BENDRASIS PROGRAMAVIMO DOKUMENTAS

EUROPOS SĄJUNGA

SOCIALINĖS APSAUGOS IR DARBO MINISTERIJA

NACIONALINĖ ŽEMĖS TARNYBA
PRIE ŽEMĖS ŪKIO MINISTERIJOS

THE BASIS OF AN EU CO-FINANCED PROJECT "QUALIFICATION IMPROVEMENT OF THE LITHUANIAN GEOGRAPHICAL INFORMATION MANAGERS" (NO BPD2004-ESF-2.2.0.-02-05/0143)

Civil servants learning programme

DISTANCE LEARNING OF GEOGRAPHIC INFORMATION INFRASTRUCTURE

Training Material

SPATIAL ANALYSIS AND MODELING

GII-07

Vilnius, 2008

MALASPINA
UNIVERSITY-COLLEGE

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

HNIT - BALTIC

Training material „Spatial analysis and modeling“ (GII-07)

Authors

1 Module -Ann Blyth, Dave Cake

2 Module - Ann Blyth, Ian Laing, Dave Cake

3 Module - Martin Andresen

4 Module - dr. Gennady Gienko and Michael Govorov

5 Module - Ian Laing, Dave Cak

Reviewed and edited by

doc. dr. Gintautas Mozgeris (Lithuanian University of Agriculture)

Reviewed by

doc. dr. Antanas Dumbrauskas (Lithuanian University of Agriculture)

From English translated and edited by

Astraneta UAB

Study material for civil servants distance learning programme “GII Distance Learning”, which has been approved by a law “RE Approval of Study Programmes for Civil Servants” (No 27V-72; Apr 17, 2007) by the State Service Department under the Ministry of the Interior has been prepared on the basis of an EU co-financed project “Qualification Improvement of the Lithuanian Geographical Information Managers” (No BPD2004-ESF-2.2.0.-02-05/0143)

Beneficiary of the project National Land Service under the Ministry of Agriculture.

Contractor of the project HNIT-BALTIC UAB, Vilnius Gediminas Technical University

Table of Contents

1 Data Exploration 4
1.1 Data Attributes 6
1.2 Querying and Selecting Vector Data 17
1.3 Querying and Selecting Raster Data 28
1.4 Summarizing and Interpreting Data 32
References 39
2 Basic Spatial Analysis 40
2.1 Topic 1: Vector Analysis Methods 41
2.2 Raster Analysis Methods 54
2.3 Generalizing Data 73
References 81
3 Spatial Statistics 82
3.1. Classical Correlation, Spatial Autocorrelation, and the Location Quotient 83
3.2. Modifiable Areal Unit Problem and the Ecological Fallacy 88
3.3. Pattern Analysis 90
3.4. Edge Effects 98
3.5. Density Estimation and Hot Spot Mapping 101
3.6. Local Spatial Statistics 107
References 112
4 Geostatistics 114
4.1. Introduction to Geostatistics 115
4.2. Classification of Geostatistics Methods 119
4.3. Interpolation, Estimation and Smoothing Methods 121
4.4. Kriging Prediction 131
4.5. Model Validation 154
4.6. Conclusion and Comparison of Geostatistical Methods 164
References 167
5 Spatial Modelling 169
5.1 Modelling Spatial Problems 170
5.2 Classification of Spatial Models 173
5.3 Model Types 175
5.4 Model Builder 177
5.5 Model Examples 179
References 191

1 Data Exploration

What makes Geographic Information Systems (GIS) unique is the ability to link data to spatial locations and query and summarize these data based on specific analysis requirements. Functionally, GIS provides a sophisticated tool for reporting the results of a database. These reports may be for an entire dataset (or table) or for a portion of the dataset (e.g., based on the results of a query or data summary). These 'data reports' can take the form of tabular summaries, graphs and maps. The following module will teach you about the different types of data and how to select and query the information within a database. In addition, the types of reports that can be generated are detailed. This module will examine data exploration in the following four topics:

Topic 1: Data attributes
Topic 2: Querying and selecting vector data
Topic 3: Querying and selecting raster data
Topic 4: Summarizing and interpreting data

It should be noted that many of the subtopics discussed detail specific technical concepts. To make these concepts easier to understand examples, based on the ArcMap (version 9.2) user interface have been provided. While these examples are based on ArcMap, conceptually they will be applicable to other GIS user interfaces.

1.0.1 Course Overview

GIS has been characterized as a set of tools for collecting, storing, analyzing and displaying geographic data. Much of the effort in GIS focuses on tasks relating to the ability to represent and describe real world objects in a digital environment. Topics in this field relate to the abstraction of features as points, lines and polygons in a GIS database, the understanding and use of coordinate systems and map projections for describing locations on the earth's surface, and the physical file formats used to represent features in a spatial database. Much of the capability of GIS software applications emphasizes these areas of interest. Several courses in this training program focus on these topics, including:

GII-01	Elements of GIS
GII-05	Geographic DBMS
GII-06	Geodesy and Cartography

While these GIS functions relate primarily a spatial inventory of features, the analysis functions of a GIS seek to help in the understanding of the patterns and processes which lie beneath the features represented in a spatial database. It is these analytical capabilities which separate GIS from related applications such as computer aided design (CAD) and automated cartography. Spatial analysis might help researchers understand a process or distribution of features, or it might help an organization make better decisions based on a more thorough understanding of the data.

Spatial analysis topics appear in several courses in this training program:

GII-01	Elements of GIS
GII-04	Applications of Geographic Information Infrastructure

GII-07 Spatial Analysis (this course)

GII-01, being the introductory GIS course, covers most GIS areas including spatial analysis in a generalized manner. As a result, some of its content will overlap with this course (GII-07) somewhat, particularly in Module 2 of this course. The intention is to have this course cover such material in more detail than was presented in GII-01. GII-04 will address highly specialized analysis techniques such as network analysis, terrain modeling and hydrological modeling as part of its discussion of specific application areas in the field of GIS. This course will concentrate on the major generalized analysis techniques. As a result, GII-04 and GII-07, while perhaps sharing some basic concepts, should not have significant overlap.

The goal of this course is to provide an introduction to the techniques used in the analysis of spatial data. There is a broad range of analytical tools in a typical GIS application, ranging from simple actions such as a measurement of distance to sophisticated models which seek to predict spatial outcomes based on existing or predicted conditions. Given this vast range of techniques for spatial analysis, most can only be covered to a limited depth.

Module 1 looks at the most basic of analytical tools, those to simply visualize or examine existing spatial data. We will discuss simple query and selection procedures, and look at elementary ways of summarizing existing data using basic statistics, summaries and thematic maps.

Module 2 examines what many would consider the mainstream analysis tools. These are the geoprocessing tools and techniques which are used constantly in any GIS workplace, such as overlay, buffer and clip operations. Tools for both raster and vector analysis will be examined in this module.

Module 3 will discuss the statistical methods applied to spatial data. These have their basis in classical statistics, but which accommodate the limitations and characteristics of spatial data. Such tools are very effective for describing and understanding large volumes of spatial data. We will look at topics such as spatial autocorrelation, pattern analysis and density estimation methods.

Module 4 addresses the process of understanding continuous geographic phenomena. This geostatistics module will explore the processes of interpolation, smoothing and prediction of observations using several techniques, such as inverse distance weighting (IDW), polynomial and spline approximations, and Kriging.

Module 5 examines the nature and use of models in spatial analysis. This module will discuss the many different types of models and examine several example models in detail.

1.1 Data Attributes

1.1.1 Location (e.g., X, Y and Z attributes)

Fundamental to spatial analysis is the concept of place – where on the earth's surface is a given feature or group of features located. The majority of spatial analysis is conducted on features that exist in two-dimensional space. More complex analyses can be conducted on three-dimensional data (e.g., features that exist above or below the earth's surface). In some instances GIS-based datasets can also consider a fourth dimension: time. An example of a multi-temporal dataset would be a land use coverage that stores land use attributes for multiple years allowing the dynamic nature of land use changes to be analyzed over time.

Data are stored in four basic formats - three of these involve the storage of spatial information: vector; raster; and triangular irregular network (TIN). In addition, you can utilise tabular data which can subsequently be related to spatial datasets.

Vector

Vector data is constructed using points. The location of these points is defined by up to three coordinates: x, y and z. At minimum, an x and y coordinate pair is required to specify a point's location. The z coordinate can represent an additional value related to a point, for example, elevation. The values of the coordinates themselves are a function of the coordinate system the data is stored in (e.g., latitude and longitude coordinates, or Universal Transverse Mercator [UTM]). Elevation values (a common z coordinate) can also be stored in a variety of units (e.g., metres or feet). A vector data model uses points (with their associated x and y coordinates) to construct spatial features in the form of points, lines and polygons (areas) (Figure 1). Points have 0 dimension, possessing only the property of location. Lines are one-dimensional, having a length property. The simplest line is defined by two points: a start and an end point. The shape of more complex lines is defined by an infinite number of points existing between the start and stop point – curved lines can be smoothed by increasing the number of points. Polygons (or area features) are closed lines and are two-dimensional. These features have the properties of area and perimeter. An area may exist alone or be connected to other features through shared boundaries. Topology defines the spatial relationships between connecting or adjacent vector features. For example, how polygons share common boundaries and how lines snap to one another (e.g., road intersections).

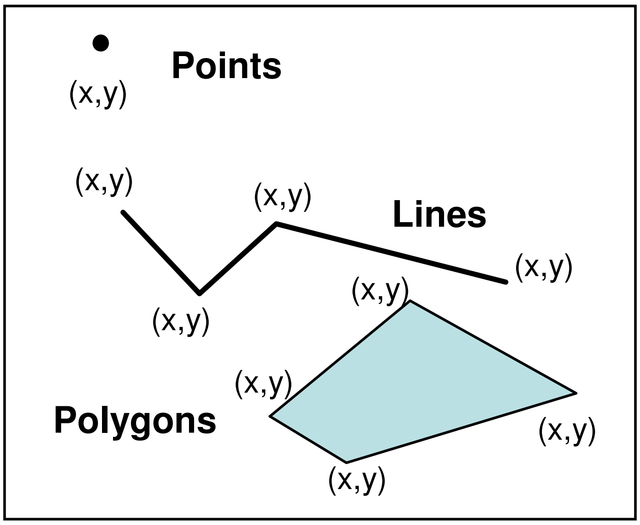


Figure 1. Vector Data: Points, Lines and Polygons

Attributes for points, lines and polygons are associated with each feature and can be stored either within the spatial dataset (a flat data structure) or in independent attribute tables which are related to the spatial features through a unique identifier.

Discrete features (individually discernable features), such as, sampling locations (usually stored as points), roads (usually stored as lines) and parcel boundaries (usually stored as polygons) are examples of data typically represented by a vector data model.

Vector data and models have their own advantages[1]:

- Data can be stored efficiently with high precision
- They require about 10% of storage space required to store same data in raster format
- Certain types of topological analysis are more efficient, or only possible, with vector
- Greater precision and accuracy
- Greater flexibility in storing and manipulating attribute data

Raster

In a raster data model, data are represented by a surface divided into a grid of regularly sized cells. Typically, a raster is a grid that has an origin (usually upper left-hand corner) and the location of each pixel is defined by this origin and an offset. A data grid is usually a rectangular shape and its size is defined by a number of rows and columns; the grid extent is calculated by multiplying the size of the grid (number of columns by number of rows) by the size of a pixel (expressed in a metric system). Although a wide variety of raster shapes are possible (e.g., triangles, hexagons) generally a series of rectangles, or more often, squares, called grid cells, are used. Each cell stores an attribute or value related to a portion of the earth's surface. A raster data model is ideally suited for storing information related to continuous features (features that are not spatially discrete) for example temperature data or elevation. Images such as, air photos and satellite images are also

[1] Which are at the same time disadvantages of raster

stored in raster format (Figure 2). Data are stored in a simple data structure based on rows and columns linked to fixed cell locations.

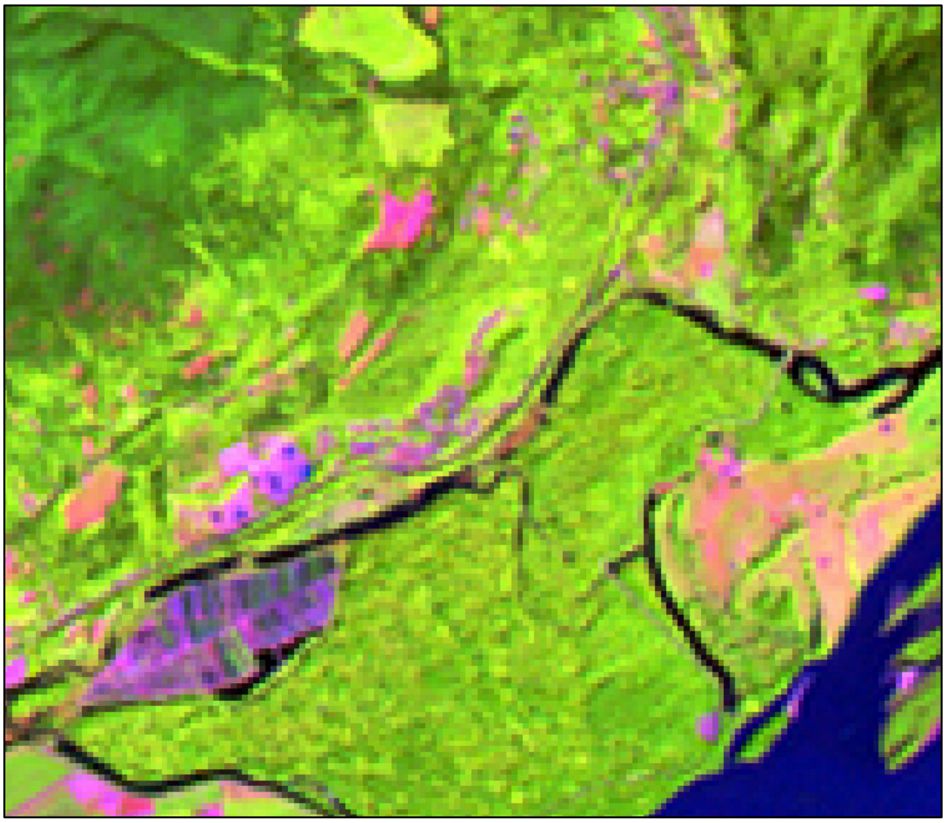
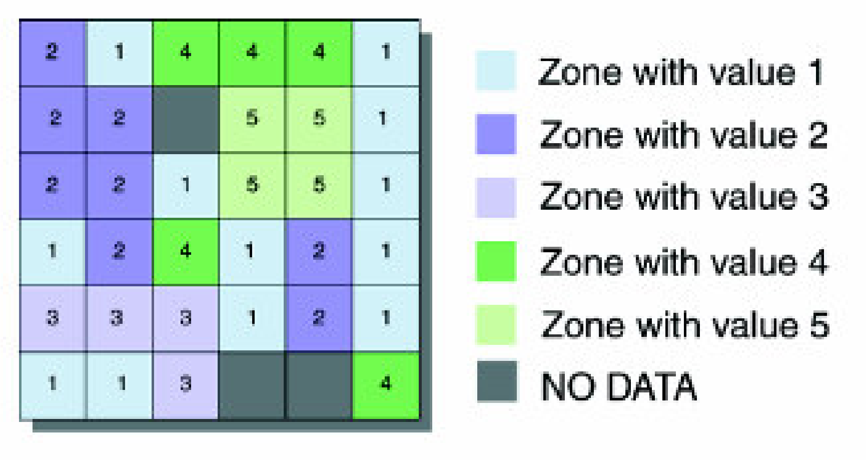


Figure 2. Example Raster Structure: Satellite Image

Each cell in the grid has a specified width and height with cell sizes ranging from a few centimetres or metres to a square kilometre. The dimension of the cells is typically a function of the resolution of the data. The cell size determines how coarsely or finely the data are represented – typically the more complex the data the more cells required to accurately represent the features. One apparent method of increasing the accuracy would be to maintain a very small cell size for all data; however, an important consideration is: the smaller the cell size the bigger the file size required to store the data and the slower the drawing time and processing speed when using the data. Typically cell sizes represent a compromise between data availability, accuracy, storage capability and processing/drawing speed. Cell size is referred to as the resolution of the data: if the cell sizes are 25 x 25 metres, the resolution of the dataset is 25 metres. The smaller the cell size, the higher the resolution and, therefore, the greater the detail of the data being displayed.

Linked to each cell is a value corresponding to the attribute being displayed (e.g., a precipitation dataset would have a rainfall amount linked to each cell or a land use coverage would have attributes such as urban, agriculture and forest linked to the cell). The values associated with each cell can be positive or negative, integer or floating-point. In some cases values of NODATA can be used to represent the absence of data (Figure 3).



2	1	4	4	4	1
2	2		5	5	1
2	2	1	5	5	1
1	2	4	1	2	1
3	3	3	1	2	1
1	1	3			4

Figure 3. Raster Data Example

There are a number of advantages of raster data and models. These include:

- Simple data model
- Multiple spatial analysis functions often simpler and faster
- Efficient for data with high spatial variability
- Efficient for low spatial variability when compressed
- Easy to integrate with satellite and remotely-sensed data

Triangulated Irregular Network

A Triangulated Irregular Network (TIN) data model is ideal for representing surfaces, for example, terrain. Data are represented in the form of a series of non-overlapping triangles drawn between irregularly spaced points (Figure 4). In the case of a terrain model, each triangle represents an area of constant slope or gradient. Due to their capability of mapping irregularly spaced data, TINs can model surfaces that vary sharply in some areas more accurately that a raster – where data is more variable, more points can be added to represent the increased variability and fewer points are required where the surface is less variable.

