSet to estimate the average wave power that winds can generate across a seascape. The fetch procedure breaks the study region into 16 quadrants and calculates the length of water over which a given wind may blow to produce waves in each quadrant. A long fetch often produces a high-energy wave capable of causing erosion and inundation along coastlines. In the model, the user enters a fetch-length threshold that determines whether a coastal segment is considered “exposed” to wind and wave erosion. If the fetch threshold is reached at the coast from two or more directions, that location is exposed, otherwise it is sheltered. For each exposed segment, the model will predict the power of locally wind-generated waves (for fetch distances smaller than 50 km) and oceanic waves (for fetch distances over 50 km) that reach the coastline.

Additionally, at a given wave height, longer wave-period waves produce more damage. Therefore, in addition to calculating fetch data, the model accesses NOAA’s sea state information point database, WAVEWATCH III. This database is found in the TerrSet installation folder under the ResFiles subfolder.

A storm surge is a rise of sea surface height that occurs offshore as a result of wind stress and atmospheric pressure associated with storm activity. High winds pushing on the ocean surface causes water to pile up, especially over shallow waters, resulting in a wave break at the shore that can cause significant damage and loss of life. Storm surge is impacted by the slope and width of a continental shelf - gentler slopes tend to produce larger storm surges, while steep slopes limit wave pile-up. If there is a continental shelf within the region of interest, it may be important to include its location so that the distance between the shelf and each segment of coastline may be calculated. If not, it may be useful to extend your area of interest so that it includes a portion of the closest continental shelf. In general, as the distance from shore to continental slope increases, the potential for storm surge is enhanced.

The geomorphology variable conveys the erodibility of different geomorphic classifications defined by Gornitz et al. (1997)⁴⁰ and USACE (2002)⁴¹. For example, rocky coasts, fjords and fiards provide excellent protection from erosion and are typically given a low

⁴⁰ Gornitz, V. M., Beaty, T.W., and R.C. Daniels. 1997. A coastal hazards database for the U.S. West Coast. ORNL/CDIAC-81, NDP-043C: Oak Ridge National Laboratory, Oak Ridge, Tennessee.

rank of vulnerability to storm damage. On the other hand, barrier and sand beaches provide minimal protection and are given a rank of 5, indicating that these types of coastlines are highly vulnerable to erosion and inundation. The model requires an image depicting the coastline, where each segment of coastline has a designated rank according to the geomorphology.

Natural habitats may also provide some protection for coastlines. Marshes, seagrass beds, kelp forests, mangroves, and coastal dunes absorb some of the energy of incoming waves, preventing a significant amount of damage to coastlines and coastal communities. Fixed, sturdy habitats that penetrate the water column such as coral reefs and mangroves tend to provide the most protection to coastlines, and are therefore given a rank of 1 for low vulnerability. Flexible or seasonal structures such as seagrass beds or kelp forests tend to dissipate wave energy and encourage sediment build-up, but are not often able to stand up to the force of storms. Therefore, these habitat types are given a rank of 4. A coast line with no habitat is given a rank of 5. High and low sand dunes (the threshold depends on the history of the study region) provide high and moderate protection, respectively. To calculate the rank for a segment of coastline, the model determines which habitat classes are within a specified search radius from the segment. The model then creates a vector that contains all the ranks from the surrounding habitats, and computes the final natural habitat exposure rank based on a formula that maximizes the accounting of beneficial services. The search radius is a user input, and depends on how far the user believes the protective capabilities extend for a particular habitat type.

Elevation, or relief, is an important variable because steep slopes and high elevation will protect coastlines better than low elevations and flat surfaces. In general, areas that are above mean sea level are at a lower risk of erosion and inundation than low-lying sites or areas at mean sea level.

A net change in sea level will impact local erosion and inundation patterns – a fact that is cause for concern in coastal communities worldwide. To include vulnerability due to sea level change, this model requires an image of the coastline where each pixel is assigned a rank value associated with sea level rise. In the table above, suggestions for rank assignment are provided for two scenarios: future sea level change, (a predicted value in feet or meters for some time in the future, which may be calculated using the Climate Change Adaptation Modeler in TerrSet) and current sea level change (a rate of change in mm/yr, which may be obtained by interpolating between National Ocean Service data stations in the United States). For future sea level change, a relative rise between -1 m and +1 m does not significantly alter erosion and is therefore given a rank of 3, but a sea-level rise greater than +1 m may increase vulnerability to flooding in some areas (rank of 5). Likewise, a sea-level decrease greater than -1 m may decrease vulnerability (rank of 1). For current sea-level change, a rate of 1.8 mm/year represents the global eustatic sea-level rise, and any rate above this level is representative of local isostatic or tectonic land motion. Therefore, a rate less than 1.8 mm/year is indicative of a very low vulnerability coastline, but rates much higher than 3.4 mm/year exceed international standards and are therefore suggestive of high vulnerability.

The final Vulnerability Index (VI) is calculated for each shoreline segment as:

$$VI = (R_{Elevation} R_{Wind} R_{Waves} R_{geomorphology} R_{Habitats} R_{Surge} R_{SeaLevel})^{1/NumVar}$$

Where R denotes the rank for each of the seven variables and $NumVar$ is the total number of variables available for calculation. Natural habitats, elevation, geomorphology of shorelines, surge potential, and sea level rise data are all optional inputs that enhance the quality of the VI but are not fundamental in the identification of vulnerable communities.

Depending on the number of inputs included in the model, the coastal vulnerability module also produces an erosion index and an inundation index, based on the formulas below.

Erosion Index:

$$EI = (R_{Waves} R_{geomorphology} R_{Habitats})^{\frac{1}{3}}$$

Inundation Index:

⁴¹ U.S. Army Corps of Engineers (USACE). 2002. U.S. Army Corps of Engineers Coastal Engineering Manual (CEM) EM 1110-2-1100 Vicksburg, Mississippi.

$$II = (R_{Elevation} R_{Wind} R_{Habitats} R_{Surge} R_{SeaLevel})^{\frac{1}{5}}$$

Outputs

The Coastal Vulnerability model produces three main index outputs, Vulnerability Index, Erosion Index and Inundation Index, however the last two are only produced if the optional variables required to compute them are included in the model.

▪ **Vulnerability Index**

This is the final vulnerability index that combines both erosion and inundation. Areas with higher values indicate a high vulnerability.

▪ **Erosion Index**

This index is a relative value representing the vulnerability of a coastal segment to erosion due to wind and waves.

▪ **Inundation Index**

This output depicts a relative index value representing the vulnerability of a coastal segment to inundation due to sea level rise and wave surge.

Another output that is not automatically displayed is the *Coast_Pop* image, which is a map of coastal population derived by overlaying the population image with the coastline image.

Note that the TerrSet implementation of the Coastal Vulnerability model is based on InVest Version 2.5.2. However the formulas for the calculation of Vulnerability Index and Erosion Index are based on the Version 2.5.6.

Marine Aquaculture

The Marine Aquaculture model can either estimate the productivity and economic value of pre-existing aquaculture systems or identify the best potential site for new aquaculture projects. The productivity and value of the ecosystem services provided by a marine fish farm or netpen is based on the total weight of fish harvested at each site and the market price of the fish. The model performs analysis for one or more harvest cycles (from the day at outplanting⁴² to the day of harvest). Harvest cycles are typically separated by a fallow period, which must be defined by the user.

The weight of an individual fish (kg) is a function of its growth rate and water temperature.⁴³ Weight over time, from the moment of outplanting to harvest, is modeled according to the equation:

$$W_{t,y,f} = [\alpha (W_{t-1,y,f})^b \cdot e^{\tau T_{t-1,f}}] + W_{t-1,y,f}$$

where a and b are growth parameters, $T_{t,f}$ is the water temperature (°C) at time t (day) and farm f , and τ is a fixed scalar (0.08 °C⁻¹) representative of biochemical rate acceleration in fish when temperature increases by 8-9 °C. Growth parameters a and b have default values of 0.038 g day⁻¹ and 0.667 (unitless) respectively, which are suitable for Atlantic salmon and may be used for other finfish with similar growth patterns. The user may adjust these parameters according to the specific species of fish, but it is

⁴² Outplanting is a term commonly used in salmon aquaculture; refers to the release of fish into an aquaculture environment after they have matured beyond their freshwater life stage.

⁴³ Stigebrandt, A. (1999). MOM (Monitoring-Ongrowing fish farms-Modeling) Turnover of Energy and Matter by Fish – a General Model with Application to Salmon. *Fisken Og Havet (Fish and Sea)*. 5.

recommended that any adjusted growth parameters be tested in the equation above estimating the time for fish to reach harvest weight. In general, growth parameters a and b should be obtained from an expert or netpen manager.

The total weight of fish produced on each farm is calculated based on the assumption of homogeneity of growth rates within the farm. This assumption is unlikely to significantly influence results because netpens are often managed such that growth rates are independent of population size, and because fish of similar weight tend to be grouped for ease of feeding and harvest. Also, it is assumed that all fish on the farm are harvested at the same time. Note that if netpens have different characteristics with respect to harvest time, outplanting weight, fish type, etc. they should be treated as individual farms and modeled accordingly.

The total weight, in kg, of processed fish (TPW) on each farm (f) during a harvest cycle (c , measured from outplanting to harvest) is a function of the product of harvest weight ($W_{th,h,f}$), the proportion of total harvest weight that is made up of headed, gutted or filleted fish after processing (fractional value, d), and the number of fish on a farm (n_f). This product is combined with a decay function that accounts for the daily mortality rate (M) expected for a particular fish species on an aquaculture farm, throughout the harvest cycle.

$$TPW_{f,c} = W_{th,h,f} d n_f e^{-M(th-to)}$$

Again, the default settings are based on estimates for Atlantic salmon: the proportion of fish remaining after processing is 0.85, and the daily natural mortality rate is 0.000137 day⁻¹. These values may be used for finfish similar to Atlantic salmon in size and harvest practices. Otherwise, the user may calculate M using empirical equations such as Pauly's M , Rikhter and Efanov's method, or from tagging data acquired from a fish farm.⁴⁴

The economic value of the total harvest for each farm is measured in terms of net present value (NPV):

$$NPV_{f,c} = TPW_{f,c} | p(1 - C) | \cdot [1/(1 - r)^t]$$

NPV is calculated as a function of the total processed weight of fish (TPW), the market price per unit weight of processed fish (p) and the fraction of the market price that is strictly attributed to benefits (C). Finally, the market value is checked by the market discount rate (r), or the interest rate used to convert future payments into present values; the U.S. government recommends an annual discount rate of 3% or 7% for many environmental projects such as those under the Clean Air Act.

Outputs

There are three image outputs from the Marine Aquaculture model:

▪ **Total Weight Harvested (kg)**

The total kilograms of fish produced in each netpen, or the potential productivity in each pixel in the landscape.

▪ **Total Number of Cycles**

The total number of cycles in the model run for each netpen or pixel in the scene. This is determined by the start date, fallow period, growth rate, and the total length of time for the analysis.

▪ **Net Present Value**

The total economic value, in thousands of USD or other currency, for each netpen, or the potential economic value for each pixel in the scene.

⁴⁴ For Pauly's M , Rikhter and Efanov's method, and more information about mortality rates and stock management: http://www.ccs.miami.edu/~fgayanilo/documents/Gayanilo_Pub_97a.pdf

Additionally, the model produces a tabular output containing harvest and revenue information for each cycle experienced by each farm.

Wave Energy

The Wave Energy model identifies hotspots of potential wave energy for siting wave energy conversion (WEC) facilities. Waves are caused by wind generated by uneven solar heating; therefore wave energy can be considered a concentrated form of solar energy. In fact, waves can generate up to 1,000 kW per meter of wave crest length with only 100W/m² of incoming solar energy!⁴⁵

The economic valuation of wave power services provided by a particular location in an oceanscape can be used to compare the costs and benefits between potential WEC sites, assisting decision makers to evaluate tradeoffs. Note that the model is only accurate for waves in deep water, where the local water-depth is greater than half of the average wave length. Typically, at 20m water depths a wave's energy drops to about 1/3 of the energy than found in deeper water.

The common measure of wave power, P , for *regular* waves is

$$P = \frac{1}{32\pi} \rho g^2 H^2 T$$

where ρ is the density of surface seawater (1,028 kg/m³), g is the gravitational constant (9.8 m/s²), T is the wave period and H is the wave height. Regular waves are defined as waves following a repeating pattern (in terms of wave period and direction) with similar conditions and characteristics. However, real ocean wave dynamics are characterized by any number of waves with different frequencies, amplitudes and directions. The wave power, in kilowatts per meter of wave crest length, transmitted by *irregular* waves in deep water is approximated as

$$P \approx \frac{1}{16} \rho g H_s^2 C_g(T_e, h)$$

Here, H_s is the significant wave height, which is typically defined as the wave height of the tallest 1/3 of the waves. C_g is the group velocity of the wave, defined as the velocity with which the overall shape of the wave's amplitudes propagates through space; it is a function of wave energy period, T_e , and local water depth, h . Wave energy period, T_e , can be approximated using peak wave period, T_p , and a particular combination of H_s and T_p represents a "sea state," or the general condition of the ocean. In this way, wave power is calculated based chiefly on sea state information.

In order to measure the amount of electricity that can be converted from wave power, the model requires information about the energy absorption performance of a particular WEC device for each sea state. Thankfully, performance data has already been collected for four widely-used WEC machines: PWP-Pelamis, Energetech-OWC, AquaBuOY, and WaveDragon. This data is provided for users in the ESM Wave Energy Tutorial folder. Power values are then calculated according to the equation above and summed across all sea states for every hour they occur to provide a total estimated power output.

Global sea state data, obtained from WAVEWATCH III model outputs⁴⁶, and ocean depth data, required as a proxy for group velocity, are also available in the ESM Wave Energy Tutorial folder. The WAVEWATCH III model output contains two key pieces of information: 1) the spatial location of each model wave site, in the form of a vector point file, and 2) an ACCDB table containing information on significant wave height.

⁴⁵ Technology White Paper on Wave Energy Potential on the U.S. Outer Continental Shelf. Minerals Management Service, U.D. Department of the Interior. 2006.

⁴⁶ For information and guidelines on the NOAA WAVEWATCH III model, see the WAVEWATCH III User Manual

A simple cost-benefit analysis is used to calculate the net present value for wave energy facilities where, for each year in the total lifespan span of a WEC project, the economic value is based on relevant benefits and costs of the facility. The market discount rate is also required for valuation, with a value typically between 3% and 7% for environmental projects. Annual benefits include the price of electricity per kWh and the annual captured wave energy in kWh. Annual costs include initial installation, annual operation and maintenance costs. Initial costs are comprised of capital cost per kW, cost of mooring lines and underwater cables, and the cost of overland transmission cables. The cost of transmission cables depends on the distance over which the energy is transmitted. Therefore, the model allows users to explore the tradeoffs between locations based on both the potential for wave energy *and* estimated costs.

Outputs

The Wave Energy model produces three outputs:

- ***Wave Power***

In kW/m

- ***Wave Energy***

In MWh/year

- ***Net present value***

In thousands of USD or other currency.

CHAPTER TEN

EARTH TRENDS MODELER

Earth observation image time series provide a critically important resource for understanding both the dynamics and evolution of environmental phenomena. As a consequence, Earth Trends Modeler (ETM) is focused on the analysis of trends and dynamic characteristics of these phenomena as evident in time series images. Further, the system is highly interactive, with the process of exploration largely being an active process. The trends and dynamics emphasized include:

- Interannual trends – trends that persist over several years or longer that may be indicative of major environmental changes. ETM provides tools for determining the presence of trends that are both linear and monotonic (non-linear) as well as their significance. In addition, trend analysis tools are provided that are resistant to the presence of outliers in short time series.
- Seasonal trends – trends in the character of the annual seasonal progression. This is a revolutionary new analytical procedure developed for ETM. The axis of the earth leads to substantial annual variation in solar energy, leading to major seasonal cycles. However, traditional multi-year trend analysis procedures consider seasonality to be a contaminant and thus intentionally reject it. This new procedure specifically seeks trends in seasonality and displays it in a dramatic manner.
- Cyclical components – ETM offers a completely new procedure for examining cyclical components in image time series. Marrying the elements of Fourier Analysis and Principal Components Analysis, ETM provides a unique form of Spatial-Temporal Spectral Analysis, called Fourier PCA, that discovers not simply cycles in isolation, but patterns of cycles that occur together. A final stage of correlation analysis also shows when these patterns were most prevalent.
- Irregular but recurrent patterns in space/time -- ETM offers a variety of tools for the analysis of recurrent patterns in space/time including a unique form of Wavelet Analysis and both standardized and unstandardized Principal Components Analysis (also known as Empirical Orthogonal Functional analysis).
- Teleconnections – ETM offers cutting-edge tools for the analysis of climate teleconnections (related patterns of variation between widely separated areas of the globe).
- Coupled modes – A special consideration in Earth System Science is the understanding of coupled modes of variability, such as the coupled ocean-atmosphere El Niño/La Niña phenomenon.

ETM offers a variety of tools for seeking and examining these modes including Canonical Correlation Analysis (CCA), Extended PCA (also known as Extended EOF), Extended EOT, MSSA (Multichannel Singular Spectrum Analysis) and MEOT (Multichannel EOT).

A Recommendation

We strongly recommend that you complete the Tutorial exercises for ETM. This is the fastest way to learn the full scope of the system. If you purchased TerrSet with the intention of working primarily in ETM, we also recommend that you complete the *Using TerrSet* Tutorial sequence, particularly Exercises 1-1 through 1-3.

Accessing ETM and its Functions / Display Recommendations

ETM can be accessed from the menu or from the Shortcut utility. Like Land Change Modeler (LCM), ETM opens in a special interface docked to the left side of TerrSet's workspace. We strongly recommend your use of a widescreen monitor (e.g., 1920 x 1200) or a dual monitor setup. However, for those with lower resolution monitors, ETM can be minimized against the left hand edge to make more space by clicking on the “-” symbol on its banner (and similarly, the “+” symbol to re-expand it).

Hate to Read the Documentation? A Power-User Alternative

Experienced users of GIS can probably figure out most aspects of the use of ETM on their own. However, there are some critical issues that expert and novice users alike need to know. Look for and pay careful attention to special sections entitled “*Critical Things to Know!*”

Critical Things to Know! – ETM in General

1. Image time series are handled in a new way, beginning with the IDRISI 16.0 Taiga Edition. A time series now consists of a pair of files. *Image series* consist of a raster group file (RGF) that contains a list of the images in their correct time sequence and a time series documentation file (TSF) that describes the nature of the series. As a first step, create the RGF. This can be done in TerrSet Explorer by selecting the images, then right-clicking within the empty space of the Files tab and selecting the Create option from the context menu. An RGF may also be created with Collection Editor, located under the File menu, or by creating it directly using the Edit module, located under the Data Entry menu. Then create the TSF metadata file. This can be done from the Project panel on the Explore tab of ETM or by selecting the Create TSF option located under the TerrSet File menu.
2. ETM also recognizes *Index series*. Index series are one-dimensional time series such as the Southern Oscillation Index. Index series also consist of a pair of files – an attribute values file (AVL) to hold the data and a TSF metadata file. The AVL files are typically created with the Edit module. For time series applications, the file consists of two columns – an ID column and a second column with the data. Columns should be separated by either one or more spaces or a tab. In this version, the ID column is ignored. The easiest way to create an AVL is to create the two columns in a spreadsheet such as Excel and then paste the columns into the Edit module of TerrSet (pasting from Excel sometimes takes time, but just wait and it will work). Save the file as an attribute values file from Edit (it knows how to document it correctly).
3. All time series are registered in ETM as part of an ETM session. ETM session files are saved as text files with an “.etm” extension. ETM projects differ from TerrSet projects, which are simply a list of folders. Any series you register with ETM will need to be located in either the TerrSet project Working Folder or one of the TerrSet project Resource Folders. We recommend that you put all series in Resource Folders. When you browse for and select a series outside of your TerrSet project, ETM automatically adds the folder to your project as a Resource Folder.
4. It is recommended that you start a project with an empty Working Folder and add series from Resource Folders. For example, let's say you're going to be looking at relationships between sea surface temperature (in a series named “SST”) and the temperature of the lower troposphere¹ (in a series named “TLT”). Let's imagine your Working Folder is named

¹ The troposphere is the portion of the atmosphere in which weather events (e.g., rain storms) occur. It extends to approximately the cruising altitude of

“d:\analysis,” your SST series is contained in a Resource Folder named “d:\sst” and the TLT series is contained in a Resource Folder named “d:\analysis\slt.” When analyses are run, ETM will put the results into a set of special subfolders of the Working Folder that bear the same name as the series. For example, all trend analyses for the SST series will go into a folder that ETM will automatically create called “d:\analysis\sst\trend.” Similarly, PCA and EOT results will go into a subfolder named “d:\analysis\sst\components” and so on. Note that in the case of your TLT series, a folder with the same name as the series already exists, so it will place the results there (i.e., “d:\analysis\slt\trend”, etc.).

5. Do not manually delete any folders or files related to your series. Use the Advanced Management option within the ETM Session Parameters panel in the Explore tab to do this. Otherwise, your project may become corrupted (you can recover from this, but it will take time).
6. Most operations have an automatic naming convention based on a user-defined prefix and a default suffix. In addition, ETM will, in these instances, supply a default prefix which is the series name. It is safe to accept this default name. For example, if you run a Theil-Sen trend analysis on a series named SST, it will automatically supply the prefix “sst” and the resulting analysis will be named “sst_ts_slope.” In cases where you want to use a different name so as not to overwrite a previous analysis (e.g., with the Linear Modeling panel), we recommend that you add additional characters to the end of the recommended prefix (although you can specify a completely new prefix if you like).
7. ETM keeps track of all your analyses so that you can re-examine and explore the results at any time using the respective panels within the Explore tab. If a series is not listed in one of these panels, it means that there are no analyses of that type that have been created yet for that series.
8. When selecting a series, ETM will display the names of all series types that apply. Thus if both index and image series are shown in the list, either can be selected.
9. Try to keep your series names short (long analysis chains can lead to very long names after all the suffixes have been added). Also, it’s tempting to create projects that contain every series you’ve ever analyzed. Resist this temptation. If anything goes wrong, there’s a lot to reconstruct. It’s better to have many ETM projects.
10. Several analytical components require an adjustment for the area each pixel occupies. If the reference system is LATLONG, an automatic adjustment is made. Otherwise, it assumes that all image series are on an equal area projection. If this is not correct, project your series images to latlong with the PROJECT module. This can be automated for the whole series by using Macro Modeler and the dynagroup option.

Tabs and Panels

ETM is organized around three tabs:

Explore
Analysis
Preprocess

Within each tab is a series of drop-down panels for tasks/analytical stages. You can have as many drop-down panels open as you wish – they are presented this way simply to accommodate varying screen resolutions.

large aircraft.

The Explore Tab

Working with time series is a highly interactive process in ETM. You will find that your need to explore series is not just a starting point, but a continuous process of discovery and validation. All panels on the Explore tab relate to this activity with one exception – the ETM Session Parameters panel in which you add or remove series from your session and specify various masks and palettes you prefer to use in the display of series.

Critical Things to Know!

1. ETM gives you the option of specifying a default mask file for each image series in the Project panel grid. Like all mask files in TerrSet, a value of 1 means a pixel contains data to be processed and a 0 means a pixel should be ignored. If you enter a mask filename in the Project panel grid, it will display as the default mask in those ETM panels containing a mask option. Note however that any mask may be specified, whether or not it was included in the ETM project.
2. The Explore Space / Time Dynamics panel allows you to examine both image series and index series. The latter are displayed as graphs, while the former are displayed as space-time *cubes*. To create a space-time cube, display the first image in the series (from TerrSet Explorer or DISPLAY Launcher). Then use one of the instant stretch options on Composer to stretch the first image. ETM will stretch all other images in the series the same way when constructing the cube. Generally, the middle stretch button should be used if the image has both negative and positive values and the left stretch button should be used otherwise. After you create the cube once, you will never need to create it again (unless you decide to use another stretch option, in which case, you must repeat the sequence above).
3. The remaining panels are only of use (and will only show series options) after running one or more of the analyses on the Analysis tab.
4. If you use the automatic vector overlay option from the Advanced Management feature of the ETM Session Parameters panel, make sure that the reference system of this overlay is identical to that of all series.

The Project Panel

The Project panel allows you to set up, modify or recall an existing project. A project consists of one or more time series and associated palettes, mask files and analyses. An ETM project is recorded in a file with an “.etm” extension which is in text format and can be edited (although we recommend using the Advanced Management feature of the Project panel for this). This ETM project file is stored in the current Working Folder. As stated earlier, we recommend that new projects be created in an empty Working Folder.

ETM uses a session concept similar to that of the Land Change Modeler. However, with ETM, the concept is taken even further. Each series is given a special Resource Folder of the same name as the series in a subfolder of the Working Folder, regardless of where the series is actually located. The subfolders have names such as:

- “trend” to hold trend analysis results
- “sta” to hold seasonal trend analysis results
- “components” to hold the results from PCA/EOF, EPCA, MSSA, EOT, EEOT and MEOT
- “fourier” to hold the results from Fourier PCA
- “linear_models” to hold linear modeling results

Thus if your Working Folder is named “d:\time_series” and you have an image series named “ndvi8203” (somewhere in your TerrSet project), then as you create analyses, you can expect to see folders being created automatically such as “d:\time_series\ndvi8203\trend” and “d:\time_series\ndvi8203\components,” etc.

Use the Add and Remove buttons on the ETM Session Parameters panel to add or remove series and specify optional default masks and palettes. For more detailed management tasks such as renaming or deleting series and analyses, click on the Advanced Management button to launch a subpanel. Note that you also can specify a vector file in the Advanced Management subpanel that should be overlaid on all displayed results.

The Explore Space / Time Dynamics Panel

This panel allows you to visually explore both index and image series. Both are available in the drop-down list. If you choose an index series, it will be displayed as a graph. Selecting an image series reveals a special four-dimensional² graphic display that will simply be referred to as the *cube*. The first time you examine an image series with this feature, you will need to create the special files required for the cube. You may encounter these if you explore a folder that contains a series – there are three of them and they have file extensions of “.bsq”, “.bil” and “.bip”.³ If you copy a series to another folder, you may wish to copy these as well to avoid having to recreate them.

An important issue in creating a visualization cube is the contrast stretch. ETM bases the stretch of all images in the series on the display minimum and display maximum settings of the first image in the series. A recommended procedure is to display the first image of the series (either from TerrSet Explorer or DISPLAY Launcher) and then use one of the instant stretch options at the bottom of Composer to stretch it and maximize the contrast⁴. Then when you click the Create/recreate visualization button, the same display parameters will be used for all other images in the series.

There are three display modes for the image series: cube, plane and sphere. The plane mode is probably the least useful, but it accurately describes the nature of the data – in reality, the viewer is cycling through three planes of the space-time continuum. The cube mode places these planes on the outer faces of the cube. The sphere mode is self-evident – it is a three-dimensional version superimposed on a sphere.

The cube can be manipulated in various ways. Grabbing it with the mouse allows you to move it spatially, while the zoom in/out buttons control the detail. A right click on the cube launches a context menu of options that are optimized for specific views. The Reset button changes it back to the default view.

Depending upon whether Time, X or Y is selected, the Play/Pause button will animate the series over that dimension (this is best appreciated in the Plane view). When the Play/Pause button is in the Pause position, use the left or right arrow keys on the keyboard to move from frame to frame. When it is in the Play position, animation is automatic according to the frame rate specified by the up/down selector in the dialog.

The Display icon for the cube allows you to look at any specific view in full resolution. When Time is selected, you are selecting a specific image out of the series. When you right-click within the panel and select either the Orient to X or Orient to Y options from the context menu, the resulting image is known as a Hovmoller⁵ or space-time plot. A particularly useful view is the Orient to Y view while the Y dimension is set to 0 (the equator). In this view, moving events (such as sea surface temperature or pressure) associated with the El Nino phenomenon will be seen as diagonal features in the graphic.

² The dimensions are 1: X, 2: Y, 3: data value, 4: time.

³ These files contain the three planes of the visualization cube. BSQ = band sequential, containing a single image for each time slice. BIL = band interleaved by line, containing an X/time slice for position in Y. BIP = band interleaved by pixel, containing a Y/time slice for each X position.

⁴ Use the middle instant stretch option on Composer for series that should be symmetric about 0 (i.e., they contain both negative and positive values) and the left-most instant stretch option otherwise.

⁵ Named after the climatologist who is widely attributed to have first proposed the utility of this view.

Displaying Secondary Series

Many of the Explore Tab panels that display graphs of index series also allow the superimposition of a second series. Use the button with the “+” symbol above the graph to do this. When a second series is superimposed, it automatically uses a secondary Y axis to make the series easier to compare. In some instances, comparison is easier if the second series is inverted. Once the secondary series is displayed, a series inversion button will also be displayed. You may use it to toggle the secondary series between the normal and inverted modes. In addition, when a secondary series is displayed, the correlation between the two series is displayed on the upper right above the graph. An input box will also indicate that this is the lag 0 correlation (i.e., they are aligned in time). To view the correlation at different lags, use the two pointer buttons on either side of the lag display to slide the secondary series relative to the original. You will find this to be an extremely useful feature.

The Explore PCA / EOT / Fourier PCA / CCA / Wavelets Panel

With the exception of the Wavelets option, this panel is used to view previously run series decomposition analyses. You must merely indicate which type of analysis to view/explore.

PCA

Principal Components Analysis (also known as Empirical Orthogonal Function analysis) decomposes an image series into a set of underlying components, ordered by the amount of variance they explain in the original series. For each component, a pair of outputs is provided – an image showing the spatial pattern of the component and a graph that shows the degree to which that pattern is present over time⁶. Several variants of PCA are offered. These include options for computing the components in both T-mode and S-mode as well as options for centering and standardization. In addition, Extended PCA (EPCA, also known as EEOF) and Multichannel Singular Spectrum Analysis (MSSA) are offered. For additional details, please see the explanatory discussion of PCA in the Analysis Tab section below.

All series for which a PCA analysis has been run will be listed in the series drop-down selector. Select a series and then the adjacent drop-down will list specific analyses. After you select the specific analysis, the component loading of the first component view will appear as a graph. To view the associated component image, click the Display icon at the upper right of the form. If your component image has both positive and negative values, it is recommended that you use the symmetric instant stretch option on Composer (the middle button) to achieve a proper visual interpretation of the component.

EOT

An Empirical Orthogonal Teleconnection analysis also produces a series of components that consist of a pair of outputs. The graph is the EOT itself while the image indicates the correlation of the series with that EOT. As with the PCA option, options are provided for Extended EOT (EEOT) and Multichannel EOT (MEOT). Please see the explanatory discussion of EOT in the Analysis Tab section below.

All series for which an EOT analysis has been run will be listed in the series drop-down selector. Select a series and then the adjacent drop-down will list specific analyses. After you select the specific analysis, the first EOT will appear as a graph. To view the associated correlation image, click the Display icon at the upper right of the form. If your component image has both positive and negative values (which is very likely), it is recommended that you use the symmetric instant stretch option on Composer (the middle button) to achieve a proper visual interpretation of the correlation image.

⁶ With T-mode analysis (a distinction that will be explained in the Analysis tab discussion) the image is the component and the graph expresses the loading – the correlation between the component image pattern and each image in the series. With S-mode, (the most commonly used mode among those that call this Empirical Orthogonal Function analysis) the graph is the component and the image contains the loadings.

Fourier PCA

Fourier PCA also produces a series of components, but in this case with an image and several possibilities of associated graphs. The default graph is a pseudo-periodogram where the X axis indicates *wave numbers* ranging from the lowest frequency of one sine wave over the whole series to the highest with $n/2$ sine waves over the whole series (where n is the total number of images in the series). The Y axis indicates the amplitude of the wave (as a component score). The associated component image is that which has this combination of waves present (the degree to which waves are present is represented by their amplitude loading). In the upper-right corner of the pseudo-periodogram is a drop-down list where you can choose to focus on just the inter-annual frequencies (those wavelengths that are longer than a year) or the sub-annual frequencies. In addition, there is a temporal loading option. The temporal loading indicates the correlation between the Fourier PCA loading image and each of the original images in the series. Please see the explanatory discussion of Fourier PCA in the Analysis tab section below.

All series for which a Fourier PCA analysis has been run will be listed in the series drop-down selector. Select a series and then the adjacent drop-down will list specific analyses. After you select the specific analysis, a periodogram of the first Fourier component will appear as a graph. To view the associated component image, click the Display icon at the upper right of the form. If your component image has both positive and negative values, it is recommended that you use the symmetric instant stretch option on Composer (the middle button) to achieve a proper visual interpretation of the component.

CCA

Canonical Correlation Analysis (CCA) computes relationships between pairs of series (such as between a series of sea surface temperature and a series of atmospheric temperature). By convention, the two series are known as the independent (X) and dependent (Y) series. However, these roles are essentially interchangeable. Both S-mode and T-mode options are available. In S-mode, the two series need to match in their temporal characteristics but are free to differ in their spatial characteristics. Conversely, in T-mode the two series need to match in their spatial characteristics but they are free to vary in their temporal specifications.

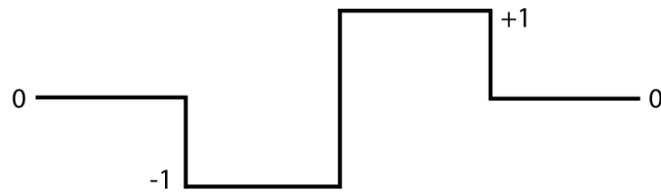
The results of a CCA analysis can be viewed from the Explore tab. Select the X and Y variates from the appropriate drop-down menus. Then select the analysis that was run from the CCA drop-down menu. Three viewing options can be selected from the Statistics drop-down menu. The Canonical Correlation option shows the correlations between corresponding canonical variates. With the Variance Explained option, it will graph the variance explained for canonical variates for the selected series (X or Y). Finally, with the Variate option, it will show the actual canonical variate as a graph for the selected series (X or Y). With this option, the *map* icon will light up indicating that you can display the loading images associated with each variate. Two loading images are displayed for each. The first is the loading (correlation) between the variate and its own series (known as the *homogenous* correlations) while the second is the loading between the variate and the other series (known as the *heterogenous* correlations). CCA is thus very effective for looking for coupled modes between series.

Wavelets

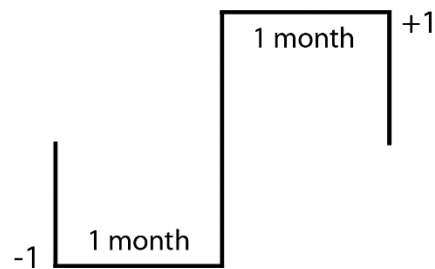
The wavelet view requires special explanation. Wavelets are used for multiple purposes. A common application is data compression. However, the application here is the search for patterns of anomalous events over time. One such pattern would be a perfect cycle, for which Fourier Analysis is a good detection tool. A perfect cycle is one which oscillates in a consistent manner over time, such as a sine wave. However, not all oscillations are perfect, nor are they necessarily sine waves or continuous over time.

A wavelet is a little wave, or perhaps better stated, a briefly appearing wave. Wavelets can be of any form. For example, one could use a sine wave as a wavelet. In practice, there are a variety of wavelets that are used for special reasons. In ETM, we have introduced an Inverse Haar wavelet that leads to a very simple form of interpretation in the context of image time series.

The Inverse Haar wavelet looks like this:



The graph represents a series of weights over four adjacent samples for a single pixel in time. However, since the weight of 0 applies to all time samples before the first and after the fourth, it's probably simpler to think of it as applying to a pair of adjacent samples over time as follows:



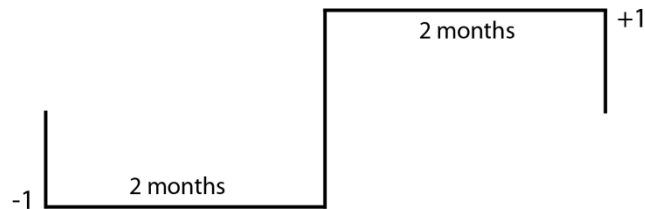
Applying the wavelet filter to any pixel allows us to determine the degree to which the wavelet is present at any moment in time. Let's assume that for a particular pixel, we have the following values over a sequence of seven months:

5 7 15 12 8 9 3

Moving the filter to the first two positions, we multiply the 5 by -1 and the 7 by +1. Then we add the results which yields +2. Now the filter is slid one time sample to the right. We multiply the 7 by -1 and the 15 by +1 which when added together yields +8. In essence, we're saying that the wavelet is present four times as strongly at this second position. Now move it to the right one more time step. Here we see something different. We multiply the 15 by -1 and the 12 by +1 yielding -3 after summation of the results. This implies that the wavelet is present in its opposite orientation with an amplitude of 3.

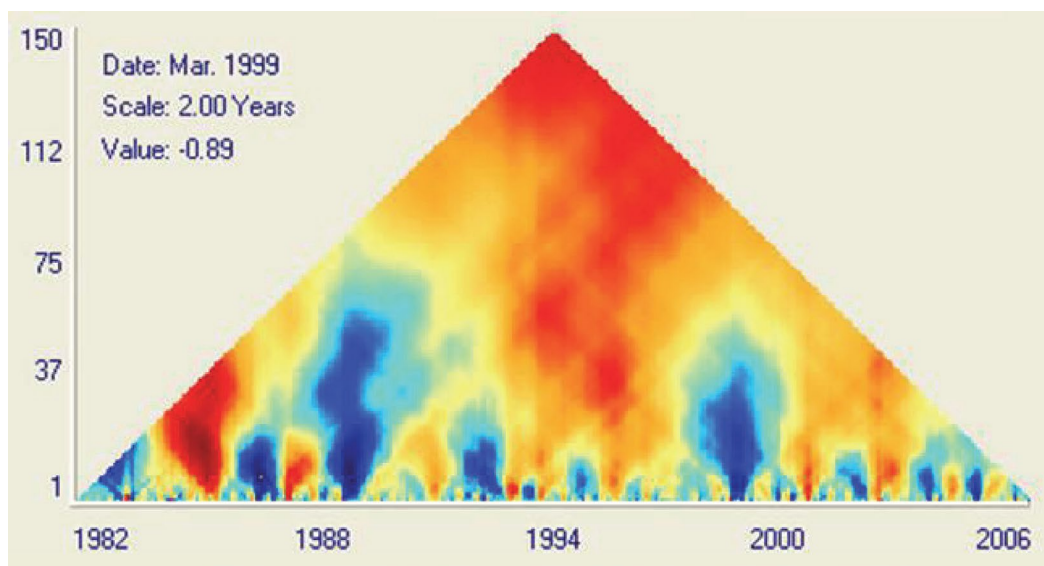
Why is this of interest? If you look at the nature of the mathematical operations undertaken at each time step, it equates to calculating the rate of change from one point of time to the next. The result of the Inverse Haar filter for the first two dates indicates that the rate of change is +2, i.e., a gain of two units. The filter thus results in an expression of monthly gains and losses.

In the next step, we increase the width of the filter to span over four months:



Applying this filter, we will first average over the months spanned by each of the two month filter sections. Thus applying this to the first four months in the series, the 5 and the 7 are averaged to become 6 and the 15 and 12 are averaged to become 13.5. Then multiplying the 6 by -1 and adding it to 1 times 13.5 yields a result of +7.5. This indicates that there was a gain of 7.5 units between the first 2 month period and the second 2 month period. Then the filter is slid just one month. The next calculation is thus on the months with values of 7, 15, 12 and 8, yielding a result of -1. Sliding this filter only a single month produces what is known as a Maximal Overlap Discrete Wavelet Transform (MODWT). The use of a MODWT filter has advantages in that it can handle any length of series (and not just those that have lengths that are a power of 2) and it significantly reduces artifacts that can be introduced by the shape of the filter⁷.

Repeating this process of increasing the number of months in each half of the Haar filter by one (thus the third filtering would calculate the gain between successive groups of three months) produces what is known as a multi-resolution analysis (often abbreviated MRA). With each level, we are changing to a progressively coarser time scale with two fewer samples at each scale (because we have no neighbors for the calculation of the first and last samples at each scale). If we array the results of our analysis in a graphic form, it forms a pyramid, with the top of the period corresponding to the coarsest scale where we are calculating the gain between the first half of the series and the last half.



In this example, we see time along the X axis and scale (expressed as the number of months in each half of the Haar filter) along the Y axis. In ETM, you can move the mouse over the diagram and it will tell you the date, the scale and the value of the wavelet transform that the color represents.

⁷ See Chapter 5 on the Maximal Overlap Discrete Wavelet Transform in Wavelet in Percival, D.B., and Walden, A.T., (2000) *Wavelet Methods for Time Series Analysis* (Cambridge University Press).

The importance of using the Inverse Haar filter is that the values it calculates have a very simple interpretation – they are simply gains in the original units. Thus in the example above based on an analysis of sea surface temperatures in the Labrador sea, the value represents a loss of 0.89 degrees Celsius (because the gain is -0.89). Notice the intense cooling that took place in 1989. At the 1 month scale, the cooling took place over only a few months (looking horizontally), but looking vertically, we see that its impact lasted for approximately 6 years (72 months). Many of the warmings and coolings are evident at both the finest and coarsest scales. They thus have a tree-like appearance in the wavelet diagram. However, not all do. In the middle of the diagram, we see evidence of warming that does not manifest itself at finer scales. This represents a very gradual warming that is lost in the higher variability that occurs at finer scales.

The Explore Temporal Profiles Panel

The temporal profiling panel allows you to examine the values from an image series for a defined region of interest. The region of interest can be defined as a circular sample region or by means of a vector feature (typically a polygon). ETM will then graph the summary values of all pixels in the sample region over time. The mean, median, minimum, maximum, range, sum or standard deviation summary values may be graphed. In addition, trend lines can be added.

Note that any profile can be saved as an index series. This can be very useful when used with the Linear Modeling tool. For example, if you have a series on precipitation anomalies and you create a profile for a region of interest, as long as the length of the series is the same, the saved profile could be used as the independent variable with a sea surface temperature series as the dependent variable to locate areas of the ocean with a similar pattern of temperature anomalies over time.

The Explore Series Relationships Panel

The Explore Series Relationships panel stores icons of linear modeling analyses already run. Select the series from the drop-down list. Only series for which linear models have been run will be listed.

Clicking on the icon will display one or more analyses, depending upon which outputs were requested when the original analysis was run. Note that only those images most typically viewed are displayed. For example, when the slope and intercept option is chosen, only the slope image is displayed. To view other images related to your analysis, use TerrSet Explorer or DISPLAY Launcher to locate them within your project and subsequently view them.

The Explore Trends Panel

The Explore Trends panel is used to explore two kinds of trends – interannual trends and seasonal trends. The interannual trends subpanel is similar to the Explore Series panel in that it displays icons for previously run analyses. Clicking an icon will redisplay that analysis.

The Seasonal Trends sub-panel needs more explanation. If you have not already done so, read the section on the Seasonal Trend Analysis Panel in the Analysis Tab below. The drop-down box lists all series for which an STA analysis has been undertaken. Clicking the Display icon will display the Phases and Amplitude images. Generally, the Amplitudes images contain the most information.

On either the Amplitudes or Phases images, colors other than neutral gray represent trends in the seasonal curve. Although a legend could technically be created for these images, they are meaningless because they represent trends in selected parameters that describe the shape of these curves. Thus this panel provides an interactive interpretation device that allows you to look at the shape of the curve and how it is changing over time.

Adjacent areas that have the same color on these images will be those that are going through similar changes in either the Amplitudes or Phases of their seasonal curves. You can draw an interactive interpretation of the trend by defining a sample region

and selecting a trend element to graph (fitted curves is the default)⁸. The sampled region can be defined either as a circular region or by selecting a vector feature. For the fitted or observed curves, the green curve shows the modeled trend for the start of the series and the red curve shows it for the end of the series. For all other options, a graph of the parameter along with the Theil-Sen median slope is presented.

A common application of STA is the examination of trends in vegetation phenology using remotely sensed vegetation index image series. As a result, an optional Green up/down output is available. *Green up* refers to the time of maximum greening while *green down* refers to the time of maximum loss of green (chlorophyll) in vegetation. The default is set at the point where the amount of greening exceeds 40% of the full trajectory from minimum green to maximum green. This can be changed to any other value desired. If selected, this output will show the date and time this threshold is passed and the net change in green up or green down over the length of the series.

The Analysis Tab

The Analysis tab is the heart of ETM. Each of the panels is devoted to an analytical process that yields results that can be viewed and reviewed from the panels on the Explore tab.

Critical Things to Know!

1. In general, it is important to consider whether the procedure to be used should be deseasoned or not. For example, the trend analysis procedures on the Series Trend Analysis panel and Linear Models should normally be applied only to deseasoned series. The STA procedure *cannot* and Fourier PCA *should not* be applied to deseasoned series. EOT and PCA are normally applied to deseasoned data, but there is nothing to prevent them from being used with data containing seasonality.
2. If there are areas of no data or background values that apply to all images in a series, the analysis will run faster if you specify a mask image.

The Series Trend Analysis Panel

The Series Trend Analysis panel is used to search for the presence of long-term trends. On the Explore tab, these are referred to as interannual trends, but in reality, this name would only be correct if the series covered more than one year.

ETM offers five types of trend analysis and one form of trend significance testing as follows:

Linearity

This procedure maps out the coefficient of determination (r^2) from a linear regression between the values of each pixel over time and a perfectly linear series. The result is a mapping of the degree to which a linear trend is present.

Linear Correlation

⁸ The number of different trend elements selected affects the speed with which ETM can calculate the necessary information. Therefore, by default, only the information necessary for the fitted curves is extracted for the sample region. You can select additional elements by using the checkboxes provided. The one you will most typically want is the observed curve information.

This maps out the Pearson Product-Moment linear correlation between the values of each pixel over time and a perfectly linear series. This is a commonly used form of trend analysis, but it is sensitive to noise in short series.

Linear Trend (OLS)

This is the slope coefficient of an Ordinary Least Squares regression between the values of each pixel over time and a perfectly linear series. The result is an expression of the rate of change per time step. Thus, if your data are monthly, it expresses the rate of change per month.

Median Trend (Theil-Sen)

This is a robust non-parametric trend operator that is highly recommended for assessing the rate of change in short or noisy series⁹. It is calculated by determining the slope between every pairwise combination and then finding the median value. For example, with a 20 year sequence of monthly data, a total of 28,680 slopes would be evaluated at every pixel. It thus takes a lot longer to calculate than the trend procedures indicated above. For long series, the result is often identical to the Linear Trend (OLS) output. However, for short or very noisy series, the result can be quite different and is more reliable. An interesting feature of the Median Trend is its *breakdown bound*. The breakdown bound for a robust statistic is the number of wild values that can occur within a series before it will be affected. For the Median Trend, the breakdown bound is approximately 29%. Thus the trends expressed in the image must have persisted for more than 29% of the length of the series (in time steps).

Monotonic Trend (Mann-Kendall)

This is a non-linear trend indicator that measures the degree to which a trend is consistently increasing or decreasing. It has a range from -1 to +1. A value of +1 indicates a trend that continuously increases and never decreases. The opposite is true when it has a value of -1. A value of 0 indicates no consistent trend. It is calculated in a similar fashion to the Median Trend. All pairwise combinations of values over time are evaluated at each pixel and a tally is made of the number that are increasing or are decreasing with time. The Mann-Kendall statistic is simply the relative frequency of increases minus the relative frequency of decreases¹⁰.

Mann-Kendall Significance

This option produces a pair of images – a significance image expressed as Z scores and a second image that expresses the probability that the observed trend could have occurred by chance. Strictly speaking, this option is expressing the significance of a Mann-Kendall trend. However, it is commonly used as a trend test for the Theil-Sen median slope operator as well.

The STA (Seasonal Trend Analysis) Panel

⁹ See Hoaglin, D.C., Mosteller, F., and Tukey, J.W., (2000) *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library Edition, (New York: Wiley).

¹⁰ With a Mann-Kendall statistic, the data series is the dependent variable and time is the independent variable. When the independent variable is something other than time, the statistic is known as Kendall's Tau. In that case, one looks at whether the two variables are both increasing or both decreasing (known as a concordance) or whether one is increasing while the other is decreasing (a discordance) between every pairwise combination of observations. Tau is then the relative frequency of concordances minus the relative frequency of discordances.

Seasonal trend analysis (STA) is a new analytical technique developed by Clark Labs¹¹. It uses two stages of time series analysis to map out trends in the shape of the seasonal curve. It can be used with any series that exhibits seasonality.

In the first stage, each year of data is submitted to a harmonic regression to yield the following shape parameters:

- An annual mean image (sometimes called Amplitude 0).
- An image expressing the amplitude of the annual cycle (a sine wave with one cycle over the year), known as Amplitude 1.
- An image expressing the phase angle of the annual cycle (an indication of where on a sine curve the beginning of the series is located). This is known as Phase 1.
- An image expressing the amplitude of a semi-annual cycle (a sine wave with two cycles over the year), known as Amplitude 2.
- An image expressing the phase angle of a semi-annual cycle (an indication of where on a sine curve the beginning of the series is located). This is known as Phase 2.

These five parameters can describe an exceptionally large family of curves. By using only two harmonics and the mean, high frequency noise and variability are rejected.

The first stage results in five images per year – one for each of the five shape parameters. Then in the second stage of the analysis, a Theil-Sen median trend is run on each of the shape parameters over the total number of years in the series. Since the median trend has a breakdown bound of 29% of the length of the series, interannual trends shorter than this length are also rejected. Thus the result of the two stages yields a focus on long-term trends in the seasonal curve while rejecting both high frequency noise and low frequency variability.

To visualize the results of this analysis, two forms of color composite are created – an Amplitudes image and a Phases image. Each is formed by assigning selected shape parameters trend images to the red, green and blue primary colors. The Amplitudes image assigns RGB to Amplitude 0, Amplitude 1 and Amplitude 2 respectively. The Phases image assigns RGB to Amplitude 0, Phase 1 and Phase 2 respectively.

The interpretation of these trend maps in terms of the seasonal curves is very difficult. As a result, ETM has a special exploration tool that allows you to visualize the curves as well as the nature of the trend. Please see the section on the Explore Trends panel above for more explanation. We also recommend that you complete the tutorial on STA to achieve a full appreciation for this tool.

The PCA (Principal Components Analysis) / EOF Panel

Principal Components Analysis (PCA) is also known as Empirical Orthogonal Function (EOF) analysis. It is a very powerful technique for the analysis of variability over space and time, and ETM offers a wide variety of variants including Extended PCA (EPCA or EEOF) and Multichannel Singular Spectrum Analysis (MSSA).

PCA / EOF

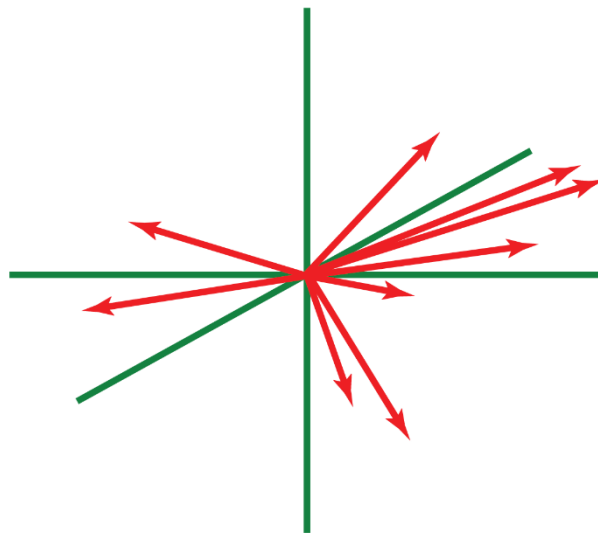
The images in a time series highly correlate with one another from one moment of time to the next. PCA transforms the series into a set of *components* that are orthogonal (i.e., independent of each other) in both time and space. They are also ordered in terms of the amount of variance that they explain from the series. In theory, one can produce as many components as there are images in the

¹¹ See: Eastman, J.R., Sangermano, F., Ghimire, B., Zhu, H., Chen, H., Neeti, N., Cao, Y., Machado, E.A., and Crema, S., (2009) Seasonal Trend Analysis of Image Time Series, *International Journal of Remote Sensing*, 30(10), 2721-2726.

original series. However, in practice, almost all the variance can be explained by only a small number of components, with the remainder expressing noise and high frequency variations.

The easiest way to understand PCA is to think about the time series of values for a single pixel across time as a vector. If you imagine that each date represents a dimension, then the series can be completely described by a single point in that space. For example, image three months (January, February, March) with mean air temperature values of 15, 18, 22, then this point would be located at position 15 on the January axis, 18 on the February axis and 22 on the March axis. The vector is then formed by joining this point with the origin of the space. Typically, of course, we need hundreds of dimensions to describe real-life series. However, this is difficult to visualize, so we will use the example of three.

The image is made up of many pixels, so we will in fact have a space occupied by many vectors (as in the figure below). The correlation between any pair of vectors is inversely proportional to the angle between them (in fact, the cosine of that angle is equal to the correlation coefficient). The first component is the average vector (i.e., a vector that is as close as possible to the entire collection of vectors). It is known as an *eigenvector* (meaning *characteristic* vector) and its length is known as the *eigenvalue*, which expresses the amount of variance it explains. The cosine of the angle between this eigenvector and each pixel vector indicates its *loading* on the component – i.e., the pixel vector's correlation with the eigenvector.



After the first component has been calculated, its effects are removed from the pixel vector field¹². The new vectors thus express the residuals after removing the effects of the first component. Then the process is repeated to extract the second component, and so on. Since each successive component is calculated on residuals, they will all thus be independent of each other. Ultimately it is possible to extract as many components as there are pixel vectors.

Efficient computation of PCA is actually done with matrix algebra and starts with a matrix of inter-relationships, expressed either as correlations or as covariances. Two modes are possible. If the procedure were to follow the logic specified (conceptually) above, the analysis would be known as S-mode and the input matrix would be a square matrix of the relationship between each pixel, over time, and every other pixel over time. Thus, for an image with a thousand pixels, the input would be a 1000 x 1000 matrix of inter-relationships. This mode is the dominant one employed by atmospheric and ocean scientists. In S-mode analysis, the component is a vector (expressed as a graph over time) and the loadings are expressed as maps of correlation coefficients.

¹² Although the actual calculation of components is not done this way, this step is equivalent to calculating the partial correlations between each pair of pixel vectors, while removing the effects of the component.

A second option would be to start with a matrix that expresses the relationship between images. This is known as T-mode analysis and is the dominant approach that has been used by geographers. With a series of 300 images, for example, the input would be a 300 x 300 matrix of correlations or covariances expressing the relationship between every image and every other image in the series. With T-mode analysis, the component is an image and the loadings are expressed as a graph (i.e., a vector).

Strictly speaking, the procedure outlined here is a *standardized* PCA since correlations express the covariance between variables standardized by their variances. It is also possible to calculate PCA's by using the variance/covariance matrix as the starting point rather than the correlation matrix, in which case it is called an *unstandardized* PCA. The difference is subtle but important. With a standardized PCA, all the variables are put on an equal footing since their variances are effectively equalized. With unstandardized PCA, variables will have weight proportional to their variance. For image time series analysis, we have generally found that standardized PCA gives the most easily interpreted results. Therefore, this is the default mode in ETM.

Another parameter of importance is what is called *centering*. Centering refers to the subtraction of a mean. In S-mode analysis, the mean of each pixel over time is subtracted during the normal processing of PCA, while in T-mode analysis, the mean of each image is subtracted from all pixels in that image. This is a natural part of the mathematics of the PCA process and is incorporated directly into the formulas you would encounter in any textbook. However, centering does have a major impact on time series analysis. Centering over space (as is normally done in T-mode) detrends the data over time while centering over time (S-mode) detrends the data over space. Detrending over space can be thought of as a process of removing the time-invariant geography. Detrending over time may seem self-evident, but the kind of trend is effectively the space-invariant pattern over time.

In interpreting components, you should always consider both the spatial and temporal patterns together. One shows you a pattern (in space or time) and the other indicates when or where it was present. Note that negative component values may be associated with negative loadings, which thus imply a positive anomaly.

PCA has many uses, but in image time series analysis, it is primarily an exploration tool. It is remarkably effective in organizing the underlying sources of variability in the data. However, components aren't always pure. If at any level in the analysis there are two or more sources of variability that have roughly equal weight, then PCA will tend to produce mixed components. The usual solution to this is known as *rotation* of the axes. There are a number of procedures for doing this that require scientific judgment. In ETM, we have elected to go a simpler route, known as Empirical Orthogonal Teleconnections (EOT). EOT produces a result similar to an oblique component rotation.

Finally, a word about masks. ETM provides the ability to specify a mask. This image should have 1's in pixels that should be included in the calculation and 0's otherwise. However, this only affects the calculation of the components. When the component images are produced, the transformation is applied to all pixels. You can decide for yourself if you wish to apply the mask to the outputs (simply multiply the components by the mask).

Note that the results from the PCA tool can be viewed on the Explore tab.

EPCA / EEOF

Extended PCA (also known as Extended EOF) allows for the analysis of multiple time series simultaneously. It is most commonly used in the search for coupled modes of variability. For example, one might use EPCA to search for coupled patterns between sea surface temperature, lower tropospheric temperature and middle tropospheric temperature. A coupled mode is a pattern that occurs across multiple series at the same time (assuming that it is not coincidental).

EPCA is running in a similar fashion to PCA except that you specify the number of series and select for each the series name and mask to use. As with PCA, the results are viewed with the Explore tab. For each extended component, a single temporal pattern will be displayed, but an image result will be displayed for each series.

Note that with EPCA, it is not necessarily the case that each component represents a coupled mode. If the values in any of the images are not strong, then it implies that that component is not strongly found in that series. Also note that ETM offers you choices between S-mode and T-mode, standardized and unstandardized, and whether to center. The meaning of these is the same as with PCA.

MSSA

Multichannel Singular Spectrum Analysis (MSSA) is a special case of EPCA. Like EPCA, MSSA analyzes multiple series simultaneously. However, in this special case, the series involved are multiple versions of the same series, but at different lags. The number of lags is known as the *embedding dimension*. For example, if the series starts in January and the embedding dimension is 3, EPCA is given three series. The first starts in January, the second starts in February and the third starts in March. The MSSA option in ETM does this all automatically and it also truncates images off the back end of series as necessary to make the lengths of the series match (in the example given, the first series would end in October, the second would end in November and the third would end in December). This collection of lagged series is known as a *phase space*—a representation of all possible states of a system.

The primary intention of MSSA is to seek patterns that evolve in both space and time. Contrast this to PCA and EOT which assume that the pattern only evolves in time. Thus MSSA can be used to search for moving phenomena. In addition MSSA is particularly adept at the analysis of oscillatory patterns. In these cases, pairs of components will be found that are highly similar but which are out of phase (i.e. they are uncorrelated at lag 0 but are highly correlated at some lag other than 0). These components are said to be in *quadrature* and are effectively *basis vectors* like the sine and cosine pair of Fourier Analysis and the X and Y axes of Cartesian space (axes that can, together, describe every possible state of the system). Note that MSSA is only capable of detecting oscillations with a period less than or equal to the embedding dimension. Also note that it is generally easier to interpret the results if the embedding dimension is specified as an odd number. Thus, if the embedding dimension is specified as 5 the results will include 5 lags including 2 time periods before, 2 time periods after, plus the time period symbolized by the temporal graph. To search for annual patterns in monthly data, the suggested embedding dimension would thus be 13.

The EOT (Empirical Orthogonal Teleconnections) Panel

The name EOT¹³ comes from its original area of application – the study of climate teleconnections. However, the technique has general utility in the exploration of image time series. It is a brute force technique that is simple to understand and produces results that are essentially identical to those one would expect from an obliquely rotated PCA. The EOT's are thus orthogonal in time but not necessarily in space. Note that like PCA, extended modes are available with EOT including Extended EOT (EEOT) and Multichannel EOT (MEOT). In addition, an experiment form T-mode EOT is now offered (TEOT).

EOT

In the default standardized implementation of EOT, pixels are treated as vectors over time. Each pixel is examined to determine the degree to which its profile over time can explain the variability of all other pixels over time. It does this by calculating the coefficient of determination (i.e., the squared correlation) between that pixel's profile and each of the other profiles. Thus with n pixels, $n-1$ correlation analyses are run. The $n-1$ coefficients of determination are then summed. This process is then repeated over all pixels in turn resulting in $n(n-1)$ correlation analyses. At the end of this process, the pixel with the highest sum of r^2 becomes the location of the first EOT. The profile for this pixel over time is thus the first EOT and the values are in the same units as the original data.

After the first EOT is found, a residual series is created, removing the effect of that EOT. Then the process is repeated all over again to find the next EOT. However, the values of the next EOT are now taken from the residual series. The units are still the same, but they

¹³ Van den Dool, H. M., Saha, S., Johansson, A., (2000) Empirical orthogonal teleconnections. *Journal of Climate*, 13:1421-1435; Van den Dool, H., (2007) *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, New York. Note that we STRONGLY recommend this book. It is exceptionally well written and provides many useful insights into image time series analysis.

represent anomalies from the first EOT. Then a residual series is taken from the residuals and the process is repeated again. Once all of the requested EOT's have been calculated (as temporal profiles) the full set is used as independent variables in a multiple regression with the original series to get a set of partial correlation images for the spatial pattern associated with each EOT.

Some important notes about the EOT procedure:

1. As a brute force technique, EOT's can take a considerable amount of time to calculate. If you take the case of a global series with a resolution of 1 degree, there are 64,800 pixels per image. If you calculate 10 EOT's, it requires almost 42 billion correlation analyses! Depending upon the resolution of your series, an analysis commonly takes hours (in fact, we commonly run them overnight).
2. Given the large amount of spatial dependence in geographic data, we generally recommend that you sample the data rather than calculate every pixel. A sampling rate of 1 means that you want to compute the EOT's using every pixel. A sampling rate of 2 implies that they will be calculated by looking at every second pixel along every second row, and so on. We recommend using an odd number as the sampling rate so that the EOT location corresponds to a specific pixel.
3. EOT also creates a vector point file showing the specific locations of each EOT. It will bear the same prefix as the other outputs for the analysis and can be found in the "components" sub-folder of your series.
4. EOT is effectively an S-mode analysis. TEOT is now offered as a T-mode equivalent.

An unstandardized variant of EOT is also provided. In this case, the coefficient of determination is weighted by the variance of each pixel profile during calculation. The difference between standardized and unstandardized EOT's is the same as it applies to PCA – the standardized version preferences the quality of the relationship expressed by the EOT (by giving equal weight to all pixels in its calculations) while unstandardized EOT is preferencing relationships with magnitude.

TEOT

As mentioned above, EOT is an S-mode analysis. It searches for locations in space where the pattern over time is highly descriptive of other patterns in time. T-mode EOT (TEOT) is the dual of EOT. It searches for images that are highly similar to other images over time. Unlike S-mode which is looking for patterns over time, TEOT is looking for patterns over space. To do so, it looks at each image in the series and correlates it with every other image. The image which has the highest correlation with all other images in the series becomes the first TEOT. As an example, consider a series of images of anomalies in sea surface temperature (SST). The first TEOT would likely look like El Nino. In fact, it would represent a single El Nino. Because a single image can have local anomalies as well as large regional values, you can select to create an average of the top n images (for example, several El Ninos). The default is 3. Once a TEOT is created, a residual series is then created with that pattern removed. The process then repeats on the residual series, and so on. Note that this procedure is experimental and is offered for comment.

Note that while EOT's are independent in time, but not in space, TEOT's are independent in space, but not in time.

CEOT

A second variant of EOT is provided that we call a Cross-EOT (Van den Dool, 2007 refers to it as EOT2). With a regular EOT we are looking for locations in a series that have good explanatory power in describing variance in other locations within the same series. With Cross-EOT we are looking for locations in one series that can best describe variance in another series. Thus, for example, you could examine a sea surface temperature series to find locations that can explain temperature anomalies on land. The only requirement is that the two series have the same length and nature (e.g., if one is monthly and has 300 images, the other must also be monthly with 300 images). Some important points with Cross-EOT:

1. BE VERY CAREFUL ABOUT SPURIOUS CORRELATIONS!!!!!! If you look at every location in one series and compare it to every location in the other, the likelihood that you will find some location that explains some part of the sequence in the other series is very high. Thus it is possible to create Cross-EOT's that are absolutely meaningless. This is a problem shared with similar techniques such as Canonical Correlation Analysis, for which a recommended practice is prefiltering the data to focus on the major elements of variability in the series. The Inverse PCA denoising filter on the Preprocessing tab can be used for this.
2. Cross-EOT produces one set of EOT graphs, but two sets of images—one for each of the two series involved.

EEOT

Extended EOT (EEOT) is the EOT equivalent of EPCA/EEOF. With EEOT multiple series are analyzed, but unlike CEOT it is not trying to find a single location where the temporal pattern in one series can explain the most variance in a second series, with EEOT it is looking for a location in any series that can explain the most variance in all of the series being considered combined. In essence, it is looking for coupled modes across series and its operation is essentially identical to that of EPCA/EEOF. Note that because EEOT is a brute force technique, it can take a very long time to calculate. Please plan accordingly. You may wish to do an initial analysis in EPCA/EEOF and then wait until you can process the data overnight or over a weekend.

MEOT

Multichannel EOT (MEOT) is the EOT equivalent of MSSA. With MEOT multiple lags of a single series are examined to find a location in any lag that can explain the most variance in all of the lags combined. It is thus identical in spirit to MSSA, but uses the logic of EOT for computation.

The number of lags is known as the *embedding dimension*. For example, if the series starts in January and the embedding dimension is 3, EPCA is given three series. The first starts in January, the second starts in February and the third starts in March. The MEOT option in ETM does this all automatically and it also truncates images off the back end of series as necessary to make the lengths of the series match (in the example given, the first series would end in October, the second would end in November and the third would end in December). This collection of lagged series is known as a *phase space* – a representation of all possible states of a system.

Like MSSA, the primary intention of MEOT is to seek patterns that evolve in both space and time. Thus MEOT can also be used to search for moving phenomena. In addition MEOT is adept at the analysis of oscillatory patterns. In these cases, pairs of MEOT's will be found that are highly similar but which are out of phase (i.e., they are uncorrelated at lag 0 but are highly correlated at some lag other than 0). These MEOT's are said to be in *quadrature* and are effectively *basis vectors* like the sine and cosine pair of Fourier Analysis and the X and Y axes of Cartesian space (axes that can, together, describe every possible state of the system). Note that MEOT is only capable of detecting oscillations with a period less than or equal to the embedding dimension. Also note that it is generally easier to interpret the results if the embedding dimension is specified as an odd number. Thus, if the embedding dimension is specified as 5 the results will include 5 lags including 2 time periods before, 2 time periods after, plus the time period symbolized by the temporal graph. To search for annual patterns in monthly data, the suggested embedding dimension would be 13. Note that because MEOT is a brute force technique, it can take a very long time to calculate. Please plan accordingly. You may wish to do an initial analysis in MSSA and then wait until you can process the data overnight or over a weekend.

Note that the results of EOT, EEOT and MEOT can be explored on the Explore tab.

The Fourier PCA Spectral Analysis Panel

This is an experimental module that was designed as a way of organizing the output of the TFA (Time Series Fourier Analysis) module. TFA decomposes a series into a set of sine waves with frequencies ranging from 1 wave over the entire series to $n/2$

complete waves over the series. This produces a large number of amplitude and phase images. Without having prior interest in specific waves, this can be a daunting image set to examine. With Fourier PCA, the amplitude images from TFA (which is called automatically by ETM) are fed into an unstandardized PCA. The components from that analysis thus indicate commonly occurring patterns of waveforms.

The loadings in Fourier PCA express the relative strength with which different frequencies are present. Thus the X-axis represents frequency indicated by the number of the harmonic (i.e., a value of 2 represents two complete waves over the entire series). This is similar in spirit to a Periodogram, but since the Y axis doesn't represent amplitude directly, we call it a *Pseudo-Periodogram*.

Because it is often difficult to understand the implication of these wave patterns, an additional analysis is added at the end. Each component is correlated against every one of the original images in the series. This generates a temporal "loading" that expresses when the pattern was present. However, it is not a true loading as in a PCA, so it is also best described as a *Pseudo-Loading*.