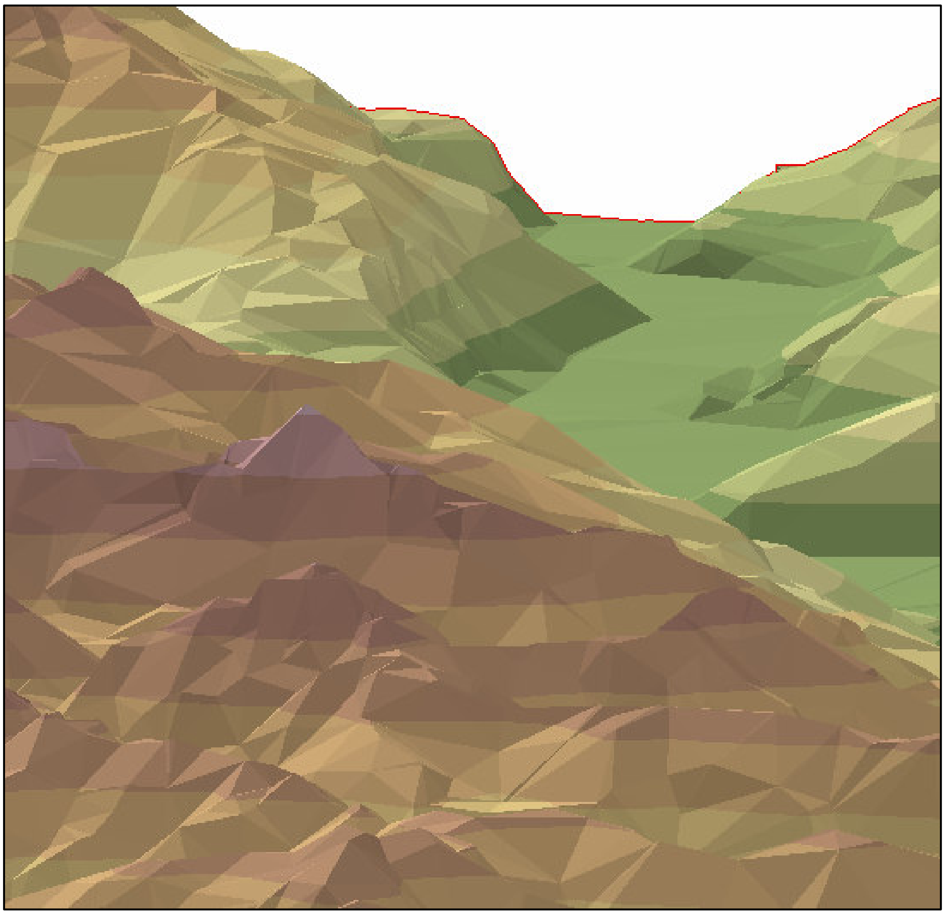


Figure 4. Example of a TIN Terrain Surface

Tabular

Tabular data are used to store attributes or descriptive information about spatial data. Typically information is stored in rows and columns (fields) in a database. The attributes may be stored in a table together with the spatial information or they can exist in separate tables that can be linked, or related to the spatial dataset through a unique identifier. GIS software packages can utilize tabular data in a variety of formats including: delimited text files, dBASE files, Microsoft Excel files, Microsoft Access files. In addition, tables can be imported from a variety of other database management software packages (e.g., Oracle).

1.1.2 Attribute types

As discussed above, attribute data describe the properties of spatial features. However, attribute data can be classified by data type. Data types vary between different GIS software packages, however typical data types include character, integer, float, decimal, single, double and string. Typical data types support within a GIS are detailed in Table 1 below.

Table 1. Typical Data Types *

Name	Field Range/Length	Application
Text	Up to 64,000 characters	Suitable for the storage of names, classes or other text-based attributes
Date	mm/dd/yyy hh:mm:ss A/PM	Dates and times
Short integer	-32,768 to 32,767	Numeric attributes without fractional values (i.e., whole numbers) within a specified range. This type of field would be useful for the storage of coded values.
Long integer	-2,147,483,648 to 2,147,486,647	Numeric attributes without fractional values with the allowed ranges (this type of field can storage a decimal value).
Single precision floating point number (Float)	Approximately -3.4E38 to 1.2E38	Numeric attributes with fractional values within the allowed ranges (this type of field can storage a decimal value).
Double precision floating point number	Approximately -2.2E308 to 1.8E3.8	Numeric attributes with fractional values within the allowed ranges (this type of field can storage a decimal value).
BLOB	Varies	Images or other multi-media files
GUID	36 characters	Customized applications requiring global identifiers

* The field names and range used are derived from the ArcMap field types but the examples would be applicable to other GIS software packages.

Numeric fields may be stored in a variety of numeric data types, for example: short integer; long integer; single precision floating point number, double precision floating point number. Each type varies in the size and method used to store numeric data values as indicated in Table 1. Text fields are used to store alphanumeric symbols (e.g., vegetation names). The date data type can store dates, times or dates and times. Images, programming code and assorted multi-media files are stored as binary large object (BLOB) data types. GlobalID and GUID data types store registry style strings that uniquely identify a feature or table row within a geodatabase and across geodatabases.

Attribute data can also be defined by measurement scale. These attribute types include categorical data (e.g., nominal, ordinal data types) and numeric data (e.g., interval, ratio and cyclic data). Each of the different data types are defined below:

- **Nominal** data describes different categories of data (e.g., land use types or vegetation). Nominal data can be numerical (e.g., phone numbers or numerical values assigned to classes of land use) but there is no implied ranking between the classes.
- **Ordinal** data implies a ranking between classes, for example, traffic volumes may be assigned the classes High, Moderate or Low. However, while ranked, ordinal data are qualitative in nature (i.e., they do not have an associated numerical value).

- **Interval** data are quantitative, having known measurements or intervals between the values. Examples of interval data include: elevation data, precipitation levels, or temperatures.

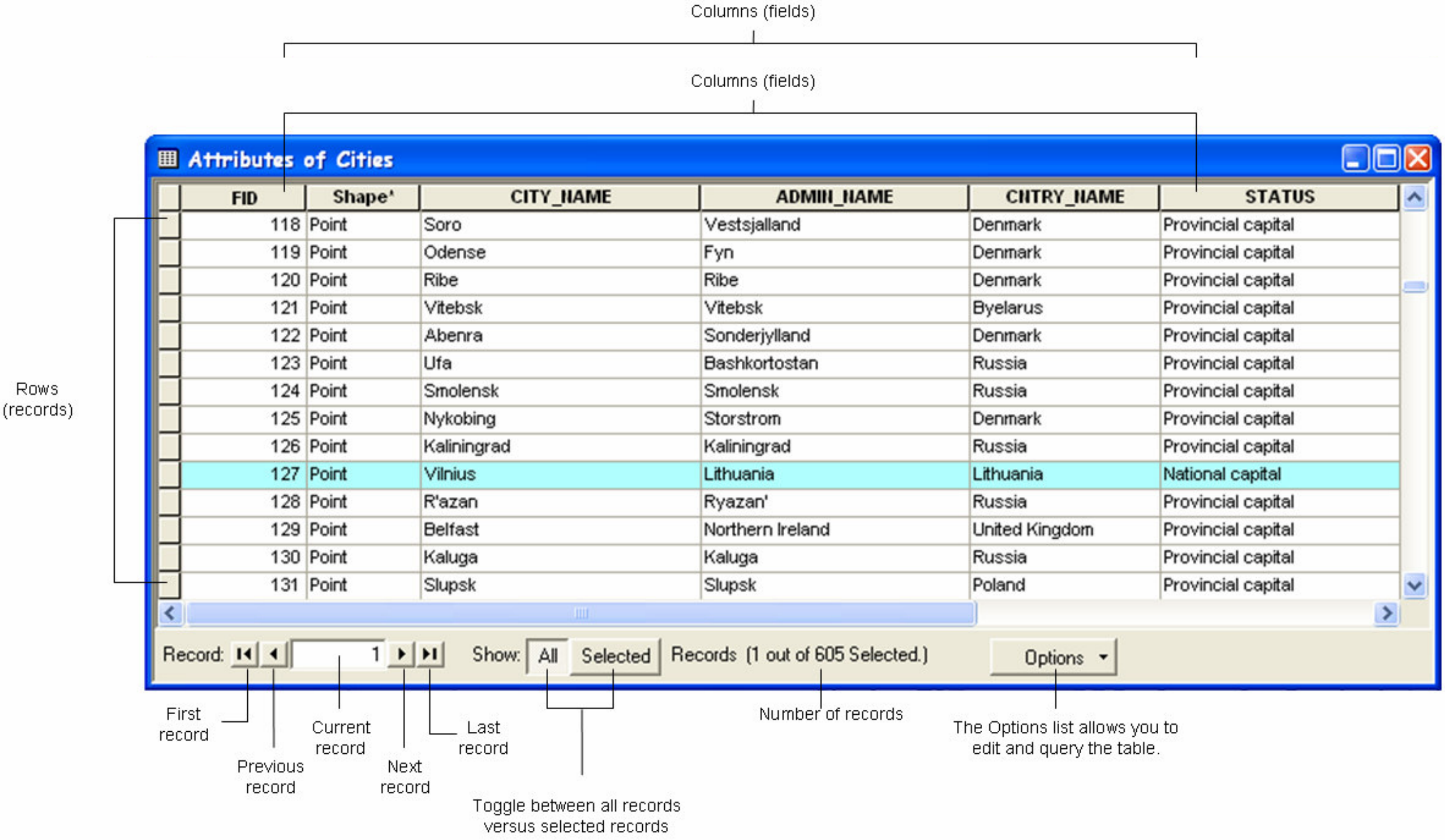
- **Ratio** data are also quantitative in nature. They are similar to interval data but they are based on a meaningful, or absolute, zero value. Density values would be an example of ratio data as densities can range from 0 to infinity.

- **Cyclic** data are measurements of attributes that represent directions or cyclic phenomena. In some cases two points on a scale can be equal. For example, in directional data 0° and 360° are equal.

The data type used to store the information associated with a given attribute is a function of the measurement scale required to represent the data. For example, nominal or ordinal data would be stored in using a character data type, whereas interval, ratio or cyclic data would be stored as an integer or decimal (float). As mentioned above, nominal data can be represented by numerical values which would typically be assigned an integer data type. However, in this instance the numbers are being used as codes that would reference a look-up table for the full values.

1.1.3 Attribute tables and data structures

Attributes are stored within attribute tables. Typically a given table will store a set of attributes for a group of features that are similar in nature. The organization of an attribute table is referred to as its data structure. Attributes tables are structured into rows and columns. Columns represent a field or attribute value or a certain type. For example, a dataset of water sampling sites might have an attribute table detailing water temperatures, pH values or sediment levels. Columns or fields within this table could include; an identifier field, minimum temperature, maximum temperature, winter pH, summer pH etc. The rows in the table represent the individual characteristics or attributes associated with each location (feature). These are also referred to as records in a database (e.g., the identifiers and the temperature and pH values associated with each location). A table has a specified number of columns (e.g., fields) but can have any number of rows. Figure 5 provides an example of how an attribute table is displayed. The example provided in the figure is based on the ArcMap user interface, however, the concepts (e.g., rows, columns and records) apply to other GIS user interfaces.



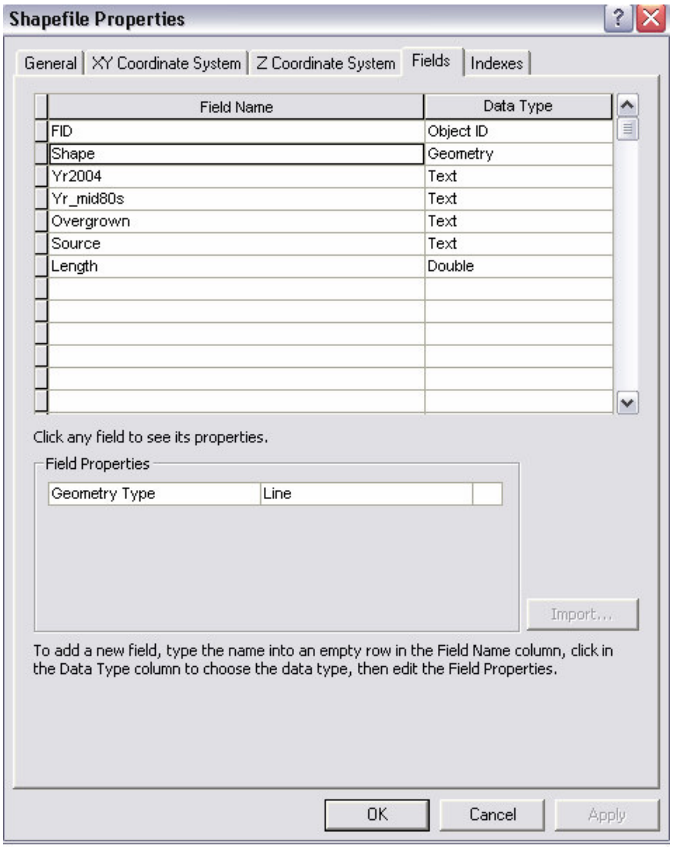
FID	Shape*	CITY_NAME	ADMIN_NAME	CNTRY_NAME	STATUS
118	Point	Soro	Vestsjalland	Denmark	Provincial capital
119	Point	Odense	Fyn	Denmark	Provincial capital
120	Point	Ribe	Ribe	Denmark	Provincial capital
121	Point	Vitebsk	Vitebsk	Byelarus	Provincial capital
122	Point	Abenra	Sonderjylland	Denmark	Provincial capital
123	Point	Ufa	Bashkortostan	Russia	Provincial capital
124	Point	Smolensk	Smolensk	Russia	Provincial capital
125	Point	Nykobing	Storstrom	Denmark	Provincial capital
126	Point	Kaliningrad	Kaliningrad	Russia	Provincial capital
127	Point	Vilnius	Lithuania	Lithuania	National capital
128	Point	R'azan	Ryazan'	Russia	Provincial capital
129	Point	Belfast	Northern Ireland	United Kingdom	Provincial capital
130	Point	Kaluga	Kaluga	Russia	Provincial capital
131	Point	Slupsk	Slupsk	Poland	Provincial capital

Figure 5. Elements of an Attribute Table

Attribute tables can be viewed, edited and queried within a GIS user interface through this kind or data browse function. In addition, tabular data related to a specific feature can also be accessed by selecting (usually by pointing at or clicking on a feature [e.g., a point, line or polygon]) in the map window. For example, ArcMap has an identify tool to perform this type of data query. To activate the Identify function left click on the identify button. In the map window, left click on the feature of interest and the tabular data for the map layers with data at that location will be displayed. You have the options of viewing data for the Top Layer, Visible Layers, Selectable Layers, All Layers, or you can pick a specific layer from a list.

1.1.4 Field properties

When you create a new table or feature class you can specify the number of fields to be included in an attribute table through a properties dialog box. Figure 6 provides an example based on the ArcMap user interface of what this type of dialog box looks like. You can also specify settings for fields, such as the field type and the maximum size of the data that can be stored in the field.



Field Name	Data Type
FID	Object ID
Shape	Geometry
Yr2004	Text
Yr_mid80s	Text
Overgrown	Text
Source	Text
Length	Double

Click any field to see its properties.

Field Properties

Geometry Type	Line

To add a new field, type the name into an empty row in the Field Name column, click in the Data Type column to choose the data type, then edit the Field Properties.

Figure 6. Shapefile Properties Dialog Box

The viewing properties of the fields associated with an existing layer can be defined. Typically you can: specify the fields to be displayed (e.g., you can toggle the field display on or off); assign an alias to a field name to make it easier to understand the contents of the field; define the formatting for numeric data (e.g., define the number of decimals displayed); and determine the primary display field.

1.1.5 Attribute joins and relates

Attributes may be stored in flat files where all the data is contained in one large table (e.g., a feature attribute table) or as a relational database. Data in a relational database are stored in multiple tables. These tables may be related to one another through unique identifiers or keys. Relational data structures reduce the level of duplication within a database. In addition, because each table is stored separately, it can be prepared, maintained and edited independently. Relational databases are also advantageous because they increase the efficiency of both data management and data processing – tables only need to be linked together when a specific query or analysis is required. In addition to these benefits, relational data structures can make it easier to exchange data – if only the attributes are changing a new version of an attribute table can be distributed without having to redistribute the spatial dataset.

There are four types of relationships possible between records within tables in a relational database:

- **one-to-one** – One record in the first table is related to one (and only one) in the second table. For example, the attribute data for a series of water quality sampling stations could be stored in a related table. There would be the same number of stations as records in the attribute table with a relationship established through a station identifier attribute common to both tables.
- **one-to-many** – One record in a table is related to many records in another table. For example, several houses may be located on the same street. The spatial feature representing the street would have a unique identifier (e.g., STREET_ID) and the house records would be related to the street via this identifier.
- **many-to-one** – Many records in a table may be related to one record in another table. For example, a land use dataset might have hundreds of land use polygons with twenty possible land use classes stored in a separate look-up table. In this example the spatial data would store a numeric code ranging from one to twenty to represent the polygon's land use. A relation between the two tables can be established based on this code. Using a one-to-many relationship in this instance is beneficial because it reduces the time required to enter the land use attributes – rather than having to type out the full land use class name each time the operator only has to enter the numeric code value. This both speeds up data entry and reduces the potential for errors (e.g., spelling mistakes) in the data.
- **many-to-many** – Many records in a table may be related to many records in another table. For example, many vegetable types might be grown on a single farm and these types of vegetables may be grown on more than one farm.

Attribute tables can be linked to the spatial layers through the join and relate functions. These are described below:

Join

A join involves appending the fields from one table to another through a relationship based on an attribute (e.g., an identifier) common to both tables. A join is usually used to attach additional attributes to a spatial data layer. Data can be joined in either a layer or a table view. While the names of the fields used to establish the join do not have to be identical, the data type does have to be the same. For example, strings are joined to strings and numeric fields to other numeric values. The join function is suitable when you are linking tables with one-to-one and many-to-one relationships. Joins are inappropriate with one-to-many relationships because only the first occurrence of a matching value is assigned to a record in the target table.

If you are in an edit session, be aware that only the columns from the source table can be changed. The data in the appended columns can not be edited. If new fields are added they are added to the target table with no effect to the join tables. The appended columns can be referenced when calculating values in the columns of the target table.

Relate

As mentioned above, joins are used to establish one-to-one or many-to-one relationships between layers and tables. The relate function can be used to establish one-to-many or a many-to-many relationships because relates are bidirectional. Related tables are connected but the tables are physically separate – data is accessed when you work with the layer's attributes. The properties of a relate (e.g., the fields and tables involved) are stored separately and when creating a relate you will be prompted to enter a relate name. The advantage of relates is that multiple files can be related, however, they can increase the time required to access and process the data.

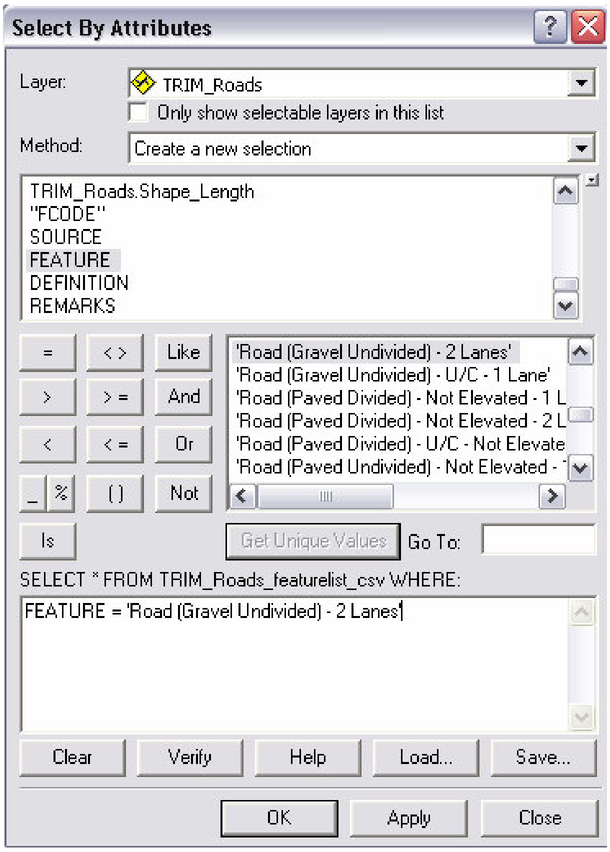
1.2 Querying and Selecting Vector Data

Querying GIS data is fundamental in retrieving pertinent data and discovering new spatial relationships. Queries are also often useful in reducing intricate or cumbersome datasets to smaller or simpler forms. They facilitate more complex interpretation or analysis. There are two methods of querying, or selecting data, that are typically available to create subset(s) of data: attribute (non-spatial); and spatial queries.

1.2.1 Select by attribute

Attribute queries are questions about the attributes (or non-spatial characteristics) of the data, for example, how many roads in a transportation layer are 2 lane gravel? Because attributes are actual information associated with features, the values stored in attribute tables often hold the most relevant answers to the questions raised in GIS analysis. Structured Query Language (SQL) is a standard interface that uses logical expressions to extract matching records (i.e. develop selection sets). The syntax for SQL queries varies between software packages, however, the following example illustrates a typical query as it would be performed in ArcMap. The example illustrates how to select a specific road type from a roads dataset.

- **Steps for selecting by attributes:**
    - From the Menu Bar in ArcMap, click *Selection* and choose *Select By Attributes*, or from an opened attribute table, click *Options* and choose *Select By Attributes*
    - In the Select By Attributes dialog box, click the *Layer* drop-down menu and choose the layer containing the features you want to select (Figure 7)
    - Click the *Method* drop-down menu and choose the selection method (see Section 5.2.3 for more information on selection methods)
    - Double-click (or type in the dialog box) the desired attribute field name
    - Click an operator button to add it to the SQL expression (e.g. = or >)
    - Click the *Get Unique Values* button to view the values for the selected field
    - Double-click on a value to add it to the SQL expression (or manually type the specific value you are looking for)
        - Note that depending upon the data type of the field you are querying, the syntax is slightly different.
        - When querying a text field, values to find are enclosed in single quotes (e.g., [FEATURE] = ‘Bridge’)
        - When querying a numeric field, values are not enclosed in any characters (e.g., [AREA] > 12.0)
    - Click *Apply* to run the query



Layer: TRIM_Roads

□ Only show selectable layers in this list

Method: Create a new selection

TRIM_Roads.Shape_Length
"FCODE"
SOURCE
FEATURE
DEFINITION
REMARKS

'Road (Gravel Undivided) - 2 Lanes'
'Road (Gravel Undivided) - U/C - 1 Lane'
'Road (Paved Divided) - Not Elevated - 1 L
'Road (Paved Divided) - Not Elevated - 2 L
'Road (Paved Divided) - U/C - Not Elevate
'Road (Paved Undivided) - Not Elevated - '

Go To:

SELECT * FROM TRIM_Roads_featurelist_csv WHERE:

FEATURE = 'Road (Gravel Undivided) - 2 Lanes'

Figure 7. Select by Attribute Dialog Box

- The number of features selected will be displayed in the lower left-hand corner of ArcMap. The features associated with the selected attributes will also be simultaneously highlighted in the map window.

Boolean Operators

Boolean (or logical) operators are used to set the conditions of the criteria, from which an evaluation of True or False is derived. Boolean operators include:

- (=) equal to
- (>) greater than
- (<) less than
- (>=) greater than or equal to
- (<=) less than or equal to
- (<>) not equal to

Boolean Connectors

Suppose you wanted to select features that satisfy two or more criteria, for example, how many 2 lane gravel roads were built in the year 2002? In this case, individual queries can be combined to answer a more complex questions. Boolean connectors (AND, OR, NOT, XOR) are used to combine these multi-part questions (Figure 8):

- AND – joins queries in order to satisfy two or more criteria
- OR – joins queries in order to satisfy either one criterion or the other (can be both)
- NOT – joins queries in order to satisfy one criterion, but not another
- XOR - joins queries in order to satisfy one criterion or the other, but not both (i.e. mutually exclusive)

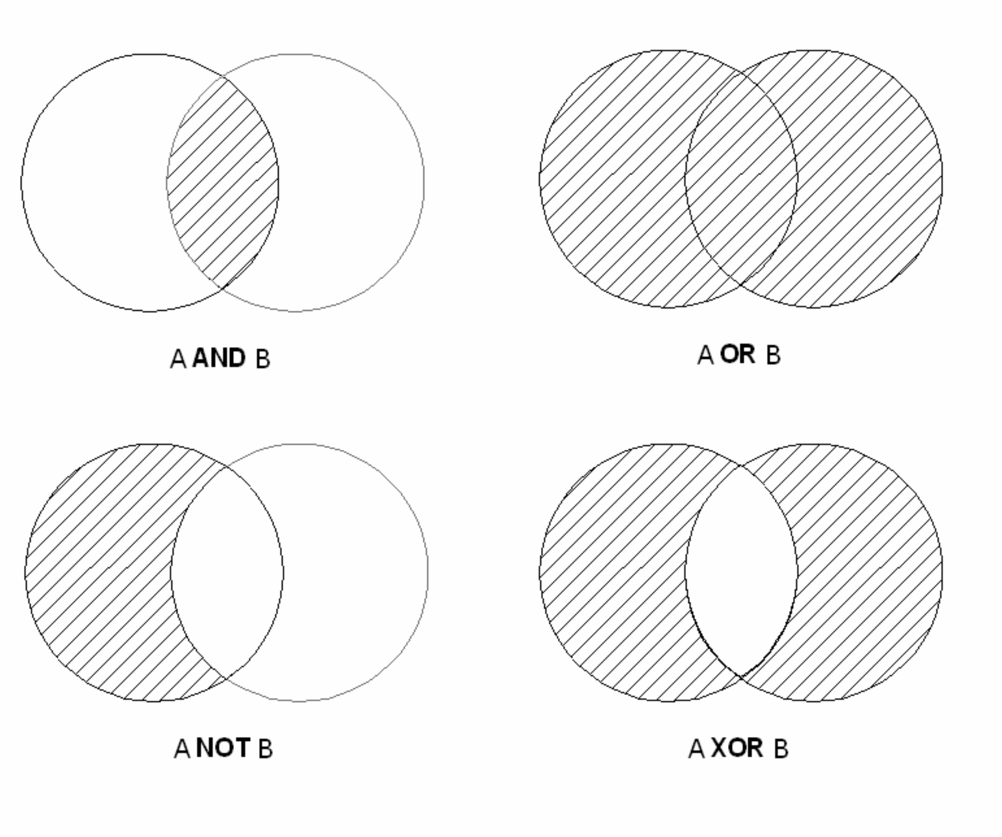
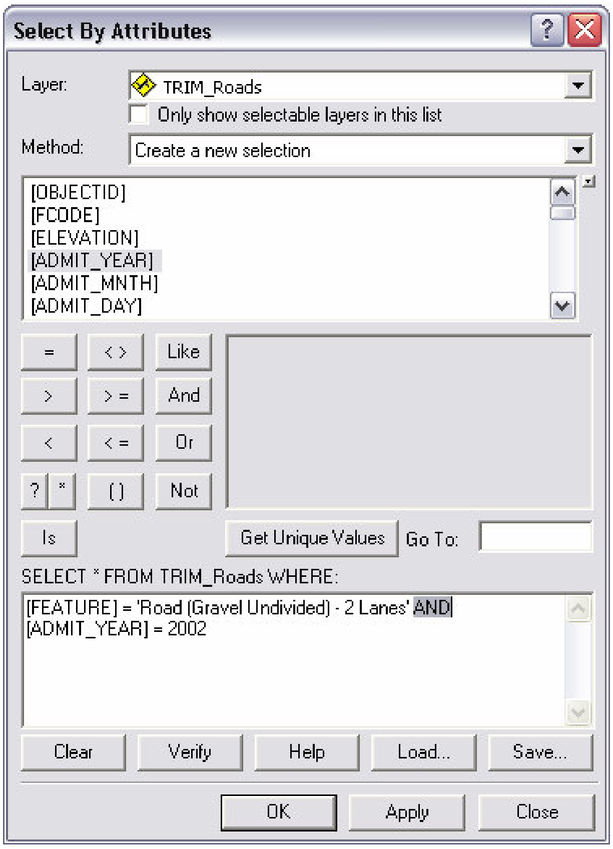


Figure 8. Boolean Connectors (Venn Diagrams)

Choosing the correct Boolean connector is important in order to correctly answer the question. Posing the question, "How many roads are 2 lane gravel AND were built in the year 2002?" will yield a different answer from the question "How many roads are 2 lane gravel OR were built in the year 2002?". The latter query will most likely identify more records because the answer only needs to satisfy one criterion, not both.

- **Steps for selecting by attributes (two or more criteria):**
  - In the Select By Attributes dialog box, double-click (or type in the dialog box) the desired attribute field name, operator, and the specific value you are looking for
  - Repeat the above step for the second criterion. Notice the two queries are combined with an *AND* Boolean connector (Figure 9).
  - Click *Apply* to run the query



Select By Attributes

Layer: TRIM_Roads

□ Only show selectable layers in this list

Method: Create a new selection

[OBJECTID]
[FCODE]
[ELEVATION]
[ADMIT_YEAR]
[ADMIT_MNTH]
[ADMIT_DAY]

Go To:

SELECT * FROM TRIM_Roads WHERE:

```
[FEATURE] = 'Road (Gravel Undivided) - 2 Lanes' AND
[ADMIT_YEAR] = 2002
```

Figure 9. Select by Attribute using a Boolean Connector

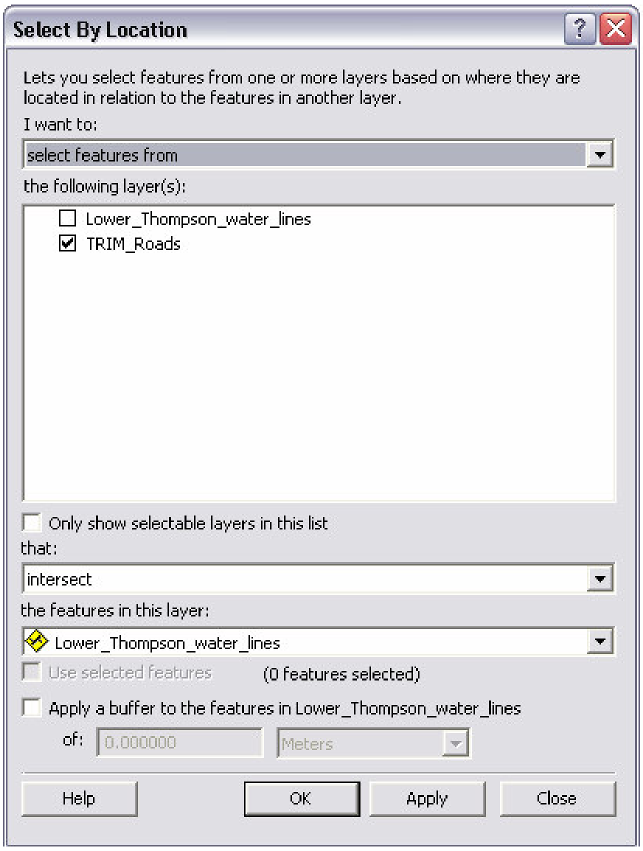
1.2.2 Spatial queries and relationships (select by location)

Spatial queries are questions about 'where' certain features exist in space and how they relate to other features. They allow you to select features based on their absolute position and/or location relative to other features in other layers. For example, a spatial query will allow you to select all the sampling sites falling inside a watershed. The simplest form of spatial query is to simply click on a position in the map view to select and/or inspect a feature.

- **Steps for selecting features by cursor:** The simplest way to query features is to click on one or, in order to capture several at a time, drag a box around an area of interest using a 'select features' pointer tool in ArcMap. To do this:
    - Ensure the target layer is checked in the Selection Tab at the bottom of the table of contents in ArcMap
    - From the Tools toolbar, click on the *Select Features* button
    - In the map window, click on an individual feature or drag a box around a group of features you wish to select
    - The features will simultaneously be selected and highlighted in the map window and attribute table

- **Steps for selecting by location:** More powerful than simply selecting features is a query that examines the spatial relationships between features in different layers (e.g., where are the roads that are crossed by streams?)

- From the Menu Bar in ArcMap, click *Selection* and choose *Select By Location*
- In the Select By Location window, click the drop-down arrow and choose a selection method (see Section 5.2.3 for more information on selection methods) (Figure 10).
- Check on the layer(s) you wish to select features from (the target layer).
- Choose from the drop-down menu, the type of spatial relationship you are looking for:
    - *Intersect* – returns any feature that geometrically shares a common part with the source feature
    - *Are within a distance of* – returns any feature within a specified buffer distance of the source feature
    - *Completely contain* – returns any feature that wholly contains a feature in the source layer
    - *Are completely within* – returns any feature wholly contained by the feature(s) in the source layer
    - *Have their centre in* – returns any feature with its centroid falling within the geometry of the source feature
    - *Share a line segment with* – returns any feature that has at least two adjacent vertices in common with the source feature
    - *Touch the boundary of* – returns any feature that touches the boundary of features in the source layer
    - *Are identical to* – returns any feature that is exactly equal (has identical vertices) to a feature in the source layer
    - *Are crossed by the outline of* – returns any feature that has at least one edge, vertex, or endpoint in common with a feature in the source layer
    - *Contain* - returns any feature that contains a feature in the source layer
    - *Are contained by* - returns any feature contained by the feature(s) in the source layer
- Choose from the bottom drop-down menu, the layer you want to relate to (the source layer)
- Click *Apply* to run the query
- As with the Select by Attributes, the selected features will be highlighted in the map window. Additionally, attributes associated with the selected features will be simultaneously selected and highlighted in the attribute table.



Select By Location

Lets you select features from one or more layers based on where they are located in relation to the features in another layer.

I want to: select features from

the following layer(s):

□ Lower_Thompson_water_lines
☑ TRIM_Roads

□ Only show selectable layers in this list

that: intersect

the features in this layer: Lower_Thompson_water_lines

□ Use selected features (0 features selected)

□ Apply a buffer to the features in Lower_Thompson_water_lines

of: 0.000000 Meters

Figure 10. Select by Location Dialog Box

1.2.3 Selection methods

A few options exist to create and modify attribute and spatial selection sets (Table 2). Though the method name differs between select by attribute and select by location, the output selection type is essentially the same:

Table 2. Selection Methods with Select by Attribute or Location

Select by attribute	**Select by location**	**Output**
create a new selection	select features from	creates a new selection set
add to current selection	add to the currently selected features in	adds records to a selection set generated from a previously executed query
remove from current selection	remove from the currently selected features in	removes records from a selection set generated from a previously executed query
select from current selection	select from the currently selected features in	selects a smaller subset of records from a selection set generated from a previously executed query

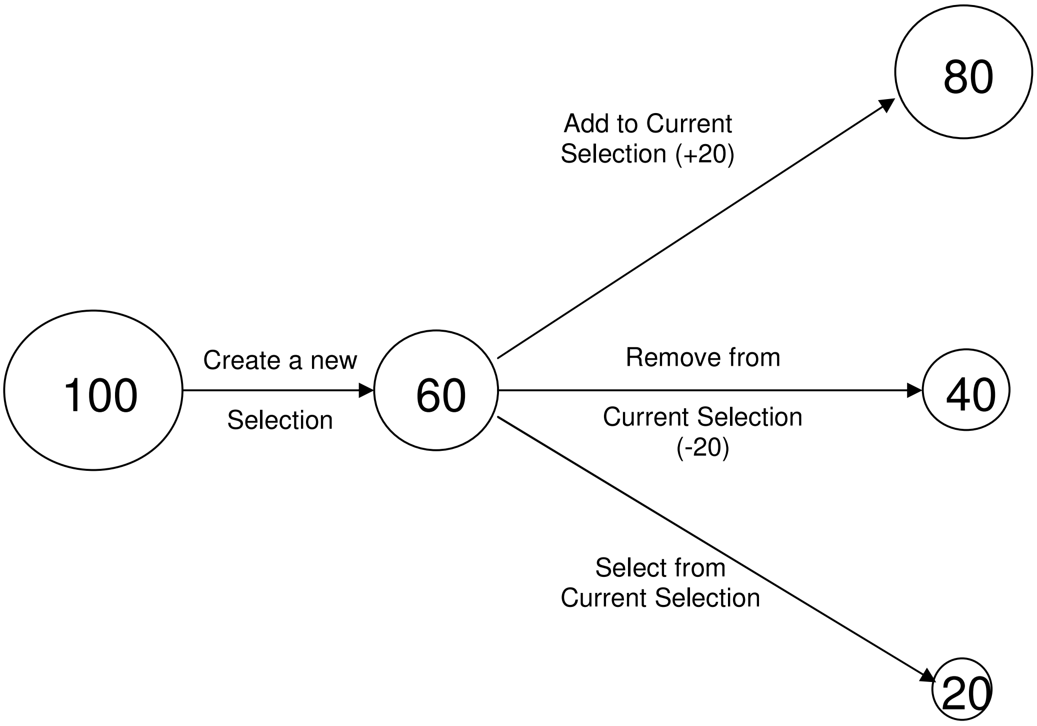

80
Add to Current Selection (+20)
100
Create a new Selection
60
Remove from Current Selection (-20)
40
Select from Current Selection
20

Figure 11. Selection Methods

After a selection set has been established, an operation can be performed to switch between the selected and unselected subsets. Note that it is important to keep track of the type of selection you are making to ensure your query generates the correct results.

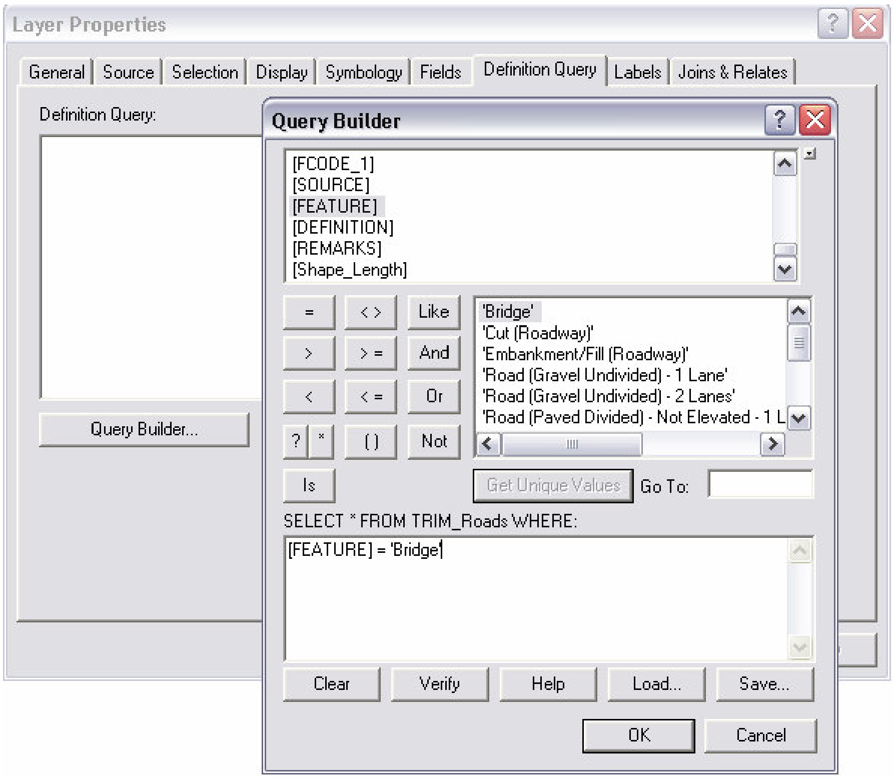
- **Steps for switching between selected and unselected subsets:**
    - From an open attribute table, click *Options* and choose *Switch Selection*. All previously selected records are now unselected, while the previously unselected records become selected.
    - To clear the selection and restore all records to unselected, click *Options* and choose *Clear Selection*

1.2.4 Definition queries

Definition queries allow you to display features in the map window (and associated attribute tables) having specific attributes. For example, as part of an inspection schedule, an engineer may wish to display and examine only bridges within a transportation layer, rather than having to view all transportation features.