Some important notes about Fourier PCA:

1. Phase information is not used in the analysis. Thus waves that occur at different times will be lumped together. This implies that the pseudo-loading is limited in its ability to fully represent the timing of a pattern. Similarly, this implies that the series may never at any one time look like the pattern portrayed. The Tutorial will help clarify this.
2. Since phase information is discarded, it is theoretically possible to detect moving phenomena. We have experimentally proved this and have been successful in detecting ocean eddies with this tool.
3. If any pair of amplitude images is perfectly, or nearly perfectly, correlated, a singular matrix is encountered and computation is not possible.
4. If you have background areas, apply a mask to remove them from consideration. This can reduce the likelihood of a singular matrix.
5. The cutoff frequency allows you to exclude high frequency waves from the analysis. The cutoff number refers to the harmonic.

Note that the results of Fourier PCA can be viewed on the Explore tab. Also note that this is an experimental procedure and you are cautioned to use it at your own risk. We welcome constructive feedback.

The CCA Panel

Canonical Correlation Analysis (CCA) computes relationships between pairs of series (such as between a series of sea surface temperature and a series of atmospheric temperature). In effect, CCA undertakes two principal components analyses, one for each series, such that the components it produces for each (known as *canonical variates*) are correlated to one another to the maximum extent possible. By convention, the two series are known as the independent (X) and dependent (Y) series. However, these roles are essentially interchangeable. Both S-mode and T-mode options are available. In S-mode, the two series need to match in their temporal characteristics but are free to differ in their spatial characteristics. Conversely, in T-mode the two series need to match in their spatial characteristics but they are free to vary in their temporal specifications.

One of the problems with CCA of image time series is that the images within each series are highly intercorrelated. This commonly leads to singular matrices such that a solution cannot be achieved. A remedy for this that has become common in atmospheric science is to apply a pre-processing step whereby a Principal Components is run on the series and the components (which are, by definition, independent) are subsequently analyzed by CCA¹⁴. This is done automatically in the ETM implementation of CCA. In fact, the terms S-

¹⁴ For a discussion of this approach using S-mode, see Barnett, T.P. and Preisendorfer, R. (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperature determined by canonical correlation analysis. American Meteorological Society, 115, 1825 - 1850. The T-mode

mode and T-mode initially refer to the nature of the pre-processing PCA that is conducted before the CCA operation. However, with centering, the selected mode effectively carries through to the CCA stage.

The results of a CCA analysis can be viewed from the Explore tab and the procedures for doing so are discussed above in the section describing that tab. The outputs of CCA include:

1. *Canonical Variates*-- the “components” produced by CCA.
2. *Canonical Correlations*-- the correlations between the Canonical Variates. CCA ensures that these are the maximum possible.
3. *Homogeneous Correlations*-- the correlations between each variate and the series from which it was computed.
4. *Heterogenous Correlations*-- the correlations between each variate and the other series.
5. *Variance Explained*-- the degree to which each variate explains the variance present in the series from which it was drawn.

Strongly coupled patterns will be those in which all of these values are high. For further information, see the tutorial sequence on CCA.

The Linear Modeling Panel

The Linear Modeling tool allows you to examine relationships between series. At this time, the dependent series is always an image series while the independent series can be either image or index series (but not mixed). It uses standard multiple regression analysis to produce its outputs and is thus subject to all of the normal caveats about regression than can be found in a standard text on multivariate statistical regression.

One of the primary uses of the Linear Modeling tool is to map the areas impacted by a particular phenomenon such as a climate teleconnection like El Nino. If you’re trying to sort out the pattern of several teleconnections, analyze them simultaneously and select the partial correlation option. This will produce a partial correlation image for each relationship with the effects of the others removed.

Note that series relationships can be analyzed at different lags. Lag 0 implies that the dependent and independent series are being compared at corresponding time steps. A negative lag shifts an independent variable to an earlier time. If you think of an event, such as the December peak of El Nino, then if your independent variable is an index to El Nino (such as the Southern Oscillation Index) you would be looking at the relationship *before* the main event. You would do this if, for example, you were looking for leading indicators of El Nino in your dependent series. Commonly, a negative lag is called a *lead* and a positive lag is simply called a *lag*.

The results of a Linear Modeling analysis can always be re-examined by going to the Explore Series Relationships panel on the first tab. Also, when the Linear Modeling tool finishes, it will show you one result as a signal that it has finished. However, in cases where you’ve selected multiple outputs (such as with partial correlation) you will need to go to the Explore Series Relationships panel to see the full set of pertinent main results. Also, as noted in the section on the Explore Series Relationships panel, some outputs (like intercept images) are not displayed by the analysis icons and will need to be displayed from TerrSet Explorer.

implementation in ETM is new and has been developed by N. Neeti

The Preprocess Tab

The Missing Data Interpolation Panel

Earth observation imagery commonly has missing data – in fact, a lot of missing data. The most common reason is because of clouds, although transmission dropouts and gaps between scans are also a source. However, many of the analytical procedures provided by ETM are sensitive to the presence of missing data. The Missing Data Interpolation panel offers a few utilities to alleviate this situation.

Missing data are identified as those that do not fall within a valid range as specified by indicating the minimum and maximum allowable values. For all options, you can elect to create a Boolean image that defines pixels that still have one or more missing values over the series. This is a good way to check on your progress. The options provided for interpolation include:

Harmonic Interpolation

This option is closely based on the procedure known as HANTS (Harmonic Analysis of Time Series) by Roerink, et al., (2000)¹⁵. The procedure is best for filling missing data in mid-latitude areas. It is not suitable for high-latitude or desert regions with long periods of uniform response (e.g., snow cover) or for tropical rainforest regions where seasonality is questionable or extremely subtle.

The basic logic of the procedure is as follows. In mid-latitude areas, the seasonal curve is very well described by the additive combination of a small set of sine waves. This can be modeled by using either Fourier Analysis or Harmonic Regression. As in the HANTS procedure, Harmonic Regression is used because it can accommodate missing data and does not require that the data be at a consistent interval, i.e.,

$$y = \alpha_0 + \sum_{n=1}^{n=2} \left\{ \alpha_n \sin\left(\frac{2\pi nt}{T}\right) + b_n \cos\left(\frac{2\pi nt}{T}\right) \right\} + e$$

where y is the series value, t is time, T is the length of the series, n is the number of harmonics to be used in the regression, e is an error term and α_0 is the mean of the series. The Julian dates associated with each step in the series cycle (contained in the .tsf file) provide the time information needed for the regression.

For each missing value in the series the harmonic regression uses a sliding window of one year, centered on the missing date. The series type will therefore dictate how many data values are used in the regression. The number of harmonics will dictate the minimum number of valid data values that must exist. For example, with the default of 2 harmonics (an annual and a semi-annual cycle), a minimum of 5 valid dates must exist in the window. The general formula is $2n+1$, where n is the number of harmonics. You may specify a higher minimum, however. The more data that go into the regression, the better the fit. You can also specify the maximum gap that can be bridged over.

In addition to the usual designation of missing values, the harmonic interpolation procedure also allows you to designate an error tolerance. After an initial fit at a location, if there is a valid value at that location and it falls outside the tolerance specified, it is considered to be noise and the pixel is replaced with the interpolated value. If you do not wish to use this feature, simply specify an impossibly high tolerance.

¹⁵ Roerink, G.J., Menenti, M., and Verhoef, W., (2000) Reconstructing cloudfree NDVI composites using Fourier analysis of time series, *International Journal of Remote Sensing*, 21, 9, 1911-1917.

Finally, note that like the original HANTS procedure, you can elect to replace every value with its interpolated estimate. This is essentially a smoothing option.

Linear Temporal Interpolation

This is perhaps the simplest of the procedures offered, but one that generally works well. When a pixel with missing data is encountered, it looks at the images before and the image after that date up to the maximum allowable data gap. If good values can be found before and after the missing data date such that the time between those two dates does not exceed the maximum allowable gap, it will then linearly interpolate a replacement value.

Spatial Interpolation

With spatial interpolation, ETM looks at the 3 by 3 neighborhood surrounding a missing pixel and replaces it with the median value. This will only fill in the pixel if the majority of the neighboring pixels have valid data.

Climatology (Temporal Median)

Data sets that express the long term average of a parameter over its cycle are often referred to as a *climatology*. In the context of missing data interpolation, it refers to replacing a missing value with the long term median value for that period in the cycle. It should be used only as a method of last resort as it will detract from evidence of trends.

Note that a best practices procedure has not been established for missing data interpolation. However, a general procedure that works well is to use one or more successive calls to the linear temporal and spatial interpolation procedures followed by a final step of climatology.

The Denoise Panel

The Denoise Panel offers several utilities for the removal of noise in image series. Options include:

Temporal Filter

The Temporal Filter option smoothes over time. It does so using a symmetric moving filter window. Filter lengths can be both odd and even numbered except in the case of either a Gaussian filter or a Maximum filter (both of which must be odd). In the case of even filters, the first and last images have a half weight each in the filtering operation. Sub-options include:

- Mean filter: the resulting pixel values are the simple averages of all time periods within the moving temporal window.
- Gaussian weighted mean: the resulting pixel values are the weighted averages of all time periods within the moving temporal window, where the weights follow a Gaussian distribution. This tends to produce a smoother series, but will typically need a wider filter window than the simple mean filter.
- Maximum value: the resulting pixel values represent the maximum value that occurs over the temporal filter window.
- Cumulative sum: the resulting pixel values represent the sum of the pixel being operated upon and all previous values that occur within the length of the temporal filter window.

- Cumulative mean: the resulting pixel values represent the average of the pixel being operated upon and all previous values that occur within the length of the temporal filter window.

Maximum Value Composite

Maximum Value Compositing is commonly used for NDVI vegetation index imagery. The logic is that the presence of clouds (even partially transparent ones) will always lead to a lower NDVI value. Thus choosing the maximum value for the time period being considered can safely be assumed to have been least affected by clouds. Note however that this logic is not necessarily applicable to other image types.

The tools necessary to create a maximum value composite exist within the Generate/Edit Series panel (Aggregate Series option, output type Maximum). Thus you are directed to use that option. Its entry is included in the Denoise panel simply to make users aware that this option exists.

Inverse PCA

Unstandardized Principal Components Analysis provides a linear transformation of a set of components in which the early components describe highly prevalent and coherent sources of variability in the series. On the other hand, the later components will contain minor sources and incoherent sources of variability such as noise. Thus an effective means of denoising data is to reverse the transformation while leaving out these noisy elements. How many should you leave out? There is no easy answer here without looking at the components quite carefully. However, most of the coherent variability in a series is contained within a surprisingly few components – e.g., the first 20 or so. For filtering for the purpose of using a Cross-EOT, some would recommend even a heavier filtering, using only perhaps the first five components in reconstructing the series.

Note that in using this option, ETM automatically computes both the forward and inverse transformations.

Also note that both S-mode and T-mode PCA are supported.

Inverse Fourier

The logic here is similar to Inverse PCA. From a forward breakdown of an image series into a set of amplitude and phase images, it is possible to reconstruct it using an inverse transformation. The higher-numbered harmonics represent increasingly high-frequency elements such as noise. A value that commonly works well is to include all the interannual cycles, the annual cycle and the semi-annual cycle. The cutoff frequency is thus going to depend on the length of the series. If, for example, you have 25 years of monthly data, interannual cycles would be the harmonics less than 25, the annual cycle would be harmonic 25 and the semi-annual cycle would be harmonic 50. Therefore, the number of harmonics to use would be 50.

The Deseason Panel

It is common that you will wish to remove seasonality from your series. For example, the trend measures on the Trend panel are intended to be used with deseasoned data. Three options are provided:

Anomalies

This is the most commonly used form of deseasoning. As a first step, ETM creates what is known as a *climatology* from the series as a whole. For example, if the data are monthly, then it creates an average January, an average February, and so on. It then goes through the series and subtracts the long term average from each month. For example, the January 1998 anomaly image would be equal to January 1998 minus the long term average of all Januaries.

Standardized Anomalies

This is a variant of regular anomalies. In calculation of the climatology (the long-term means), it also calculates the standard deviation of values for each member of the climatology. Then when the anomaly images are created (by subtracting the long term mean) the result is divided by the standard deviation to create a standardized anomaly (a z-score). In this new system, a value of 0 would mean that it has a value equal to the long term mean, a value of +1 would mean that it is 1 standard deviation of the long term mean, and so on.

Temporal Filter

This option is identical to the use of temporal filtering for denoising. However, the implication is that the filter length will be long enough that it will filter out seasonality as well as noise. ETM will suggest a filter length for each option.

The Generate / Edit Series Panel

This panel offers a number of helpful utilities for generating new series or editing existing ones. Note that procedures which generate new series will ask you to select a series that can be used as a template regarding all series parameters such as the number of images, the interval type (e.g., monthly), start and end date, etc.

Linear Index Series

Generates an index series with a perfectly linear sequence with values that start from the specified start value and increase by the specified increment with each time step. A series of this type may be useful as an independent variable representing a trend in Linear Modeling.

Sin Index Series

Generates an index series with the specified number of sine waves over the entire series. The series starts at the phase angle specified. A series of this type may be useful as an independent variable in Linear Modeling (typically paired with a Cosine series).

Cos Index Series

Generates an index series with the specified number of cosine waves over the entire series. The series starts at the phase angle specified. A series of this type may be useful as an independent variable in Linear Modeling (typically paired with a Sine series).

Lagged Series

Generates a lagged version of the input series. Specifying a lag other than 0 will start the series with the position indicated, starting the count from 0. For example, for a monthly series that starts in January, specifying a lag of 3 will have the series start from April and end with the normal end of the series. To pair an unlagged series of the same length, use the *Truncated Series* option and remove from the end of the series the same number of images as specified for the lag. Note – for negative lags truncate from the end of the series.

Truncated Series

Generates a truncated version of the input series – i.e., a shortened series by removing images either from the end or the beginning.

Supplemented Series

This is used to add additional data on either the beginning or the end of a series. Note that the series to be added (even if it contains only one image) must be correctly documented (in its .tsf file) with regards to the start and end dates of the series, series type, etc.

Skip Factor Series

A skip factor series is used to extract all members of a specific position in a cycle. For example, given a monthly series that starts in January, using a *start position* of 1, a *take factor* of 1, and a *skip factor* of 11 will produce a new series consisting of all January images. Note that the take and skip factors must add up to the number of items in a full cycle. Thus, for example, using a start factor of 6, a take factor of 3 and a skip factor of 9 will yield a series consisting of the summer months (JJA) of each year.

Rename Series Images

This utility renames the images within a series and not the series itself. To be consistent with operating system file ordering, the new names will be composed of the prefix, followed by the cycle number (e.g., year), followed by the position (e.g., month). This option is great for renaming series that do not sort properly by the operating system or that are consecutively numbered without reference to the cycle or position.

Aggregate Series

This option aggregates series to a coarser or similar temporal resolution. Depending upon the nature of the series, one or several conversion options may be available (a small number do not have any conversion options because of peculiarities in their nature). For cases where a mean is used and values must be taken from more than one input image, a weighted mean is used.

Subset Series

This option creates a new series as a spatial subset of another series. Essentially, it performs the equivalent of the WINDOW operation on every member of the original series.

The Detrend Prewhitened Panel

This panel contains several utilities for removing trends and handling serial correlation (correlation between successive values over time).

Detrend (Linear)

This option removes the presence of a linear trend in the series. It does this by extracting the residuals from a linear regression between the series (the dependent variable) and time (the independent variable).

Detrend (Difference Series)

This option creates a new series where values express the difference between values of the original series and the value in the previous time step of that series. If the series is first order autoregressive (i.e., the correlation between values and immediately preceding values accounts for all successive correlations), this procedure will remove serial correlation.

Trend Preserving Prewhitening

Prewhitening refers to the removal of serial correlation in the error (noise) component of a series. The procedure performed here assumes that the series can be described as:

$$Y_t = a + bt + X_t$$

where

$$X_t = \rho X_{t-1} + e_t$$

where a is an intercept, b is the trend slope, t is time, X_t is a lag-1 red noise process, ρ is the serial correlation and e is a white noise error term. The trend preserving prewhitening in ETM uses the procedure described by Wang and Swail (2001)¹⁶ to remove the red noise component yielding a new series (W_t) that can be described by:

$$W_t = a' + bt + e'_t$$

where

$$a' = \frac{a + \rho b}{1 - \rho}$$

and

$$e'_t = \frac{e_t}{1 - \rho}$$

An iterative procedure is used to estimate the true serial correlation ρ and the trend-preserving prewhitened series is calculated as:

$$W_t = \frac{Y_t - \rho Y_{t-1}}{1 - \rho}$$

¹⁶ Wang, X.L., and V.R. Swail, (2001). Changes of extreme wave heights in northern hemisphere oceans and related atmospheric circulation regimes, *Journal of Climate*, 14, 2204-2221.

This prewhitened series has the same trend as the original series, but with no serial correlation (Wang and Swail (2001). Note that this prewhitening method decreases the sample size by one which is recovered using the Prais-Winsten transformation (Kmenta, 2004)¹⁷.

Durbin Watson

The Durbin Watson option computes the Durbin Watson statistic of serial correlation over time for each pixel. For index series, a tabulation is made of the slope and intercept of the best fit linear trend, along with the Durbin Watson statistic. Its value always lies between 0 and 4. A value of 2 indicates no serial autocorrelation. A Durbin Watson statistic less than 2 indicates evidence of a positive serial correlation and a statistic greater than 2 indicates evidence of a negative serial autocorrelation. Critical values for the Durbin Watson statistic can be found in standard statistical texts.

Cochrane-Orcutt

The Cochrane-Orcutt transformation¹⁸ transforms the dependent variable and each of the independent variables in order to remove serial correlation. It estimates the first order serial correlation ρ by calculating the correlation between the residuals and the residuals at lag 1. The transformation of *both* the dependent and independent variables is calculated as:

$$X^* = X_1 - \rho X_1$$

As with Prewhitening, the Prais-Winsten transformation (Kmenta, 2004) is used to estimate the initial value of the transformed variables. Note that it is assumed that all variables have been deseasoned such that the expected intercept is 0. In this case, the transformed intercept is also 0. Therefore the intercept term drops out of the transformation and the user can use the transformed dependent and independent variables with the Linear Modeling tool to complete the analysis.

¹⁷ Kmenta, J. 2004. *Elements of Econometrics*, The University of Michigan Press.

¹⁸ Cochrane D, Orcutt GH (1949) Application of Least Squares Regression to Relationships Containing Auto-correlated Error Terms. *J Amer Statistical Assoc* 44:32-61

CHAPTER ELEVEN

CLIMATE CHANGE ADAPTATION MODELER

The Climate Change Adaptation Modeler (CCAM) is intended as a toolbox to address the growing challenge of adapting to a rapidly changing climate. It includes tools for climate scenario generation using NCAR's MAGICC and SCENGEN models, crop suitability modeling using FAO's EcoCrop database, a sea level rise impact modeler, a tool for downscaling using the change-factor logic, and a tool to generate bioclimatic variables. CCAM offers other tools for ingesting climate data.

MAGICC/SCENGEN

The MAGICC and SCENGEN models, within the Climate Adaptation Modeler, work in conjunction with one another and allow you to explore global patterns of climate change. First, MAGICC (Model for the Assessment of Greenhouse-gas Induce Climate Change) is run to predict temperature and sea level change up to the year 2100 based on several management scenarios. Then, SCENGEN (SCENario GENerator) uses the results from MAGICC using a single or ensemble of climate models to map how temperature and precipitation vary globally. In essence, it produces an image of how the globe will react to climate change in accordance with coinciding anthropogenic behaviors and management plans.

MAGICC and SCENGEN are developed and maintained by the National Center for Atmospheric Research (NCAR) and incorporated into TerrSet with their acknowledgment. To use the MAGICC and SCENGEN models in CCAM, it is required that you download and install the stand-alone products¹. To access the stand-alone program and full documentation for MAGICC and SCENGEN, visit: <http://www.cgd.ucar.edu/cgi-bin/cas/magicc.cgi>.

MAGICC: Model Global Warming and Sea Level Rise

MAGICC is an upwelling-diffusion, energy-balance climate model that produces global mean temperature and oceanic thermal expansion outputs. This climate model is linked with various gas-cycle models which account for climate feedbacks to the carbon cycle by keeping track of the concentrations of the main greenhouse gases. The MAGICC software relies on the 5.3 version of MAGICC used within the IPCC Fourth Assessment Report, Working Group 1 (AR4).

¹ See the Help documentation for MAGICC for complete installation instructions.

The user is able to define much of the information contributing to the model's results including the selection of scenarios of greenhouse gas and sulfur dioxide emissions, specifying the implications of uncertainties within the carbon cycle, and noting the scale of aerosol forcing, climate system's sensitivity to external forcings, and the variability in the ocean's thermohaline circulation. Also, the user has the ability to modify the intervals and years in which climate predictions occur and are referenced.

MAGICC includes a total of 49 different emission scenarios (see table below). Thirty five of the scenarios are based on the no-climate policy SRES, Special Report on Emissions Scenarios (Nakicenovic and Swart, 2000)²; six of these are illustrative scenarios. Each scenario fits to one of four classifications; these can be divided into A1, A2, B1 and B2. From these four groups there are several more breakdowns into the individual emission scenarios. Many of these separations relate to the modeling approach which was used; we include six modeling approaches:

1. Asian Pacific Integrated Model (AIM) from the National Institute of Environmental Studies in Japan (Morita et al., 1994)³.
2. Atmospheric Stabilization Framework Model (ASF) from ICF Consulting in the US (Lashof and Tirpak, 1990⁴; Pepper et al., 1998⁵; Sankovski et al., 2000⁶).
3. Integrated Model to Assess the Greenhouse Effect (IMAGE) from the National Institute for Public Health and Hygiene in the Netherlands (Alcamo et al., 1998⁷; de Vries et al., 1994⁸, 1999⁹, 2000¹⁰), also used along with the Dutch Bureau for Economic and Public analysis WorldScan model (deJong and Zalm, 1991¹¹) in the Netherlands.
4. Multiregional Approach for Resource and Industry Allocation (MARIA) from the Science University of Tokyo in Japan (Mori and Takahashi, 1999¹²; Mori, 2000¹³).
5. Model for Energy Supply Strategy Alternatives and their General Environmental Impact (MESSAGE) from the International Institute of Applied Systems Analysis in Austria (Messner and Strubegger, 1995; Riahi and Roehrl, 2000).

² Nakicenovic, N., and Swart, R., Eds., 2000, Special Report on Emissions Scenarios. Cambridge University Press, Cambridge, UK, 570 pp.

³ Morita, Y., Y. Matsuoka, M. Kainuma, and H. Harasawa, 1994: AIM - Asian Pacific integrated model for evaluating policy options to reduce GHG emissions and global warming impacts. In Global Warming Issues. In Global Warming Issues in Asia. S Bhattacharya et al (ed.), AIT Bangkok, pp. 254-273.

⁴ Lashof, D., and Tirpak, D.A., 1990: Policy Options for Stabilizing Global Climate. 21P-2003. U.S. Environmental Protection Agency, Washington, D.C.

⁵ Pepper, W.<J, Barbour, W. Sankovski, A., and Brasz B., 1998: No-policy greenhouse gas emission scenarios: revisiting IPCC, 1992. Environmental Science & Policy, 1, 289-312.

⁶ Sankovski, A., W. Barbour, and W. Pepper, 2000: Quantification of the IS99 emission scenario storylines using the atmospheric stabilization framework (ASF). Technological Forecasting and Social Change, 63(2-3) (in press).

⁷ Alcamo J., E. Kreileman, M. Krol, R. Leemans, J. Bollen, J. van Minnen, M. Schaefer, S. Toet, and B. de Vries, 1998: Global modelling of environmental change: an overview of IMAGE 2.1. In Global change scenarios of the 21st century. Results from the IMAGE 2.1 Model. J. Alcamo, R. Leemans, E. Kreilman (eds.), Elsevier Science, Kidlington, Oxford, pp. 3-94.

⁸ De Vries, H.J.M., J.G.J. Olivier, R.A. van den Wijngaart, G.J.J. Kreileman, and A.M.C. Toet, 1994: Model for calculating regional energy use, industrial production and greenhouse gas emissions for evaluating global climate scenarios. Water, Air, Soil Pollution, 76, 79-131.

⁹ De Vries, B., M. Janssen, and A. Beusen, 1999: Perspectives on global energy futures - simulations with the TIME model. Energy Policy, 27, 477-494.

¹⁰ De Vries, B., J. Bollen, L. Bouwman, M. den Elzen, M. Janssen, and E. Kreileman, 2000: Greenhouse gas emissions in an equity-, environment-and service-oriented world: An IMAGE-based scenario for the next century. Technological Forecasting & Social Change, 63(2-3). (In press).

¹¹ De Jong, A., and G. Zalm, 1991: Scanning the Future: A long-term scenario study of the world economy 1990-2015. In Long-term Prospects of the World Economy. OECD, Paris, pp. 27-74.

¹² Mori, S., and M. Takahashi, 1999: An integrated assessment model for the evaluation of new energy technologies and food productivity. International Journal of Global Energy Issues, 11(1-4), 1-18.

¹³ Mori, S., 2000: The development of greenhouse gas emissions scenarios using an extension of the MARIA model for the assessment of resource and energy technologies. Technological Forecasting & Social Change, 63(2-3). (In press).

6. Mini Climate Assessment Model (MiniCAM) from the Pacific Northwest National Laboratory in the USA (Edmonds et al., 1994, 1996a, 1996b).

Scenario	Scenario Type	Description
350NFB	Policy	Stabilized CO2 concentrations at 350 ppm along the NFB pathway
450NFB	Policy	Stabilized CO2 concentrations at 450 ppm along the NFB pathway
450OVER	Policy	CO2 concentrations above 450 ppm along the NFB pathway
550NFB	Policy	Stabilized CO2 concentrations at 550 ppm along the NFB pathway
650NFB	Policy	Stabilized CO2 concentrations at 650 ppm along the NFB pathway
750NFB	Policy	Stabilized CO2 concentrations at 750 ppm along the NFB pathway
A1ASF	SRES	High energy demand, global cooperation & new alternative technologies; ASF modeling
A1B-AIM	SRES Illustrative	High energy demand, global cooperation & new alternative technologies; AIM modeling; balance of technologies and supply
A1B-NEW	SRES Illustrator	High energy demand, global cooperation & new alternative technologies; using new modeling; balance of technologies and
A1CAI	SRES	High energy demand, global cooperation & new alternative technologies; AIM modeling; reliance on coal energy
A1CME	SRES	High energy demand, global cooperation & new alternative technologies; MESSAGE modeling; reliance on coal energy
A1CMI	SRES	High energy demand, global cooperation & new alternative technologies; MiniCAM modeling; reliance on coal energy
A1FI-MI	SRES Illustrator	High energy demand, global cooperation & new alternative technologies; MiniCAM modeling; fossil fuel intensive, combination
A1GAI	SRES	High energy demand, global cooperation & new alternative technologies; AIM modeling; reliance on unconventional oil and gas
A1GME	SRES	High energy demand, global cooperation & new alternative technologies; MESSAGE modeling; reliance on unconventional oil and
A1MA	SRES	High energy demand, global cooperation & new alternative technologies; IMAGE modeling
A1MES	SRES	High energy demand, global cooperation & new alternative technologies; MESSAGE modeling
A1MIN	SRES	High energy demand, global cooperation & new alternative technologies; MiniCAM modeling
A1TAI	SRES Illustrator	RES Illustrative High energy demand, global cooperation & new alternative technologies; MESSAGE modeling; non-fossil
A1T-MES	SRES Illustrator	High energy demand, global cooperation & new alternative technologies; AIM modeling; non-fossil fuel and enhanced energy
A1V1MI	SRES	High energy demand, global cooperation & new alternative technologies; MiniCAM modeling; demographic, economic & energy
A1V2MI	SRES	High energy demand, global cooperation & new alternative technologies; MiniCAM modeling; demographic, economic & energy
A2A1MI	SRES	High energy demand, new alternative technologies & intermediate focus between global and local; MiniCAM modeling
A2AIM	SRES	Uneven economic growth & regional consolidation; AIM modeling
A2-ASF	SREF Illustrator	Uneven economic growth & regional consolidation; ASF modeling
A2GIM	SREF	Uneven economic growth & regional consolidation; IMAGE modeling; reliance on gas energy
A2MES	SREF	Uneven economic growth & regional consolidation; MESSAGE modeling
A2MIN	SREF	Uneven economic growth & regional consolidation; MiniCAM modeling
AVE	SREF	Average of all emission scenarios
B1AIM	SREF	Global cooperation & economic, social, and environmental sustainability; AIM modeling
B1HIME	SREF	Global cooperation & economic, social, and environmental sustainability; MESSAGE modeling; higher energy demand

B1HIM1	SREF	Global cooperation & economic, social, and environmental sustainability; MiniCAM modeling; higher energy demand
B1-IMA	SRES Illustrative	Global cooperation & economic, social, and environmental sustainability; IMAGE modeling
B1MES	SREF	Global cooperation & economic, social, and environmental sustainability; MESSAGE modeling
B1MIN	SREF	Global cooperation & economic, social, and environmental sustainability; MiniCAM modeling
B1TME	SREF	Global cooperation & economic, social, and environmental sustainability; MESSAGE modeling; non-fossil fuel and enhanced
B2AIM	SREF	Regional consolidation & economic, social, and environmental sustainability; AIM modeling
B2ASF	SREF	Regional consolidation & economic, social, and environmental sustainability; ASF modeling
B2HIMI	SREF	Regional consolidation & economic, social, and environmental sustainability; MiniCAM modeling; higher energy demand
B2IMA	SREF	Regional consolidation & economic, social, and environmental sustainability; IMAGE modeling
B2-MES	SREF	Regional consolidation & economic, social, and environmental sustainability; MESSAGE modeling
B2MIN	SRES	Regional consolidation & economic, social, and environmental sustainability; MiniCAM modeling
P50	Baseline	Average of all of the SRES scenarios
WRE350	Policy	Stabilized CO ₂ concentrations at 350 ppm along WRE pathway
WRE450	Policy	Stabilized CO ₂ concentrations at 450 ppm along WRE pathway
WRE550	Policy	Stabilized CO ₂ concentrations at 550 ppm along WRE pathway
WRE650	Policy	Stabilized CO ₂ concentrations at 650 ppm along WRE pathway
WRE750	Policy	Stabilized CO ₂ concentrations at 750 ppm along WRE pathway

Although there are many subsets of the A1 group, all of the emission scenarios within this family show specific qualities. This group describes a future in which global populations and communities become integrated and economic and social differences between these groups decrease. This future also includes rapidly increasing world economic prosperity and populations, peaking in 2050 and declining after the mid-century. Energy demand increases and technological advancements quickly match this demand. These alternative technological advancements are differentiated into specific A1 scenarios: fossil intensive (A1FI), non-fossil energy sources (A1T) and balanced across all sources (A1B).

Unlike the A1 grouping, A2 scenarios are less interested in global cooperation and more focused on development on a regional scale. With that said, economic growth and technological advancements are slower and fragmented as regions develop separately and within the input of others. The economic gap between developing and industrialized populations widens, unlike in the A1 and B1 scenarios.

Scenarios within B1 are similar to those in A1 in the sense that they describe a future world in which global populations are intertwined and populations peak and decline after 1950. The difference here though is that there is a focus on clean and sustainable energy production. The dominant variation in these scenarios is the emphasis on being environmentally and socially mindful in order to create a future of equality and sustainability.

The final scenario grouping is B2, a world with a regional, as opposed to global, focus and an interest in a sustainable future of environmental protection and social equality. B2 shows the same trend as A2 in directing attention to stronger and independent communities. These scenarios are marked by gradually increasing populations and economic development, but slower technological advancements.

The NFB and WRE scenarios within MAGICC lead to stabilizing CO₂ concentrations. WRE scenarios will stabilize only if carbon cycle feedbacks are included, and NFB scenarios will only stabilize if these same feedbacks are not included. If the proper selection is not made for these feedbacks, then the former scenarios will show lower concentrations than the stabilization value and the latter

scenarios will be higher than the stabilization value. This points to the effect of these climate feedbacks. In the descriptions of each of these families of scenarios we mentioned specific pathways, NFB and WRE pathways. The NFB pathway is ocean-only and does not include any feedbacks. In comparison, the WRE pathway is one in which there is a gradual switch away from carbon-producing fuels .

MAGICC also provides for the Carbon Cycle Model input allowing the user to specify a low, mid or high range of carbon cycle uncertainties, unrelated to climate. Low amounts of uncertainty are linked with higher concentrations of CO₂ emissions and vice versa. The value associated with high is 0.4 GtC (Gigatonnes of Carbon) per year, 1.1 GtC for mid, and 1.8 GtC for low. These values are taken from estimates from the IPCC Second Assessment Report and refer to the change in the mean value in the 1980s for net land-use change CO₂ emissions. Although there are more recent IPCC reports, the original values continue to agree with CO₂ concentrations estimated in more updated reports. The user is also able to include the effect of carbon cycle climate feedbacks if so desired; including these feedbacks leads to higher concentrations of CO₂. There remain many concerns about the validity of these concentrations; each carbon cycle model is associated with a different feedback value (Joos et al., 2001 , Kheshgi et al., 2003). While the default is to have feedbacks included, the user can choose to disregard feedbacks if they judge them to be unimportant.

Variation in thermohaline circulation is an important determinant of temperature change and thermal expansion in the ocean. The default setting for thermohaline circulation is variable, representing a case where thermohaline circulation slows as the earth's temperature increases. Choosing the default setting, variable, denotes a situation where the slowing of the thermohaline circulation occurs at a rate equal to the median of thermohaline change for seven specific models within SCENGEN, the seven models used to calculate MAGICC. The user can also choose the constant option, referring to no slow-down in the thermohaline circulation.

Another setting is the aerosol forcing which are taken from the IPCC Third Assessment Report. There is a choice of three settings: high, mid or low. Aerosol forcings occur through four channels: direct, indirect, biospheric and fossil plus organic carbonaceous (FOC). Direct forcing is the direct effect of sulfate aerosols derived from combustion of fossil fuels; MAGICC estimates this amount to be -0.4Wm⁻². Indirect forcing is given a value of -0.8 Wm⁻² and it represents indirect effects of sulfate aerosols. Biospheric forcing aerosols occur as a result of burning organic and carbonaceous material; assumedly why this value corresponds to total land-use change emissions. The default value for biospheric forcing is -0.2 Wm⁻². This final aerosol forcing, FOC, is the forcing that reflects fossils in addition to organic and carbonaceous material. This default value is 0.1Wm⁻². By selecting the mid option, the user is choosing an aerosol forcing of -1.3Wm⁻². Similarly, low means a value of -0.8 Wm⁻² and high means a value of -1.8Wm⁻².