- **Steps for adding a definition query:**
    - Right-click the desired layer in the table of contents and choose *Properties*.
    - Click the Definition Query tab.
    - Type an SQL expression or click *Query Builder* (Figure 12).
        - Double-click (or type in the dialog box) the desired attribute field name

    - Note that field names are delimited differently depending upon the source format of the feature class.
    - When querying a feature class from a file geodatabase, shapefile or coverage, enclose field names in double quotes (e.g., “AREA”)
    - When querying a personal geodatabase feature class, enclose field names in square brackets (e.g., [AREA])
    - When querying an SDE feature class, fields are not enclosed at all (e.g., AREA)
  - Click an operator button to add it to the SQL expression (e.g. =)
  - Click *Get Unique Values* to view the values for the selected field
  - Double-click on a value to add it to the SQL expression (or manually type the specific value you are looking for)
    - Note that depending upon the data type of the field you are querying, the syntax is slightly different.
    - When querying a text field, values to find are enclosed in single quotes (e.g., [FEATURE] = ‘Bridge’)
    - When querying a numeric field, values are not enclosed in any characters (e.g., [AREA] > 12.0)
  - Click *OK*



[FCODE_1]
[SOURCE]
[FEATURE]
[DEFINITION]
[REMARKS]
[Shape_Length]

'Bridge'
'Cut (Roadway)'
'Embankment/Fill (Roadway)'
'Road (Gravel Undivided) - 1 Lane'
'Road (Gravel Undivided) - 2 Lanes'
'Road (Paved Divided) - Not Elevated - 1 L

SELECT * FROM TRIM_Roads WHERE:

[FEATURE] = 'Bridge'

Figure 12. Definition Query Dialog Box

- Click *Apply* to activate the query.
- To display all the features again, delete the query

1.2.5 Viewing selection sets

Once a query has been used to create a selection set, it is then valuable to view the newly created subset of data. Features selected by attribute or location appear highlighted (in cyan) in the map window (Figure 13).

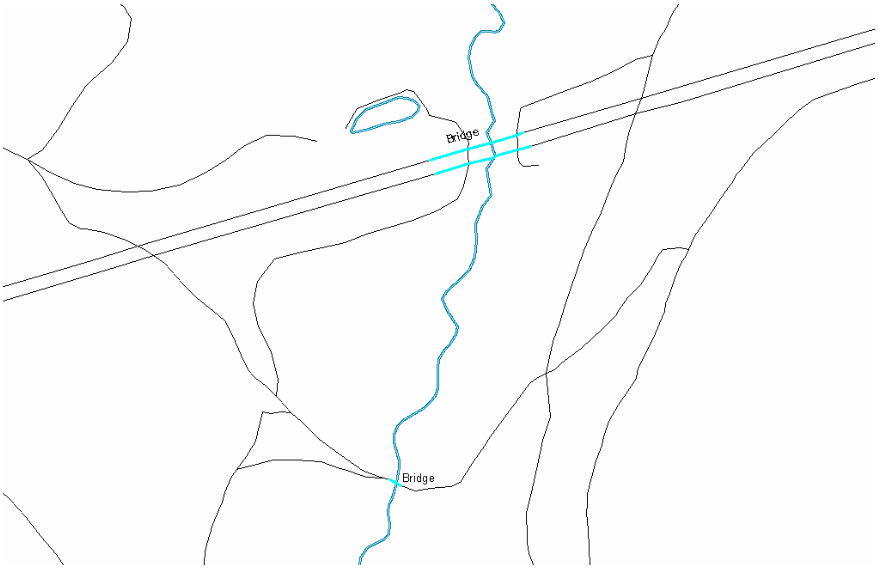
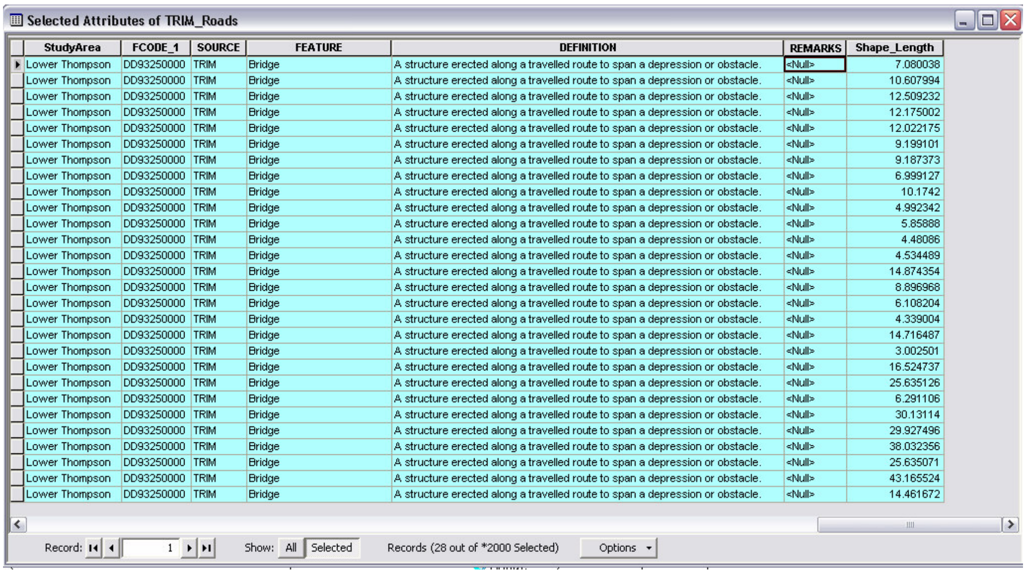


Figure 13. Viewing a Selection Set in a Map Window

It is also possible to view and subsequently analyse selected non-spatial data in an associated attribute table. These records will be highlighted in the attribute table but you can also view just the selected records.

- **Steps for viewing a selection set in an attribute table:**
    - Right-click a layer in the table of contents and choose Open Attribute Table
    - Click on the *Selected* button at the bottom of the attribute table to display only the selected features (records) (Figure 14).



StudyArea	FCODE_1	SOURCE	FEATURE	DEFINITION	REMARKS	Shape_Length
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	7.080038
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	10.607994
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	12.509232
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	12.175002
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	12.022175
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	9.199101
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	9.187373
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	6.999127
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	10.1742
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	4.992342
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	5.85888
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	4.48086
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	4.534489
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	14.874354
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	8.896968
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	6.108204
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	4.339004
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	14.716487
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	3.002501
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	16.524737
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	25.635126
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	6.291106
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	30.13114
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	29.927496
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	38.032356
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	25.635071
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	43.165524
Lower Thompson	DD93250000	TRIM	Bridge	A structure erected along a travelled route to span a depression or obstacle.	<Null>	14.461672

Records (28 out of *2000 Selected)

Figure 14. Viewing a Selection Set in an Attribute Table

1.2.6 Copying (extracting) selected data

To perform more detailed analyses on a portion of a dataset it may be beneficial to permanently isolate a subset of data from the original dataset. This is achieved by extracting (or exporting) the selected features to a new layer or attributes to a new table. It is also often useful to copy selected data to another existing layer or table in order to merge a subset of the original file with another dataset.

- **Steps for exporting selected attributes to a new table:**
    - Open the attribute table for the feature class you will query
    - Create a selection of records in a table, either by selecting features manually or by using the *Select by Attributes* function
    - Click *Options* and choose *Export*
    - Click the Export drop-down menu and choose *Selected Records*
    - Click the *Browse* button and navigate to the existing folder or geodatabase you wish to export to
    - Choose the type of table you wish to export to from the Save as type drop-down menu (e.g. geodatabase table, dbase, text file)
    - Type a name for the new table to be created and click *Save* and then *OK*
    - The newly created table can then be viewed and the data analysed in other programs such as MS Excel or Access. This is useful for summarizing and interpreting tabular data.

- **Steps for exporting selected features to a new layer:**
    - Create a selection of features
    - In the table of contents, right-click the layer with the selected features and choose *Data→ Export Data*
    - Specify the desired coordinate system using the radio buttons

  - Click the *Browse* button and navigate to the desired geodatabase or folder
  - Choose the type of featureclass you wish to export to from the Save as type drop-down menu (e.g. geodatabase featureclass or shapefile)
  - Type a name for the new layer to be created and click *Save* and then *OK*
  - The newly created featureclass can then be added to ArcMap for further analysis

- **Steps for copying features from one layer to another existing layer:**
  - Start Editing in ArcMap by clicking *Editor* and choosing *Start Editing* – ensure the correct layer is set as the target layer
  - Create a selection of features from the source layer
  - Click the *Copy* button on the toolbar
  - Click the Paste button on the toolbar
  - The selected feature from the source layer will be copied to the target layer
  - Click *Editor* and choose *Save Edits* and then *Stop Editing*

1.3 *Querying and Selecting Raster Data*

As with vector layers, raster data can be queried by attribute or spatial location in order to extract subsets of information for more focused analysis.

1.3.1 Extracting raster data

Select by Attribute

Some raster datasets contain attribute tables. This might be the case where a satellite image has been classified to create a raster definition of land use, for example. Where a raster attribute table exists, users may select cells of the raster using the *Select by Attribute* dialog, using a similar method as that used for vector data. In the raster world, this function might be useful in identifying and summarizing the area covered by lakes in a land cover dataset (Figure 15).

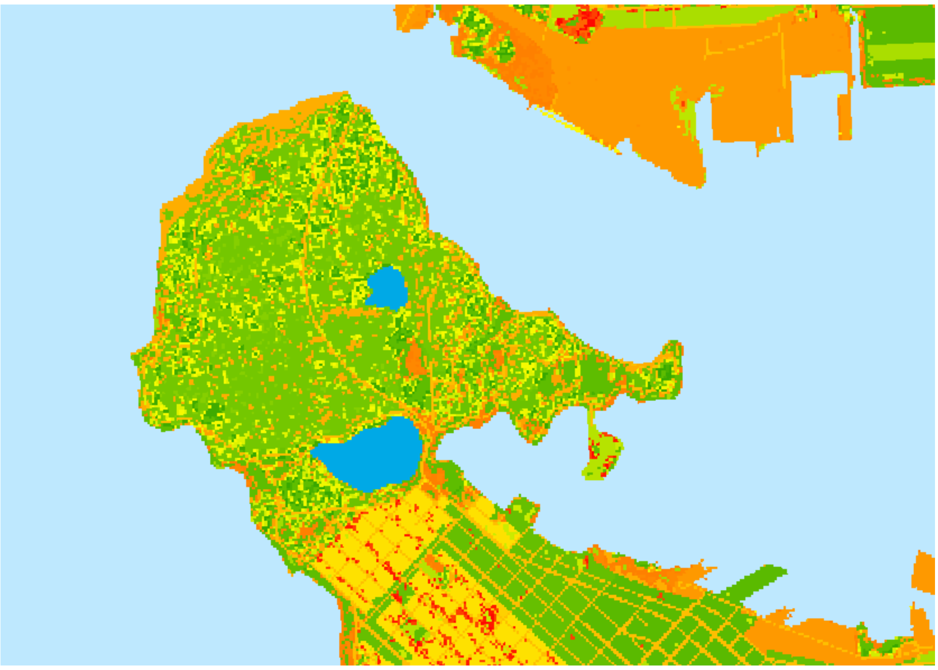


Figure 15. Select by Raster Attribute

Area = # of cells with Land Cover = 'Lake' x cell area
= 53,619 x 225m^2 = 12,064,275 m^2 (1,206 ha)

- **Steps for selecting by attributes:**
  - From an open raster attribute table, click *Options* and choose *Select By Attributes*
  - In the Select By Attributes window, click the *Method* drop-down menu and choose the selection method (see Section 5.2.3 for more information on selection methods)
  - Double-click (or type in the dialog box) the desired attribute field name (with rasters, this will normally be the coded 'value' field which represents a real world property such as elevation or land-use)
  - Click an operator button to add it to the SQL expression (e.g., = or >)
  - Click *Get Unique Values* to view the values for the selected field

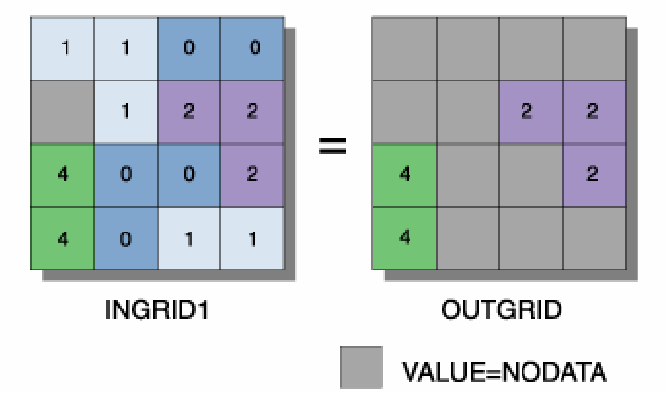
- Double-click on a value to add it to the SQL expression (or manually type the specific value you are looking for - in single quotations)
- Click *Apply* to run the query

Extract by Attribute

Extracting raster cells by an attribute is accomplished by executing a ‘where’ clause query which results in a new raster. A ‘where’ clause is a conditional statement that establishes the desired criteria of the query. Consider a wildlife biologist modelling goat habitat, where the animals are only found on steep slopes. We can extract all cells from a slope raster with a value greater than or equal to 20% to easily identify potential goat habitat. In this case the where clause would be ‘[SLOPE] >= 20’ (the square brackets [ ] surrounding the slope raster is normal syntax for rasters in ArcMap).

- **Steps for extracting by attributes (Extract by Attributes Tool):**
  - Open ArcToolbox by clicking on the red toolbox button in the toolbar
  - Choose *Spatial Analyst Tools* → *Extraction* → *Extract by Attributes*
  - Create a ‘where’ clause statement to extract the cell values you wish to isolate
  - Cells in the output raster that satisfy the criteria retain their cell value from the original dataset, while cells not meeting the specified query are assigned NoData (Figure 16).

Select(SlopeGrid, ‘SLOPE >= 2’)*



1	1	0	0
	1	2	2
4	0	0	2
4	0	1	1

=

		2	2
4			2
4			

*Slope class 2 = 20% and 4 = 40%.

Figure 16. Extract by Raster Attribute

Like vector queries, raster queries can use Boolean connectors (AND, OR, NOT) to combine two or more criteria (or rasters) into one logical expression. Typically these compound expressions work to combine multiple raster datasets and create an output raster. For example, the goat habitat model from above might be modified to include an elevation raster, where only areas with slopes greater than or equal to 20% AND elevations higher than 1000 metres constitute suitable goat habitat. In this case the statement would be ‘([SLOPE] >= 20) AND ([ELEV] > 1000).

Select by Location

Selecting by location in the raster world allows the user to extract cells based on location and investigate spatial relationships between disparate layers. Suppose a city planner wanted to summarise the area of forested land cover in an urban area, but only within certain municipal parks. GIS can aid in this investigation by using a park polygon layer to extract forested areas from a city-wide land use raster. Restricting the data to the parkland would facilitate summarising the data according to the requirements of the city planner.

There are a number of ways to extract raster cells by their location, the majority of which use a specified geometric shape to exclude or include individual or groups of cells in a raster dataset.

Extract by Mask – allows the user to extract raster data with a polygon feature class. The polygon is applied as a mask and only the raster cells falling inside the mask are processed.

Extract by Shapes

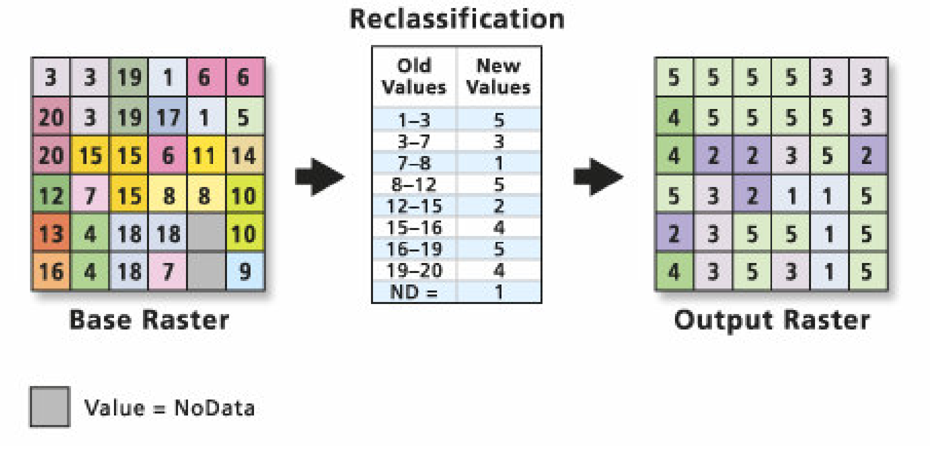
> Extract by Points – uses a list of coordinate values (x, y values representing points) to output only the cells of a raster situated at these point locations.
> Extract by Circle – uses the centre coordinate and radius of a circle to output the cells of a raster situated inside (or outside) of the circle.
> Extract by Polygon/Rectangle – uses a list of coordinate values defining an area to output the cells of a raster situated either inside or outside of the area.

1.3.2 Reclassification

Raster reclassification (also known as recoding or transforming) allows you to simplify or aggregate data within a raster dataset. For example, if you have a dataset with 10 tree species values and you want to group all tree species into a single class, the reclassification function will allow you to do this.

- **Steps for reclassifying a raster dataset:**
    - Activate the Spatial Analyst extension by clicking on *Tools* and choosing *Extensions*. Place a check next to Spatial Analyst and click *Close*.
    - Add the Spatial Analyst toolbar by right clicking anywhere in the toolbar area and choosing *Spatial Analyst* from the list
    - Choose the target raster from the *Layer* drop-down menu (e.g., LandUse)
    - Click on *Spatial Analyst* and choose *Reclassify…*
    - In the Reclassification dialogue box, enter new values for output raster in the right-hand column
    - Click *OK* to execute the reclassification

An example of grouping cells with common characteristics to create a simplified raster (e.g., combining continuous slope values into fewer slope classes [0-10% = 1, 11-20% = 2, etc.]) is provided in Figure 17.