MAGICC also includes information for vertical diffusion which is the rate of mixing in the ocean that determines the amount of heat circulated into deeper waters. The default value, 2.3cm²/s, was chosen as the median diffusion value for the same seven models used for thermohaline circulation calculations.

Another important variable specifically important for oceanic thermal expansion, and therefore in turn sea level rise, is the rate of ice melt. The IPCC Third Annual Report has several estimates for the range of ice melt; MAGICC averages the median values for all of these ranges. The level of ice melt can be set to either high, low or medium. High and low ice melt is calculated by concatenating the two-sigma high and low limits for non-expansion terms.

Climate sensitivity is included as an input variable because it accounts for the earth's temperature response to a doubling of the CO₂ concentration. The IPCC Third Annual Report calculates a median value of 2.6°C and past IPCC reports predict a range of 1.5°C to 4.5°C. MAGICC uses a default value of 3.0°C but it is possible to enter a user-specific value, although we recommend a value between 0.5°C and 5.5°C.

The final parameter needed is the selection of the model to run. MAGICC was first calibrated to fit with the temperature and expansion rates of seven models and a 1%/year CO₂ change (from CMIP2 database); these seven were included in the IPCC Third Annual Report for simulating global-mean temperature and sea level rise. They are GFDL, CSIRO, HadCM3, HadCM2, ECH4-OPYC, PCM, and CSM. MAGICC ran each of these for all emission scenarios. By selecting one of the above MAGICC will use the known parameters, specific for that model. This allows the user to simulate temperature and sea-level rise for a single model. Select user to send to MAGICC user-defined parameters that will overwrite the default parameters for the seven models.

In the output parameters section of MAGICC, the user must specify years for reference, prediction and interval. MAGICC does not include any reference years beyond 1990, and the model cannot make climate predictions past 2100. The user also has the ability to modify the interval for which the climate model should run; the default is set to 5.

Outputs

After running MAGICC the user is left with a single output graph of temperature change and sea level rise. The SCENGEN section of the model leaves the user with eight global images, four for precipitation and four for temperature, an average of absolute change, a new mean climate (with aerosols) using a model-mean baseline, a mean climate (with aerosols) using an observed baseline, and scaled changes (aerosols only).

SCENGEN: Generate Climate Scenarios

The SCENGEN component of CCAM produces images revealing the spatial distribution of climate change from a database of CMIP3/AR4 Atmosphere/Ocean General Circulation Model (AOGCM) data. It does this by combining the global-mean temperature values produced in MAGICC with a pattern scaling method (Santer et al., 1990). This pattern scaling method separates components of future climate change into global-mean elements and spatial-pattern elements. Spatial-pattern elements are further split into greenhouse gases and aerosols. These components are "normalized" and then weighted, summed and scaled to match the global-mean temperature per year emission scenario and climate model provided by MAGICC. This scaling refers to the use of one or more AOGCMs to predict the effect of greenhouse gases on the climate.

SCENGEN relies upon the most recent results produced by MAGICC. The first step in using SCENGEN is to choose the scale: monthly, seasonal, annual or climatological. There is an option to select the pattern scaling method, linear (default) or exponential. It should be noted that linear scaling can cause some issues for both precipitation and temperature predictions due to variation in these variables not occurring equally year round or from low to high latitudes. Exponential scaling is calculated using:

Where D is the scaled percentage change for the increase in average global temperatures and a is the normalized percentage change at a single grid point. When D is very small the equation is essentially a linear scaling. While exponential scaling can be beneficial, in this case by preventing impossible outcomes, like negative precipitation, if a is large it can create improbably large changes.

The scenario year specified by the user is central year at which climate data in the 30 years around this year will be averaged to create an estimate for that specific year.

SCENGEN includes 20 AOGCMs that could be included for modeling representations of the pattern of greenhouse-gas-induced climate. Regardless of the type and number of models, each is given equal weight. We recommend the use of more than one model to improve accuracy. If you wish to be more specific about the most important models in your scenario and which ones are best to include, we recommend first running SCENGEN using all 20 models. Then navigate to the file OUTLIERS.OUT within the `c:\SG53\SCEN-53\engine\imout` folder. This file contains a table where the first column shows each of the models and the second column shows their correlation. A higher correlation means a stronger model. You can then choose to re-run SCENGEN, excluding the models that have the lowest correlation values.

There is an option to include the spatial effects of aerosols. Running without this option can demonstrate the role of aerosols in the climate system. Also, there is an option for drift correction. Selecting not to include drift, the model will use the difference between the climate state at its initial state and the end of its perturbed state. Selecting drift correction calculates the difference between the perturbed state and a control state at the same time. The drift correction should be used if a model has any spatial drift based on the assumption that this drift will be equivalent to the difference between the perturbed and control settings. Using both also is an option, then SCENGEN does two calculations. It will produce an image that represents the average normalized drift, per 1°C average global warming. This is done by averaging the values for no drift correction and drift correction. SCENGEN will also calculate the

difference between having no corrections and having drift corrections included; it does this by simply subtracting the former from the latter. This produces a text file output representing correlation, where a very low value would represent a high correlation and vice versa, indicating whether or not drift is a contributing factor in the climate prediction. This text file can be found by navigating to the folder C:\SG53\SCEN-53\engine\scengen and opening the file DRIFT.OUT.

Outputs

There are two graphs and generally eight raster outputs from the MAGICC/SCENGEN model. The four outputs listed below are the prefixes used for both mean temperature and precipitation outputs. When the climatology option is chosen within SCENGEN, there will be additional outputs using the prefixes.

- ***ABSDEL***

Average of absolute changes.

- ***ABS-MOD***

New mean state, when including aerosols, using a model-mean baseline.

- ***ABS-OBS***

New mean state, when including aerosols, using an observed baseline.

- ***AEROSOL***

Scaled change field, aerosols only

Sea Level Rise Impact

The Sea Level Rise Impact model relies on a previously developed TerrSet module, PCLASS, probability reclassification. PCLASS evaluates each pixel in an image and assesses the probability of that data cell exceeding a user-defined threshold value, based on a user-defined level of uncertainty. In contrast to the hard Boolean image produced from the reclassification module RECLASS, PCLASS produces a soft probability image where values fall between zero and one. PCLASS provides a better tool for modeling projected sea level rise as it can take into account uncertainty in both future projections and in the maps themselves.

Similar to RECLASS, PCLASS has a threshold value around which probabilities are predicted. In this model, the threshold value represents the predicted ocean height. This value is user-defined, given within the input "Projected sea level rise". The user also provides a Root Mean Square Error (RMSE) value for the projected sea level rise. The model works by calculating the area under a normal bell curve, defined by the threshold value, with the RMSE standing in for standard deviations. For a more extensive discussion on PCLASS, see the section Database Uncertainty and Decision Risk.

Output

Accounting for the error in the future prediction and the level of uncertainty associated with the DEM image, the result of the Sea Level Rise Impact model is a probability map representing the likelihood of an area being submerged.

Crop Climatic Suitability Modeling

The Crop Climatic Suitability Modeling (CCSM) model in CCAM creates global suitability distribution maps for selected plant species. Plant species and their climate parameters are taken from the EcoCrop crop species database developed by the Food and Agricultural Organization of the United Nations (FAO) (<http://ecocrop.fao.org>). The model relies on temperature and precipitation data for its calculations and closely follows the methodology outlined by Ramirez-Villegas et al. (2011). The FAO database used in CCSM is located in the resfiles\CCAM folder under the TerrSet application folder.

The CCSM model requires monthly climatology datasets for precipitation, minimum temperatures and mean temperatures. TerrSet provides global datasets for each of these parameters developed by WorldClim. They can be found in the CCAM tutorial folder. But the user can supply and data.

The model separates each year into 12 potential growing seasons, one for each month. It then assesses the level of suitability of that month for a plant's growth based on various temperature and precipitation values. In terms of temperature values, the model uses an absolute temperature at which plants will be killed (KTMP), a minimum temperature at which plants will grow (TMIN), a minimum average temperature at which plants will grow optimally (TOPMIN), a maximum average temperature at which plants will grow optimally (TOPMAX) and a maximum average temperature at which plants will stop growing (TMAX). If a pixel has an average minimum temperature value 4°C or less above KTMP then it assumes that the KTMP value will be reached at some point during that month so that the month will no longer be considered suitable (0% suitable). If the temperature is not in the 4°C range of the KTMP value then the other temperature values are used to determine overall temperature suitabilities. TMIN, TOPMIN, TOPMAX and TMAX combine in the following fashion for each month for each individual pixel:

The model also uses precipitation as a suitability variable. These suitability calculations are similar to those for temperature, but there is no absolute minimum, or killing precipitation level. The suitability evaluation is not broken into monthly segments, but rather a crop is given a suitability assessment based on the entire year's precipitation. The growing season is essentially the entire year. Similar to the temperature parameters, the precipitation parameters are the minimum rainfall during the year (RMIN), optimal minimum rainfall during the year (ROPMIN), optimal maximum rainfall during the year (ROPMAX) and maximum rainfall during the year (RMAX). All precipitation values are given in millimeters.

TMIN/RMIN and TMAX/RMAX establish the absolute range. A temperature value outside this range is not suitable. TOPMIN/ROPMIN and TOPMAX/ROPMAX establish the optimum range. A pixel with a value within this range has a suitability value of 100%. If a pixel has a value outside the optimal range but within the absolute range, then the suitability value ranges from 1-99%.

Temperature suitabilities are calculated using the following algorithm:

TSUIT_i is the temperature suitability index for month *i*; TMIN, TOPMIN, TOPMAX and TMAX are dependent on the specific crop; aT1 is the intercept and mT1 is the slope of the regression curve between [TMIN, 0] and [TOPMIN, 0]; aT2 is the intercept and mT2 is the slope of the regression curve between [TOPMAX, 0] and [TMAX, 0]; TMIN-P_i is the minimum temperature of month *i* in site *P*; TMEAN-P_i is the mean temperature of month *i* in site *P*; TKILL-M is the killing temperature plus 4°C. TSUIT_i is the lowest monthly suitability score of all consecutive months within the crop's growing season.

Precipitation suitabilities are calculating using the following algorithm:

RSUIT is the precipitation suitability value; RTOTAL-P is the total precipitation for the entire year at site *P*; RMIN, ROPMIN, ROPMAX, and RMAX are dependent on the specific crop; aR1 is the intercept and mR1 is the slope of the regression curve between [RMIN, 0] and [ROPMIN, 0]; aR2 is the intercept and mR2 is the slope of the regression curve between [ROPMAX, 0] and [RMAX, 0].

Crop suitability can be evaluated based on either temperature, precipitation, or both. Choosing both temperature and precipitation the model can aggregate the two suitabilities either through, 1) taking the minimum of the two scores, or 2) taking the product of the two suitability scores.

Bioclimatic Variables

The Bioclimatic Variables model in CCAM generates 19 bioclimatic variable images. Bioclimatic variables are derived either from minimum and maximum temperature and precipitation data or from average temperature and precipitation data. The goal is to produce images that are more biologically meaningful, especially for input into habitat risk assessment and species modeling. The variables generated represent annual trends, seasonality and extreme environmental factors. They are generated annually using relatively simple calculations with the data provided.

Outputs

There are 19 possible outputs from the Bioclimatic Variable model. The number of output images depends on whether or not the user enters minimum/maximum temperatures or average temperatures. If using the latter input then the model will not generate the second (bio2) or third (bio3) outputs.

- ***OutputPrefix_bio1***

Annual Mean Temperature

- ***OutputPrefix_bio2***

Mean Diurnal Range (Mean of monthly (max temp - min temp))

- ***OutputPrefix_bio3***

Isothermality ($((\text{bio2}/\text{bio7}) * 100)$)

- ***OutputPrefix_bio4***

Temperature Seasonality (standard deviation*100)

- ***OutputPrefix_bio5***

Max Temperature of Warmest Month

- ***OutputPrefix_bio6***

Min Temperature of Coldest Month

- ***OutputPrefix_bio7***

Temperature Annual Range (bio5-bio6)

- ***OutputPrefix_bio8***

Mean Temperature of Wettest Quarter

- ***OutputPrefix_bio9***

Mean Temperature of Driest Quarter

- ***OutputPrefix_bio10***

Mean Temperature of Warmest Quarter

- ***OutputPrefix_bio11***

Mean Temperature of Coldest Quarter

- ***OutputPrefix_bio12***

Annual Precipitation

- ***OutputPrefix_bio13***

Precipitation of Wettest Month

- ***OutputPrefix_bio14***

Precipitation of Driest Month

- ***OutputPrefix_bio15***

Precipitation Seasonality (Coefficient of Variation)

- ***OutputPrefix_bio16***

Precipitation of Wettest Quarter

- ***OutputPrefix_bio17***

Precipitation of Driest Quarter

- ***OutputPrefix_bio18***

Precipitation of Warmest Quarter

- ***OutputPrefix_bio19***

Precipitation of Coldest Quarter

NetCDF Import

The NetCDF import is a utility that converts NetCDF data (version 3.6.1 or earlier) to TerrSet. NetCDF stands for Network Common Data Form. The NETCDF module is used to retrieve time series files, which have at least three dimensional data, i.e., the time, latitude and longitude. Dimensions other than these cannot be retrieved unless its length equals to one.

Gaussian to LatLong Transformations

The Gaussian to Latlong Transformation utility converts data formatted within a Gaussian grid coordinate systems to a latitude and longitude coordinate system. Most climate modelers work with data in the Gaussian grid format (essentially a vector point format) which facilitates the high analytical demands of global climate modeling. In theory, this format is not a true raster format, even though the data is visualized as a raster in TerrSet when it is viewed. What is important to the climate modelers is not the raster surface, but the point locations that they represent. The Gaussian to Latlong Transformation model takes this "point" file and converts it to a true raster file. The main difference between a Gaussian grid and regular geographic coordinate systems is that, while longitudes are equally spaced, latitudes decrease towards the poles in the Gaussian format.

Downscale Scenario

The Downscale Scenario model is used to convert low resolution images to high resolution images. The model produces anomalies, or relies on user input anomalies, to convert fine resolution baseline images to fine resolution future images. The idea of the model is that the user contains coarse and fine resolution images for a specific year in time, but contains a coarse resolution image for a future point in time that one is interested in having at a finer resolution.

The model subtracts the coarse baseline input image from the coarse future input image to produce an anomaly image for the change in conditions between these specific years. This anomaly image matches the coarse resolution of the two images which represents the change in conditions between the years of the two input images. There is an option to use climatology inputs where the model will subtract each baseline image from its respective future image within the raster group file. The user can bypass this step by simply entering an already produced anomaly image.

The model then sums the anomaly image with the input fine resolution baseline image. This summation takes the known difference between the baseline and the future, although coarse, and adds this difference to the same baseline image, although containing a finer resolution. For this reason, the resolution of the fine resolution baseline image, determines the resolution of all fine resolution outputs. The method described here is adapted from Tabor and Williams 2010.

APPENDIX 1

ELLIPSOID PARAMETERS

ELLIPSOID NAME	MAJOR SEMI-AXIS	MINOR SEMI-AXIS
Airy	6377563.396	6356256.909
Modified Airy	6377340.189	6356034.448
Australian National	6378160.000	6356774.719
Average Terrestrial System 1997	6378135	6356750.305
Bessel 1841 (Ethiopia, Indonesia, Japan, Korea)	6377397.155	6356078.963
Bessel 1841 (Namibia)	6377483.865	6356165.383
Bessel Modified	6377492.018	6356173.509
Clarke 1858	20926348	20855233
Clarke 1866	6378206.400	6356583.800
Clarke 1866 Michigan	20926631.531	20855688.674
Clarke 1880	6378249.145	6356514.870
Clarke 1880 Benoit	6378300.789	6356566.435
Clarke 1880 IGN	6378249.2	6356515

Clarke 1880 (SGA 1922)	6378249.2	6356514.997
Everest - India 1830	6377276.345	6356075.413
Everest - India 1956	6377301.24	6356100.23
Everest - Pakistan	6377309.61	6356108.57
Everest - Sabah and Sarawak	6377298.56	6356097.55
Everest - West Malaysia 1969	6377295.66	6356094.67
Everest - West Malaysia and Singapore 1948	6377304.063	6356103.039
Everest (1830 definition)	20922931.8	20853374.58
Everest 1830 (1962 definition)	6377301.243	6356100.23
Everest 1830 (1967 definition)	6377298.556	6356097.55
Everest 1830 (1975 definition)	6377299.151	6356098.145
Fischer 1960	6378166.000	6356784.280
Modified Fischer 1960	6378155.000	6356773.320
Fischer 1968	6378150.000	6356768.340
GEM 10C	6378137	6356752.314
GRS 1967	6378160	6356774.516
GRS 1980	6378137.000	6356752.314
Helmert 1906	6378200.000	6356818.170

Hough	6378270.000	6356794.343
Indonesian 1974	6378160.00	6356774.50
International 1924	6378388.000	6356911.946
Krassovsky 1940	6378245.000	6356863.019
NWL 9D	6378145	6356759.769
OSU86F	6378136.2	6356751.517
OSU91A	6378136.3	6356751.617
Plessis 1817	6376523	6355862.933
SGS 85	6378136.000	6356751.300
South American 1969	6378160.000	6356774.719
Sphere	6371000	6371000
Struve 1860	6378298.3	6356657.143
War Office	6378300.583	6356752.27
WGS 60	6378165.000	6356783.290
WGS 66	6378145.000	6356759.770
WGS 72	6378135.000	6356750.520
WGS 84	6378137.000	6356752.314

APPENDIX 2

DATUM PARAMETERS

The following table contains the constants required for the Molodensky Datum Transformation procedure. The ΔX , ΔY and ΔZ values are the three values (in that order) that should be specified in the "Delta WGS84" field of the Reference System Parameter File. These values represent the three-dimensional difference in position of the datum ellipsoid from that of WGS84. The values listed here were taken from the European Petroleum Survey Group Database.

DATUM	LOCATION	ELLIPSOID	DX	DY	DZ
ABIDJAN 1987	Cote D'Ivoire (Ivory Coast)	Clarke 1880	-124.76	53	466.79
ADINDAN	MEAN FOR Ethiopia, Sudan	Clarke 1880	-166	-15	204
	Burkina Faso		-118	-14	218
	Cameroon		-134	-2	210
	Ethiopia		-165	-11	206
	Mali		-123	-20	220
	Senegal		-128	-18	224
	Sudan		-161	-14	205
AFGOOYE	Somalia	Krassovsky 1940	-43	-163	45
AIN EL ABD 1970	Bahrain Island	International 1924	-150	-250	-1
	Saudi Arabia		-143	-236	7
AMERICAN SAMOA 1962	American Samoa Islands	Clarke 1866	-115	118	426
AMERSFOORT	Netherlands	Bessel 1841	593.16	26.15	478.54

		National			
AUSTRALIAN GEODETIC 1984	Australia & Tasmania	Australian National	-134	-48	149
AYABELLE LIGHTHOUSE	Djibouti	Clarke 1880	-79	-129	145
BATAVIA	Indonesia (Sumatra)	Bessel 1841	-377	681	-50
BELLEVUE (IGN)	Efate & Erromango Islands	International 1924	-127	-769	472
BERMUDA 1957	Bermuda	Clarke 1866	-73	213	296
BISSAU	Guinea - Bissau	International 1924	-173	253	27
BOGOTA OBSERVATORY	Colombia	International 1924	307	304	-318
BUKIT RIMPAH	Indonesia (Bangka & Belitung Islands)	Bessel 1841	-384	664	-48
CAMP AREA ASTRO	Antarctica (McMurdo Camp Area)	International 1924	-104	-129	239
CAMPO INCHAUSPE	Argentina	International 1924	-148	136	90
CANTON ASTRO 1966	Phoenix Islands	International 1924	298	-304	-375
CAPE	South Africa	Clarke 1880	-136	-108	-292
CAPE CANAVERAL	Bahamas, Florida	Clarke 1866	-2	151	181
CARTHAGE	Tunisia	Clarke 1880	-263	6	431

CHATHAM ISLAND ASTRO 1971	New Zealand (Chatham Island)	International 1924	175	-38	113
CHTRF95	Liechtenstein, Switzerland	GRS1980	0	0	0
CHUA ASTRO	Paraguay	International 1924	-134	229	-29
CORREGO ALEGRE	Brazil	International 1924	-206	172	-6
DABOLA	Guinea	Clarke 1880	-83	37	124
DECEPTION ISLAND	Deception Island, Antarctica	Clarke 1880	260	12	-147
DJAKARTA (BATAVIA)	Indonesia (Sumatra)	Bessel 1841	-377	681	-50
DOMINICA 1945	Dominica	Clarke 1880	725	685	536
DOS 1968	New Georgia Islands (Gizo Island)	International 1924	230	-199	-752
EASTER ISLAND 1967	Easter Island	International 1924	211	147	111
EGYPT 1907	Egypt	Helmert 1906	-130	110	-13
EST92	Estonia	GRS80	0.055	-0.541	-0.185
ETRF89	Europe	GRS80	0	0	0

EUROPEAN 1950	MEAN FOR Austria, Belgium, Denmark, Finland, France, West Germany, Gibraltar, Greece, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland	International 1924	-87	-98	-121
	MEAN FOR Austria, Denmark, France, West Germany, Netherlands, Switzerland		-87	-96	-120
	MEAN FOR Iraq, Israel, Jordan, Lebanon, Kuwait, Saudi Arabia, Syria		-103	-106	-141
	Cyprus		-104	-101	-140
	Egypt		-130	-117	-151
	England, Channel Islands, Ireland, Scotland, Shetland Islands		-86	-96	-120
	Finland, Norway		-87	-95	-120
	France		-84	-97	-117
	Greece		-84	-95	-130
	Iran		-117	-132	-164
	Italy (Sardinia)		-97	-103	-120
	Italy (Sicily)		-97	-88	-135
	Malta		-107	-88	-149
	Portugal, Spain		-84	-107	-120
	Tunisia		-112	-77	-145
	United Kingdom UKCS offshore east of 6 deg. west		-89.5	-93.8	-123.1
EUROPEAN 1979	MEAN FOR Austria, Finland, Netherlands, Norway, Spain, Sweden, Switzerland	International 1924	-86	-98	-119
Fahud					

FD58	Iran (Kangan district)	Clarke 1880	-241.54	- 163.64	396.06
FORT THOMAS 1955	Nevis, St. Kitts (Leeward Islands)	Clarke 1880	-7	215	225
GAN 1970	Republic of Maldives	International 1924	-133	-321	50
GEODETIC DATUM 1949	New Zealand	International 1924	84	-22	209
GGRS87	Greece	GRS 1980	-199.87	74.79	246.62
GRACIOSA BASE SW 1948	Azores (Faial, Graciosa, Pico, Sao Jorge, Terceira)	International 1924	-104	167	-38
GRENADA 1953	Grenada	Clarke 1880	72	213.7	93
GUAM 1963	Guam	Clarke 1866	-100	-248	259
GUNUNG SEGARA	Indonesia (Kalimantan)	Bessel 1841	-403	684	41
GUX 1 ASTRO	Guadalcanal Island	International 1924	252	-209	-751
HERAT NORTH	Afghanistan	International 1924	-333	-222	114
HJORSEY 1955	Iceland	International 1924	-73	46	-86
HONG KONG 1963	Hong Kong	International 1924	-156	-271	-189
HU-TZU-SHAN	Taiwan	International 1924	-637	-549	-203

INDIAN	Bangladesh	Everest 1830	282	726	254
INDIAN	India, Nepal	Everest 1956	295	736	257
INDIAN	Pakistan	Everest	283	682	231
INDIAN 1954	Thailand	Everest 1830	217	823	299
INDIAN 1960	Vietnam (near 16°N)	Everest 1830	198	881	317
	Con Son Island (Vietnam)		182	915	344
INDIAN 1975	Thailand	Everest 1830	209	818	290
INDONESIAN 1974	Indonesia	Indonesian 1974	-24	-15	5
IRELAND 1965	Ireland	Modified Airy	506	-122	611
ISTS 061 ASTRO 1968	South Georgia Islands	International 1924	-794	119	-298
ISTS 073 ASTRO 1969	Diego Garcia	International 1924	208	-435	-229
JOHNSTON ISLAND 1961	Johnston Island	International 1924	189	-79	-202
KANDAWALA	Sri Lanka	Everest 1830	-97	787	86
KERGUELEN ISLAND 1949	Kerguelen Island	International 1924	145	-187	103
KERTAU 1948	West Malaysia & Singapore	Everest 1948	-11	851	5
KUSAIE ASTRO 1951	Caroline Islands, Federal States of Micronesia	International 1924	647	1777	-1124
L. C. 5 ASTRO 1961	Cayman Brac Island	Clarke 1866	42	124	147

LEIGON	Ghana	Clarke 1880	-130	29	364
LIBERIA 1964	Liberia	Clarke 1880	-90	40	88
LUZON	Philippines (Excluding Mindanao)	Clarke 1866	-133	-77	-51
	Philippines (Mindanao)		-133	-79	-72
MAHE 1971	Mahe Island	Clarke 1880	41	-220	-134
MASSAWA	Ethiopia (Eritrea)	Bessel 1841	639	405	60
MERCHICH	Morocco	Clarke 1880	31	146	47
MIDWAY ASTRO 1961	Midway Island	International 1924	912	-58	1227
MINNA	Cameroon	Clarke 1880	-81	-84	115
	Nigeria		-92	-93	122
MONTSERRAT ISLAND ASTRO 1958	Montserrat (Leeward Islands)	Clarke 1880	174	359	365
M'PORALOKO	Gabon	Clarke 1880	-74	-130	42
NAHRWAN	Oman (Masirah Island)	Clarke 1880	-247	-148	369
	Saudi Arabia		-243	-192	477
	United Arab Emirates		-249	-156	381
NAPARIMA BWI	Trinidad & Tobago	International 1924	-10	375	165

NORTH AMERICAN 1927	MEAN FOR Antigua, Barbados, Barbuda, Caicos Islands, Cuba, Dominican Republic, Grand Cayman, Jamaica, Turks Islands	Clarke 1866	-3	142	183
	MEAN FOR Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua		0	125	194
	MEAN FOR Canada		-10	158	187
	MEAN FOR Continental US (CONUS)		-8	160	176
	MEAN FOR CONUS (East of Mississippi River) Including Louisiana, Missouri, Minnesota		-9	161	179
	MEAN FOR CONUS (West of Mississippi River)		-8	159	175
	Alaska (Excluding Aleutian Islands)		-5	135	172
	Aleutian Islands (East of 180°W)		-2	152	149
	Aleutian Islands (West of 180°W)		2	204	105
	Bahamas (Except San Salvador Island)		-4	154	178
	Bahamas (San Salvador Island)		1	140	165
	Canada (Alberta, British Columbia)		-7	162	188
	Canada (Manitoba, Ontario)		-9	157	184
	Canada (New Brunswick, Newfoundland, Nova Scotia, Quebec)		-22	160	190
	Canada (Northwest Territories, Saskat- chewan)		4	159	188
	Canada (Yukon)		-7	139	181
	Canal Zone		0	125	201
	Cuba		-9	152	178
	Greenland (Hayes Peninsula)		11	114	195
			-12	130	190

	Mexico				
NORTH AMERICAN 1983	Alaska (Excluding Aleutian Islands), Canada, Central America, CONUS, Mexico	GRS 80	0	0	0
	Aleutian Islands		-2	0	4
	Hawaii		1	1	-1
NORTH SAHARA 1959	Algeria	Clarke 1880	-186	-93	310

OBSERVATORIO METER- EO 1939	Azores (Corvo & Flores Islands)	International 1924	-425	-169	81
OLD EGYPTIAN 1907	Egypt	Helmert 1906	-130	110	-13
OLD HAWAIIAN	MEAN FOR Hawaii, Kauai, Maui, Oahu Hawaii Kauai Maui Oahu	Clarke 1866	61 89 45 65 58	-285 -279 -290 -290 -283	-181 -183 -172 -190 -182
OMAN	Oman	Clarke 1880	-346	-1	224
ORD. SURVEY GREAT BRITIAN 1936	MEAN FOR England, Isle of Man, Scotland, Shetland Islands, Wales England England, Isle of Man, Wales Scotland, Shetland Islands Wales	Airy	375 371 371 384 370	-111 -112 -111 -111 -108	431 434 434 425 434
PICO DE LAS NIEVES	Canary Islands	International 1924	-307	-92	127
PITCAIRN ASTRO 1967	Pitcairn Island	International 1924	185	165	42
POINT 58	MEAN FOR Burkina Faso & Niger	Clarke 1880	-106	-129	165
POINTE NOIRE 1948	Congo	Clarke 1880	-148	51	-291
PORTO SANTO 1936	Porto Santo, Madeira Islands	International 1924	-499	-249	314

PROVISIONAL S. AMERICAN 1956	MEAN FOR Bolivia, Chile, Colombia, Ecuador, Guyana, Peru, Venezuela	International 1924	-288	175	-376
			-270	188	-388
	Bolivia		-270	183	-390
	Chile (Northern, Near 19°S)		-305	243	-442
	Chile (Southern, Near 43°S)		-282	169	-371
	Colombia		-278	171	-367
	Ecuador		-298	159	-369
	Guyana		-279	175	-379
	Peru		-295	173	-371
	Venezuela				
PROVISIONAL S. CHILEAN 1963	Chile (South, Near 53°S) (Hito XVIII)	International 1924	16	196	93
PUERTO RICO	Puerto Rico, Virgin Islands	Clarke 1866	11	72	-101
PULKOVO 1942	Russia	Krassovsky 1940	28	-130	-95
QATAR NATIONAL	Qatar	International 1924	-128	-283	22
QORNOQ	Greenland (South)	International 1924	164	138	-189
REUNION	Mascarene Islands	International 1924	94	-948	-1262
ROME 1940	Italy (Sardinia)	International 1924	-225	-65	9
S-42 (PULKOVO 1942)	Hungary	Krassovsky 1940	28	-121	-77
SANTO (DOS) 1965	Espirito Santo Island	International 1924	170	42	84

SAO BRAZ	Azores (Sao Miguel, Santa Maria Islands)	International 1924	-203	141	53
SAPPER HILL 1943	East Falkland Island	International 1924	-355	21	72
SCHWARZECK	Namibia	Bessel 1841 (Namibia)	616	97	-251
SELVAGEM GRANDE 1938	Salvage Island	International 1924	-289	-124	60
SGS 85	Soviet Geodetic System 1985	SGS 85	3	9	-9
S-JTSK	Czechoslovakia (prior to 1 Jan. 1993)	Bessel 1841	589	76	480
SOUTH AMERICAN 1969	MEAN FOR Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Trinidad & Tobago, Venezuela	South American 1969	-57	1	-41
	Argentina		-62	-1	-37
	Bolivia		-61	2	-48
	Brazil		-60	-2	-41
	Chile		-75	-1	-44
	Colombia		-44	6	-36
	Ecuador (Excluding Galapagos)		-48	3	-44

	Islands)		-47	26	-42
	Ecuador (Baltra, Galapagos)		-53	3	-47
	Guyana		-61	2	-33
	Paraguay		-58	0	-44
	Peru		-45	12	-33
	Trinidad & Tobago		-45	8	-33
	Venezuela				
SOUTH ASIA	Singapore	Modified Fischer 1960	7	-10	-26
TANANARIVE OBSERVATORY 1925	Madagascar	International 1924	-189	-242	-91
TIMBALAI 1948	Brunei, East Malaysia (Sabah, Sarawak)	Everest (Sabah & Sarawak)	-679	669	-48
TOKYO	MEAN FOR Japan, Okinawa, South Korea	Bessel 1841	-148	507	685
	Japan		-148	507	685
	Okinawa		-158	507	676
			-146	507	687

	South Korea				
TRISTAN ASTRO 1968	Tristan da Cunha	International 1924	-632	438	-609
VOIROL 1960	Algeria	Clarke 1880	-123	-206	219
VITI LEVU 1916	Fiji (Viti Levu Island)	Clarke 1880	51	391	-36
WAKE-ENIWETOK 1960	Marshall Islands	Hough	102	52	-38
WAKE ISLAND ASTRO 1952	Wake Atoll	International 1924	276	-57	149
WGS 1972	Global Definition	WGS 72	0	0	0
YACARE	Uruguay	International 1924	-155	171	37
ZANDERIJ	Surinam	International 1924	-265	120	-358

APPENDIX 3

SUPPLIED REFERENCE SYSTEM PARAMETER FILES

Geodetic (Latitude/Longitude)

A single REF file named LATLONG is supplied for geodetic coordinates. This file is based on the WGS84 datum. For other datums, use the COPY function in TerrSet Explorer to copy this file to another name and then use Metadata in TerrSet Explorer or Edit along with the data in Appendices 1 and 2 to enter the new datum information.

Universal Transverse Mercator (UTM)

160 REF files are supplied for the UTM system—60 for the northern hemisphere using the WGS84 datum, 60 for the southern hemisphere using the WGS84 datum, 20 for North America based on NAD27 and 20 for North America based on NAD83. The northern WGS84 group has names ranging from UTM-01N to UTM-60N, while the southern WGS84 group has names ranging from UTM-01S to UTM-60S. The NAD27 group (covering zones 1-20) has names ranging from US27TM01 to US27TM20 while those for NAD83 range from US83TM01 to US83TM20. Note that the North American groups all use a North American mean value for the Molodensky constants.

For other datums, the module UTMREF may be used to create the reference system parameter file. In addition, several very specific instances of the UTM system are available in the Miscellaneous group listed later in this appendix.

US State Plane Coordinate System 1927

REF files are supplied for all US State Plane Coordinate Systems based on the Transverse Mercator and Lambert Conformal Conic projections for NAD27 and NAD83. The following table lists these files for the NAD27 datum. The *Projection* column indicates the projection upon which the system is based (L=Lambert Conformal Conic / TM=Transverse Mercator).