3	3	19	1	6	6
20	3	19	17	1	5
20	15	15	6	11	14
12	7	15	8	8	10
13	4	18	18		10
16	4	18	7		9

Old Values	New Values
1–3	5
3–7	3
7–8	1
8–12	5
12–15	2
15–16	4
16–19	5
19–20	4
ND =	1

5	5	5	5	3	3
4	5	5	5	5	3
4	2	2	3	5	2
5	3	2	1	1	5
2	3	5	5	1	5
4	3	5	3	1	5

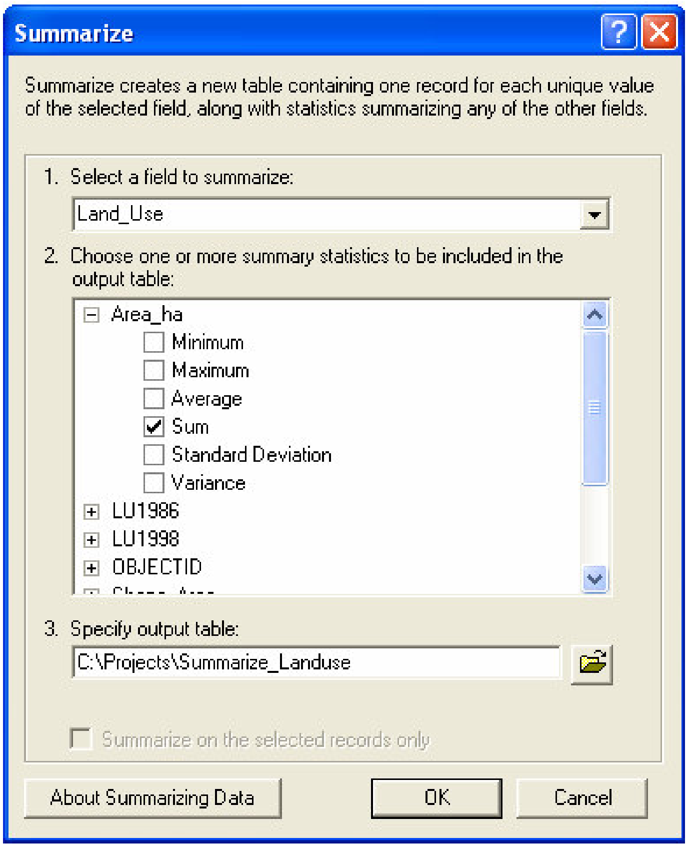
Figure 17. Reclassification by Grouping Values

1.4 Summarizing and Interpreting Data

1.4.1 Summarizing data

The summarize function allows you to classify data based on an attribute. In other words, you can organize data in different ways based on specific requirements. Summarize allows you to generate summary statistics (e.g., counts, average, minimum and maximum values) for your data. For example, a land use dataset consisting of hundreds of polygons with ten potential land use classes can be summarized to tell you the total area of each class. In this example, a new table would be created containing a column with each of the land uses listed and an area column containing the sum of the area of all the polygons falling within that class.

To summarize data in ArcMap open an attribute table view and then right-click on the name of the field you want to summarize. Select (by left-clicking) on the summarize function. This will present the summarize dialog box (Figure 18). The name of the field you selected to summarize will appear in the Select field to summarize step. If you want to select a different field just pull down the list and select one of the field names. The second step in the dialog box prompts you to select the summary statistics to be included in the output table. In the land use example we want to sum the total area of each land use class so you would tick (by left-clicking) on the Sum check box. The final step is to specify the name of the output table. If you have any records selected (e.g., the results of a query) the summarize function gives you the option of summarizing all the records or only the selected records. When you have finished making the selection, click on the OK button and then click Yes when prompted to add the new table.



Summarize

Summarize creates a new table containing one record for each unique value of the selected field, along with statistics summarizing any of the other fields.

1. Select a field to summarize: Land_Use

2. Choose one or more summary statistics to be included in the output table:
   - ⊟ Area_ha
     - □ Minimum
     - □ Maximum
     - □ Average
     - ☑ Sum
     - □ Standard Deviation
     - □ Variance
   - ⊞ LU1986
   - ⊞ LU1998
   - ⊞ OBJECTID

3. Specify output table: C:\Projects\Summarize_Landuse

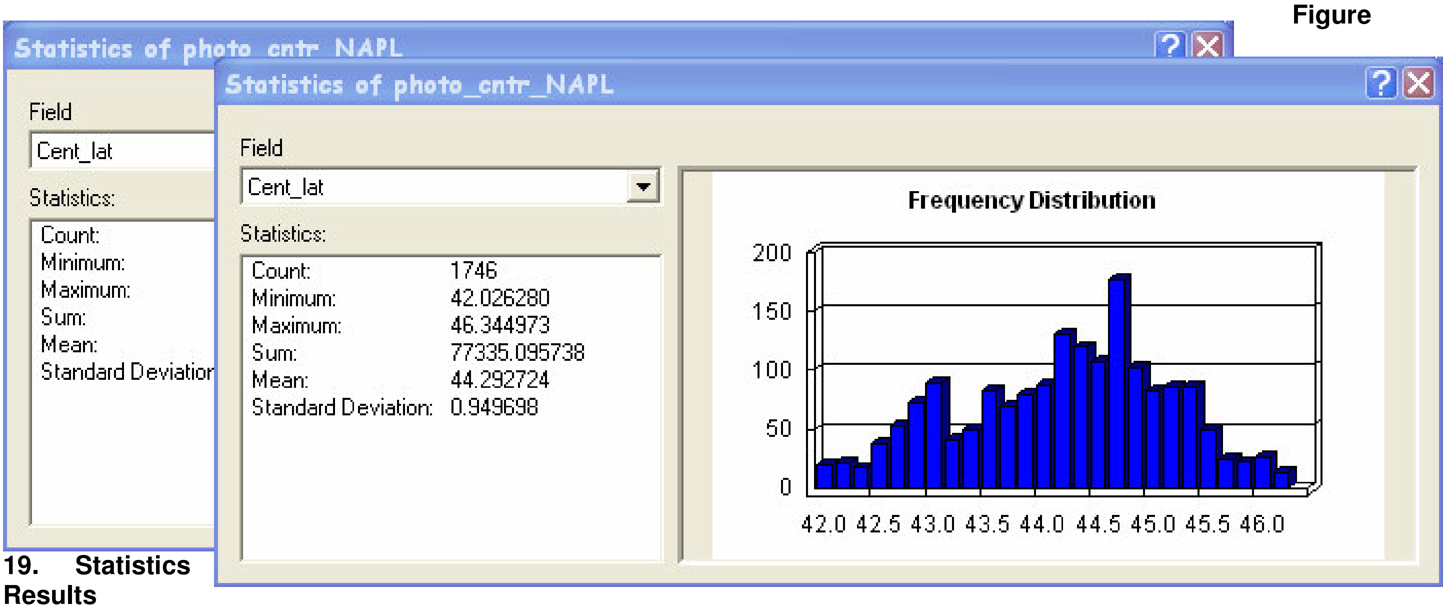
□ Summarize on the selected records only

Figure 18. Summarize Dialog Box

1.4.2 Statistics

A GIS will allow you to calculate statistics describing the contents of numeric fields. You can view a count of the number of records, the minimum and maximum values for the field, the sum of the values, a mean and the standard deviation. In addition, many interfaces will display a histogram providing a frequency distribution for the values in the field. The dialog box will allow you to generate statistics for other field in the table by pulling down the field dropdown list and selecting another field name.

To calculate statistics for a field in ArcMap open an attribute table view and then right-click on the name of the field you want to describe. Selecting Statistics from the list will display a dialog box (Figure 19). Note that statistics can only be generated for numeric type fields.



Statistics of photo_cntr_NAPL

Field

Cent_lat

Statistics:

Count: 1746
Minimum: 42.026280
Maximum: 46.344973
Sum: 77335.095738
Mean: 44.292724
Standard Deviation: 0.949698

Statistics of photo_cntr_NAPL

Field

Cent_lat

Statistics:

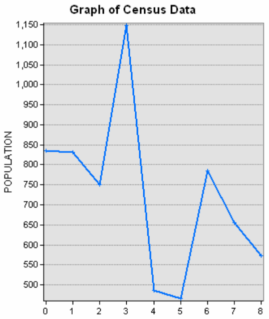
Count:
Minimum:
Maximum:
Sum:
Mean:
Standard Deviatior

Figure 19. Statistics Results Example

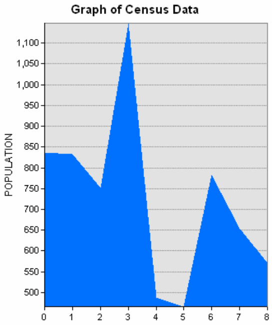
1.4.3 Graphs

Graphs allow you to examine and summarize the data in a format that is often easier to understand than tabular data because they allow you to visualize the data in different ways. In map layouts they can be used to show additional information related to information on the map or present the same information in a different way. Using our land use example, we could generate a map where each land use class is displayed in a unique colour and then include a graph that provides a summary of the total area of each land use class. The map window and the graph would compliment one another, giving the reader more details concerning the information being displayed. There are a number of different types of graphs (both two- and three-dimensional) to choose from. Each graph type has display properties that can be adjusted based on the type of data you are displaying and the way you want to present the data. Various types of graphs available are described below, with some simple examples of common graph types:

- **Line** – Line graphs display data in lines on an x, y grid. One or more lines may be in the graph. Line graphs are useful for displaying trends in data along a continuous scale. Changes in population rates or gross domestic product (GDP) over a series of years would be displayed effectively using this type of graph.

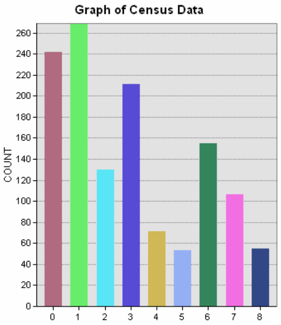


- **Polar** – A polar graph is similar to a line graph but it displays angular data (in degrees or radians) on a circular grid. They are useful for displaying the results of mathematical formulas.

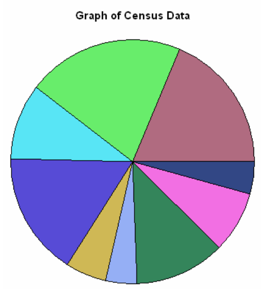


- **Area** – An area graph is similar to a line graph in that one or more lines are displayed on and x, y grid and are also useful for showing trends in values. The shading in area graphs can more effectively emphasize differences in quantities

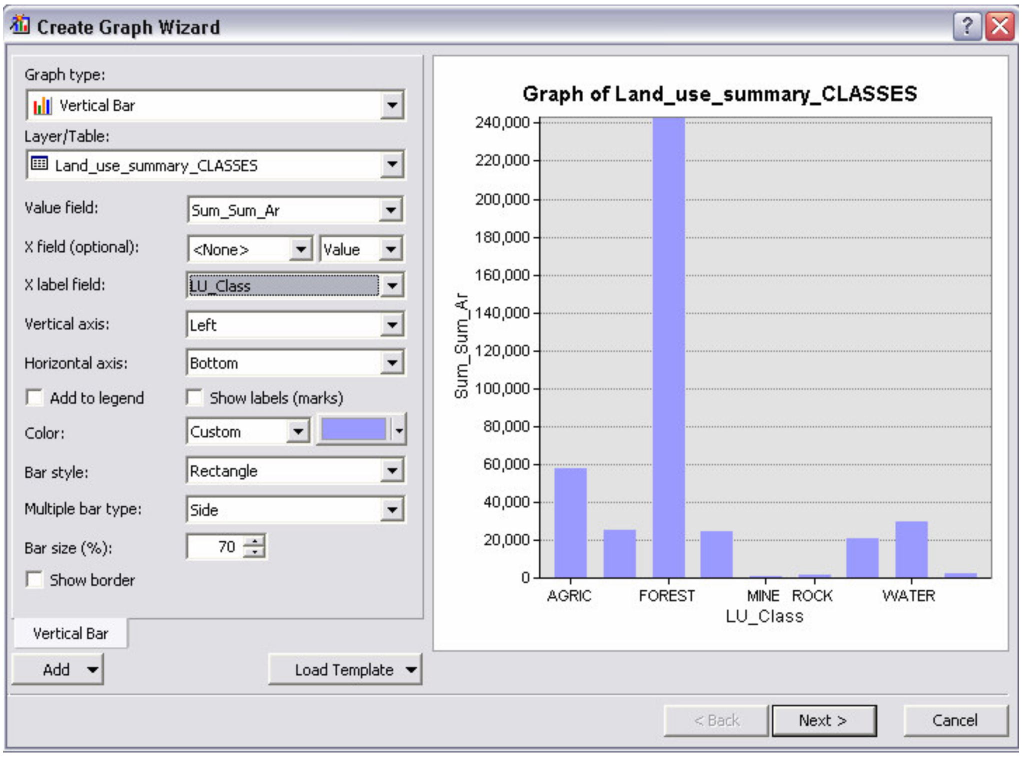
- **Scatter** – A scatter plot uses symbols (e.g., crosses) to plot x, y (and potentially z) values based on the attributes in the dataset. Scatter plots allow multiple variables to be displayed effectively and may allow associations between variables to be examined more effectively. For example, if we wanted to examine the potential association between annual income rates and life expectancy these two variables could be represented using a scatter plot.
- **Bubble** – Bubble graphs are similar to scatter plots but they allow you to plot three variables in two dimensions. Rather than using uniformly sized symbols a bubble plot uses symbols (or bubbles) that are proportionally sized to portray the values associated with the third variable. Using the scatter graph example of comparing income to life expectancy we could compare the third value of weight to examine potential correlations between the variables.
- **Bar and Column** – Bar or column charts are also referred to as histograms. They group data into equal intervals (represented as classes) and use either bars or columns to depict the number or frequency of values in each class. These types of graphs are useful for showing trends in values (e.g., monthly temperature or precipitation values).



- **High-Low-Close** – A high-low-close graph displays a range of y values (as a vertical bar) at each x value. Horizontal crossbars are placed on the vertical bar to represent highs and lows in the data. This type of graph can be used to depict fluctuations in the values of stocks over the course of the day – plotting the opening, high, low and closing prices.
- **Pie Chart** – Pie charts can be two- or three-dimensional. They display data in a circle, or pie, where wedges represent different proportions or ratios in the data. The proportion of different land use classes could be effectively displayed using a pie chart.



When creating a graph you should determine the variable(s) you want to graph and then select a graph type that will effectively display the data. To create a graph in ArcMap and add it to a layout select the *Graph* option from the Tools menu and click on *Create*. This will display the first page of the Graph Wizard where you will be prompted to select a graph type (Figure 20). The example that follows shows how to generate a column graph depicting land cover classes



Graph type: Vertical Bar

Layer/Table: Land_use_summary_CLASSES

Value field: Sum_Sum_Ar

X field (optional): <None> Value

X label field: LU_Class

Vertical axis: Left

Horizontal axis: Bottom

□ Add to legend □ Show labels (marks)

Color: Custom

Bar style: Rectangle

Multiple bar type: Side

Bar size (%): 70

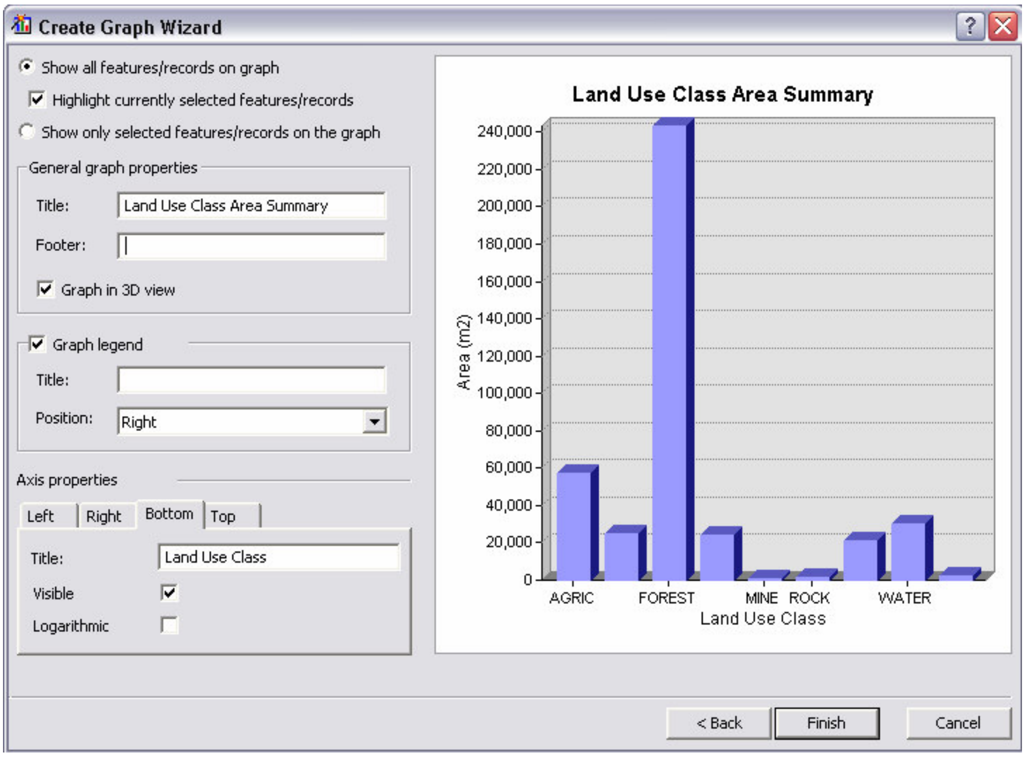
□ Show border

Figure 20. Graph Wizard Step 1 of 2

Use the graph wizard to select the appropriate layer (or table) and one of the styles of column graphs. Once you have decided upon the graph type/style and the data fields, click on the Next button to progress to the second screen in the wizard (Figure 21). You can specify the general graph properties such as title names and x and y axis labels. By clicking the check boxes, you can instruct the software to use either the all records in the file or only the selected records.

- **Steps for creating a graph:**

1. Under the *Graph type* drop-down menu, choose the type of graph you want to create
2. Under the *Layer/Table* drop-down menu, choose the layer or table you want to graph
3. Choose the data field you wish to graph from the *Value field* drop-down menu and specify an x field if desired
4. Click the check box if you wish to create a legend to accompany the graph
5. Change the bar style and colour if desired
6. Click Next
7. Select the radio button next to all features or selected features
8. Enter a title for the graph
9. Click the Tabs under Axis properties to adjust the visibility and title names for each axis
10. Click Finish



Create Graph Wizard

● Show all features/records on graph

☑ Highlight currently selected features/records

○ Show only selected features/records on the graph

General graph properties

Title: Land Use Class Area Summary

Footer:

☑ Graph in 3D view

☑ Graph legend

Title:

Position: Right

Axis properties

Title: Land Use Class

Visible ☑

Logarithmic □

Figure 21. Graph Wizard Step 2 of 2

Self-Study Questions

1. What are the advantages of the raster data model over the vector data model?

2. What data type (e.g., text, long integer, etc.) would be most appropriate to store each of the following attributes of a land parcel feature class?

ParcelOwner	(name of the person owning the land)
ParcelArea	(area, in square metres, of the parcel)
PurchaseDate	(self-explanitory)
NumStructures	(the number of buildings on the parcel)

3. A land use attribute (e.g., agriculture, forest, urban, etc.) is an example of which level of measurement (nominal, ordinal, interval, ratio or cyclic)?

4. Describe the difference between a Join and a Relate operation in ArcMap.

5. Describe a query which would select all LandParcels with a *LandUse* classification of Agricultural and an *AreaHa* (area measured in hectares) larger than 10 hectares.