State	File name	Title	Projection
-------	-----------	-------	------------

Alabama	SPC27AL1	Alabama State Plane Coordinate System Eastern Zone	TM
	SPC27AL2	Alabama State Plane Coordinate Western Zone	TM
Alaska	SPC27AK0	Alaska State Plane Coordinate System Zone 10	TM
	SPC27AK2	Alaska State Plane Coordinate System Zone 2	TM
	SPC27AK3	Alaska State Plane Coordinate System Zone 3	TM
	SPC27AK4	Alaska State Plane Coordinate System Zone 4	TM
	SPC27AK5	Alaska State Plane Coordinate System Zone 5	TM
	SPC27AK6	Alaska State Plane Coordinate System Zone 6	TM
	SPC27AK7	Alaska State Plane Coordinate System Zone 7	TM
	SPC27AK8	Alaska State Plane Coordinate System Zone 8	TM
	SPC27AK9	Alaska State Plane Coordinate System Zone 9	TM
Arizona	SPC27AZ1	Arizona State Plane Coordinate System Eastern Zone	TM
	SPC27AZ2	Arizona State Plane Coordinate System Central Zone	TM
	SPC27AZ3	Arizona State Plane Coordinate System Western Zone	TM
Arkansas	SPC27AR1	Arkansas State Plane Coordinate System Northern Zone	L
	SPC27AR2	Arkansas State Plane Coordinate System Southern Zone	L
California	SPC27CA1	California State Plane Coordinate System Zone I	L
	SPC27CA2	California State Plane Coordinate System Zone II	L

	SPC27CA3	California State Plane Coordinate System Zone III	L
	SPC27CA4	California State Plane Coordinate System Zone IV	L
	SPC27CA5	California State Plane Coordinate System Zone V	L
	SPC27CA6	California State Plane Coordinate System Zone VI	L
	SPC27CA7	California State Plane Coordinate System Zone VII	L
Colorado	SPC27CO1	Colorado State Plane Coordinate System Northern Zone	L
	SPC27CO2	Colorado State Plane Coordinate System Central Zone	L
	SPC27CO3	Colorado State Plane Coordinate System Southern Zone	L
Connecticut	SPC27CT1	Connecticut State Plane Coordinate System Zone 1	L
Delaware	SPC27DE1	Delaware State Plane Coordinate System Zone 1	TM
Florida	SPC27FL1	Florida State Plane Coordinate System Eastern Zone	TM
	SPC27FL2	Florida State Plane Coordinate System Western Zone	TM
	SPC27FL3	Florida State Plane Coordinate System Northern Zone	TM
Georgia	SPC27GA1	Georgia State Plane Coordinate System Eastern Zone	TM
	SPC27GA2	Georgia State Plane Coordinate System Western Zone	TM
Hawaii	SPC27HI1	Hawaii State Plane Coordinate System Zone 1	TM
	SPC27HI2	Hawaii State Plane Coordinate System Zone 2	TM
	SPC27HI3	Hawaii State Plane Coordinate System Zone 3	TM

	SPC27HI4	Hawaii State Plane Coordinate System Zone 4	TM
	SPC27HI5	Hawaii State Plane Coordinate System Zone 5	TM
Idaho	SPC27ID1	Idaho State Plane Coordinate System Eastern Zone	TM
	SPC27ID2	Idaho State Plane Coordinate System Central Zone	TM
	SPC27ID3	Idaho State Plane Coordinate System Western Zone	TM
Illinois	SPC27IL1	Illinois State Plane Coordinate System Eastern Zone	TM
	SPC27IL2	Illinois State Plane Coordinate System Western Zone	TM
Indiana	SPC27IN1	Indiana State Plane Coordinate System Eastern Zone	TM
	SPC27IN2	Indiana State Plane Coordinate System Western Zone	TM
Iowa	SPC27IA1	Iowa State Plane Coordinate System Northern Zone	L
	SPC27IA2	Iowa State Plane Coordinate System Southern Zone	L
Kansas	SPC27KS1	Kansas State Plane Coordinate System Northern Zone	L
	SPC27KS2	Kansas State Plane Coordinate System Southern Zone	L
Kentucky	SPC27KY1	Kentucky State Plane Coordinate System Northern Zone	L
	SPC27KY2	Kentucky State Plane Coordinate System Southern Zone	L
Louisiana	SPC27LA1	Louisiana State Plane Coordinate System Northern Zone	L
	SPC27LA2	Louisiana State Plane Coordinate System Southern Zone	L
	SPC27LA3	Louisiana State Plane Coordinate System Offshore Zone	L

Maine	SPC27ME1	Maine State Plane Coordinate System Eastern Zone	TM
	SPC27ME2	Maine State Plane Coordinate System Western Zone	TM
Maryland	SPC27MD1	Maryland State Plane Coordinate System Zone 1	L
Massachusetts	SPC27MA1	Massachusetts State Plane Coordinate System Mainland Zone	L
	SPC27MA2	Massachusetts State Plane Coordinate System Island Zone	L
Michigan	SPC27MI1	Current Michigan State Plane Coordinate System Northern Zone	L
	SPC27MI2	Current Michigan State Plane Coordinate System Central Zone	L
	SPC27MI3	Current Michigan State Plane Coordinate System Southern Zone	L
	SPC27MI4	Old Michigan State Plane Coordinate System Eastern Zone	TM
	SPC27MI5	Old Michigan State Plane Coordinate System Central Zone	TM
	SPC27MI6	Old Michigan State Plane Coordinate System Western Zone	TM
Minnesota	SPC27MN1	Minnesota State Plane Coordinate System Northern Zone	L
	SPC27MN2	Minnesota State Plane Coordinate System Central Zone	L
	SPC27MN3	Minnesota State Plane Coordinate System Southern Zone	L
Mississippi	SPC27MS1	Mississippi State Plane Coordinate System Eastern Zone	TM
	SPC27MS2	Mississippi State Plane Coordinate System Western Zone	TM
Missouri	SPC27MO1	Missouri State Plane Coordinate System Eastern Zone	TM
	SPC27MO2	Missouri State Plane Coordinate System Central Zone	TM

	SPC27MO3	Missouri State Plane Coordinate System Western Zone	TM
Montana	SPC27MT1	Montana State Plane Coordinate System Northern Zone	L
	SPC27MT2	Montana State Plane Coordinate System Central Zone	L
	SPC27MT3	Montana State Plane Coordinate System Southern Zone	L
Nebraska	SPC27NE1	Nebraska State Plane Coordinate System Northern Zone	L
	SPC27NE2	Nebraska State Plane Coordinate System Southern Zone	L
Nevada	SPC27NV1	Nevada State Plane Coordinate System Eastern Zone	TM
	SPC27NV2	Nevada State Plane Coordinate System Central Zone	TM
	SPC27NV3	Nevada State Plane Coordinate System Western Zone	TM
New Hampshire	SPC27NH1	New Hampshire State Plane Coordinate System Zone 1	TM
New Jersey	SPC27NJ1	New Jersey State Plane Coordinate System Zone 1	TM
New Mexico	SPC27NM1	New Mexico State Plane Coordinate System Eastern Zone	TM
	SPC27NM2	New Mexico State Plane Coordinate System Central Zone	TM
	SPC27NM3	New Mexico State Plane Coordinate System Western Zone	TM
New York	SPC27NY1	New York State Plane Coordinate System Eastern Zone	TM
	SPC27NY2	New York State Plane Coordinate System Central Zone	TM
	SPC27NY3	New York State Plane Coordinate System Western Zone	TM
	SPC27NY4	New York State Plane Coordinate System Long Island Zone	L

North Carolina	SPC27NC1	North Carolina State Plane Coordinate System Zone 1	L
North Dakota	SPC27ND1	North Dakota State Plane Coordinate System Northern Zone	L
	SPC27ND2	North Dakota State Plane Coordinate System Southern Zone	L
Ohio	SPC27OH1	Ohio State Plane Coordinate System Northern Zone	L
	SPC27OH2	Ohio State Plane Coordinate System Southern Zone	L
Oklahoma	SPC27OK1	Oklahoma State Plane Coordinate System Northern Zone	L
	SPC27OK2	Oklahoma State Plane Coordinate System Southern Zone	L
Oregon	SPC27OR1	Oregon State Plane Coordinate System Northern Zone	L
	SPC27OR2	Oregon State Plane Coordinate System Southern Zone	L
Pennsylvania	SPC27PA1	Pennsylvania State Plane Coordinate System Northern Zone	L
	SPC27PA2	Pennsylvania State Plane Coordinate System Southern Zone	L
Puerto Rico & Virgin Islands	SPC27PR1	Puerto Rico & Virgin Islands State Plane Coordinate System Zone 1	L
	SPC27PR2	Puerto Rico & Virgin Islands State Plane Coordinate System Zone 2 (St. Croix)	L
Rhode Island	SPC27RI1	Rhode Island State Plane Coordinate System Zone 1	TM
Samoa	-none-	-not supported-	L
South Carolina	SPC27SC1	South Carolina State Plane Coordinate System Northern Zone	L
	SPC27SC2	South Carolina State Plane Coordinate System Southern Zone	L
South Dakota	SPC27SD1	South Dakota State Plane Coordinate System Northern Zone	L

	SPC27SD2	South Dakota State Plane Coordinate System Southern Zone	L
Tennessee	SPC27TN1	Tennessee State Plane Coordinate System Zone 1	L
Texas	SPC27TX1	Texas State Plane Coordinate System Northern Zone	L
	SPC27TX2	Texas State Plane Coordinate System North Central Zone	L
	SPC27TX3	Texas State Plane Coordinate System Central Zone	L
	SPC27TX4	Texas State Plane Coordinate System South Central Zone	L
	SPC27TX5	Texas State Plane Coordinate System Southern Zone	L
Utah	SPC27UT1	Utah State Plane Coordinate System Northern Zone	L
	SPC27UT2	Utah State Plane Coordinate System Central Zone	L
	SPC27UT3	Utah State Plane Coordinate System Southern Zone	L
Vermont	SPC27VT1	Vermont State Plane Coordinate System Zone 1	TM
Virginia	SPC27VA1	Virginia State Plane Coordinate System Northern Zone	L
	SPC27VA2	Virginia State Plane Coordinate System Southern Zone	L
Washington	SPC27WA1	Washington State Plane Coordinate System Northern Zone	L
	SPC27WA2	Washington State Plane Coordinate System Southern Zone	L
West Virginia	SPC27WV1	West Virginia State Plane Coordinate System Northern Zone	L
	SPC27WV2	West Virginia State Plane Coordinate System Southern Zone	L
Wisconsin	SPC27WI1	Wisconsin State Plane Coordinate System Northern Zone	L

	SPC27WI2	Wisconsin State Plane Coordinate System Central Zone	L
	SPC27WI3	Wisconsin State Plane Coordinate System Southern Zone	L
Wyoming	SPC27WY1	Wyoming State Plane Coordinate System Eastern Zone	TM
	SPC27WY2	Wyoming State Plane Coordinate System East Central Zone	TM
	SPC27WY3	Wyoming State Plane Coordinate System West Central Zone	TM
	SPC27WY4	Wyoming State Plane Coordinate System Western Zone	TM

US State Plane Coordinate System 1983

REF files are supplied for all US State Plane Coordinate Systems based on the Transverse Mercator and Lambert Conformal Conic projections for NAD27 and NAD83. These are located in the GEOREF subdirectory of your TerrSet program directory. The following table lists these files for the NAD83 datum. The *Proj* column indicates the projection upon which the system is based (L=Lambert Conformal Conic / TM=Transverse Mercator). Note that Samoa did not make the change from NAD27 to NAD83.

State	File name	Title	Proj
Alabama	SPC83AL1	Alabama State Plane Coordinate System Eastern Zone	TM
	SPC83AL2	Alabama State Plane Coordinate System Western Zone	TM
Alaska	SPC83AK0	Alaska State Plane Coordinate System Zone 10	TM
	SPC83AK2	Alaska State Plane Coordinate System Zone 2	TM
	SPC83AK3	Alaska State Plane Coordinate System Zone 3	TM
	SPC83AK4	Alaska State Plane Coordinate System Zone 4	TM
	SPC83AK5	Alaska State Plane Coordinate System Zone 5	TM

	SPC83AK6	Alaska State Plane Coordinate System Zone 6	TM
	SPC83AK7	Alaska State Plane Coordinate System Zone 7	TM
	SPC83AK8	Alaska State Plane Coordinate System Zone 8	TM
	SPC83AK9	Alaska State Plane Coordinate System Zone 9	TM
Arizona	SPC83AZ1	Arizona State Plane Coordinate System Eastern Zone	TM
	SPC83AZ2	Arizona State Plane Coordinate System Central Zone	TM
	SPC83AZ3	Arizona State Plane Coordinate System Western Zone	TM
Arkansas	SPC83AR1	Arkansas State Plane Coordinate System Northern Zone	L
	SPC83AR2	Arkansas State Plane Coordinate System Southern Zone	L
California	SPC83CA1	California State Plane Coordinate System Zone I	L
	SPC83CA2	California State Plane Coordinate System Zone II	L
	SPC83CA3	California State Plane Coordinate System Zone III	L
	SPC83CA4	California State Plane Coordinate System Zone IV	L
	SPC83CA5	California State Plane Coordinate System Zone V	L
	SPC83CA6	California State Plane Coordinate System Zone VI	L
Colorado	SPC83CO1	Colorado State Plane Coordinate System Northern Zone	L
	SPC83CO2	Colorado State Plane Coordinate System Central Zone	L
	SPC83CO3	Colorado State Plane Coordinate System Southern Zone	L

Connecticut	SPC83CT1	Connecticut State Plane Coordinate System Zone 1	L
Delaware	SPC83DE1	Delaware State Plane Coordinate System Zone 1	TM
Florida	SPC83FL1	Florida State Plane Coordinate System Eastern Zone	TM
	SPC83FL2	Florida State Plane Coordinate System Western Zone	TM
	SPC83FL3	Florida State Plane Coordinate System Northern Zone	TM
Georgia	SPC83GA1	Georgia State Plane Coordinate System Eastern Zone	TM
	SPC83GA2	Georgia State Plane Coordinate System Western Zone	TM
Hawaii	SPC83HI1	Hawaii State Plane Coordinate System Zone 1	TM
	SPC83HI2	Hawaii State Plane Coordinate System Zone 2	TM
	SPC83HI3	Hawaii State Plane Coordinate System Zone 3	TM
	SPC83HI4	Hawaii State Plane Coordinate System Zone 4	TM
	SPC83HI5	Hawaii State Plane Coordinate System Zone 5	TM
Idaho	SPC83ID1	Idaho State Plane Coordinate System Eastern Zone	TM
	SPC83ID2	Idaho State Plane Coordinate System Central Zone	TM
	SPC83ID3	Idaho State Plane Coordinate System Western Zone	TM
Illinois	SPC83IL1	Illinois State Plane Coordinate System Eastern Zone	TM
	SPC83IL2	Illinois State Plane Coordinate System Western Zone	TM
Indiana	SPC83IN1	Indiana State Plane Coordinate System Eastern Zone	TM

	SPC83IN2	Indiana State Plane Coordinate System Western Zone	TM
Iowa	SPC83IA1	Iowa State Plane Coordinate System Northern Zone	L
	SPC83IA2	Iowa State Plane Coordinate System Southern Zone	L
Kansas	SPC83KS1	Kansas State Plane Coordinate System Northern Zone	L
	SPC83KS2	Kansas State Plane Coordinate System Southern Zone	L
Kentucky	SPC83KY1	Kentucky State Plane Coordinate System Northern Zone	L
	SPC83KY2	Kentucky State Plane Coordinate System Southern Zone	L
Louisiana	SPC83LA1	Louisiana State Plane Coordinate System Northern Zone	L
	SPC83LA2	Louisiana State Plane Coordinate System Southern Zone	L
	SPC83LA3	Louisiana State Plane Coordinate System Offshore Zone	L
Maine	SPC83ME1	Maine State Plane Coordinate System Eastern Zone	TM
	SPC83ME2	Maine State Plane Coordinate System Western Zone	TM
Maryland	SPC83MD1	Maryland State Plane Coordinate System Zone 1	L
Massachusetts	SPC83MA1	Massachusetts State Plane Coordinate System Mainland Zone	L
	SPC83MA2	Massachusetts State Plane Coordinate System Island Zone	L
Michigan	SPC83MI1	current Michigan State Plane Coordinate System Northern Zone	L
	SPC83MI2	current Michigan State Plane Coordinate System Central Zone	L
	SPC83MI3	current Michigan State Plane Coordinate System Southern Zone	L

	SPC83MI4	old Michigan State Plane Coordinate System Eastern Zone	TM
	SPC83MI5	old Michigan State Plane Coordinate System Central Zone	TM
	SPC83MI6	old Michigan State Plane Coordinate System Western Zone	TM
Minnesota	SPC83MN1	Minnesota State Plane Coordinate System Northern Zone	L
	SPC83MN2	Minnesota State Plane Coordinate System Central Zone	L
	SPC83MN3	Minnesota State Plane Coordinate System Southern Zone	L
Mississippi	SPC83MS1	Mississippi State Plane Coordinate System Eastern Zone	TM
	SPC83MS2	Mississippi State Plane Coordinate System Western Zone	TM
Missouri	SPC83MO1	Missouri State Plane Coordinate System Eastern Zone	TM
	SPC83MO2	Missouri State Plane Coordinate System Central Zone	TM
	SPC83MO3	Missouri State Plane Coordinate System Western Zone	TM
Montana	SPC83MT1	Montana State Plane Coordinate System Single Zone	L
Nebraska	SPC83NE1	Nebraska State Plane Coordinate System Single Zone	L
Nevada	SPC83NV1	Nevada State Plane Coordinate System Eastern Zone	TM
	SPC83NV2	Nevada State Plane Coordinate System Central Zone	TM
	SPC83NV3	Nevada State Plane Coordinate System Western Zone	TM
New Hampshire	SPC83NH1	New Hampshire State Plane Coordinate System Zone 1	TM
New Jersey	SPC83NJ1	New Jersey State Plane Coordinate System Zone 1	TM

New Mexico	SPC83NM1	New Mexico State Plane Coordinate System Eastern Zone	TM
	SPC83NM2	New Mexico State Plane Coordinate System Central Zone	TM
	SPC83NM3	New Mexico State Plane Coordinate System Western Zone	TM
New York	SPC83NY1	New York State Plane Coordinate System Eastern Zone	TM
	SPC83NY2	New York State Plane Coordinate System Central Zone	TM
	SPC83NY3	New York State Plane Coordinate System Western Zone	TM
	SPC83NY4	New York State Plane Coordinate System Long Island Zone	L
North Carolina	SPC83NC1	North Carolina State Plane Coordinate System Zone 1	L
North Dakota	SPC83ND1	North Dakota State Plane Coordinate System Northern Zone	L
	SPC83ND2	North Dakota State Plane Coordinate System Southern Zone	L
Ohio	SPC83OH1	Ohio State Plane Coordinate System Northern Zone	L
	SPC83OH2	Ohio State Plane Coordinate System Southern Zone	L
Oklahoma	SPC83OK1	Oklahoma State Plane Coordinate System Northern Zone	L
	SPC83OK2	Oklahoma State Plane Coordinate System Southern Zone	L
Oregon	SPC83OR1	Oregon State Plane Coordinate System Northern Zone	L
	SPC83OR2	Oregon State Plane Coordinate System Southern Zone	L
Pennsylvania	SPC83PA1	Pennsylvania State Plane Coordinate System Northern Zone	L
	SPC83PA2	Pennsylvania State Plane Coordinate System Southern Zone	L

Puerto Rico & Virgin Islands	SPC83PR1	Puerto Rico & Virgin Islands State Plane Coordinate System Zone 1	L
	SPC83PR2	Puerto Rico & Virgin Islands State Plane Coordinate System Zone 2 (St. Croix)	L
Rhode Island	SPC83RI1	Rhode Island State Plane Coordinate System Zone 1	TM
Samoa	-none-	-not supported-	L
South Carolina	SPC83SC1	South Carolina State Plane Coordinate System, single zone	L
South Dakota	SPC83SD1	South Dakota State Plane Coordinate System Northern Zone	L
	SPC83SD2	South Dakota State Plane Coordinate System Southern Zone	L
Tennessee	SPC83TN1	Tennessee State Plane Coordinate System Zone 1	L
Texas	SPC83TX1	Texas State Plane Coordinate System Northern Zone	L
	SPC83TX2	Texas State Plane Coordinate System North Central Zone	L
	SPC83TX3	Texas State Plane Coordinate System Central Zone	L
	SPC83TX4	Texas State Plane Coordinate System South Central Zone	L
	SPC83TX5	Texas State Plane Coordinate System Southern Zone	L
Utah	SPC83UT1	Utah State Plane Coordinate System Northern Zone	L
	SPC83UT2	Utah State Plane Coordinate System Central Zone	L
	SPC83UT3	Utah State Plane Coordinate System Southern Zone	L
Vermont	SPC83VT1	Vermont State Plane Coordinate System Zone 1	TM
Virginia	SPC83VA1	Virginia State Plane Coordinate System Northern Zone	L

	SPC83VA2	Virginia State Plane Coordinate System Southern Zone	L
Washington	SPC83WA1	Washington State Plane Coordinate System Northern Zone	L
	SPC83WA2	Washington State Plane Coordinate System Southern Zone	L
West Virginia	SPC83WV1	West Virginia State Plane Coordinate System Northern Zone	L
	SPC83WV2	West Virginia State Plane Coordinate System Southern Zone	L
Wisconsin	SPC83WI1	Wisconsin State Plane Coordinate System Northern Zone	L
	SPC83WI2	Wisconsin State Plane Coordinate System Central Zone	L
	SPC83WI3	Wisconsin State Plane Coordinate System Southern Zone	L
Wyoming	SPC83WY1	Wyoming State Plane Coordinate System Eastern Zone	TM
	SPC83WY2	Wyoming State Plane Coordinate System East Central Zone	TM
	SPC83WY3	Wyoming State Plane Coordinate System West Central Zone	TM
	SPC83WY4	Wyoming State Plane Coordinate System Western Zone	TM

Gauss-Kruger

The Gauss-Kruger reference system is used primarily in the countries of the former Soviet Union and Eastern Bloc. REF files are included for zones 1-32. All use the Pulkovo 1942 datum and Krassovsky 1940 ellipsoid. The names of these files are GK01_P42.ref through GK32_P42.ref. The Gauss-Kruger projection is identical to the Transverse Mercator projection. (Note that the alternate spelling “Gauss-Krueger” is also acceptable in a REF file.)

Miscellaneous

Note that TerrSet also includes miscellaneous regional and local reference system files. A list can be found within the TerrSet Help System, in the Notes section of the PROJECT module.

APPENDIX 4

ERROR PROPAGATION FORMULAS

Arithmetic Operations

In the formulas below, S refers to RMS error. Formulas are presented for each of the arithmetic operations performed by the OVERLAY and SCALAR modules in TerrSet. In OVERLAY operations, S_x would refer to the RMS error in Map X, S_y refers to the RMS error in Map Y, and S_z refers to the RMS error in the final map produced, Map Z. In SCALAR operations, K refers to a user-defined constant. Often, error is computed as a uniform value for the entire resulting map. However, in some cases, the formula depends upon the values in corresponding cells of the input maps. These are referred to as X and Y. In these instances, the error would vary over the face of the map and would thus need to be computed separately for each cell. Note that these formulas assume that the input maps are uncorrelated with each other.

Overlay Add / Subtract

$$(e.g., Z = X + Y \text{ or } Z = X - Y) \quad S_z = \sqrt{S_x^2 + S_y^2}$$

Overlay Multiply / Divide

$$(e.g., Z = X \times Y \text{ or } Z = \frac{X}{Y}) \quad S_z = \sqrt{(S_x^2 \cdot Y^2) + (S_y^2 \cdot X^2)}$$

Scalar Add / Subtract

$$(e.g., Z = X + k \text{ or } Z = X - k) \quad S_z = S_x \text{ i.e., no change}$$

Scalar Multiply

$$(e.g., Z = X \times k) \quad S_z = S_x \times k$$

Scalar Divide

$$(e.g., Z = \frac{X}{k}) \quad S_z = \frac{S_x}{k}$$

Scalar Exponentiate

$$(e.g., Z = X^k) \quad S_z = \sqrt{k^2 \cdot X^{(2(k-1))} \cdot S_x^2}$$

Logical Operations

For Boolean operations, logical errors may be expressed by the proportion of cells that are expected to be in error (e) in the category being overlaid. Since a Boolean overlay requires two input maps, the error of the output map will be a function of the errors in the two input maps and the logic operation performed as follows:

Logical AND

$$e_z = e_x + (1 - e_x) \times e_y \quad \text{or equivalently} \quad e_z = e_x + e_y - (e_x \times e_y)$$

Logical OR

$$e_z = e_x \times e_y$$

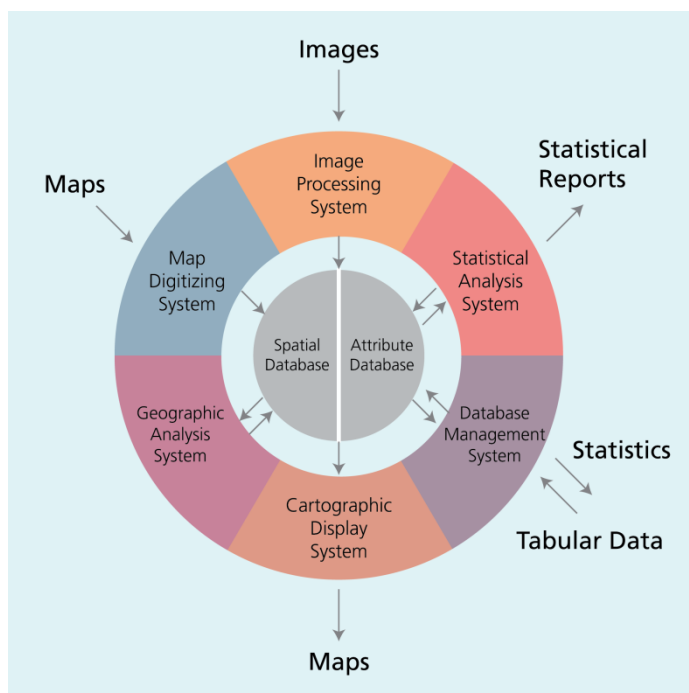
APPENDIX 5

INTRODUCTION TO GIS

A Geographic Information System (GIS) is a computer-assisted system for the acquisition, storage, analysis and display of geographic data. Today, a variety of software tools are available to assist this activity. However, they can differ from one another quite significantly, in part because of the way they represent and work with geographic data, but also because of the relative emphasis they place on these various operations. In this chapter, we will explore these differences as a means of understanding the special characteristics of the TerrSet system.

Components of a GIS

Although we think of a GIS as a single piece of software, it is typically made up of a variety of different components. The figure below gives a broad overview of the software components typically found in a GIS. Not all systems have all of these elements, but to be a true GIS, an essential group must be found.



Spatial and Attribute Database

Central to the system is the database—a collection of maps and associated information in digital form. Since the database is concerned with earth surface features, it can be seen to be comprised of two elements—a spatial database describing the geography (shape and position) of earth surface features, and an attribute database describing the characteristics or qualities of these features. Thus, for example, we might have a property parcel defined in the spatial database and qualities such as its landuse, owner, property valuation, and so on, in the attribute database.

In some systems, the spatial and attribute databases are rigidly distinguished from one another, while in others they are closely integrated into a single entity—hence the line extending only half-way through the middle circle of figure. TerrSet is of the type that integrates the two components into one. However, it also offers the option of keeping some elements of the attribute database quite separate. This will be explored further below when we examine techniques for the digital representation of map data.

Cartographic Display System

Surrounding the central database, we have a series of software components. The most basic of these is the Cartographic Display System. The cartographic display system allows one to take selected elements of the database and produce map output on the screen or some hardcopy device such as a printer or plotter. The range of cartographic production capabilities among GIS software systems is great. Most provide only very basic cartographic output, and rely upon the use of high quality publication software systems for more sophisticated production needs such as color separation.

TerrSet allows for highly interactive and flexible on-screen cartographic composition, including the specification of multiple data layers, customization and positioning of map elements such as annotation, scale bars, insets and so forth, and customized color and symbol sets. TerrSet map compositions may be saved for later display, printed to Windows-compatible devices, and exported in a variety of common desktop publishing formats.

Software systems that are only capable of accessing and displaying elements of the database are often referred to as Viewers or Electronic Atlases.

Map Digitizing System

After cartographic display, the next most essential element is a Map Digitizing System. With a map digitizing system, one can take existing paper maps and convert them into digital form, thus further developing the database. In the most common method of digitizing, one attaches the paper map to a digitizing tablet or board, then traces the features of interest with a stylus or puck according to the procedures required by the digitizing software. Many map digitizing systems also allow for editing of the digitized data.

Scanners may also be used to digitize data such as aerial photographs. The result is a graphic image, rather than the outlines of features that are created with a digitizing tablet. Scanning software typically provides users with a variety of standard graphics file formats for export. These files are then imported into the GIS. TerrSet supports import of TIF and BMP graphics file formats.

Digitizing packages, Computer Assisted Design (CAD), and Coordinate Geometry (COGO) are examples of software systems that provide the ability to add digitized map information to the database, in addition to providing cartographic display capabilities.

Database Management System

The next logical component in a GIS is a Database Management System (DBMS). Traditionally, this term refers to a type of software that is used to input, manage and analyze attribute data. It is also used in that sense here, although we need to recognize that spatial database management is also required. Thus, a GIS typically incorporates not only a traditional DBMS, but also a variety of utilities to manage the spatial and attribute components of the geographic data stored.

With a DBMS, it is possible to enter attribute data, such as tabular information and statistics, and subsequently extract specialized tabulations and statistical summaries to provide new tabular reports. However, most importantly, a DBMS provides us with the ability to analyze attribute data. Many map analyses have no true spatial component, and for these, a DBMS will often function quite well. For example, we might use the system to find all property parcels where the head of the household is single but with one or more child dependents, and to produce a map of the result. The final product (a map) is certainly spatial, but the analysis itself has no spatial qualities whatsoever. Thus, the double arrows between the DBMS and the attribute database in the figure above signify this distinctly non-spatial form of data analysis.

In TerrSet, a DBMS is provided by Database Workshop. One can perform analyses in Database Workshop, then immediately apply the results to the proper spatial data, viewing the results as a map. In addition to Database Workshop, an extensive set of program modules is also available for spatial and attribute data management.

Software that provides cartographic display, map digitizing, and database query capabilities are sometimes referred to as Automated Mapping and Facilities Management (AM/FM) systems.

Geographic Analysis System

Up to this point, we have described a very powerful set of capabilities—the ability to digitize spatial data and attach attributes to the features stored, to analyze these data based on those attributes, and to map out the result. Indeed, there are a variety of systems on the market that have just this set of abilities, many of which will call themselves a GIS. But useful as this is, such a set of capabilities does not necessarily constitute a full GIS. The missing component is the ability to analyze data based on truly spatial characteristics. For this we need a Geographic Analysis System.

With a Geographic Analysis System, we extend the capabilities of traditional database query to include the ability to analyze data based on their location. Perhaps the simplest example of this is to consider what happens when we are concerned with the joint occurrence of features with different geographies. For example, suppose we want to find all areas of residential land on bedrock types associated with high levels of radon gas. This is a problem that a traditional DBMS simply cannot solve because bedrock types and landuse divisions do not share the same geography. Traditional database query is fine as long as we are talking about attributes belonging to the same features. But when the features are different, it cannot cope. For this we need a GIS. In fact, it is this ability to compare different features based on their common geographic occurrence that is the hallmark of GIS. This analysis is accomplished through a process called *overlay*, thus named because it is identical in character to overlaying transparent maps of the two entity groups on top of one another.

Like the DBMS, the Geographic Analysis System is seen in the figure above to have a two-way interaction with the database—the process is distinctly analytical in character. Thus, while it may access data from the database, it may equally contribute the results of that analysis as a new addition to the database. For example, we might look for the joint occurrence of lands on steep slopes with erodible soils under agriculture and call the result a map of soil erosion risk. This risk map was not in the original database, but was derived based on existing data and a set of specified relationships. Thus the analytical capabilities of the Geographic Analysis System and the DBMS play a vital role in extending the database through the addition of knowledge of relationships between features.

While overlay is still the hallmark of GIS, computer-assisted geographic analysis has matured enormously over the past several years. However, for now it is sufficient to note that it is this distinctly geographic component that gives a true GIS its identity. In TerrSet, these abilities are extensive and form the foundation of the software system.

Image Processing System

In addition to these essential elements of a GIS—a cartographic display system, a map digitizing system, a database management system and a geographic analysis system—some software systems also include the ability to analyze remotely sensed images and provide specialized statistical analyses. TerrSet is of this type. Image processing software allows one to take raw remotely sensed imagery (such as Landsat or SPOT satellite imagery) and convert it into interpreted map data according to various classification procedures. In recognition of its major importance as a technique for data acquisition, TerrSet offers a broad set of tools for the computer-assisted interpretation of remotely sensed data.

Statistical Analysis System

For statistical analysis, TerrSet offers both traditional statistical procedures as well as some specialized routines for the statistical analysis of spatial data. Geographers have developed a series of specialized routines for the statistical description of spatial data, partly because of the special character of spatial data, but also because spatial data pose special problems for inferences drawn from statistical procedures.

Decision Support System

While decision support is one of the most important functions of a GIS, tools designed especially for this are relatively few in most GIS software. However, TerrSet includes several modules specifically developed to aid in the resource allocation decision making process. These include modules that incorporate error into the process, help in the construction of multi-criteria suitability maps under varying levels of tradeoff, and address allocation decisions when there are multiple objectives involved. Used in conjunction with the other components of the system, these modules provide a powerful tool for resource allocation decision makers.

Map Data Representation

The way in which the software components mentioned above are combined is one aspect of how Geographic Information Systems vary. However, an even more fundamental distinction is how they represent map data in digital form.

A Geographic Information System stores two types of data that are found on a map—the geographic definitions of earth surface features and the attributes or qualities that those features possess. Not all systems use the same logic for achieving this. Nearly all, however, use one or a combination of both of the fundamental map representation techniques: *vector* and *raster*.

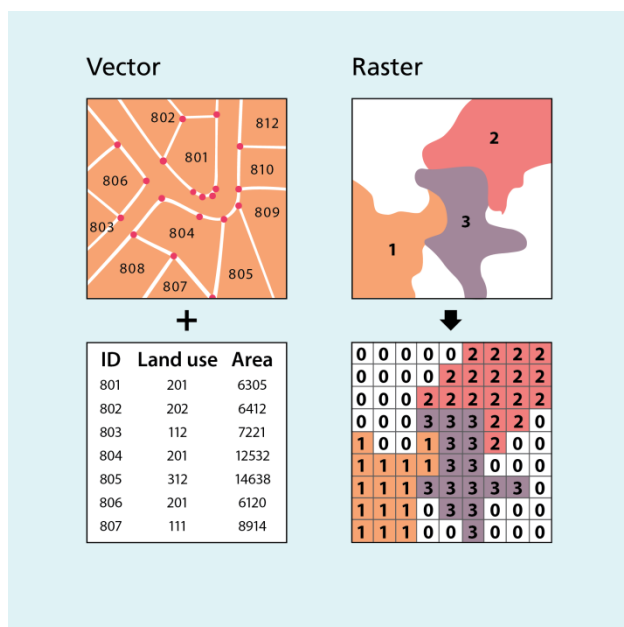
Vector

With vector representation, the boundaries or the course of the features are defined by a series of points that, when joined with straight lines, form the graphic representation of that feature. The points themselves are encoded with a pair of numbers giving the X and Y coordinates in systems such as latitude/longitude or Universal Transverse Mercator grid coordinates. The attributes of features are then stored with a traditional database management (DBMS) software program. For example, a vector map of property parcels might be tied to an attribute database of information containing the address, owner's name, property valuation and landuse. The link between these two data files can be a simple identifier number that is given to each feature in the map (figure below).