References

1. Chang, K.T. *Introduction to Geographic Information Systems*. McGraw-Hill, 2006.

2. Chrisman, N. *Exploring Geographic Information Systems*. John Wiley and Sons, 1997.

3. Heywood, I., Cornelius, S. and Carver, S. An *Introduction to Geographical Information Systems*. Pearson Education, 2002,

2 Basic Spatial Analysis

Spatial analysis in GIS allows us to turn data into information and create new data (derivative datasets) by manipulating existing spatial features and their related attributes. GIS packages are equipped with a variety of analysis functions that allow us to manipulate both vector and raster data formats. These functions can be thought of as a set of tools for spatial analysis, and in fact several GIS applications use this “toolbox” analogy in describing the geoprocessing functions available.

Tasks performed by a GIS analyst will typically involve making use of several of these analytical tools. For example, a simple analytical problem might be to determine the amount of agricultural activity within 500m of streams, perhaps as a means of quantifying riparian disturbance, or water quality degradation. To answer such a question, an analyst must buffer the streams by 500m, overlay the buffers with agriculture land use polygons, and then quantify the resulting intersection. This module seeks to enumerate and define many of the common analysis tools which are available in most commercial GIS applications, which can then be combined to resolve specific analytical problems.

For organisational reasons, these functions are divided into the following topics:

Topic 1: Vector analysis methods
Topic 2: Raster analysis methods
Topic 3: Generalizing data

2.1 Topic 1: Vector Analysis Methods

2.1.1 Extraction

Extracting portions of data is an effective means of isolating specific areas for further processing or data analysis. Similar to queries and selection sets, extraction functions can be used to reduce the size of datasets and/or facilitate more complex interpretation. While the development of queries and selection sets also will allow you to isolate portions of a dataset, extraction techniques differ in that these portions of data are isolated in a permanent way - through the creation of new data layers. GIS software packages provide a suite of tools to extract data, the most useful being, clip, select, split and erase.

Clip

Working much like a cookie-cutter, clip allows you to intersect two feature layers to extract a portion of a dataset (the input layer) based on the spatial extent of another dataset (the clip layer). The clip function creates a new data layer (output) consisting of the features of the input layer that fall within the extent of the clip layer (Figure 1).

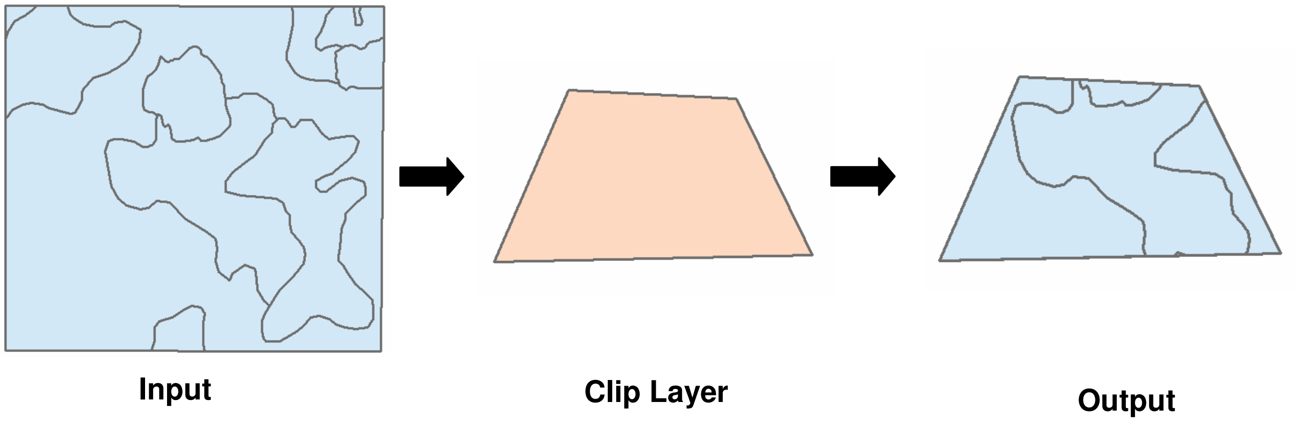


Figure 1. Clip Example

Clip is useful for developing a subset of features from a series of existing data layers to match a common boundary, for example the boundary of a study area or a jurisdictional boundary (e.g., a province, county, state, or municipal boundary). For example, an urban planner might wish to look at a street network layer, but only those streets falling within a certain municipal boundary. Clipping would be useful in order to permanently extract the street features matching the extent of the municipal boundary.

The input layer to be clipped may contain points, lines or polygons; however, because the element of area is required, the clip layer must be a polygon. The field names and attributes of the features in the output layer's table are maintained (i.e., they are identical to those of the input table). One potential exception to this rule are area, length and perimeter fields, which, depending on the software and/or data format being used, may or may not automatically recalculate. The values of any features intersected by the clip boundary may require updating to reflect the change in area.

Split

Split is used to divide an input layer into two or more independent layers, based on geographically corresponding features in a split layer. Similar to the clip function, the input layer may consist of point, line or polygon features, however, the split layer must be a polygon to define the areal extent of the analysis. The features in the input layer are broken up along the boundaries of the split layer features as illustrated in Figure 2.

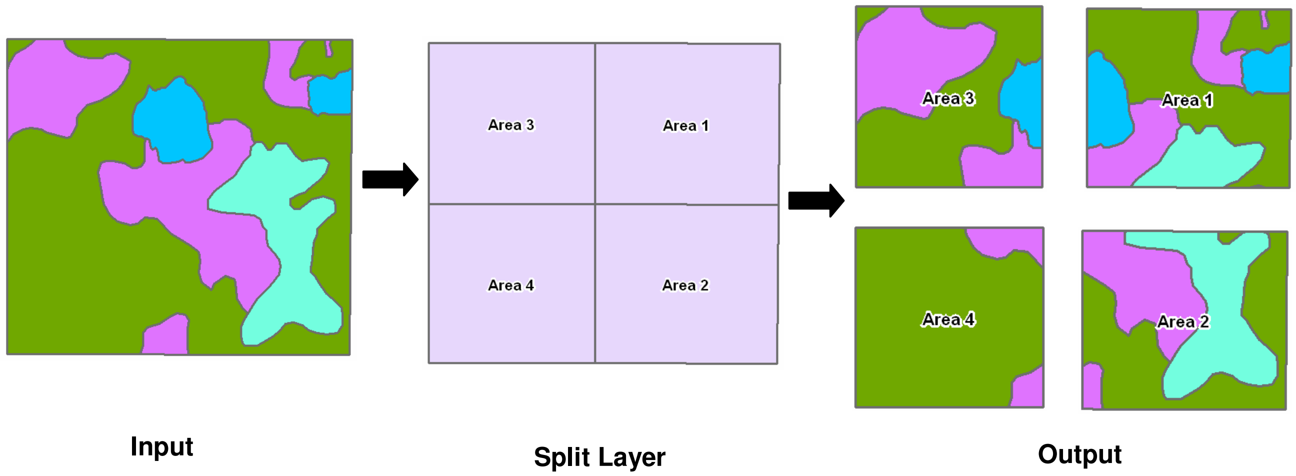


Figure 2. Spilt Example

Splitting is essentially a means of simultaneously executing a series of clips along boundaries defined by the split layer. The number of output layers is dictated by the number of split features (i.e., if the split file contains four polygons, the input file will be split into four separate files). Each resulting output layer will be named with the unique attribute value present within the selected field from the split layer. As with clipping, the field names and attributes of the input table are maintained in the output layers. The split function would be useful for dividing a large coverage into jurisdictional areas, for example, the zoning data associated with a city could be divided based on municipal boundaries or a national map series could be developed by dividing topographic data based on a defined grid.

Select

The Select tool may be used to create a new layer containing features extracted from an input layer. This is achieved through the execution of a user-defined query expression to select a subset of the data; these selected features are permanently extracted to a new output layer (Figure 3).

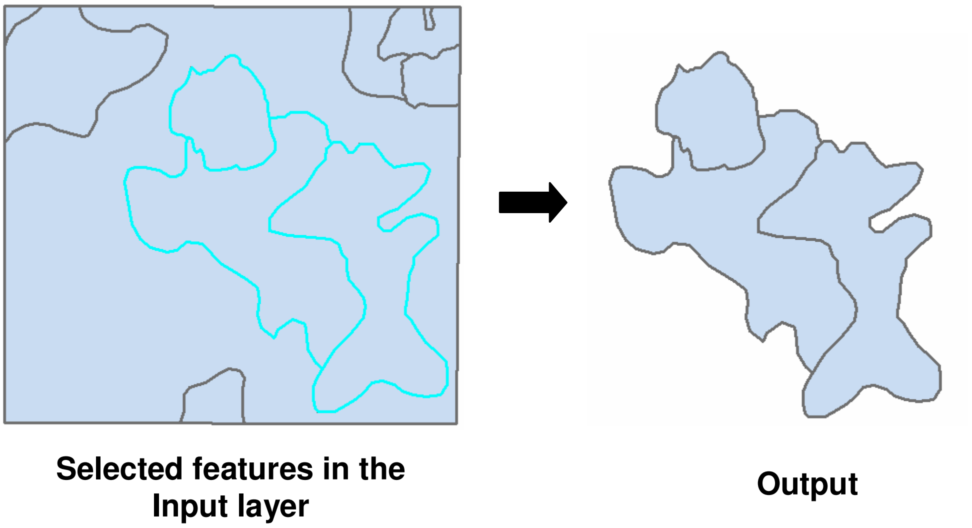


Figure 3. Select Example

To build on the example above, the urban planner might wish to look at only double-line streets in the particular municipality of interest. In this case, he or she would execute a selection query to extract only those desired features to a new layer.

Erase

Erase creates a new output layer by discarding features from the input layer that fall within the area extent of the erase (overlay) layer (Figure 4). The input layer can be points, lines, or polygons; however, because the element of area is required, the erase layer must be a polygon. Features in the output layer will be of the same geometry type as the features in the input layer. Examples of how the erase function could be used include:

- In a map layout, erase can be used to develop a mask to allow only those features falling within a given area (e.g., a study area boundary) to be displayed.
- In a suitability analysis, erase could be used to apply suitability rules. For example, if potential sites have to have a 200 metre setback from wetlands then wetland features can be buffered by 200 metres and the buffer polygon used as the erase layer to remove potential sites falling within this zone from consideration.

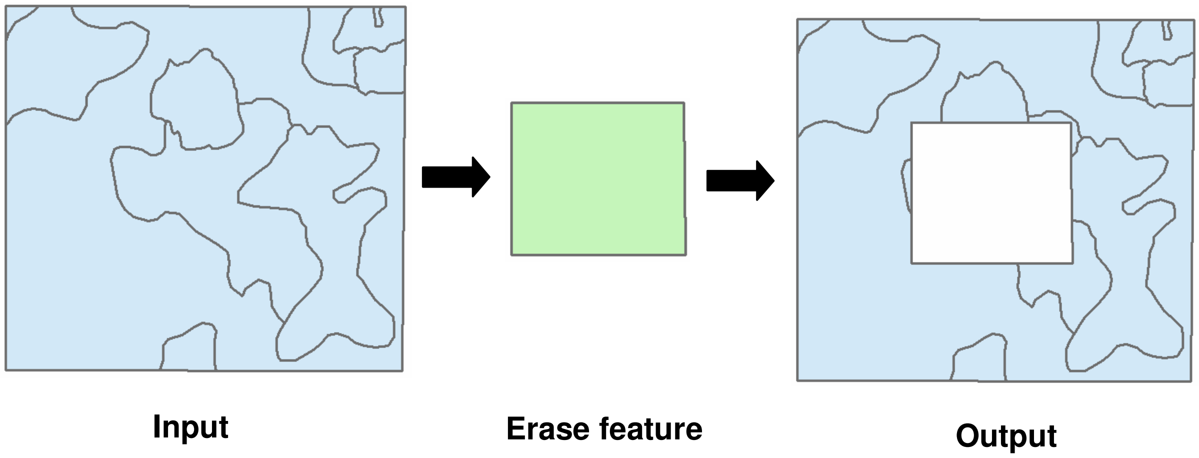


Figure 4. Erase Example

2.1.2 Overlay

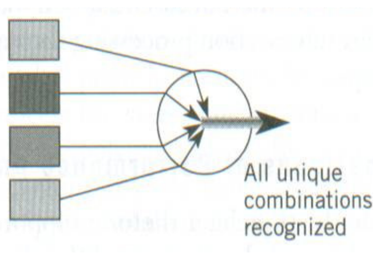
Central to GIS analysis is the integration of data to reveal the relationship(s) between two or more data sources. Overlay is one method of integrating information as it combines the spatial and attribute data from two or more input layers to create a new output layer. The spatial form of the new layer is shaped by the geometric intersection of the input and overlay features. Generally, the overlay of two or more layers results in a more complex output layer, with more polygons, intersections and/or line segments than what is present in the input layers.

Each feature in the newly created output layer contains a combination of attributes from the input layers. Overlay functions, when associated with geometrical (or ‘physical’) overlays of data layers, are implemented by certain mathematical operations – both arithmetic and logical. Arithmetical operations commonly used, but not limited to, include addition, subtraction, division and multiplication. Logical operations are aimed at finding where specified conditions occur and use logical operands such as: AND; OR; >; and <.

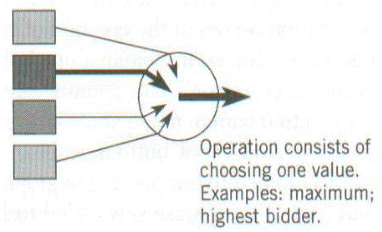
As discussed later in this module, methods for overlaying vector data differ from those of raster data related methods. However, basically vector-based methods do not generalize the data but, due to relatively more intensive processing requirements, may be more appropriate for smaller or sparser datasets. Raster analysis methods generalize the data based on the largest cell size found among the input layers. Raster-based grid calculations are, however, often faster and easier.

Four basic rules for combining the attributes of several layers can be applied to overlay analyses, and these are presented in Figure 5Figure below. While these rules are presented here with the vector analysis methods, they are equally relevant to a discussion of raster overlay methods.

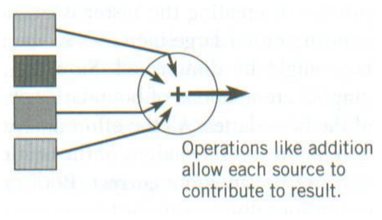
1. Enumeration Rule: Each attribute is preserved in the output layer and all unique combinations are recognized. For example, a soils layer, vegetation layer and precipitation layer are overlaid yielding a derivative coverage with a unique polygon for each possible combination.



2. Dominance Rule: One value wins – the dominant (e.g., highest) value is the only one value assigned. For example, an overlay based on a series of sensitivity layers would assign the highest sensitivity value to each derivative polygon.



3. Contributory Rule: Each attribute value contributes to the result - each source layer contributes to the result. For example, environmental sensitivity could be calculated based on the sum of a set of input layers: wildlife habitat sensitivity; riparian sensitivity; slope; and proximity to human disturbance.



4. Interaction Rule: A pair of values contribute to the result (i.e., decisions in each step may differ)

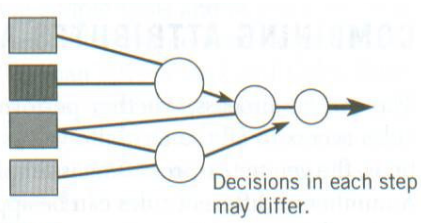


Figure 5. Overlay Rules

Three main types of vector overlay exist:

- **point-in-polygon** – The point features, which maintain their spatial location and attribute integrity in the output layer, are also assigned the attributes of the polygon they fall within. This type of overlay might allow an association between meteorological stations (the met station point layer) and vegetation types (the forest polygon layer) to be identified (Figure 6). The output layer would be the meteorological station point file with the addition of a vegetation type attribute.

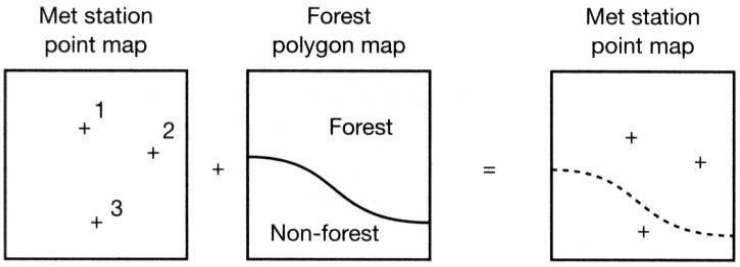


Figure 6. Point in Polygon Overlay Example

- **line-in-polygon** – The line features, which maintain their spatial location and attribute integrity in the resulting output layer, are assigned the attributes of the polygon they fall within. This type of overlay would allow you to determine the vegetation types (derived from the forest polygon layer) associated with each line of a road layer (the road line layer) (Figure 7). Because a line may overlap multiple polygon types, the output layer (the road line map) will generally contain more line segments than the input layer.

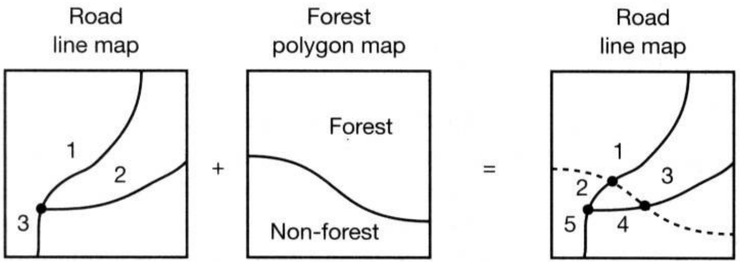


Figure 7. Line in Polygon Overlay Example

- **polygon-on-polygon** – The polygon geometries from the input and overlay layers combine to create a new set of polygons where each new polygon maintains the attributes from both input layers (Figure 8). This type of overlay might be used to find the association between slope and avalanche chutes. Polygon-on-polygon is the most common of the vector overlay methods.

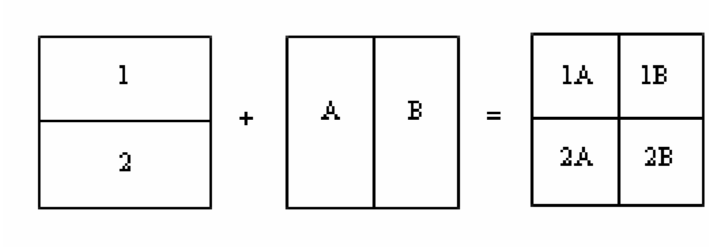


Figure 8. Polygon on Polygon Overlay Example

When performing a polygon-on-polygon overlay, there are several ways of combining the two sets of polygons. These are as follows, with detailed discussions following:

- Identity
- Intersect
- Symmetrical Difference
- Union
- Update

Identity

Identity is an overlay function that produces an output layer with the same area extent as the input layer (Figure 9). All input features and attributes are maintained. The operation also preserves the geometry and attributes of the overlay (or identity) layer that fall within the input layer's extent. The input layer may contain points, lines or polygons, but the identity features (overlay) must be polygons. An example of when the identity function could be used would be if you wanted to identify those roads located beneath 1,000 metres elevation. The input layer would be the roads layer, the identity layer a polygonal coverage representing all areas lower than 1,000 metres. The resulting output layer would be populated with an attribute identifying those roads below 1,000 metres. We are then also able to conclude that the roads without an elevation attribute are above the 1,000 metre contour.