Raster

The second major form of representation is known as raster. With raster systems, the graphic representation of features and the attributes they possess are merged into unified data files. In fact, we typically do not define features at all. Rather, the study area is subdivided into a fine mesh of grid cells in which we record the condition or attribute of the earth's surface at that point (figure below). Each cell is given a numeric value which may then represent either a feature identifier, a qualitative attribute code or a quantitative attribute value. For example, a cell could have the value "6" to indicate that it belongs to District 6 (a feature identifier), or that it is covered by soil type 6 (a qualitative attribute), or that it is 6 meters above sea level (a quantitative attribute value). Although the data we store in these grid cells do not necessarily refer to phenomena that can be seen in the environment, the data grids themselves can be thought of as *images* or *layers*, each depicting one type of information over the mapped region. This information can be made visible through the use of a raster display. In a raster display, such as the screen on your computer, there is also a grid of small cells called *pixels*. The word pixel is a contraction of the term *picture element*. Pixels can be made to vary in their color, shape or grey tone. To make an image, the cell values in the data grid are used to regulate directly the graphic appearance of their corresponding pixels. Thus, in a raster system, the data directly controls the visible form we see.

Raster versus Vector



Raster systems are typically data intensive (although good data compaction techniques exist) since they must record data at every cell location regardless of whether that cell holds information that is of interest or not. However, the advantage is that geographical space is uniformly defined in a simple and predictable fashion. As a result, raster systems have substantially more analytical power than their vector counterparts in the analysis of continuous space¹ and are thus ideally suited to the study of data that are continuously changing over space such as terrain, vegetation biomass, rainfall and the like. The second advantage of raster is that its structure closely matches the architecture of digital computers. As a result, raster systems tend to be very rapid in the evaluation of problems that involve various mathematical combinations of the data in multiple layers. Hence they are excellent for evaluating environmental models such as soil erosion potential and forest management suitability. In addition, since satellite imagery employs a raster structure, most raster systems can easily incorporate these data, and some provide full image processing capabilities.

While raster systems are predominantly analysis oriented, vector systems tend to be more database management oriented. Vector systems are quite efficient in their storage of map data because they only store the boundaries of features and not that which is inside those boundaries. Because the graphic representation of features is directly linked to the attribute database, vector systems usually allow one to roam around the graphic display with a mouse and query the attributes associated with a displayed feature, such as the distance between points or along lines, the areas of regions defined on the screen, and so on. In addition, they can produce simple thematic maps of database queries, such as one showing all sewer line sections over one meter in diameter installed before 1940.

Compared to their raster counterparts, vector systems do not have as extensive a range of capabilities for analyses over continuous space. They do, however, excel at problems concerning movements over a network and can undertake the most fundamental of GIS operations that will be sketched out below. For many, it is the simple database management functions and excellent mapping capabilities that make vector systems attractive. Because of the close affinity between the logic of vector representation and traditional map production, vector systems are used to produce maps that are indistinguishable from those produced by traditional means. As a result, vector systems are very popular in municipal applications where issues of engineering map production and database management predominate.

Raster and vector systems each have their special strengths. As a result, TerrSet incorporates elements from both representational techniques. Though it is primarily a raster analytical system, TerrSet does employ vector data structures as a major form of map data display and exchange. In addition, fundamental aspects of vector database management are also provided.

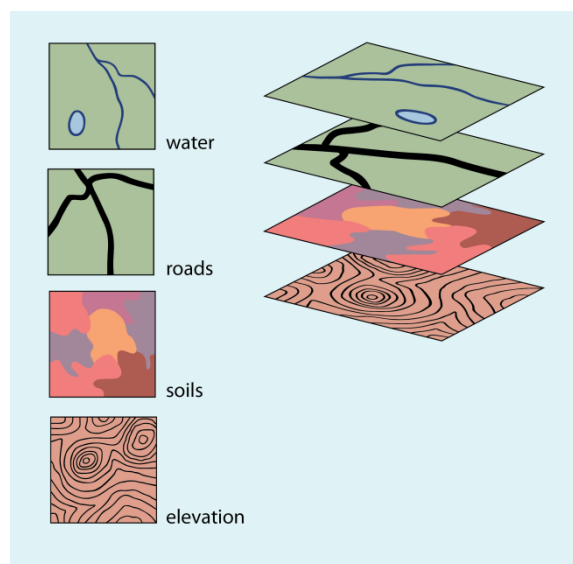
Geographic Database Concepts

Organization

Whether we use a raster or vector logic for spatial representation, we begin to see that a geographic database—a complete database for a given region—is organized in a fashion similar to a collection of maps (figure below). Vector systems may come closest to this

¹ The basic data structure of vector systems can best be described as a network. As a result, it is not surprising to find that vector systems have excellent capabilities for the analysis of network space. Thus the difference between raster and vector is less one of inherent ability than one of the difference in the types of space they describe.

logic with what are known as *coverages*—map-like collections that contain the geographic definitions of a set of features and their associated attribute tables. However, they differ from maps in two ways. First, each will typically contain information on only a single feature type, such as property parcels, soils polygons, and the like. Second, they may contain a whole series of attributes that pertain to those features, such as a set of census information for city blocks.



Raster systems also use this map-like logic, but usually divide data sets into unitary *layers*. A layer contains all the data for a single attribute. Thus one might have a soils layer, a roads layer and a landuse layer. A few raster systems, including TerrSet, can link a feature identifier layer (a layer that contains the identifiers of the features located at each grid cell) with attribute tables. More commonly, separate layers exist for each attribute and on-screen displays and paper maps are produced from these, either singly or in combination.

Although there are subtle differences, for all intents and purposes, raster layers and vector coverages can be thought of as simply different manifestations of the same concept—the organization of the database into elementary map-like themes. Layers and coverages differ from traditional paper maps, however, in an important way. When map data are encoded in digital form (digitized), scale differences are removed. The digital data may be displayed or printed at any scale. More importantly, digital data layers that were derived from paper maps of different scales, but covering the same geographic area, may be combined.

In addition, many GIS packages, including TerrSet, provide utilities for changing the projection and reference system of digital layers. This allows multiple layers, digitized from maps having various projections and reference systems, to be converted to a common system.

With the ability to manage differences of scale, projection and reference system, layers can be merged with ease, eliminating a problem that has traditionally hampered planning activities with paper maps. It is important to note, however, that the issue of *resolution* of the information in the data layers remains. Although features digitized from a poster-sized world map *could* be combined in a GIS with features digitized from a very large scale local map, such as a city street map, this would normally not be done. The level of accuracy and detail of the digital data can only be as good as that of the original maps.

Georeferencing

All spatial data files in a GIS are georeferenced. Georeferencing refers to the location of a layer or coverage in space as defined by a known coordinate referencing system. With raster images, a common form of georeferencing is to indicate the reference system (e.g., latitude/longitude), the reference units (e.g., degrees) and the coordinate positions of the left, right, top and bottom edges of the image. The same is true of vector data files, although the left, right, top and bottom edges now refer to what is commonly called the *bounding rectangle* of the coverage—a rectangle which defines the limits of the mapped area.² This information is particularly important in an integrated GIS such as TerrSet since it allows raster and vector files to be related to one another in a reliable and meaningful way. It is also vital for the referencing of data values to actual positions on the ground.

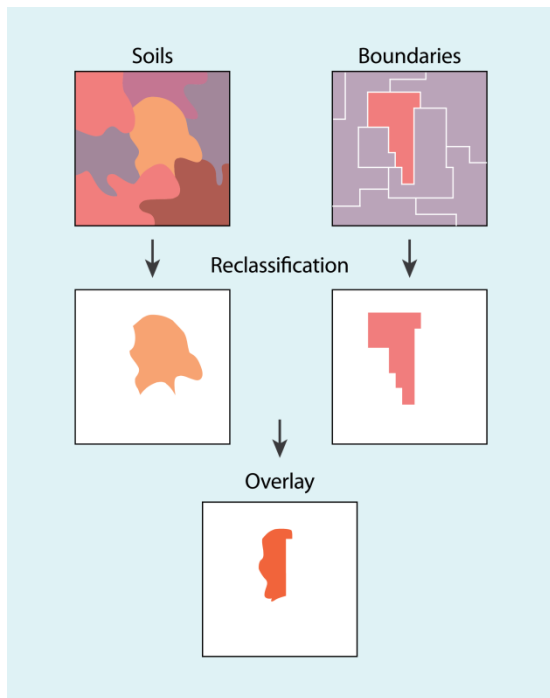
Georeferencing is an extremely important consideration when using GIS. Therefore a separate chapter later in this volume treats this topic in detail.

² The bounding rectangle is defined by the study region of interest and does not necessarily refer to the actual minimum and maximum coordinates in the data file.

Analysis in GIS

The organization of the database into layers is not simply for reasons of organizational clarity. Rather, it is to provide rapid access to the data elements required for geographic analysis. Indeed, the *raison d'être* for GIS is to provide a medium for geographic analysis.

The analytical characteristics of GIS can be looked at in two ways. First, one can look at the tools that GIS provides. Then one can look at the kinds of operations that GIS allows. Regardless of whether we are using a raster or a vector system, we will tend to find that the tools fall into four basic groups and that the operations undertaken fall into three.



Analytical Tools

Database Query

The most fundamental of all tools provided by a GIS are those involved with Database Query. Database query simply asks questions about the currently-stored information. In some cases, we query by location—*what landuse is at this location?* In other cases, we query by attribute—*what areas have high levels of radon gas?* Sometimes we undertake simple queries such as those just illustrated, and at other times we ask about complex combinations of conditions—*show me all wetlands that are larger than 1 hectare and that are adjacent to industrial lands.*

In most systems, including TerrSet, these query operations are undertaken in two steps. The first step, called a *reclassification*, creates a new layer of each individual condition of interest (see figure). For example, consider a query to find residential areas on bedrock associated with high levels of radon gas. The first step would be to create a layer of residential areas alone by reclassifying all landuse codes into only two—a 1 whenever an area is residential and a 0 for all other cases. The resulting layer is known as a

Boolean layer since it shows only those areas that meet the condition (1 = true, residential) and those that don't (0 = false, not residential). Boolean layers are also called *logical* layers since they show only true/false relationships. They are also sometimes called *binary* layers since they contain only zeros and ones. We will avoid using that term, however, since it also describes a particular kind of data storage format. Here we will call them Boolean layers.

Once the residential layer has been created, a geology layer is then also reclassified to create a Boolean layer showing areas with bedrock associated with high levels of radon gas. At this point we can combine the two conditions using an overlay operation (figure above). As mentioned previously, it is only a GIS that can combine conditions such as this that involve features with different geographies. Typically, an overlay operation in GIS will allow the production of new layers based on some logical or mathematical combination of two or more input layers. In the case of database query, the key logical operations of interest are the AND and OR relational operators, also known as the INTERSECTION and UNION operations respectively. Here we are looking for cases of residential land AND high radon gas—the logical intersection of our two Boolean layers.

Map Algebra

The second set of tools that a GIS will typically provide is that for combining map layers mathematically. Modeling requires the ability to combine layers according to various mathematical equations. For example, we might have an equation that predicts mean annual temperature as a result of altitude. Or, as another example, consider the possibility of creating a soil erosion potential map based on

factors of soil erodability, slope gradient and rainfall intensity. Clearly, we need the ability to modify data values in our map layers by various mathematical operations and transformations and to combine factors mathematically to produce the final result.

The Map Algebra tools will typically provide three different kinds of operations:

1. the ability to mathematically modify the attribute data values by a constant (i.e., scalar arithmetic);
2. the ability to mathematically transform attribute data values by a standard operation (such as the trigonometric functions, log transformations and so on);
3. the ability to mathematically combine (such as add, subtract, multiply, divide) different data layers to produce a composite result.

This third operation is simply another form of overlay—mathematical overlay, as opposed to the logical overlay of database query.

To illustrate this, consider a model for snow melt in densely forested areas:³

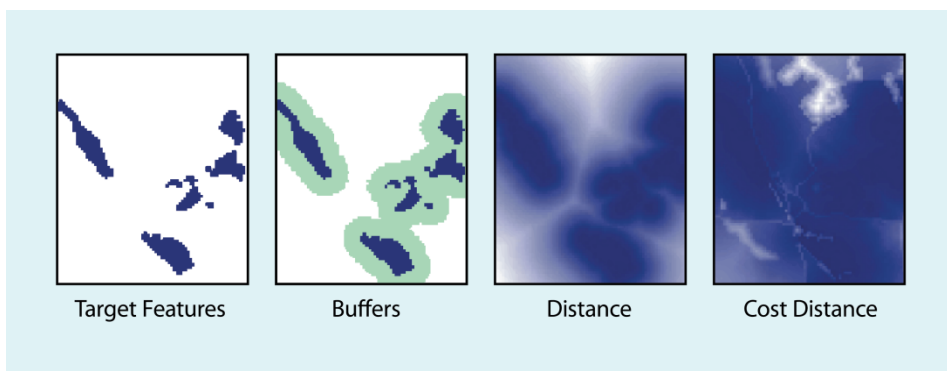
$$M = (0.19T + 0.17D)$$

where M is the melt rate in cm/day, T is the air temperature and D is the dewpoint temperature. Given layers of the air temperatures and dewpoints for a region of this type, we could clearly produce a snow melt rate map. To do so would require multiplying the temperature layer by 0.19 (a scalar operation), the dewpoint layer by 0.17 (another scalar operation) and then using overlay to add the two results. While simple in concept, this ability to treat map layers as variables in algebraic formulas is an enormously powerful capability.

Distance Operators

The third tool group provided by GIS consists of the Distance Operators. As the name suggests, these are a set of techniques where distance plays a key role in the analysis undertaken. Virtually all systems provide the tools to construct buffer zones—areas within a specified distance of designated target features. Some can also evaluate the distance of all locations to the nearest of a set of designated features, while others can even incorporate frictional effects and barriers in distance calculations (see figure below).

When frictional effects are incorporated, the distance calculated is often referred to as a *cost distance*. This name is used because movement through space can be considered to incur costs, either in money, time or effort. Frictions increase those costs. When the costs of movement from one or more locations are evaluated for an entire region, we often refer to the result as a *cost surface*. In this



case, areas of low cost (presumably near to the starting point) can be seen as valleys and areas of high cost as hills. A cost surface thus has its lowest point(s) at the starting location(s) and its highest point(s) at the locations that are farthest away (in the sense of the greatest accumulated cost).⁴

There may be cases in which frictions do not affect the cost of movement the same

³ Equation taken from Dunne, T., and Leopold, L.B., (1978) *Water in Environmental Planning*, (W.H. Freeman and Co.: San Francisco), 480.

⁴ It should be noted here that a cost surface as just described can only be evaluated with a raster system. For vector systems, the closest equivalent would be cost distances evaluated over a network. Here we see a simple, but very powerful, illustration of the differences between raster and vector systems in how they conceive of space.

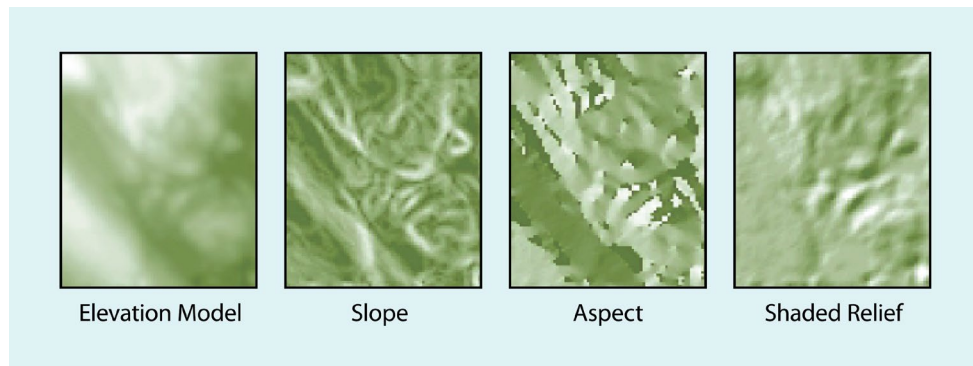
way in all directions. In other words, they act *anisotropically*. For example, going up a steep slope might incur a cost that would be higher than going down the same steep slope. Thus the direction of movement through the friction is important, and must be taken into account when developing the cost surface. TerrSet provides modules to model this type of cost surface which are explained in detail in the **Anisotropic Cost Analysis** chapter.

Given the concept of a cost surface, Geographic Information Systems also commonly offer *least-cost path analysis*—another important distance operation. As the name suggests, our concern is to evaluate the least-cost path between two locations. The cost surface provides the needed information for this to be evaluated.

Regardless of how distance is evaluated, by straight line distance or by cost distance, another commonly provided tool is *allocation*. With allocation, we assign locations to the nearest of a set of designated features. For example, we might establish a set of health facilities and then wish to allocate residents to their nearest facility, where "near" might mean linear distance, or a cost distance such as travel time.

Context Operators

Finally, most Geographic Information Systems provide a variety of Context Operators (also known as *neighborhood* or *local operators*). With context operators, we create new layers based on the information on an existing map and the context in which it is found. One of the simplest examples of this is *surface analysis* where we use a digital elevation model to produce a slope layer by examining the heights of locations in comparison to the heights of neighboring locations. In a similar fashion, the aspect (the direction of maximum downward slope) can also be evaluated. We might also position an artificial light source and calculate a shaded relief model. These context operator products and the elevation model from which they were derived are illustrated in the figure below.



A second good example of a context operator is a *digital filter*. Digital filters operate by changing values according to the character of neighboring values. For example, a surface of terrain heights can be smoothed by replacing values with the average of the original height and all neighboring heights. Digital filters have a broad range of applications in GIS and remote sensing, ranging from noise removal to the visual enhancement of images.

Because of their simple and uniform data structure, raster systems tend to offer a broad range of context operators. In TerrSet, for example, these include surface analysis and digital filtering, identification of contiguous areas, watershed analysis, viewshed analysis (an evaluation of all areas in view of one or more designated features) and a special supply/demand modeling procedure where demands are satisfied by taking supplies in a radial fashion from neighboring locations.

Analytical Operations

Given these basic tools, a broad range of analytical operations can be undertaken. However, it would appear that most of these fall into one of three basic groups: Database Query, Derivative Mapping and Process Modeling.

Database Query

With database query, we are simply selecting out various combinations of variables for examination. The tools we use are largely the database query tools previously discussed (hence the name), but also include various measurement and statistical analysis procedures. The key thing that distinguishes this kind of analysis is that we have taken out no more than we have put into the system. While we may extract combinations we have never examined before, the system provides us with no new information—we are simply making a withdrawal from a data bank we have built up.

One of the key activities in database query is pattern seeking. Typically we are looking for spatial patterns in the data that may lead us to hypothesize about relationships between variables.

Derivative Mapping

With derivative mapping, we combine selected components of our database to yield new derivative layers. For example, we might take our digital elevation data to derive slope gradients, and then take our slope data and combine it with information on soil type and rainfall regime to produce a new map of soil erosion potential. This new map then becomes an addition to our growing database.

How is it that we can create new data from old? Unlike database query where we simply extracted information that was already in the database, with derivative mapping we take existing information and add to it something new—*knowledge of relationships* between database elements. We can create a soil erosion potential map using a digital elevation layer, a soils layer and a rainfall regime layer, only if we know the relationship between those factors and the new map we are creating. In some cases, these relationships will be specified in logical terms (such as creating a suitability map for industrial location based on the condition that it be on existing forest land, outside protective buffers around wetlands and on low slopes) and we will use our database query tools. In other cases, however, these relationships will be specified in mathematical terms and we will rely heavily on the map algebra tools. Regardless, the relationships that form the model will need to be known.

In some cases, the relationship models can be derived on logical or theoretical grounds. However, in many instances it is necessary that the relationships be determined by empirical study. Regression analysis, for example, is one very common way in which empirical testing is used to develop a mathematical relationship between variables. If one takes the soil erosion example, one might set up a series of test sites at which the soil erosion is measured along with the slope, soil type and rainfall data. These sample points would then be used to develop the equation relating soil erosion to these variables. The equation would then be used to evaluate soil erosion potential over a much broader region.

Process Modeling

Database query and derivative mapping make up the bulk of GIS analysis undertaken today. However, there is a third area that offers incredible potential—Process or Simulation Modeling.

With process modeling, we also bring something new to the database—*knowledge of process*. Process refers to the causal chain by which some event takes place. For example, a simple model of fuel-wood demand satisfaction might run as follows:

1. Take all the wood you need (if you can) from your present location.
2. If your demand is satisfied or if you have traveled more than 10 kilometers from home, go to step 4.
3. If your demand is not met, move to an immediately adjacent location not already visited and repeat step 1.
4. Stop.

Process modeling is a particularly exciting prospect for GIS. It is based on the notion that in GIS, our database doesn't simply represent an environment, *it is an environment!* It is a surrogate environment, capable of being measured, manipulated and acted

upon by geographic and temporal processes. Our database thus acts as a laboratory for the exploration of processes in a complex environment. Traditionally, in science, we have had to remove that complexity in order to understand processes in isolation. This has been an effective strategy and we have learned much from it. However, technologies such as GIS now offer the tools to reassemble those simple understandings in order to gain an understanding and appreciation of how they act in the full complexity of a real environmental setting. Often even very simple understandings yield complex patterns when allowed to interact in the environment.

A different sort of process, the decision making process, may also be supported and in some ways modeled with the use of GIS. GIS technology is becoming more important as a tool for decision support. Indeed, even the simplest database query results may prove to be invaluable input to the decision maker. However, the more complex *process* of decision making, in which decision makers often think in terms of multiple criteria, soft boundaries (non-Boolean) and levels of acceptable risk, may also be modeled using GIS. TerrSet provides a suite of decision support modules to help decision makers develop more explicitly rational and well-informed decisions. The **Decision Making** chapter discusses this important use of GIS in detail and provides illustrative case examples.

Despite its evident attraction, process modeling, both in environmental processes and decision making, is still a fairly uncommon activity in GIS. The reason is quite simple. While more and more modeling tools are becoming available in the GIS, it is not uncommon that the process of interest requires a capability not built into the system. These cases require the creation of a new program module. Many systems are not well set up for the incorporation of user-developed routines. TerrSet, however, has been designed so programs in any computer language may be merged into the system and called from the TerrSet interface.

The Philosophy of GIS

GIS has had an enormous impact on virtually every field that manages and analyzes spatially distributed data. For those who are unfamiliar with the technology, it is easy to see it as a magic box. The speed, consistency and precision with which it operates is truly impressive, and its strong graphic character is hard to resist. However, to experienced analysts, the philosophy of GIS is quite different. With experience, GIS becomes simply an extension of one's own analytical thinking. The system has no inherent answers, only those of the analyst. It is a tool, just like statistics is a tool. It is a tool for thought.

Investing in GIS requires more than an investment in hardware and software. Indeed, in many instances this is the least issue of concern. Most would also recognize that a substantial investment needs to be placed in the development of the database. However, one of the least recognized yet most important investments is in the analysts who will use the system. The system and the analyst cannot be separated—one is simply an extension of the other. In addition, the process of incorporating GIS capabilities into an institution requires an investment in long-term and organization-wide education and training.

In many ways, learning GIS involves learning to think—learning to think about patterns, about space and about processes that act in space. As you learn about specific procedures, they will often be encountered in the context of specific examples. In addition, they will often have names that suggest their typical application. But resist the temptation to categorize these routines. Most procedures have many more general applications and can be used in many novel and innovative ways. Explore! Challenge what you see! What you will learn goes far beyond what this or any software package can provide.

APPENDIX 6

INTRODUCTION TO REMOTE SENSING AND IMAGE PROCESSING

Of all the various data sources used in GIS, one of the most important is undoubtedly that provided by remote sensing. Through the use of satellites, we now have a continuing program of data acquisition for the entire world with time frames ranging from a couple of weeks to a matter of hours. Very importantly, we also now have access to remotely sensed images in digital form, allowing rapid integration of the results of remote sensing analysis into a GIS.

The development of digital techniques for the restoration, enhancement and computer-assisted interpretation of remotely sensed images initially proceeded independently and somewhat ahead of GIS. However, the raster data structure and many of the procedures involved in these *Image Processing Systems* (IPS) were identical to those involved in raster GIS. As a result, it has become common to see IPS software packages add general capabilities for GIS, and GIS software systems add at least a fundamental suite of IPS tools. TerrSet is a combined GIS and image processing system that offers advanced capabilities in both areas.

Because of the extreme importance of remote sensing as a data input to GIS, it has become necessary for GIS analysts (particularly those involved in natural resource applications) to gain a strong familiarity with IPS. Consequently, this chapter gives an overview of this important technology and its integration with GIS. The Image Processing exercises in the **Tutorial** illustrate many of the concepts presented here.

Definition

Remote sensing can be defined as any process whereby information is gathered about an object, area or phenomenon without being in contact with it. Our eyes are an excellent example of a remote sensing device. We are able to gather information about our surroundings by gauging the amount and nature of the reflectance of visible light energy from some external source (such as the sun or a light bulb) as it reflects off objects in our field of view. Contrast this with a thermometer, which must be in contact with the phenomenon it measures, and thus is not a remote sensing device.

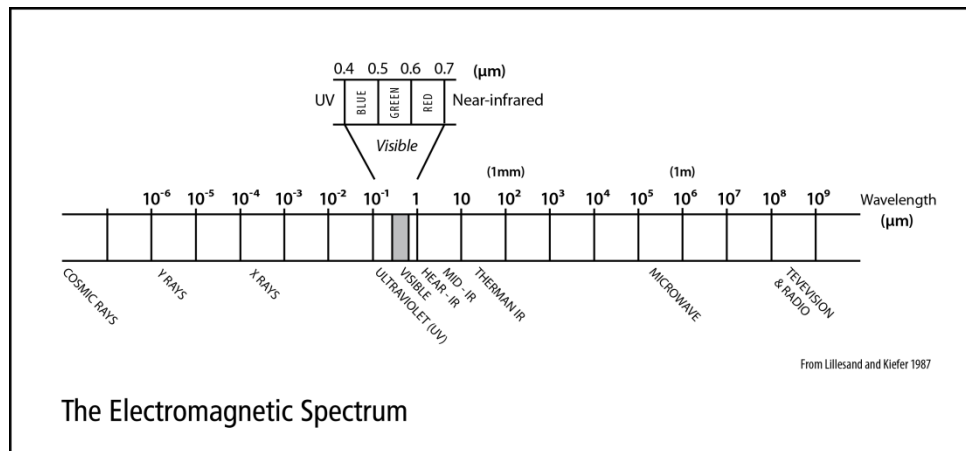
Given this rather general definition, the term *remote sensing* has come to be associated more specifically with the gauging of interactions between earth surface materials and electromagnetic energy. However, any such attempt at a more specific definition becomes difficult, since it is not always the natural environment that is sensed (e.g., art conservation applications), the energy type is not always electromagnetic (e.g., sonar) and some procedures gauge natural energy emissions (e.g., thermal infrared) rather than interactions with energy from an independent source.

Fundamental Considerations

Energy Source

Sensors can be divided into two broad groups—*passive* and *active*. Passive sensors measure ambient levels of existing sources of energy, while active ones provide their own source of energy. Most remote sensing is done with passive sensors, for which the sun is the major energy source. The earliest example of this is photography. With airborne cameras we have long been able to measure and record the reflection of light off earth features. While aerial photography is still in use, newer solid-state technologies have extended capabilities for viewing in the visible and near infrared wavelengths to include longer wavelength solar radiation as well. However, not all passive sensors use energy from the sun. Thermal infrared and passive microwave sensors both measure natural earth energy emissions. Thus, the passive sensors are simply those that do not themselves supply the energy being detected.

By contrast, active sensors provide their own source of energy. The most familiar form of this is flash photography. However, in environmental and mapping applications, the best example is RADAR. RADAR systems emit energy in the microwave region of the electromagnetic spectrum (figure below). The reflection of that energy by earth surface materials is then measured to produce an image of the area sensed.



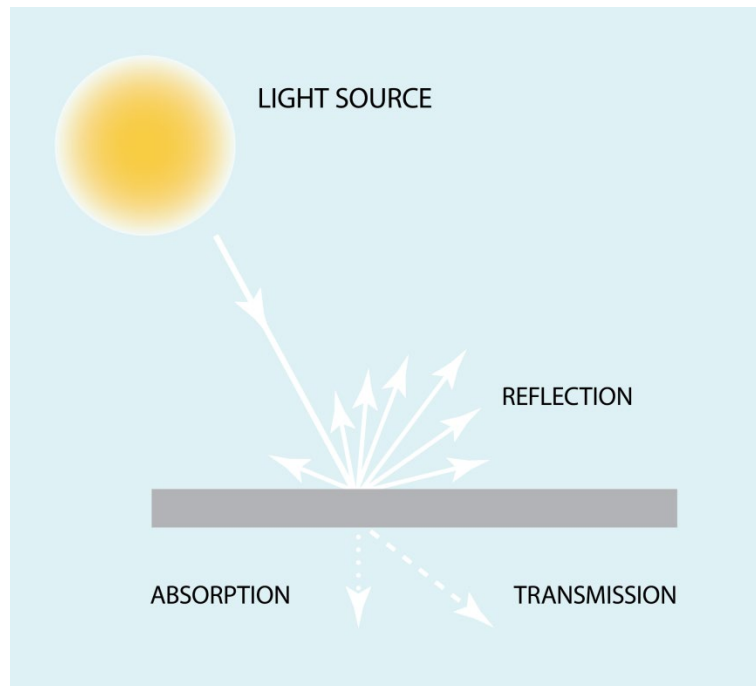
Wavelength

As indicated, most remote sensing devices make use of electromagnetic energy. However, the electromagnetic spectrum is very broad and not all wavelengths are equally effective for remote sensing purposes. Furthermore, not all have significant interactions with earth surface materials of interest to us. The figure above illustrates the electromagnetic spectrum. The atmosphere itself causes significant absorption and/or scattering of the very shortest wavelengths. In addition, the glass lenses of many sensors also cause significant absorption of shorter wavelengths such as the ultraviolet (UV). As a result, the first significant *window* (i.e., a region in which energy can significantly pass through the atmosphere) opens up in the visible wavelengths. Even here, the blue wavelengths undergo substantial attenuation by atmospheric scattering, and are thus often left out in remotely sensed images. However, the green, red and near infrared (IR) wavelengths all provide good opportunities for gauging earth surface interactions without significant interference by the atmosphere. In addition, these regions provide important clues to the nature of many earth surface materials. Chlorophyll, for example, is a very strong absorber of red visible wavelengths, while the near infrared wavelengths provide important clues to the structures of plant leaves. As a result, the bulk of remotely sensed images used in GIS-related applications are taken in these regions.

Extending into the middle and thermal infrared regions, a variety of good windows can be found. The longer of the middle infrared wavelengths have proven to be useful in a number of geological applications. The thermal regions have proven to be very useful for

monitoring not only the obvious cases of the spatial distribution of heat from industrial activity, but a broad set of applications ranging from fire monitoring to animal distribution studies to soil moisture conditions.

After the thermal IR, the next area of major significance in environmental remote sensing is in the microwave region. Several important windows exist in this region and are of particular importance for the use of active radar imaging. The texture of earth surface materials causes significant interactions with several of the microwave wavelength regions. This can thus be used as a supplement to information gained in other wavelengths, and also offers the significant advantage of being usable at night (because as an active system it is independent of solar radiation) and in regions of persistent cloud cover (since radar wavelengths are not significantly affected by clouds).



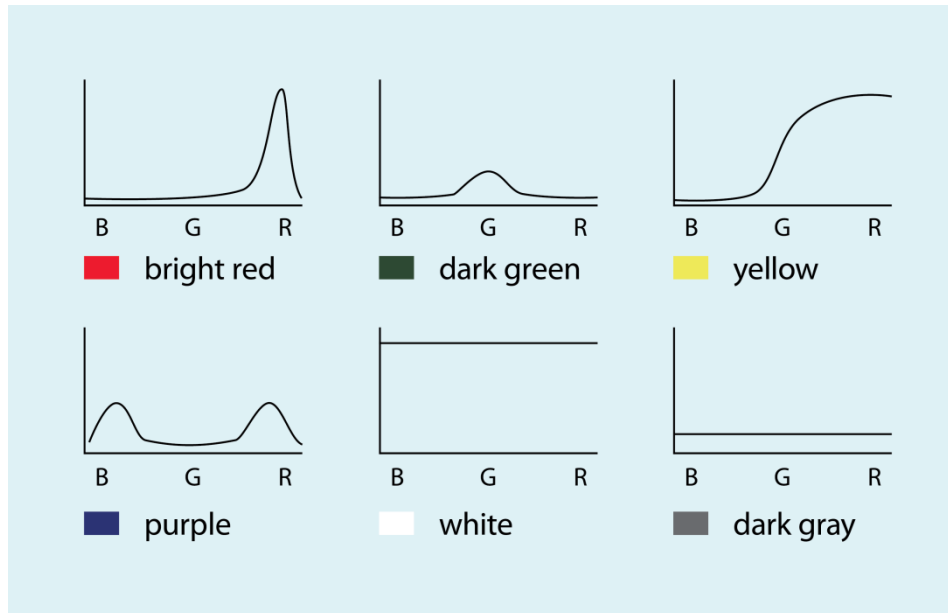
Interaction Mechanisms

When electromagnetic energy strikes a material, three types of interaction can follow: reflection, absorption and/or transmission (figure above). Our main concern is with the reflected portion since it is usually this which is returned to the sensor system. Exactly how much is reflected will vary and will depend upon the nature of the material and where in the electromagnetic spectrum our measurement is being taken. As a result, if we look at the nature of this reflected component over a range of wavelengths, we can characterize the result as a *spectral response pattern*.

Spectral Response Patterns

A spectral response pattern is sometimes called a *signature*. It is a description (often in the form of a graph) of the degree to which energy is reflected in different regions of the spectrum. Most humans are very familiar with spectral response patterns since they are equivalent to the human concept of color. For example, the figure below shows idealized spectral response patterns for several familiar colors in the visible portion of the electromagnetic spectrum, as well as for white and dark grey. The bright red reflectance pattern, for example, might be that produced by a piece of paper printed with a red ink. Here, the ink is designed to alter the white light that shines upon it and absorb the blue and green wavelengths. What is left, then, are the red wavelengths which reflect off the

surface of the paper back to the sensing system (the eye). The high return of red wavelengths indicates a bright red, whereas the low return of green wavelengths in the second example suggests that it will appear quite dark.



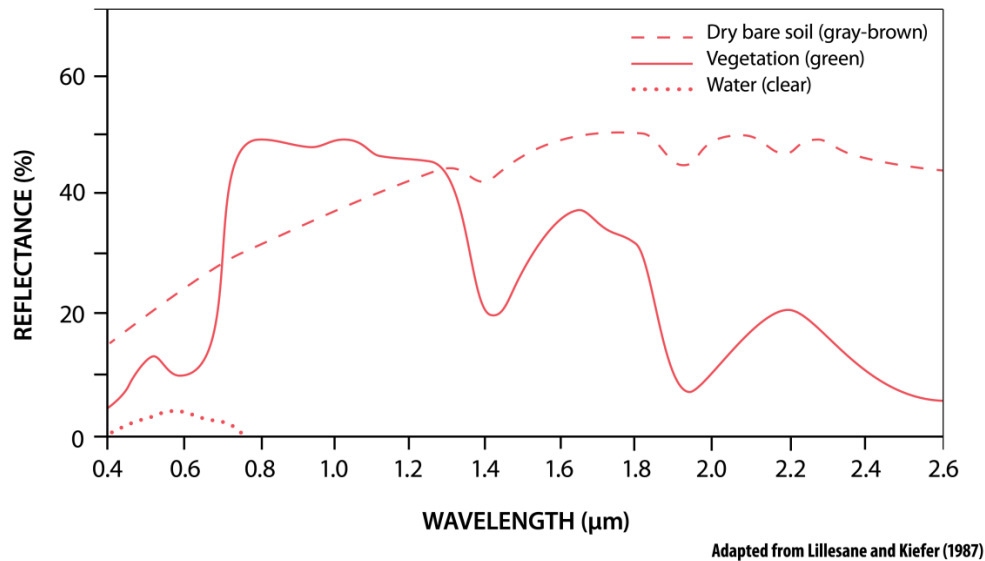
The eye is able to sense spectral response patterns because it is truly a multi-spectral sensor (i.e., it senses in more than one place in the spectrum). Although the actual functioning of the eye is quite complex, it does in fact have three separate types of detectors that can usefully be thought of as responding to the red, green and blue wavelength regions. These are the *additive primary colors*, and the eye responds to mixtures of these three to yield a sensation of other hues. For example, the color perceived by the third spectral response pattern in figure would be a yellow—the result of *mixing* a red and green. However, it is important to recognize that this is simply our phenomenological perception of a spectral response pattern. Consider, for example, the fourth curve. Here we have reflectance in both the blue and red regions of the visible spectrum. This is a bimodal distribution, and thus technically not a specific hue in the spectrum. However, we would perceive this to be a purple! Purple (a color between violet and red) does not exist in nature (i.e., as a hue—a distinctive dominant wavelength). It is very real in our perception, however. Purple is simply our perception of a bimodal pattern involving a non-adjacent pair of primary hues.

In the early days of remote sensing, it was believed (more correctly *hoped*) that each earth surface material would have a distinctive spectral response pattern that would allow it to be reliably detected by visual or digital means. However, as our common experience with color would suggest, in reality this is often not the case. For example, two species of trees may have quite a different coloration at one time of the year and quite a similar one at another.

Finding distinctive spectral response patterns is the key to most procedures for computer-assisted interpretation of remotely sensed imagery. This task is rarely trivial. Rather, the analyst must find the combination of spectral bands and the time of year at which distinctive patterns can be found for each of the information classes of interest.

For example, the figure below shows an idealized spectral response pattern for vegetation along with those of water and dry bare soil. The strong absorption by leaf pigments (particularly chlorophyll for purposes of photosynthesis) in the blue and red regions of the visible portion of the spectrum leads to the characteristic green appearance of healthy vegetation. However, while this *signature* is distinctively different from most non-vegetated surfaces, it is not very capable of distinguishing between species of vegetation—most will have a similar color of green at full maturation. In the near infrared, however, we find a much higher return from vegetated surfaces because of scattering within the fleshy mesophyll layer of the leaves. Plant pigments do not absorb energy in this region, and thus the scattering, combined with the multiplying effect of a full canopy of leaves, leads to high reflectance in this region of the spectrum. However, the extent of this reflectance will depend highly on the internal structure of leaves (e.g., broadleaf versus needle). As a result, significant differences between species can often be detected in this region. Similarly, moving into the middle

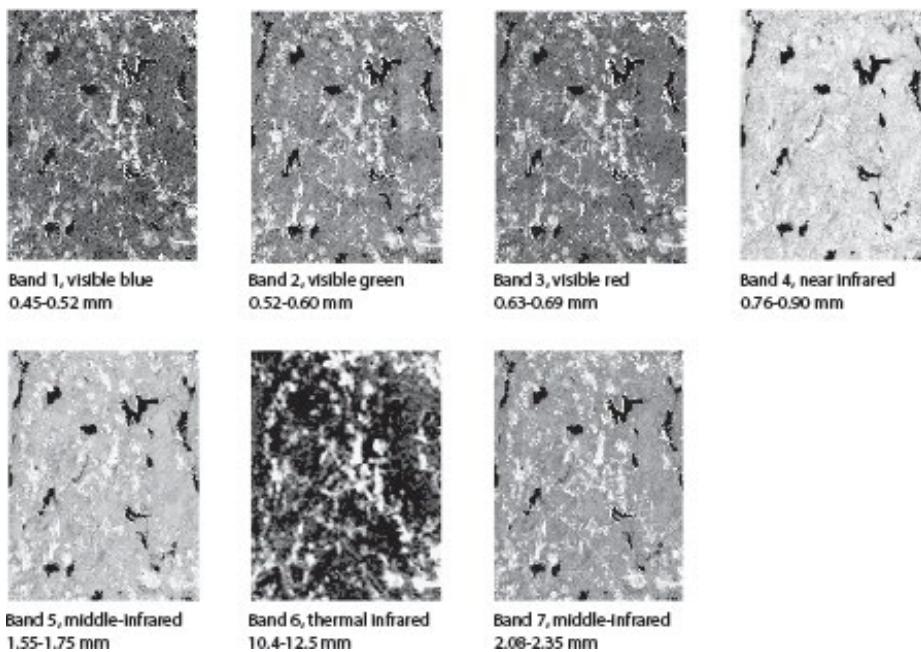
infrared region we see a significant dip in the spectral response pattern that is associated with leaf moisture. This is, again, an area where significant differences can arise between mature species. Applications looking for optimal differentiation between species, therefore, will typically involve both the near and middle infrared regions and will use imagery taken well into the development cycle.



Multispectral Remote Sensing

In the visual interpretation of remotely sensed images, a variety of image characteristics are brought into consideration: color (or tone in the case of panchromatic images), texture, size, shape, pattern, context, and the like. However, with computer-assisted interpretation, it is most often simply color (i.e., the spectral response pattern) that is used. It is for this reason that a strong emphasis is placed on the use of multispectral sensors (sensors that, like the eye, look at more than one place in the spectrum and thus are able to gauge spectral response patterns), and the number and specific placement of these spectral *bands*.

The figure below illustrates the spectral bands of the Landsat Thematic Mapper (TM) system. The Landsat satellite is a commercial system providing multi-spectral imagery in seven spectral bands at a 30-meter resolution.



It can be shown through analytical techniques, such as Principal Components Analysis, that in many environments, the bands that carry the greatest amount of information about the natural environment are the near infrared and red wavelength bands. Water is strongly absorbed by infrared wavelengths and is thus highly distinctive in that region. In addition, plant species typically show their greatest differentiation here. The red area is also very important because it is the primary region in which chlorophyll absorbs energy for photosynthesis. Thus, it is this band which can most readily distinguish between vegetated and non-vegetated surfaces.

Given this importance of the red and near infrared bands, it is not surprising that sensor systems designed for earth resource monitoring will invariably include these in any particular multispectral system. Other bands will depend upon the range of applications envisioned. Many include the green visible band since it can be used, along with the other two, to produce a traditional false color composite—a full color image derived from the green, red, and infrared bands (as opposed to the blue, green, and red bands of natural color images). This format became common with the advent of color infrared photography, and is familiar to many specialists in the remote sensing field. In addition, the combination of these three bands works well in the interpretation of the cultural landscape as well as natural and vegetated surfaces. However, it is increasingly common to include other bands that are more specifically targeted to the differentiation of surface materials. For example, Landsat TM Band 5 is placed between two water absorption bands and has thus proven very useful in determining soil and leaf moisture differences. Similarly, Landsat TM Band 7 targets the detection of hydrothermal alteration zones in bare rock surfaces. By contrast, the AVHRR system on the NOAA series satellites includes several thermal channels for the sensing of cloud temperature characteristics.

Hyperspectral Remote Sensing

In addition to traditional multispectral imagery, some systems such as AVIRIS and MODIS, can capture *hyperspectral* data. These systems cover a similar wavelength range to multispectral systems, but in much narrower bands. This dramatically increases the number of bands (and thus precision) available for image classification (typically tens and even hundreds of very narrow bands). Moreover, hyperspectral signature libraries have been created in lab conditions and contain hundreds of signatures for different types of landcovers, including many minerals and other earth materials. Thus, it should be possible to match signatures to surface materials with great precision. However, environmental conditions and natural variations in materials (which make them different from standard library materials) make this difficult. In addition, classification procedures have not been developed for hyperspectral

data to the degree they have been for multispectral imagery. As a consequence, multispectral imagery still represents the major tool of remote sensing today.

Sensor/Platform Systems

Given recent developments in sensors, a variety of platforms are now available for the capture of remotely sensed data. Here we review some of the major sensor/platform combinations that are typically available to the GIS user community.

Aerial Photography

Aerial photography is one of the oldest method of remote sensing. Cameras mounted in light aircraft flying between 200 and 15,000 m capture a large quantity of detailed information. Aerial photos provide an instant visual inventory of a portion of the earth's surface and can be used to create detailed maps. Aerial photographs commonly are taken by commercial aerial photography firms which own and operate specially modified aircraft equipped with large format (23 cm x 23 cm) mapping quality cameras. Aerial photos can also be taken using small format cameras (35 mm and 70 mm), hand-held or mounted in unmodified light aircraft.

Camera and platform configurations can be grouped in terms of oblique and vertical. Oblique aerial photography is taken at an angle to the ground. The resulting images give a view as if the observer is looking out an airplane window. These images are easier to interpret than vertical photographs, but it is difficult to locate and measure features on them for mapping purposes.

Vertical aerial photography is taken with the camera pointed straight down. The resulting images depict ground features in plan form and are easily compared with maps. Vertical aerial photos are always highly desirable, but are particularly useful for resource surveys in areas where no maps are available. Aerial photos depict features such as field patterns and vegetation which are often omitted on maps. Comparison of old and new aerial photos can also capture changes within an area over time.

Vertical aerial photos contain subtle displacements due to relief, tip and tilt of the aircraft and lens distortion. Vertical images may be taken with overlap, typically about 60 percent along the flight line and at least 20 percent between lines. Overlapping images can be viewed with a stereoscope to create a three-dimensional view, called a *stereo model*.

Large Format Photography

Commercial aerial survey firms use light single or twin engine aircraft equipped with large-format mapping cameras. Large-format cameras, such as the Wild RC-10, use 23 cm x 23 cm film which is available in rolls. Eastman Kodak, Inc., among others, manufactures several varieties of sheet film specifically intended for use in aerial photography. Negative film is used where prints are the desired product, while positive film is used where transparencies are desired. Print film allows for detailed enlargements to be made, such as large wall-sized prints. In addition, print film is useful when multiple prints are to be distributed and used in the field.

Small Format Photography

Small-format cameras carried in chartered aircraft are an inexpensive alternative to large-format aerial photography. A digital camera or lenses, light aircraft or drones are required, along with some means to process the acquired digital imager.

Oblique photos can be taken with a hand-held camera in any light aircraft; vertical photos require some form of special mount on the aircraft.

Small-format aerial photography has several drawbacks. Light unpressurized aircraft are typically limited to altitudes below 4000 m. As film size is small, sacrifices must be made in resolution or area covered per frame. Because of distortions in the camera system,

small-format photography cannot be used if precise mapping is required. Nonetheless, small-format photography can be very useful for reconnaissance surveys and can also be used as point samples.

Color Photos

Normal color photos are produced from a composite of three digital layers with intervening filters that act to isolate, in effect, red, green, and blue wavelengths separately. With color infrared, these wavelengths are shifted to the longer wavelengths to produce a composite that has isolated reflectances from the green, red and near infrared wavelength regions. However, because the human eye cannot see infrared, a false color composite is produced by making the green wavelengths appear blue, the red wavelengths appear green, and the infrared wavelengths appear red.

Older aerial photographs are not in a format that can immediately be used in digital analysis. It is possible to scan photographs with a scanner and thereby create multispectral datasets either by scanning individual band images, or by scanning a color image and separating the bands. However, the geometry of aerial photographs (which have a central perspective projection and differential parallax) is such that they are difficult to use directly. More typically they require processing by special photogrammetric software to rectify the images and remove differential parallax effects.

Aerial Videography

Light, portable, inexpensive video cameras and recorders can be carried on drones and aircraft. In addition, a number of smaller aerial mapping companies offer videography as an output option. By using several cameras simultaneously, each with a filter designed to isolate a specific wavelength range, it is possible to isolate multispectral image bands that can be used individually, or in combination in the form of a color composite. For use in digital analysis, special graphics hardware boards known as *frame grabbers* can be used to freeze any frame within a continuous video sequence and convert it to digital format, usually in one of the more popular exchange formats such as TIF. Like small-format photography, aerial videography cannot be used for detailed mapping, but provides a useful overview for reconnaissance surveys, and can be used in conjunction with ground point sampling.

Satellite-Based Scanning Systems

Photography has proven to be an important input to visual interpretation and the production of analog maps. However, the development of satellite platforms, the associated need to telemeter imagery in digital form, and the desire for highly consistent digital imagery have given rise to the development of solid-state scanners as a major format for the capture of remotely sensed data. The specific features of these systems vary (including, in some cases, the removal of a true scanning mechanism). However, in the discussion which follows, an idealized scanning system is presented that is highly representative of current systems in use.

The basic logic of a scanning sensor is the use of a mechanism to sweep a small field of view (known as an *instantaneous field of view*—IFOV) in a west to east direction at the same time the satellite is moving in a north to south direction. Together this movement provides the means of composing a complete raster image of the environment.

A simple scanning technique is to use a rotating mirror that can sweep the field of view in a consistent west to east fashion. The field of view is then intercepted with a prism that can spread the energy contained within the IFOV into its spectral components. Photoelectric detectors (of the same nature as those found in the exposure meters of commonly available photographic cameras) are then arranged in the path of this spectrum to provide electrical measurements of the amount of energy detected in various parts of the electromagnetic spectrum. As the scan moves from west to east, these detectors are polled to get a set of readings along the east-west scan. These form the columns along one row of a set of raster images—one for each detector. Movement of the satellite from north to south then positions the system to detect the next row, ultimately leading to the production of a set of raster images as a record of reflectance over a range of spectral bands.

There are many satellite systems in operation today that collect imagery that is subsequently distributed to users. Several of the most common systems include AVHRR, EOS, ERS, IRS, IKONOS, JERS, LANDSAT, SPOT, Quickbird, RADARSAT, and many more. Each type of satellite data offers specific characteristics that make it appropriate for a particular application.

In general, there are two characteristics that may help guide the choice of satellite data: *spatial resolution* and *spectral resolution*. The spatial resolution refers to the size of the area on the ground that is summarized by one data value in the imagery. This is the Instantaneous Field of View (IFOV) described earlier. Spectral resolution refers to the number and width of the spectral bands that the satellite sensor detects. In addition, issues of cost and imagery availability must also be considered.

Digital Image Processing

Overview

As a result of solid-state multispectral scanners and other raster input devices, we now have available digital raster images of spectral reflectance data. The chief advantage of having these data in digital form is that they allow us to apply computer analysis techniques to the image data—a field of study called *Digital Image Processing*.

Digital Image Processing is largely concerned with four basic operations: image restoration, image enhancement, image classification, and image transformation. Image restoration is concerned with the correction and calibration of images in order to achieve as faithful a representation of the earth surface as possible—a fundamental consideration for all applications. Image enhancement is predominantly concerned with the modification of images to optimize their appearance to the visual system. Visual analysis is a key element, even in digital image processing, and the effects of these techniques can be dramatic. Image classification refers to the computer-assisted interpretation of images—an operation that is vital to GIS. Finally, image transformation refers to the derivation of new imagery as a result of some mathematical treatment of the raw image bands.

In order to undertake the operations listed in this section, it is necessary to have access to image processing software. TerrSet is one such system. While it is known primarily as a GIS software system, it also offers a full suite of image processing capabilities.

Image Restoration

Remotely sensed images of the environment are typically taken at a great distance from the earth's surface. As a result, there is a substantial atmospheric path that electromagnetic energy must pass through before it reaches the sensor. Depending upon the wavelengths involved and atmospheric conditions (such as particulate matter, moisture content and turbulence), the incoming energy may be substantially modified. The sensor itself may then modify the character of that data since it may combine a variety of mechanical, optical and electrical components that serve to modify or mask the measured radiant energy. In addition, during the time the image is being scanned, the satellite is following a path that is subject to minor variations at the same time that the earth is moving underneath. The geometry of the image is thus in constant flux. Finally, the signal needs to be telemetered back to earth, and subsequently received and processed to yield the final data we receive. Consequently, a variety of systematic and apparently random disturbances can combine to degrade the quality of the image we finally receive. Image restoration seeks to remove these degradation effects.

Broadly, image restoration can be broken down into the two sub-areas of *radiometric restoration* and *geometric restoration*.

Radiometric Restoration

Radiometric restoration refers to the removal or diminishment of distortions in the degree of electromagnetic energy registered by each detector. A variety of agents can cause distortion in the values recorded for image cells. Some of the most common distortions for which correction procedures exist include:

uniformly elevated values, due to atmospheric haze, which preferentially scatters short wavelength bands (particularly the blue wavelengths);

striping, due to detectors going out of calibration;

random noise, due to unpredictable and unsystematic performance of the sensor or transmission of the data; and

scan line drop out, due to signal loss from specific detectors.

It is also appropriate to include here procedures that are used to convert the raw, unitless relative reflectance values (known as digital numbers, or DN) of the original bands into true measures of reflective power (radiance).

See the section on **Image Restoration** for a more detailed discussion of radiometric restoration and how it can be implemented in TerrSet.

Geometric Restoration

For mapping purposes, it is essential that any form of remotely sensed imagery be accurately registered to the proposed map base. With satellite imagery, the very high altitude of the sensing platform results in minimal image displacements due to relief. As a result, registration can usually be achieved through the use of a systematic *rubber sheet* transformation process¹ that gently warps an image (through the use of polynomial equations) based on the known positions of a set of widely dispersed control points. This capability is provided in TerrSet through the module RESAMPLE.

With aerial photographs, however, the process is more complex. Not only are there systematic distortions related to tilt and varying altitude, but variable topographic relief leads to very irregular distortions (differential parallax) that cannot be removed through a rubber sheet transformation procedure. In these instances, it is necessary to use photogrammetric rectification to remove these distortions and provide accurate map measurements². Failing this, the central portions of high altitude photographs can be resampled with some success.

RESAMPLE is a module of major importance, and it is essential that one learn to use it effectively. Doing so also requires a thorough understanding of reference systems and their associated parameters such as datums and projections. The chapter on **Georeferencing** provides an in-depth discussion of these issues.

Image Enhancement

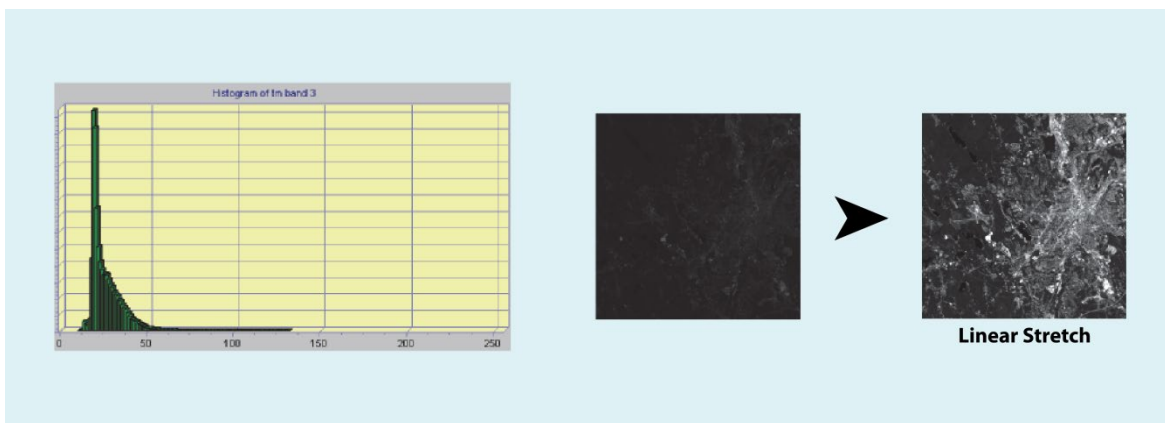
¹ Satellite-based scanner imagery contains a variety of inherent geometric problems such as skew (caused by rotation of the earth underneath the satellite as it is in the process of scanning a complete image) and scanner distortion (caused by the fact that the instantaneous field of view (IFOV) covers more territory at the ends of scan lines, where the angle of view is very oblique, than in the middle). With commercially-marketed satellite imagery, such as Landsat, IRS and SPOT, most elements of systematic geometric restoration associated with image capture are corrected by the distributors of the imagery. Thus, for the end user, the only geometric operation that typically needs to be undertaken is a rubber-sheet resampling in order to rectify the image to a map base. Many commercial distributors will perform this rectification for an additional fee.

² Photogrammetry is the science of taking spatial measurements from aerial photographs. In order to provide a full rectification, it is necessary to have *stereoscopic* images—photographs which overlap enough (e.g., 60% in the along-track direction and 10% between flight lines) to provide two independent images of each part of the landscape. Using these stereoscopic pairs and ground control points of known position and height, it is possible to fully re-create the geometry of the viewing conditions, and thereby not only rectify measurements from such images, but also derive measurements of terrain height. The rectified photographs are called *orthophotos*. The height measurements may be used to produce digital elevation models.

Image enhancement is concerned with the modification of images to make them more suited to the capabilities of human vision. Regardless of the extent of digital intervention, visual analysis invariably plays a very strong role in all aspects of remote sensing. While the range of image enhancement techniques is broad, the following fundamental issues form the backbone of this area:

Contrast Stretch

Digital sensors have a wide range of output values to accommodate the strongly varying reflectance values that can be found in different environments. However, in any single environment, it is often the case that only a narrow range of values will occur over most areas. Grey level distributions thus tend to be very skewed. Contrast manipulation procedures are thus essential to most visual analyses. The figure below shows TM Band 3 (visible red) and its histogram. Note that the values of the image are quite skewed. The right image of the figure shows the same image band after a linear stretch between values 12 and 60 has been applied. In TerrSet, this type of contrast enhancement may be performed interactively through Composer's Layer Properties while the image is displayed. This is normally used for visual analysis only—original data values are used in numeric analyses. New images with stretched values are produced with the module STRETCH.



Composite Generation

For visual analysis, color composites make fullest use of the capabilities of the human eye. Depending upon the graphics system in use, composite generation ranges from simply selecting the bands to use, to more involved procedures of band combination and associated contrast stretch. The figure below shows several composites made with different band combinations from the same set of TM images. The TerrSet module COMPOSITE is used to construct three-band 24-bit composite images for visual analysis.

Digital Filtering

One of the most intriguing capabilities of digital analysis is the ability to apply digital filters. Filters can be used to provide edge enhancement (sometimes called *crispening*), to remove image blur, and to isolate lineaments and directional trends, to mention just a few. The TerrSet module FILTER is used to apply standard filters and to construct and apply user-defined filters.

Pansharpening

Pansharpening is the process of merging lower resolution multispectral imagery with the higher resolution panchromatic image. Typically, the panchromatic band associated with most systems, e.g., SPOT, IKONOS or QuickBird, are captured across the visible range of the spectrum affording them a higher resolution, commensurately giving them better detail in shape and texture. But what they gain in clarity penalizes them in regards to their spectral properties, unlike the multispectral bands. Merging the two results in increasing the resolution of the multispectral imagery while preserving its spectral information.



In TerrSet the module PANSHARPEN is used to pansharpen multispectral images using the panchromatic band.

Image Classification

Image classification refers to the computer-assisted interpretation of remotely sensed images. The procedures involved are treated in detail in the chapter titled **Classification of Remotely Sensed Imagery**. This section provides a brief overview.

Although some procedures are able to incorporate information about such image characteristics as texture and context, the majority of image classification is based solely on the detection of the spectral signatures (i.e., spectral response patterns) of landcover classes. The success with which this can be done will depend on two things: 1) the presence of distinctive signatures for the landcover classes of interest in the band set being used; and 2) the ability to reliably distinguish these signatures from other spectral response patterns that may be present.

There are two general approaches to image classification: *supervised* and *unsupervised*. They differ in how the classification is performed. In the case of supervised classification, the software system delineates specific landcover types based on statistical characterization data drawn from known examples in the image (known as training sites). With unsupervised classification, however, clustering software is used to uncover the commonly occurring landcover types, with the analyst providing interpretations of those cover types at a later stage.

Supervised Classification

The first step in supervised classification is to identify examples of the information classes (i.e., landcover types) of interest in the image. These are called *training sites*. The software system is then used to develop a statistical characterization of the reflectances for each information class. This stage is often called *signature analysis* and may involve developing a characterization as simple as the mean or the range of reflectances on each band, or as complex as detailed analyses of the mean, variances and covariances over all bands.

Once a statistical characterization has been achieved for each information class, the image is then classified by examining the reflectances for each pixel and making a decision about which of the signatures it resembles most. There are several techniques for making these decisions, called *classifiers*. Most image processing software will offer several, based on varying decision rules. TerrSet offers a wide range of options falling into three groups, depending upon the nature of the output desired and the nature of the input bands.

Hard Classifiers

The distinguishing characteristic of hard classifiers is that they all make a definitive decision about the landcover class to which any pixel belongs. TerrSet offers a host of supervised classifiers in this group. Some like parallelepiped (PIPED), minimum distance to means (MINDIST), maximum likelihood (MAXLIKE), linear discriminant analysis (FISHER), Radial Basis Function Neural Network (RBFNN) and Fuzzy ARTMAP neural network only output one hard classified map. Others can output both a hard and soft outputs. These include: multi-layer perceptron (MLP) neural network, self-organizing map (SOM) neural network, k-nearest neighbor (KNN), and classification tree analysis (CTA). All these classifiers differ only in the manner in which they develop and use a statistical characterization of the training site data. Of the list, the maximum likelihood procedure is unquestionably the most widely used classifier in the classification of remotely sensed imagery.

A distinctive hard classifier is the segmentation classification routine in TerrSet. The module SEGCLASS classifies imagery using a majority rule algorithm that is applied to image segments created by the module SEGMENTATION.

Soft Classifiers

Contrary to hard classifiers, soft classifiers do not make a definitive decision about the landcover class to which each pixel belongs. Rather, they develop statements of the degree to which each pixel belongs to each of the landcover classes being considered. Thus, for example, a soft classifier might indicate that a pixel has a 0.72 probability of being forest, a 0.24 probability of being pasture, and a 0.04 probability of being bare ground. A hard classifier would resolve this uncertainty by concluding that the pixel was forest. However, a soft classifier makes this uncertainty explicitly available, for any of a variety of reasons. For example, the analyst might conclude that the uncertainty arises because the pixel contains more than one cover type and could use the probabilities as indications of the relative proportion of each. This is known as *sub-pixel* classification . Alternatively, the analyst may conclude that the uncertainty arises because of unrepresentative training site data and therefore may wish to combine these probabilities with other evidence before *hardening* the decision to a final conclusion.

TerrSet offers many soft classifiers. Along with those listed above as hybrid hard and soft classifiers, also included are classifiers that only output soft images: Bayesian (BAYCLASS), Mahalanobis typicalities (MAHALCLASS), Dempster-Shafer belief (BELCLASS) , linear spectral unmixing (UNMIX), and fuzzy (FUZCLASS) classifiers. The results from these five can be hardened using the module HARDEN. The difference between them relates to the logic by which uncertainty is specified—Bayesian, Dempster-Shafer, and Fuzzy Sets. In addition, the system supplies a variety of additional tools specifically designed for the analysis of sub-pixel mixtures (e.g., UNMIX, BELCALC and MAXSET).

Hyperspectral Classifiers

All of the classifiers mentioned above operate on multispectral imagery—images where several spectral bands have been captured simultaneously as independently accessible image components. Extending this logic to many bands produces what has come to be known as *hyperspectral* imagery.

Although there is essentially no difference between hyperspectral and multispectral imagery (i.e., they differ only in degree), the volume of data and high spectral resolution of hyperspectral images does lead to differences in the way that they are handled. TerrSet provides special facilities for creating hyperspectral signatures either from training sites or from libraries of spectral response patterns developed under lab conditions (HYPERSIG) and an automated hyperspectral signature extraction routine (HYPERAUTOSIG). These signatures can then be applied to any of several hyperspectral classifiers: spectral angle mapper (HYPERSAM), minimum distance to means (HYPERMIN), linear spectral unmixing (HYPERUNMIX), orthogonal subspace projection (HYPEROSP), and absorption area analysis (HYPERABSORB). An unsupervised classifier (see next section) for hyperspectral imagery (HYPERUSP) is also available.

Unsupervised Classification

In contrast to supervised classification, where we tell the system about the character (i.e., signature) of the information classes we are looking for, unsupervised classification requires no advance information about the classes of interest. Rather, it examines the data and breaks it into the most prevalent natural spectral groupings, or clusters, present in the data. The analyst then identifies these clusters as landcover classes through a combination of familiarity with the region and ground truth visits. For example, the system might identify classes for asphalt and cement which the analyst might later group together, creating an information class called pavement.

TerrSet includes several unsupervised techniques. The module CLUSTER performs classification based on a set of input images using a multi-dimensional histogram peak technique. While attractive conceptually, unsupervised classification has traditionally been hampered by very slow algorithms. However, the CLUSTER procedure provided in TerrSet is extraordinarily fast and can thus be used iteratively in conjunction with ground truth data to arrive at a very strong classification. With suitable ground truth and accuracy assessment procedures, this tool can provide a remarkably rapid means of producing quality landcover data on a continuing basis.

The KMEANS module is a true K-means clustering routine with several cluster rules (random seed, random partition, and diagonal axis) and stopping criteria thresholds. Two of TerrSet's neural network modules also allow for unsupervised classification, self-organizing map (SOM) and Fuzzy ARTMAP. ISODATA and CHAINCLUSTER routines are included also.

In addition to the above-mentioned techniques, two modules bridge both supervised and unsupervised classifications. ISOCLUST uses a procedure known as *Self-Organizing Cluster Analysis* to classify up to 7 raw bands with the user specifying the number of clusters to process. The procedure uses the CLUSTER module to initiate a set of clusters that seed an iterative application of the MAXLIKE procedure, each stage using the results of the previous stage as the training sites for this supervised procedure. The result is an unsupervised classification that converges on a final set of stable members using a supervised approach (hence the notion of "self-organizing"). MAXSET is also, at its core, a supervised procedure. However, while the procedure starts with training sites that characterize individual classes, it results in a classification that includes not only these specific classes, but also significant (but unknown) mixtures that might exist. Thus the end result has much the character of that of an unsupervised approach.

Accuracy Assessment

A vital step in the classification process, whether supervised or unsupervised, is the assessment of the accuracy of the final images produced. This involves identifying a set of sample locations (such as with the SAMPLE module) that are visited in the field. The landcover found in the field is then compared to that which was mapped in the image for the same location. Statistical assessments of accuracy may then be derived for the entire study area, as well as for individual classes (using ERRMAT).

In an iterative approach, the error matrix produced (sometimes referred to as a *confusion matrix*), may be used to identify particular cover types for which errors are in excess of that desired. The information in the matrix about which covers are being mistakenly included in a particular class (*errors of commission*) and those that are being mistakenly excluded (*errors of omission*) from that class can be used to refine the classification approach.

Image Transformation

Digital Image Processing offers a limitless range of possible transformations on remotely sensed data. TerrSet includes many transformations including: Principal Components Analysis (PCA) allowing standardized and unstandardized, centered and normalized, and T-mode versus S-mode transformations; Canonical Components Analysis (CANCOMP); Canonical Correlation Analysis (CANCOR); Minimum Noise Fraction (MNF) that maximizes the signal to noise ratio; Temporal Fourier Analysis (TFA) that performs harmonic analysis on temporal images; Color Space Transformation (COLSPACE); texture calculations (TEXTURE); blackbody thermal transformations (THERMAL); green vegetation indices (VEGINDEX); and Tasseled Cap (TASSCAP). In addition, a wide variety of ad hoc transformations (such as image ratioing) that can be most effectively accomplished with the Image Calculator

utility. Two transformations are mentioned specifically below (VEGINDEX and PCA), because of their special significance in environmental monitoring applications.

Vegetation Indices

There are a variety of vegetation indices that have been developed to help in the monitoring of vegetation. Most are based on the very different interactions between vegetation and electromagnetic energy in the red and near infrared wavelengths. Refer back to the earlier figure showing an idealized spectral response pattern, which includes a generalized spectral response pattern for green broad leaf vegetation. As can be seen, reflectance in the red region (about 0.6 - 0.7 μ) is low because of absorption by leaf pigments (principally chlorophyll). The infrared region (about 0.8 - 0.9 μ), however, characteristically shows high reflectance because of scattering by the cell structure of the leaves. A very simple vegetation index can thus be achieved by comparing the measure of infrared reflectance to that of the red reflectance.