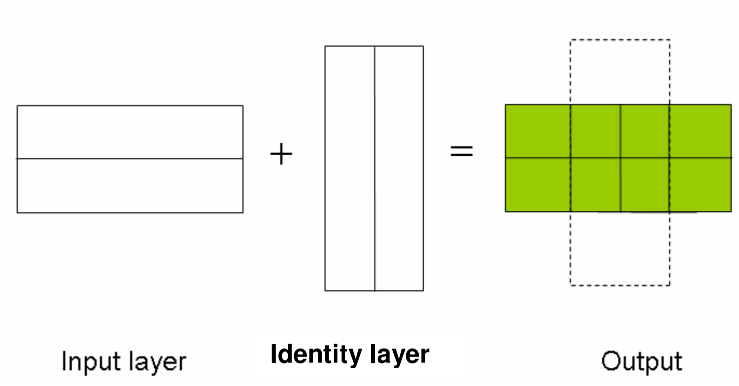


Figure 9. Identity Example

Intersect

Intersect creates an output layer by preserving only those features (or portions of features) that are common to both inputs (Figure 10). All features in the output layer contain attribute data from both of the input layers. The inputs may contain different geometry types (points, lines or polygons), but normally the overlay input is a polygon layer. The output geometry can only be of the lowest common geometry of the input layers - a point input combined with a polygon overlay will produce a point output with the points containing attributes from the polygon they fall within. An example of when the intersect function could be used would be if you wanted to identify only those roads located beneath 1,000 metres elevation. The input layers would be the roads layer and a polygonal coverage representing all areas beneath 1,000 metres. The resulting output layer would be a subset of roads – only the features located below the 1,000 metre contour.

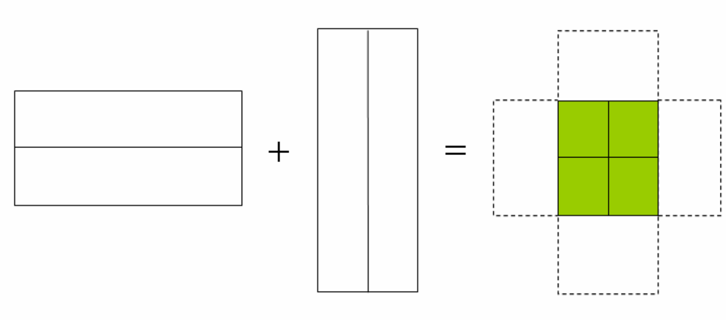


Figure 10. Intersect Example

Symmetrical Difference

The symmetrical difference tool creates an output layer that preserves those features (or portions of features) that are not common to features in the other input layer - portions from the inputs that do not overlap (Figure 11). Both input layers must have polygon geometry in order to execute the operation. The symmetrical difference function would be used when you want to remove areas of overlap between layers, for example, if you had a zone of influence polygon around a chemical plant you could identify those residences potentially not impacted by potential chemical emissions.

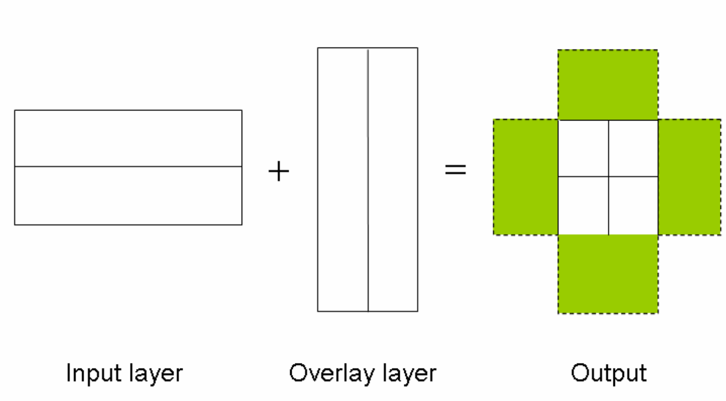


Figure 11. Symmetrical Difference Example

Union

Union combines and maintains all features and attributes from both input and overlay layers (Figure 12). Both input layers must have polygon geometry in order to execute the operation. Using our chemical plant example, a union would allow you to identify those residences falling both inside and outside the zone of influence of the plant.

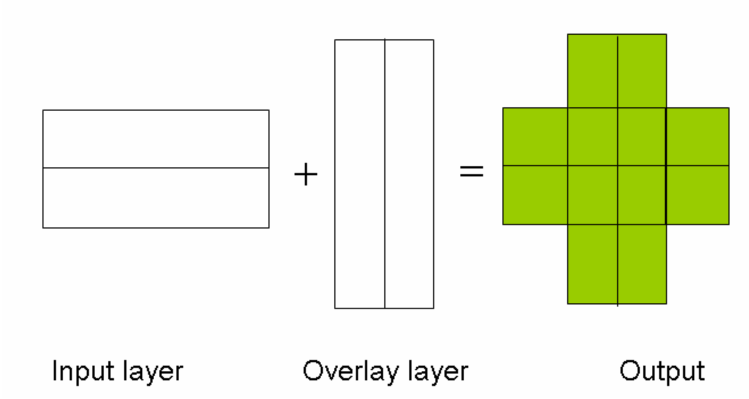


Figure 12. Union Example

Update

The update tool is used to create an output layer by erasing and replacing the features and attributes of an input layer with those from an overlapping update layer. Those portions of the input layer that are not overlapped by features in the update layer are not affected, and hence, are preserved in their original state in the output layer. This operation might be useful in updating a land use layer with a more current forest harvest layer. The regions of outdated land use, for example, where new harvesting has occurred, would be erased and replaced by harvest polygons. Caution should be taken during this operation as differences between layers (arising from dissimilar data sources with differences in quality and scale) can result in mismatched boundaries and/or slivers (see below) in the output layer.

One of the typical errors arising from overlaying polygon layers involves the generation of **slivers** in the output layers. Slivers are very small polygons created along shared boundaries during the overlay of inputs (Figure 13). In an overlay analysis, this problem can be the result of overlaying two files of different scales. Slivers can also be due to digitizing errors, non-precise geo-referencing, or data export. The Eliminate function may be used to help operators remove slivers resulting from polygon overlay operations.

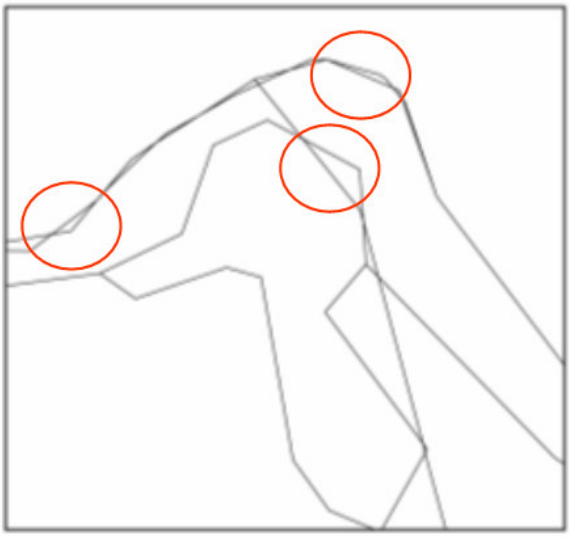


Figure 13. Slivers Example

Eliminate

The eliminate command can be used to remove slivers from a layer resulting from an overlay analysis or buffering (see Section 1.1.3). The sliver polygons are merged with neighbouring polygons. The user can select whether the sliver polygons merge with the largest adjacent polygon or the polygon with the largest shared boundary. Typically, eliminate would be run subsequent to an overlay analysis - a query would be conducted to select those polygons in the layer beneath a specified area threshold (i.e., very small polygons that would be slivers), this selection set would then be used to identify the polygons to be eliminated. Care should be taken when using the eliminate function as it can result in substantial alterations to a data layer. For example, slivers can often develop along coastlines when multiple layers are overlaid. Running an eliminate command would therefore potentially alter the coastline in the resultant coverage. Care should also be taken when establishing the area threshold during the development of the selection set because as if you establish too large a threshold potentially smaller polygons that are ‘real’ data will be merged with their neighbouring polygons.

2.1.3 Proximity

Proximity is a spatial relationship concept which corresponds to the geographic “nearness” of features. This allows us to select features located within a certain distance of other specified features, or to create new features by expanding a feature’s extent. For example, we may wish to find all hotels within 10 kilometres of the Vilnius Cathedral located in the city centre. Buffers and other proximity-based analyses help us answers these types of questions.

Buffer

Building on the notion of proximity, buffering creates a zone of inclusion or exclusion by creating an area around existing point, line and polygon features based on a specified distance. The GIS software extends a line in all directions around the features until a solid polygon has been formed and, finally, a new layer containing the buffer results is created. Point buffers are circular in shape, while the form of line and polygon buffers is defined by the geometries of the input features (Figure 14).

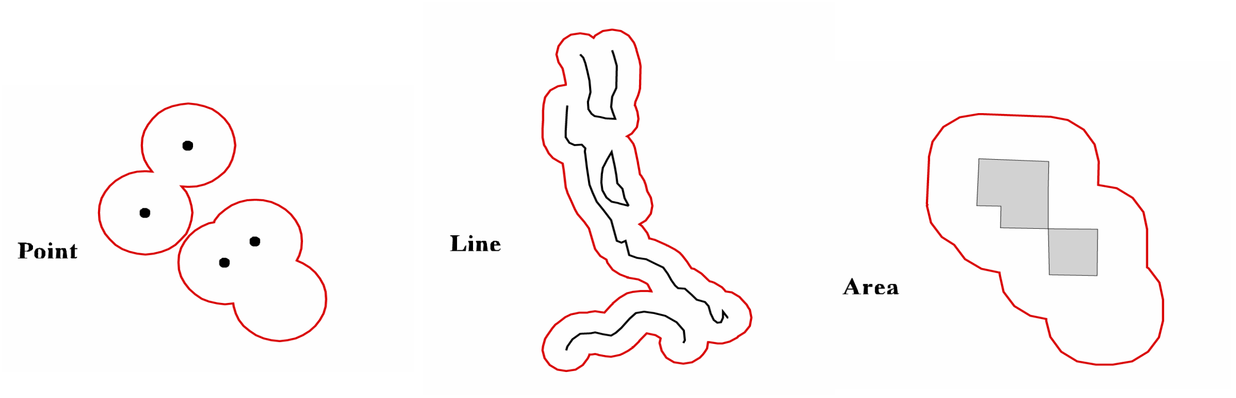


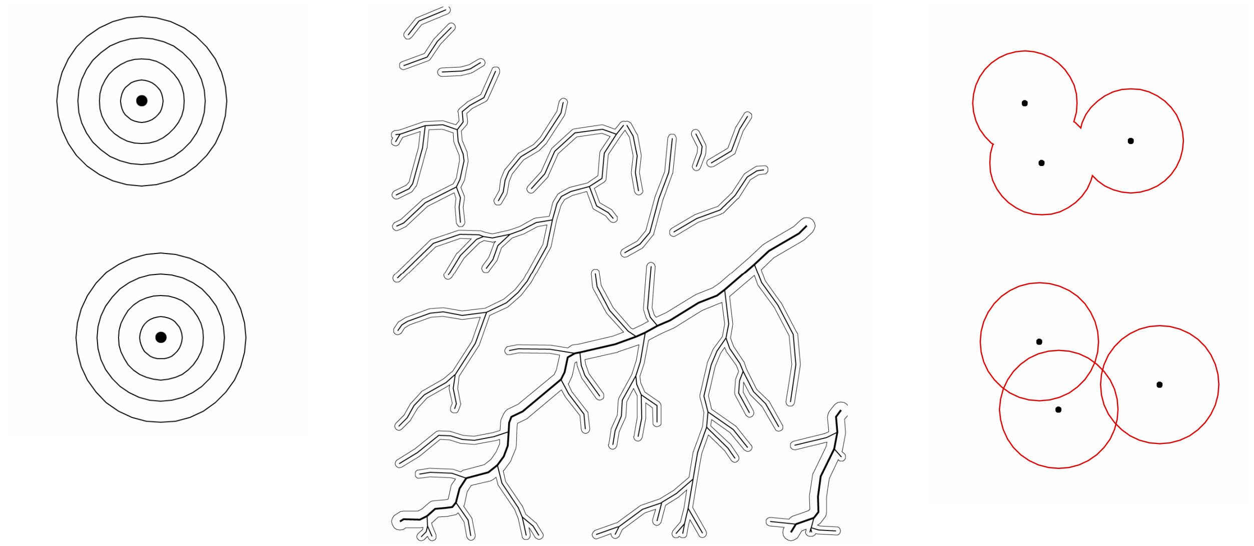
Figure 14. Types of Buffers

With line buffers, the buffer zone can be created on both sides of the line, or on either only the left or right side. In the case of polygon buffers, users may choose how the buffer zone is created, depending on where the area of interest lies:

- only the area outside of the polygon is buffered;
- the buffer zone includes the area outside of the polygon plus the entire area of the original polygon; or
- the buffer area is created both inside and outside of the polygon boundary.

A newly-created buffer feature may be used in several ways. It may form a new feature, such as a riparian zone, or a road polygon based on the stream or road centreline. It might also form the basis of a proximity analysis, by then overlaying the buffer with another input layer to find features which intersect the buffer.

The size of buffers can be defined by a variable distance, thus allowing individual features to have different buffer widths; normally, each particular buffer width would be based on distance value in an associated attribute table. For example, using a hydrology network to create varying riparian zones representing restricted logging areas, larger rivers might be buffered by a greater width than lower order streams and creeks. Related to varying buffer size, is the notion of creating multiple buffers for features. It is possible to create multiple rings around features to delineate zones depicting some hierarchy or changing level of influence. This might be useful in defining zones of varying noise level around machinery in a manufacturing plant. When creating buffers, the option also exists to dissolve the resultant adjacent polygon buffers (i.e., merging overlapping areas) (Figure 15). We will discuss dissolving in more detail in section 5.3.1.



Multiple zones (or rings)

Varying buffer size defined by distance

Dissolved (top) and un-dissolved buffers (bottom)

Figure 15. Buffer Variables

To avoid potential errors, it is important to know the units of measurement for a particular dataset (e.g., metres versus kilometres). Because buffering will operate on a selection set, ensuring the desired features are selected prior to initiating the buffer operation may also be of importance.

Near

Near calculates the distance between each point in an input layer and the nearest point or location along a line in another layer (also called the Near Feature layer). The resultant distance values are recorded in the input layer's attribute table. A fire department might use this tool to determine the closest water hydrant (near feature layer) for each school (input layer) in a certain district.

Point Distance

Point distance determines the distances between each point feature in an input layer to all points in another layer, within a specified search radius. The results are recorded in an output table, containing fields for the feature's unique identifier and distance values. Using our hydrant example, the fire department might wish to expand their search to determine the distances separating each school from all water hydrants within a specified search radius.

2.1.4 Statistics

Frequency

Frequency produces a table summarizing the unique codes (and their frequency) for a specified set of fields from an input feature layer or table. This might be useful to a vegetation ecologist interested in determining the frequency of all the plant types within a particular study area.

Summary Statistics

The summary statistics tool calculates one or more of the following statistics on numeric fields in an attribute table: sum, mean, maximum, minimum, range, standard deviation, first and last. The resulting output table can be saved in a variety of formats, including; dbase, or as a personal geodatabase table.

2.2 Raster Analysis Methods

2.2.1 Analysis Options

Prior to running analyses on a raster dataset it is important to establish the parameters of the analysis environment, which include the spatial extent of the analysis and the output cell size.

Analysis Extent

It may be necessary to perform an analysis on only a portion of a larger raster dataset, the area of interest may be defined by specified minimum and maximum map coordinates (x, y) which define a rectangle. Alternatively, some combination of input raster datasets may be used to define the analysis extent based on multiple inputs. In this case, either the union or intersect of the rasters defines the area. With a union setting, the analysis extent encompasses the entire area of all input rasters. Using the intersect option results in an analysis extent equal to only the area of overlap between all input raster datasets (i.e. the minimum of the inputs).

Masks

Setting a mask allows you to conduct analyses on a selected set of cells, and hence, is another method to define the extent of analysis. Only those cells that are identified by the mask will be considered when running the analysis. Two methods exist for setting analysis masks:

1. By attribute: Selecting rows in the attribute table of a raster allows you to restrict analysis to particular raster values. For example, a wildlife ecologist may wish to examine only those areas above a certain elevation to assess bear denning habitat.
2. By area: An existing feature layer (point, line or polygon) or raster dataset defines the spatial extent of analysis. Only those cells falling within the mask extent will be processed during analysis. For instance, the same ecologist might wish to limit the denning habitat analysis to areas within a park boundary and therefore a polygon, defining the extent of the park, would be used as a mask.

Cell Size

Establishing an output raster cell size is heavily influenced by the resolution or cell size of the input datasets. As a guideline, the output cell size should be equal to, or larger than, the coarsest cell size of the input raster datasets (known as the maximum of inputs). This ensures the resolution of the output raster is consistent with that of the least accurate (the coarsest) input dataset. Using a cell size finer than that of the input rasters does not ‘improve’ the accuracy of data. Other options available include:

- **Minimum of inputs** – The output cell size equals the input raster with the smallest cell size. Take caution with this option because, as mentioned above, refining cell sizes of coarser inputs does not result in accurate higher resolution output.
- **As specified** – Allows you to specify an output cell size.
- **Same as Layer** – Allows you to specify the output resolution equal to the cell size of an input raster.

2.2.2 Surface Analysis

GIS allows us to accurately represent, model and analyze terrain with 3-Dimensional (3-D) and raster analysis tools. The following surface analysis tools provide insight into the shape of landforms and reveal surface patterns that may not normally be apparent from a raster-based digital elevation model (DEM) (Figure 16) or vector-based triangulated irregular network (TIN) (Figure 17).

Such terrain tools are discussed in detail in the course GII-04, but are presented here in summary form for completeness.