Although a number of variants of this basic logic have been developed, the one which has received the most attention is the *normalized difference vegetation index* (NDVI). It is calculated in the following manner:

$$\text{NDVI} = (\text{NIR} - \text{R}) / (\text{NIR} + \text{R})$$

where NIR = Near Infrared

and R = Red

This kind of calculation is quite simple for a raster GIS or image processing software system, and the result has been shown to correlate well with ground measurements of biomass. Although NDVI needs specific calibration to be used as an actual measure of biomass, many agencies have found the index to be useful as a relative measure for monitoring purposes. For example, the United Nations Food and Agricultural Organization (FAO) Africa Real Time Information System (ARTEMIS) and the USAID Famine Early Warning System (FEWS) programs both use continental scale NDVI images derived from the NOAA- AVHRR system to produce vegetation index images for the entire continent of Africa every ten days.³

While the NDVI measure has proven to be useful in a variety of contexts, many alternative indices have been proposed to deal with special environments, such as arid lands. TerrSet offers a wide variety of these indices (19) in the VEGINDEX module. The section on **Vegetation Indices** offers a detailed discussion of their characteristics and potential application.

Principal Components Analysis

Principal Components Analysis (PCA) is a linear transformation technique related to Factor Analysis. Given a set of image bands, PCA produces a new set of images, known as components, that are uncorrelated with one another and are ordered in terms of the amount of variance they explain from the original band set.

PCA has traditionally been used in remote sensing as a means of data compaction. For a typical multispectral image band set, it is common to find that the first two or three components are able to explain virtually all of the original variability in reflectance values. Later components thus tend to be dominated by noise effects. By rejecting these later components, the volume of data is reduced with no appreciable loss of information.

Given that the later components are dominated by noise, it is also possible to use PCA as a noise removal technique. The output from the PCA module in TerrSet includes the coefficients of both the forward and backward transformations. By zeroing out the

³ An archive dataset of monthly NDVI images for Africa is available on CD from Clark Labs. The Africa NDVI data CD contains monthly NDVI maximum value composite images (1982-1999), average and standard deviation of monthly NDVI images for each month over the same time period, monthly NDVI anomaly images, and ancillary data (DEM, landuse and landcover, country boundaries and coast line) for Africa in IDRISI format. Contact Clark Labs for more information.

coefficients of the noise components in the reverse transformation, a new version of the original bands can be produced with these noise elements removed.

Recently, PCA has also been shown to have special application in environmental monitoring. In cases where multispectral images are available for two dates, the bands from both images are submitted to a PCA as if they all came from the same image. In these cases, changes between the two dates tend to emerge in the later components. More dramatically, if a time series of NDVI images (or a similar single-band index) is submitted to the analysis, a very detailed analysis of environmental changes and trends can be achieved. In this case, the first component will show the typical NDVI over the entire series, while each successive component illustrates change events in an ordered sequence of importance. By examining these images, along with graphs of their correlation with the individual bands in the original series, important insights can be gained into the nature of changes and trends over the time series. The vertical application Earth Trends Modeler in TerrSet is a specially tailored to facilitate the processing and analysis of time series data, including the incorporation of PCA and many other techniques.

Conclusions

Remotely sensed data is important to a broad range of disciplines. This will continue to be the case and will likely grow with the greater availability of data promised by an increasing number of operational systems. The availability of this data, coupled with the computer software necessary to analyze it, provides opportunities for environmental scholars and planners, particularly in the areas of landuse mapping and change detection, that would have been unheard of only a few decades ago.

The inherent raster structure of remotely sensed data makes it readily compatible with raster GIS. Thus, while TerrSet provides a wide suite of image processing tools, they are completely integrated with the broader set of raster GIS tools the system provides.

APPENDIX 7

DATABASE DEVELOPMENT

As illustrated in the **Introduction to GIS** chapter, the database is at the center of the Geographic Information System. In fact, it provides fuel for the GIS. The tasks of finding, creating, assembling and integrating these data may collectively be termed *database development*. While it is seldom the most interesting part, database development can easily consume 80 to 90 percent of the time and resources allocated to any project. This chapter discusses many of the issues encountered in database development and presents a summary of the database development tools included in TerrSet. Also presented are techniques and tips for importing raster data, particularly satellite imagery.

Collecting Data

The first stage in database development is typically the identification of the data layers necessary for the project. While it is tempting to acquire every available data layer that coincides with the study area, it is usually more efficient to identify what is needed prior to assessing what is available. This helps to avoid the problems of data-driven project definitions and encourages a question-driven approach.

In addition to identifying the themes that are necessary (e.g., elevation, landcover), it is also important to determine what resolution (precision) and what level of accuracy are required. These issues will be discussed more thoroughly below.

Once the necessary data requirements for the project have been specified, the search for data can begin. There are five main ways to get data into the database:

1. Find data in digital format and import;
2. Find data in hard-copy format and digitize;
3. Collect data yourself in the field, then enter it;
4. Substitute an existing data layer as a surrogate;
5. Derive new data from existing data.

Find Data in Digital Format and Import

Where to Find Data

In many countries, governmental organizations provide data as a service. In the United States, these data are often free for electronic download or are available for a small fee. The United States Geological Survey, the National Oceanic and Atmospheric Administration

and the Census Bureau are good examples of governmental agencies in the United States that have a mandate to create and provide digital data. All these agencies have Web sites that provide information about the availability of digital data. In addition, it is becoming more and more common for states (or provinces) to have a agency or department that is responsible for creating, maintaining and distributing GIS data layers. In the state of Massachusetts, for example, the state office MASSGIS has this mandate. For many, the Web is fast becoming the preferred method for finding and acquiring government-provided data.

Commercial companies also provide digital data. These data might be generic sets of commonly used base layers, such as transportation networks, or they might be customized to the needs of specific users. Commercial companies often begin with free government data and "add value" by converting it to a specific software format or updating it.

With the proliferation of GIS and image processing technologies, many non-governmental and academic institutions may also possess data layers that would be useful to your project. While these organizations seldom create data specifically to share with others, they will often make available what they have collected. Therefore, it is usually worthwhile to identify other researchers or institutions working in the same study area to inquire about their data holdings. Electronic discussion forums also provide a venue to request data layers.

Once you have located electronic data useful to your project, you need to import it into the software system you are using. Sound simple? Ideally, these format conversions are trivial. In reality, it is not uncommon to find that getting from what you have to what you want is not straightforward.

Data Formats

You must also determine the file format of the data. In general terms, you should know which category it falls within. For example, a data layer of Superfund sites could be stored as a raster, vector, database or spreadsheet data file. More specifically, you will need to know if it is in a particular software format, an agency format, or some sort of data interchange format. TerrSet provides for direct import of a number of standard raster, vector and attribute data formats. A powerful utility in TerrSet is GDALIDRISI. This utility incorporates the opensource GDAL program that translates many raster formats to the IDRISI file format.

If the file format you have is not directly supported by TerrSet import routines, then you may still be able to use it. Raster data tends to have a simpler structure than vector data and it is often the case that a raster file can be imported into TerrSet using a low-level tool, such as GENERICRASTER, to re-order the pixels. Vector data are more complex, and unless the structure is extremely simplistic, it is unlikely that you will be able to import vector data using low-level tools.

You must also consider data compression when acquiring data. Compression allows files to be stored using less memory than they would normally require. There are two types of compression of which to be aware. The first is compression that is used by the GIS software itself. It is best to avoid sharing files that are compressed with software-specific routines between different GIS software systems.

The second type of compression, sometimes referred to as "zipping", is used for files that are not currently being used. This compression is used to make the transfer and storage of files easier. The files must be uncompressed or "unzipped" prior to use.

Information about importing specific file formats is available in the TerrSet Help System.

Find Data in Hard Copy Format and Digitize

If the data you need is in a hard-copy format, you will need to digitize it (i.e., make it digital) to bring it into your GIS database. Some hard-copy data might be digitized by simply typing it into an ASCII editor or a database table. For example, you might have tabular field notes from sample survey sites that can be typed directly into Database Workshop.

More commonly, hard-copy data is in the form of a map, an orthophoto, or an aerial photograph. If you want to extract features from a map, such as elevation contours, well-head locations or park boundaries, then you will digitize these as vector features using a digitizing tablet. (Alternatively, features plainly visible on a digital orthophoto can be captured with on-screen digitizing, which is discussed below.)

A digitizing tablet (or digitizing board) contains a fine mesh of wires that define a Cartesian coordinate system for the board. Most boards have 1000 wires per inch, which results in a maximum resolution of 1/1000 inch. The user attaches the hard copy map to the digitizing board, then traces features on the map with a digitizing puck (a mouse-like device) or stylus (a pen-like device). The digitizing tablet senses the X,Y positions of the puck as the features are traced and communicates these to the digitizing software.

Digitizing software packages vary tremendously in ease of use and capability. Most digitizing software will allow the user to "register" the map on the digitizing tablet. This process establishes the relationship between the tablet's Cartesian coordinates and the coordinate system of the paper map. The software compares the tablet coordinates and the map coordinates for a set of control points and then derives a best-fit translation function. This function is then applied to all the coordinates sent from the board to the software. If your digitizing software does not provide this translation, the RESAMPLE module in TerrSet may be used to transform the tablet coordinates to map coordinates after digitizing.

In addition to digitizing using a digitizing tablet, scanners can be used to digitize hard-copy images such as maps or aerial photos. Unlike digitizing tablets that produce vector data, scanners produce raster data. Also, scanners do not capture distinct features but measure the relative reflectance of light across a document according to a user-specified resolution, normally specified as dots per inch. Each dot becomes a pixel in the resulting image. Sensors detect the reflection of red, green, and blue (or grey level) and record these as digital values for each pixel. The scanned image is then imported to the GIS software as a raster image. Scanned images are normally imported to TerrSet through either the .BMP or .TIF file formats.

Once scanned, features such as roads may be extracted into a vector file format using specialized software. However, this process is often too costly or inaccurate, and it requires very clean hard-copy sources for scanning. Another method to extract vector features from scanned raster images is with on-screen digitizing.

With on-screen digitizing, sometimes referred to as "heads-up" digitizing, the scanned source map or photo is displayed on screen and features are digitized using a standard mouse. The RESAMPLE module may be used to georegister the image before digitizing. TerrSet allows you to capture features as vector format files from a displayed raster image through on-screen digitizing.

Collect Data Yourself in the Field then Enter It

For many research projects, it is necessary to go into the field and collect data. When collecting data in the field for use in a GIS, it is imperative to know the location of each data point collected. Depending upon the nature of the project and level of accuracy required, paper maps may be used in conjunction with physical landmarks (e.g., roads and buildings) to determine locations. However, for many projects, traditional surveying instruments or Global Positioning System (GPS) devices are necessary to accurately locate data points.

Locational coordinates from traditional surveys are typically processed in Coordinate Geometry (COGO) software and may then be transformed into a GIS-compatible vector file type.

For many GIS applications, however, GPS devices may provide a less expensive alternative. The Global Positioning System is composed of 274 US Department of Defense satellites orbiting the Earth at approximately 20,000 kilometers. Each satellite continuously transmits a time and location signal. A GPS receiver processes the satellite's signals and calculates its position.¹ The level of error in the position depends on the quality (and price) of the receiver, atmospheric conditions and other variables. Some units provide display of locations only, while others record locational information electronically for later download to a computer.

¹ Several GPS companies (e.g., Trimble, Magellan) have excellent Web sites explaining how positions are calculated from the satellite signals.

TerrSet includes a GPS link. When this is active, the position recorded by the GPS receiver is displayed on the active display window and is updated as the position changes. The geodetic coordinates produced by the receiver are automatically projected to the reference system of the display. This GPS link is intended primarily to display and save waypoints and routes.

Most GPS data processing software supports several GIS export formats. Shapefiles are commonly used as a bridge to TerrSet. If this formats are not available or if the data set is small enough to be entered by hand, creation of an TerrSet vector file is easily accomplished through the use of a text editor like the Edit module in TerrSet.

Additionally, GPS software typically exports to several types of ASCII text files. The XYZIDRIS module in TerrSet imports one of the more common types. The input to XYZIDRIS must be a space- or comma-delimited ASCII file in which numeric X, Y, and Z values (where Z = elevation or a numeric identifier) are separated by one or more spaces and each line is terminated by a CR/LF pair. XYZIDRIS produces an IDRISI vector point file.

Substitute an Existing Data Layer as a Surrogate

At times, there is simply no way to find or create a data layer. In these cases, it may be possible to substitute existing data as a surrogate. For example, suppose an analysis requires powerline location information, but a powerlines data file is not available, and you don't have time or funds to collect the data in the field. You know, however, that in your study area, powerlines generally follow paved roads. For the purposes of your analysis, if the potential level of error introduced is acceptable, you may use a paved roads layer as a surrogate for powerlines.

Derive New Data from Existing Data

New data layers may also be derived from existing data. This is referred to as derivative mapping and is the primary way in which a GIS database grows. With derivative mapping, some knowledge of relationships is combined with existing data layers to create new data layers. For example, if an image of slopes is needed, but none exists, you could derive the slope image (using the TerrSet SURFACE module) from a digital elevation model, if available. Similarly, if you need an image of the relative amount of green biomass on the ground, you might derive such an image from the red and infrared bands of satellite imagery.

Another common form of derivative mapping is the interpolation of a raster surface (e.g., an elevation model or temperature surface) from a set of discrete points or isolines using TIN modeling or Geostatistics, both of which are available in TerrSet. See the **Surface Interpolation** section for more information about this form of derivative mapping.

Considerations

Resolution

In all database development tasks, no matter the source of data, there are several issues to consider. The first consideration is the resolution of the data. How much detail do you need? Resolution affects storage and processing time if too fine and limits the questions you can ask of the data if too coarse.

Resolution may refer to the spatial, temporal or attribute components of the database. Spatial resolution refers to the size of the pixel in raster data or the scale of the map that was digitized to produce vector data. Temporal resolution refers to the currency of the data and whether the images in the time series are frequent enough and spaced appropriately. Attribute resolution refers to the level of detail captured by the data values. This might be exemplified by the difference between a landcover map with a single forest class, a landcover map with hardwood and softwood forest classes, and a landcover map with many different forest classes.

Accuracy

Accuracy is the second consideration. While accuracy does have a relationship to resolution, it is also a function of the methods and care with which the data were collected and digitized. Unfortunately, it is not always easy to evaluate the accuracy of data layers. However, most government mapping agencies do have mapping standards that are available. The TerrSet metadata (documentation) file structure includes fields for accuracy information, but many other file formats carry no accuracy information. In addition, even when such information is reported, the intelligent use of accuracy information in GIS analysis is still an avenue of active research. The capabilities of TerrSet in this area are discussed in the section on **Decision Support**.

Georeferencing

The third consideration is georeferencing. Data layers from various sources will often be georeferenced using different reference systems.² For display and GIS analysis, all data layers that are to be used together must use the same reference system. Providing that the details of the reference systems are known, they can usually be converted with the TerrSet module PROJECT.

However, it may be possible to download graphic files from the Web, for example, that are not georeferenced. Also, while TerrSet supports many projections, you may find data that are in an unsupported projection. In both cases, you will not be able to use PROJECT. It is sometimes possible to georeference an unreferenced file or a file with an unsupported projection through resampling (the TerrSet module RESAMPLE) if points of known locations can be found on the unreferenced image. If data is in a projection that is not supported by TerrSet, another alternative is to use a third-party software that recognized that projection. Use that third-party software to reproject the data to a projection supported by TerrSet and then import the data. All users are encouraged to read the chapter on **Georeferencing** for more information about this key issue. Also, see the section on Data Integration below.

Cost

Finally, the fourth consideration in database development is cost. This must be assessed in terms of both time and money. Other considerations include whether the data will be used once or many times, the accuracy level necessary for the (and future) uses of the data, and how often it must be updated to remain useful. Do not underestimate the amount of labor required to develop a good database. As stated in the introduction, database development quite commonly consumes a large portion of the labor allocated to a GIS project.

General Import Tips

TerrSet requires files to be in Intel format. This typically is only an issue when integer data are acquired from Macintosh, UNIX, workstation and mainframe platforms which use the Motorola format. The byte order is different between the two formats. The GENERICRASTER module can be used to reverse the byte order of an integer file.

Tools for Import

Files are sometimes in an unspecified format. The following modules are useful for inspecting files and changing file structures.

GDALIDRISI: A general utility for importing raster data to the IDRISI file format.

Edit: Displays and edits an ASCII file. Vector data are often ASCII.

² A reference system is defined by a projection, datum and grid system. See the chapter on **Georeferencing** for more information.

CONVERT: Converts raster and vector ASCII or binary files.

TerrSet Explorer Show Structure: Displays byte-level contents of any file (in ASCII or binary format). Used to view data files for presence and size of headers, as well as to check files for carriage returns (CR's) and line feed (LF's).

CRLF: Adds or removes carriage returns and line feeds. Used with import modules that require specific fixed record lengths (e.g., DLG's, DEM's).

VAR2FIX: Converts variable length ASCII files to fixed length. Used in conjunction with CRLF.

TerrSet Explorer Metadata Utility: Creates and updates documentation files for data files already in TerrSet format.

GENERICRASTER: Imports, converts and swaps byte order on single or multi-banded raster data in band-interleaved-by-line (BIL), band-interleaved-by-pixel (BIP) or band sequential (BSQ) format.

SSTIDRIS: Converts raster images entered in from a spreadsheet program or any ASCII grid format into a TerrSet image.

Importing Satellite Imagery

TerrSet includes several special import routines for specific satellite image formats, such as Landsat, HDF-EOS and SPOT. If a specific import routine isn't available for the format of your data, the generic import tools available in TerrSet may be used to bring the satellite imagery into TerrSet. When preparing to import satellite data using the generic tools, there are three primary concerns: the location and contents of the header information (the information that describes the data), the size of the data in bytes, and the format in which the data are stored.

The first task is to locate and read the header information. Commonly, the header will be stored as a separate file in ASCII format or it will be distributed as paper documentation. Use the Edit module to read this data. If, as a third possibility, the header is attached to the data file, it will be stored in binary format because satellite imagery is usually distributed in binary format. In this case, you can try to read the information from the header with the Show Structure utility of the TerrSet Explorer which displays binary data as ASCII text. If you know the number of rows and columns in the image, but not the header size, there is a way to figure this out. If the image is in byte binary format, the number of bytes in the image should equal the number of pixels in the image. To find the number of pixels in the image, multiply the number of rows by the number of columns. The extra bytes in the data file should be the size of the header. If the file contains 2-byte integer data, the number of bytes in the image equals rows multiplied by columns multiplied by 2. You can determine the exact file size of a file by using the Properties command in Windows Explorer. For example, if an image has 512 rows and 512 columns, and the data type is binary, the file size with no header must be 262,144. As another example, if we know the file has 512 rows and 512 columns, and the file size is 262,656, we can assume that the header size is 512 bytes.

The contents of the header information will be presented in a variety of formats and with varying levels of detail. This section will tell you what to look for in the header, but it cannot anticipate the format in which you will find that information.

There are three pieces of information you must find in the header (the answers to 1 and 2 will determine the import routine you will use):

- a. the data file format;
- b. whether a header file is attached to the data file, and its size; and
- c. the number of rows and columns in the data file.

Most satellite imagery is distributed in one of two file formats: band sequential (BSQ) and band-interleaved-by-line (BIL). Band sequential format is the same as TerrSet image format. The data values are stored as a single series of numbers and each band of

imagery is contained in a separate data file. SPOT panchromatic³ and Landsat TM and MSS satellite imagery are commonly distributed in this format. Band-interleaved-by-line (BIL) format stores all the bands of multi-spectral imagery in a single data file. The bands are combined by storing the first row of data from the first band of imagery, followed by the first row of data from the second band and so on for every band of imagery stored in the data file. The data file then continues with the second row of data from the first band, the second row of data from the second band and so on. SPOT multispectral imagery (SPOT-XS) is commonly distributed in BIL format. A third format, band-interleaved-by-pixel (BIP), is uncommon but used with older datasets. The figures below illustrate the three formats.

SATELLITE IMAGERY FORMATS



Band Sequential (BSQ) with three bands 4 columns by 4 rows



Band-interleaved-by-line (BIL)
with three bands 4 columns by 12 rows



Band-interleaved-by-pixel (BIP)
with three bands 12 columns by 4 rows

With BSQ format files, the following two import routines apply:

1. If a header is attached to the data file, use the GENERICRASTER module, specify BSQ and no header. GENERICRASTER will remove the header, rename the data file to have an .RST extension, and create a documentation file. GENERICRASTER will ask for the number of rows and columns, the minimum information necessary to create the documentation file. GENERICRASTER will ask for additional information as well, such as the reference system parameters which you should be able to find in the header. When in doubt, you can try the following values: Minimum X=0, Maximum X=1 (or # of columns), Minimum Y=0, Maximum Y=1 (or # of rows), Reference System=Plane.
2. If the header file is separate from the data file, then run GENERICRASTER and use the BSQ option and specify one band. Then specify the appropriate reference information as in the step above. Again, you must know the number of columns and rows in order to complete this step, but you should also enter any reference system information that is available from the header. If the values are unknown, try entering the values noted in the previous paragraph, with the addition of Reference Units=meters and Unit Distance=1.

For all BIL files, with or without a header, the following import routine is used:

If a header is attached, determine its size (in bytes) and then use the GENERICRASTER module. GENERICRASTER will pull apart and create individual TerrSet format image and documentation files for each band of imagery.

There are some common errors during the import process that can give the user some useful information for correction. If the header size is incorrectly specified using GENERICRASTER, the image will not be correctly imported. Because too much or too little is taken out of the file as a header, the actual data left behind is either incomplete or contains extra information. This can lead to an error

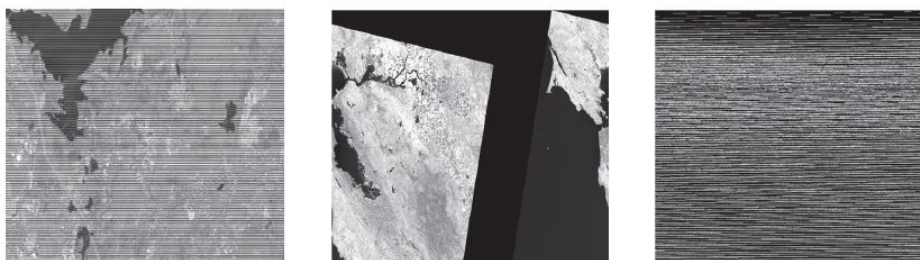
³ The documentation for the SPOT panchromatic band will indicate that it is in BIL format, but it contains only one band, so there is no interleaving. This is the same as BSQ format.

message informing you that the image is incorrectly sized, or the image will display improperly (e.g., data from the right side may appear moved to the left side). You will need to determine the correct header size and try again.

If an image seems to successfully import, but the display of the image is staggered or skewed to one side, it is possible that you have incorrectly specified the number of rows and columns in the documentation file. Recheck your documentation file for the correct information or use the method discussed above to determine the correct rows and columns based on the image and header size. If an image seems to successfully import but the display of the image contains a venetian blind striping effect, there are two possible causes. If the striping is vertical, try importing the image as a BIP format file using GENERICRASTER. If the striping is horizontal, try importing as a BIL format file. The first figure below left illustrates an import when the incorrect number of bands is specified during a BIL import. The figure in the center illustrates a common result when the header is not calculated correctly. The figure on the right demonstrates a result when the incorrect number of columns is specified.

A common error associated with importing integer data results from data originating from the UNIX environment. If after successfully importing integer data the result shows very maximum and minimum values in the image, most likely the bytes need to be swapped. Run the GENERICRASTER routine again but check the option to Swap byte order. The byte swapping option is used for the conversion of UNIX and Macintosh files that have the low byte/high byte order reversed from Intel-based systems.

Common errors during import



Importing GIS Data

Much of the information presented above regarding the import of satellite imagery also applies to the import of raster GIS data. Specific import routines are also available in TerrSet for many common agency and software-specific GIS formats. The procedures for importing common GIS data formats are found in the Help System. Search the Help System index for the name of the format you would like to import. In many cases, GIS data can be saved in a format that can be imported into TerrSet, such as ERDAS Imagine, ESRI ArcRaster, or GEOTIFF files.

Data Integration

Requirements for Data Integration

Data integration refers to the use of various data layers together in display or analysis. There are several parameters that must match if data layers are to be used together. First, the reference systems must match.⁴ This is important because we are often combining separate themes based on their spatial characteristics. A roads vector layer can be properly overlayed on a raster satellite image, for example, only if both layers have the same reference system. A particular X,Y coordinate pair in the vector file must represent exactly the same place on the ground as that same X,Y coordinate pair in the raster image.

⁴ The reference system projection, datum and grid system must all match for the reference system to match.

For raster layers to be used in image-to-image operations, such as OVERLAY, the extent of the images, the number of rows and columns in the images, and the pixel resolution of the images must all match. Note that if any two of these three conditions are met, the third is met automatically.

The steps used to integrate data layers, especially if data is transformed, should be recorded in the lineage field of the documentation file of each data layer. For example, it is trivial to make smaller pixels from larger pixels, and this is often done to achieve data integration. However, it is important to note that the information carried by those smaller pixels still represents data gathered at a coarser resolution. Such information should be recorded in the new image's documentation file so users will be aware that the apparent resolution exceeds that of the information.

Tools for Data Integration

The modules PROJECT and RESAMPLE are the primary tools available in TerrSet for georeferencing and changing reference systems. The chapter on **Georeferencing** details the differences between these two modules and when they should be employed. In addition, the **Tutorial** includes exercises on the use of RESAMPLE and PROJECT.

The WINDOW module is used to change the extent of raster images. It works by saving a subset of a larger image as a new image.

Changing the number of rows and columns in an image and/or changing the image pixel resolution can be accomplished with the modules CONTRACT, EXPAND and PROJECT. CONTRACT and EXPAND work by thinning or duplicating pixels in an image by an integer multiple. For example, an image with 100 rows and 100 columns and a pixel resolution of 30 meters could be contracted to an image of 25 rows and 25 columns by using a contraction factor of 4. Similarly, an image of 100 rows and 100 columns could be transformed to an image of 1000 rows and 1000 columns by using EXPAND and an expansion factor of 10.

If it is necessary to contract or expand an image by a non-integer factor (e.g., 1.5), CONTRACT and EXPAND cannot be used. In these cases, the easiest method to use is PROJECT. Project to the same reference system that the file is currently in, enter the new number of rows and columns, and choose the resampling technique that best suits your data and application. In general, if quantitative data are used, then the bilinear resampling type should be chosen, but if qualitative data are used, the nearest neighbor resampling type should be chosen.

Conclusions

Database development is seldom quick and easy. The quality of the database, in many ways, limits the quality of any resulting analysis to be carried out with the data. There are a significant number of technical pitfalls that are typically encountered during database development and it is easy to become overwhelmed by these. However, despite these difficulties, it is important to maintain a project-driven approach to GIS analysis and to limit the influence data constraints can have on the definition of the problems to be addressed.

APPENDIX 8

IMAGE PAIRWISE COMPARISONS

With pairwise comparisons we can break down the techniques according to whether they are suitable for quantitative or qualitative data. Quantitative data has values that indicate an amount or measurement, such as NDVI, rainfall or reflectance. Qualitative data has values that indicate different categories, such as census tract ID's or landuse classes.

Quantitative Data

Image Differencing

With quantitative data, the simplest form of change analysis is *image differencing*. In TerrSet, this can be achieved with the OVERLAY module through a simple subtraction of one image from the other. However, a second stage of analysis is often required since the difference image will typically contain a wide range of values. Both steps are included in the module IMAGEDIFF, which produces several common image difference products: a simple difference image (later - earlier), a percentage change image (later-earlier/earlier), a standardized difference image (Z-scores), or a classified standardized difference image (z-scores divided into 6 classes). Mask images that limit the study area may also be specified.

Care must be taken in choosing a threshold to distinguish true change from natural variability in any of these difference images. There are no firm guidelines for this operation. A commonly used value for the threshold is 1 standard deviation (STD) (i.e., all areas within 1 STD are considered non-change areas and those beyond 1 STD in either the positive or negative direction are considered change areas), but this should be used with caution. Higher values may be more appropriate and in some cases natural breaks in a histogram of the simple difference or percentage change images may be more sensible as a basis for choosing the threshold values.

Image Ratioing

While image differencing looks at the absolute difference between images, *image ratioing* looks at the relative difference. Again, this could be achieved with OVERLAY using the ratio option. However, because the resulting scale of relative change is not symmetric about 1 (the no change value), it is recommended that a logarithmic transformation be undertaken before thresholding the image. The module IMAGERATIO offers both a simple ratio and a log ratio result.

Regression Differencing

A third form of differencing is called *regression differencing*. This technique should be used whenever it is suspected that the measuring instrument (e.g., a satellite sensor) has changed its output characteristics between the two dates being compared. Here the earlier image is used as the independent variable and the later image as the dependent variable in a linear regression. The intercept and slope of this regression expresses the offset and gain required to adjust the earlier image to have comparable measurement characteristics to the later. In effect, we create a predicted later image in which the values are what we would expect if there were no change other than the offset and gain caused by the changes in the sensor. The equation is:

$$\text{predicted later image} = (\text{earlier image} * \text{gain}) + \text{offset}$$

With the sensor differences accounted for, the predicted later image and the actual later image may then be analyzed for change. Note that this technique requires that the overall numeric characteristics of the two images be equal except for sensor changes. The technique may not be valid if the two images represent conditions that are overall very different between the two dates.

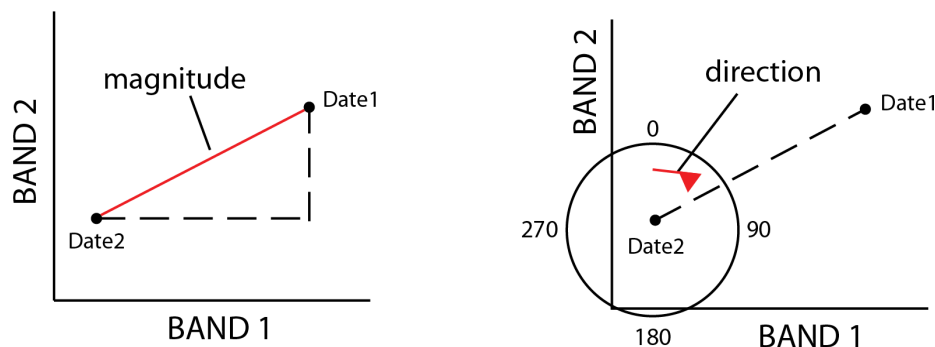
The module CALIBRATE automates the image adjustment process. The input image (the one to calibrate) is used as the independent variable and the reference image is used as the dependent variable in the regression. The output image is adjusted to the characteristics of the reference image and thus can be used in a standard comparison operation (such as IMAGEDIFF or IMAGERATIO) with any image also based on this reference, including the reference image itself.

Note that CALIBRATE also offers options to adjust an image by entering offset and gain values or by entering mean and standard deviation values.

Change Vector Analysis

Occasionally, one needs to undertake pairwise comparisons on multi-dimensional images. For example, one might wish to undertake a change analysis between two dates of satellite imagery where each is represented by several spectral bands. To do so, *change vector analysis* can be used. With change vector analysis, difference images are created for each of the corresponding bands. These difference images are then squared and added. The square root of the result represents the magnitude of the change vector. All these operations can be carried out with the Image Calculator, or a combination of TRANSFORM and OVERLAY. The resulting image values are in the same units as the input images (e.g., dn).

When only two bands (for each of the two dates) are involved, it is also possible to create a direction image (indicating the direction of change in band space). The module CVA calculates both magnitude and direction images for 2-band image pairs. The figures below illustrate these calculations. The magnitude image is in the same units as the input bands and is the distance between the Date 1 and Date 2 positions. The direction image is in azimuths measured clockwise from a vertical line extending up from the Date 2 position.



Qualitative Data

Crosstabulation / Crossclassification

With qualitative data, CROSSTAB should be used for change analysis between image pairs and there are several types of output that can be useful. The crosstabulation table shows the frequencies with which classes have remained the same (frequencies along the diagonal) or have changed (off-diagonal frequencies). The Kappa Index of Agreement (KIA) indicates the degree of agreement between the two maps, both in an overall sense and on a per-category basis. Finally, the crossclassification image can readily be reclassified into either a change image or an agreement image. Note that the numeric values of data classes must be identical on both maps for the output from CROSSTAB to be meaningful.

APPENDIX 9

PREDICTIVE MODELING WITH GEOMOD

Predictive Change Modeling

In some cases, knowing the changes that have occurred in the past may help predict future changes. A suite of modules in TerrSet has been developed to provide the basic tools for predictive landcover change modeling. These tools include Land Change Modeler and GEOMOD. Land Change Modeler is covered extensively in the next chapter, but briefly, it is used to assess historical land cover data and to use that assessment to predict future scenarios. It does this by assessing past land transition information and incorporating environmental variable maps that might drive or explain change to create future scenarios and layers of transition potential maps. Although the underlying logic is different, GEOMOD is similar in that it can simulate the transition from one land use state to another, however, Land Change Modeler can act upon many transitions at once.

GEOMOD

GEOMOD is a land use change simulation model that predicts the transition from one landuse state to another landuse state, i.e., the location of grid cells that change over time from one state to another. GEOMOD simulates the change between exactly two categories, state 1 and state 2. For example, GEOMOD could be used to predict areas likely to change from forest (state 1) to non-forest (state 2) over a given time. The simulation can occur either forward or backward in time. The simulation is based on:

- specification of the beginning time, ending time and time step for the simulation,
- an image showing the location of landuse states 1 and 2 at the beginning time,
- an optional mask image distinguishing between areas in and out of the study region,
- an optional image of stratification that shows the study area divided into regions, where each region is a stratum,
- a decision whether to constrain the simulated change to the border between state 1 and state 2,
- a map of suitability for the transition to landuse state 2,
- the anticipated quantity of landuse states 1 and 2 at the ending time.

An additional option allows for the creation of an image of environmental impact for each time step of the simulated land change. GEOMOD can also generate a map of cumulative impact for the entire duration of the simulation. These maps of impact show the

magnitude of change to an environmental resource at the locations of simulated landuse change. For example, these maps could show carbon emissions that result from the conversion from forest to non-forest.

A full discussion of GEOMOD along with an extensive white paper can be found in the Help.