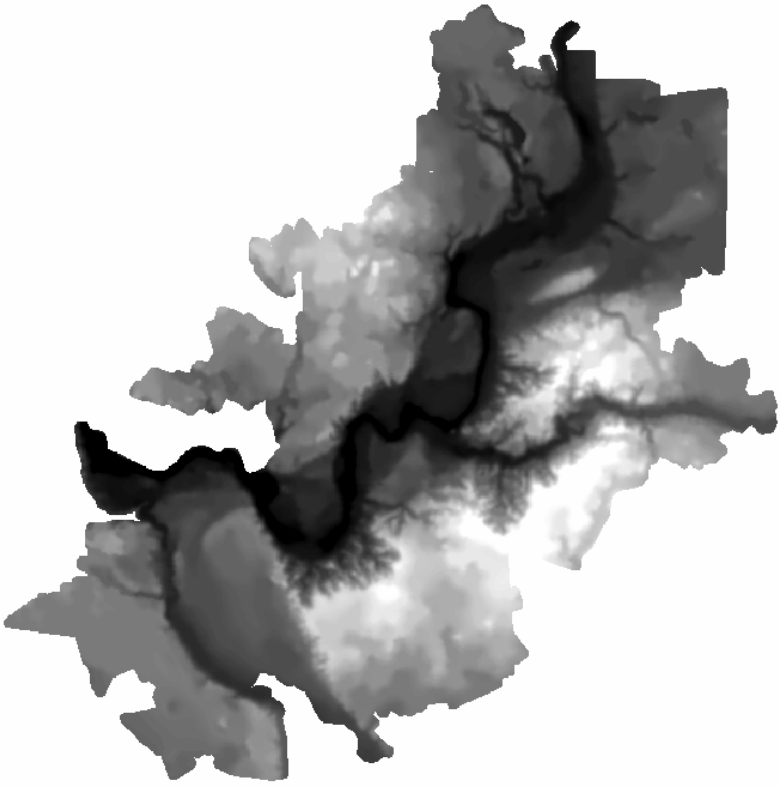


Figure 16. DEM for Vilnius

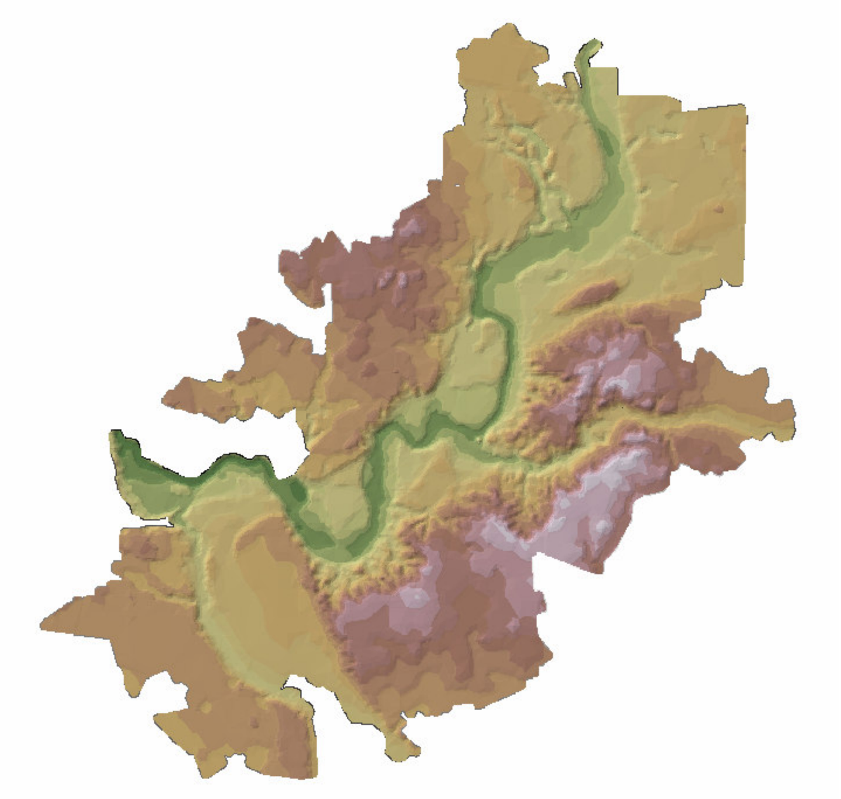


Figure 17. TIN Terrain Model of Vilnius

Slope

Slope is a measure of the maximum rate of change in elevation at a particular surface location. Although it can be expressed in either degrees or percent, both are simply variations on the evaluation of the rise in elevation over the run (distance on the ground). An output raster is created by calculating the slope for each cell in a raster DEM or each facet in a TIN, depending on the format of the input dataset. In Figure 18

, we are able to view and identify the areas of gentle, moderate and steep slope by theming the raster on grouped slope values (e.g., 0-2°, 2-4°, etc.). For instance, we can see the steeper areas (regions coloured yellow & red) occur along the banks of Vilnia and Neris rivers. In order to mitigate the potential threat of landslides, a structural engineer might develop and examine this type of slope information.

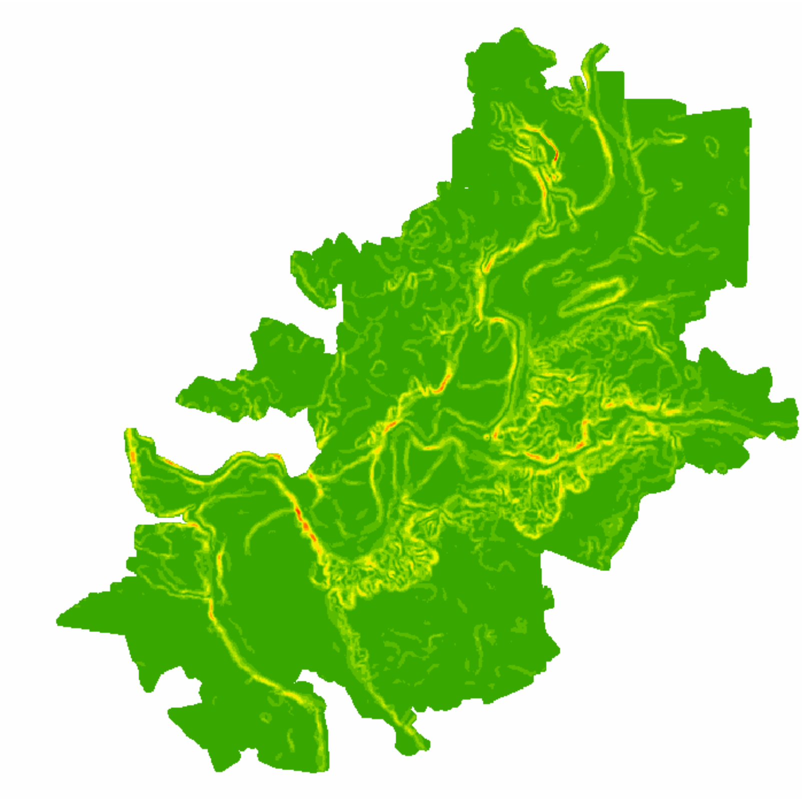


Figure 18. Slope Map of Vilnius

Aspect

Aspect is defined as the directional measure of the maximum rate of change in elevation; essentially, it identifies the cardinal direction a certain slope is facing (e.g., north, south, east, west). Aspect is measured clockwise from the north (0°) and is expressed in degrees, ranging from 0° to 359.9° (Figure 19).

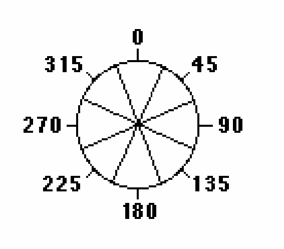


Figure 19. Aspect – Measured in Degrees - Clockwise from North

As with a slope calculation, an output raster is created by determining the aspect for each cell in a raster DEM or each facet in a TIN, depending on the format of the input dataset. Cells with a slope equal to zero are assigned an aspect of -1. An aspect layer reveals patterns in terrain not visible in a simple DEM or even slope map, as evident in Figure 20.

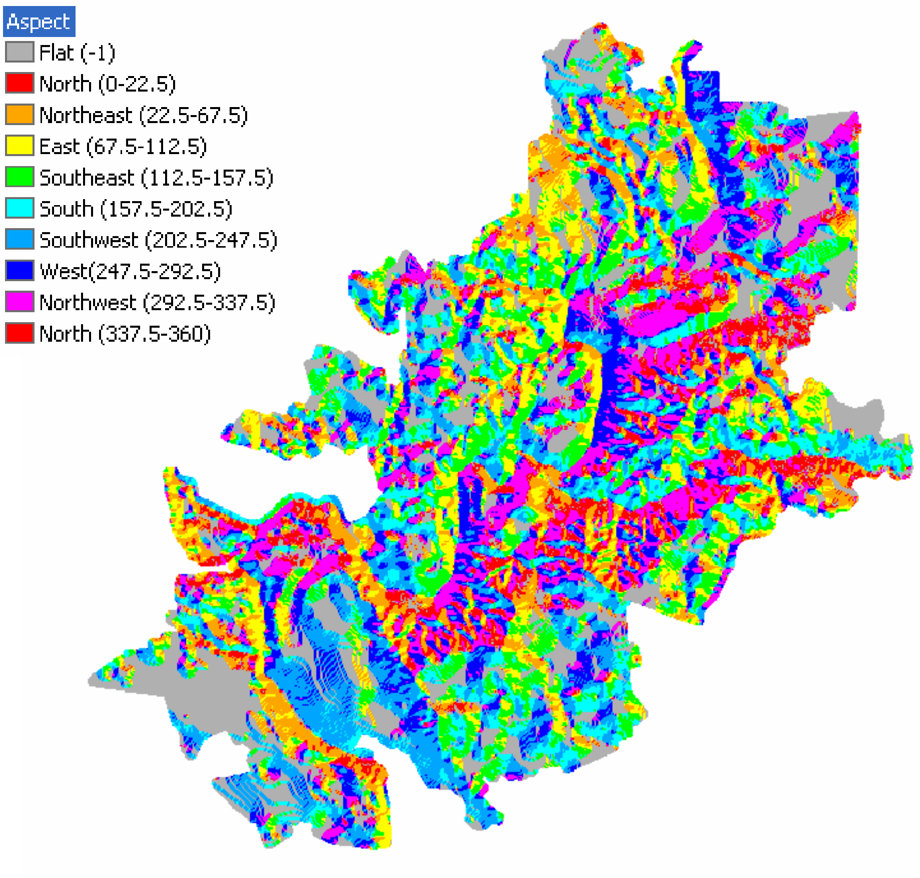


Figure 20. Aspect Map of Vilnius

A hydrologist looking to establish locations for snow pack sampling might wish to examine only north and northeast slopes in an aspect layer.

As with any raster dataset, you must consider resolution when creating a slope or aspect raster. As a general rule, the output cell size should never be smaller than the cell size of the input raster.

Contour

The contour tool is used to create a layer showing contours, or lines of equal elevation (also called isolines) (Figure 21). Contour maps allow you to identify and view regions of equal elevation and, by examining the spacing and position of isolines, they enable you to infer where steep slopes, cliffs, river valleys and ridge lines occur. Contour lines are spaced closely together in areas of steep terrain and further apart in flatter areas. In the areas surrounding water, the peaks of contour lines indicate the upstream direction of streams and rivers. Important in the creation of contours is the selection of a contour interval – the distance in elevation between adjacent contour lines. A higher resolution DEM or TIN will be able to produce a tighter contour interval than a surface with a coarse cell size.

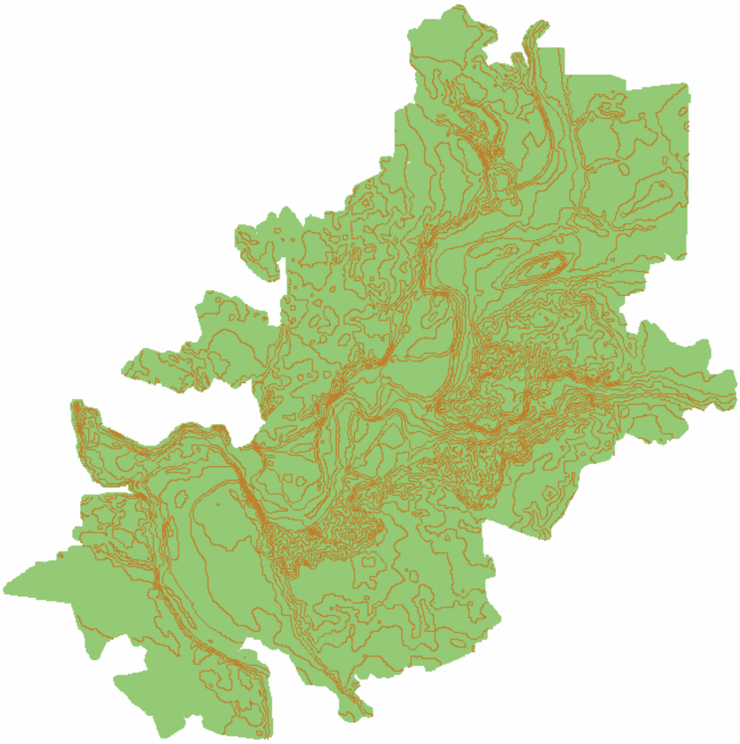


Figure 21. Contour Map of Vilnius

A contour layer could be used to limit the search for potential habitat for wolves above certain elevation.

Hill Shade

The hill shade tool creates a shaded relief raster from an elevation grid or TIN. By employing an illumination and shadowing of a surface layer, a hill shade can be very effective in representing relief or terrain, as it gives the impression of a three-dimensional landscape (Figure 22). Four factors combine to create a hill shade:

1. Azimuth of the Sun – the direction of incoming light measured clockwise in degrees from north (0° - 360°);
2. Altitude of the Sun – the angle of the illuminating source measured in degrees above the horizon (0° - 90°);
3. Slope of the Surface – the slope of the cell or facet from the input DEM or TIN respectively;
4. Aspect of the Surface – the aspect of the cell or facet from the input DEM or TIN respectively.

Each cell in the output hill shade is assigned an illumination value (ranging from 0 (black) to 255 (white) that when viewed simultaneously, gives the appearance of 3-D terrain.

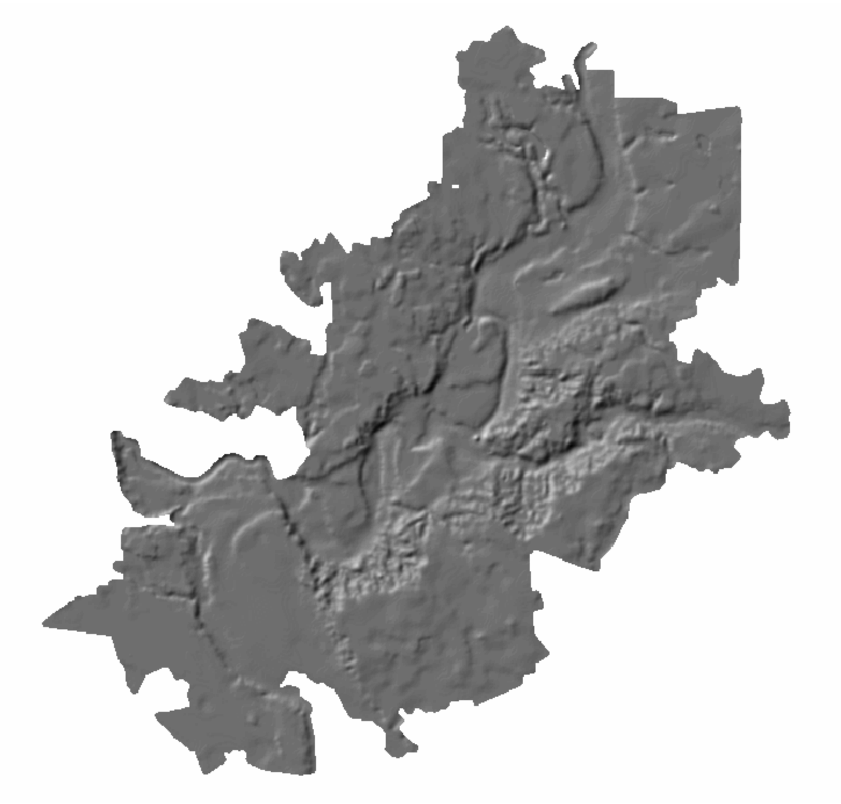


Figure 22. Hill Shade of Vilnius

Often, a hill shade is used as a backdrop for thematic maps to convey a sense of terrain, without overwhelming the other, more pertinent map information. Polygon or raster layers can be displayed using a transparency setting, thereby allowing the hill shade to be displayed 'underneath' the data layers.

Viewshed

The viewshed tool answers the question, 'what features or regions of a surface layer are visible from one or more vantage points?' Two input datasets are required to run a viewshed analysis; the first being a point layer with one or more viewpoints and the second a DEM or TIN surface representing a terrain model.

Viewshed analysis works on the notion of 'line-of-sight' - a line connecting a viewpoint to a target. A feature that separates a viewpoint from a target (e.g., a mountain) will render that target invisible. Conversely, if no feature on the surface is blocking the view from an observation point to a target, then that target is visible. In the case of a viewshed, the target is, in fact, every cell or facet of the input surface layer. The output raster summarizes all the possible line-of-sight permutations and classifies each cell as either 'visible' or 'not visible' (Figure 23). By creating a viewshed, a logging company could minimize the visual impact of their forest harvest operations by restricting harvest in areas that can be seen from a community or highway.

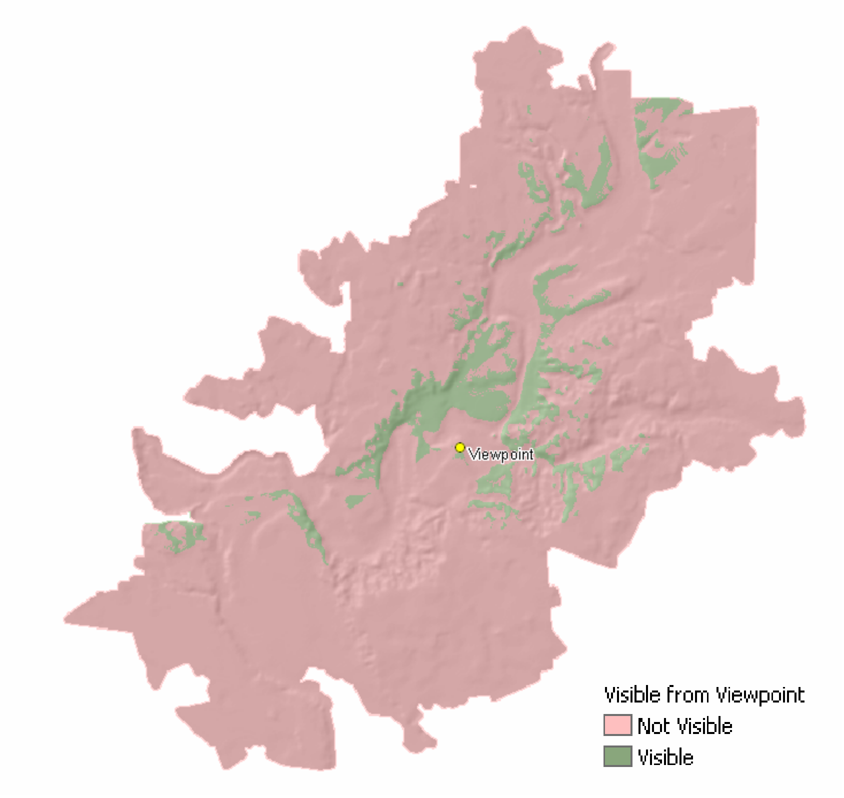


Figure 23. Example of a Viewshed Analysis

Options are available control the execution of a viewshed analysis and restrict or modify the results. Some examples available in ArcGIS are detailed below. These variables can be established as numeric fields within the attribute table of the observation layer and contain values that serve as observation constraints:

- SPOT: A z-value (height) which overrides the value normally derived from the surfaces z-value.
- OFFSETA: A value specifying a vertical distance to be added to the z-value of the observation point (e.g., 1.5 metres added to a point to represent the average height of a person standing in that location).
- OFFSETB: A value specifying a vertical distance to be added to the z-value of each target cell in the surface (e.g., this could be used to simulate average vegetation height).
- AZIMUTH1: The first of two values (in degrees) used to limit the scan of a viewshed to a certain direction.
- AZIMUTH2: The second of two values (in degrees) used to limit the scan of a viewshed to a certain direction.
- VERT1: A value used to establish the upper vertical angle limit of a viewshed analysis.
- VERT2: A value used to establish the lower vertical angle limit of a viewshed analysis.
- RADIUS1: a value used to establish the minimum distance in which to begin a viewshed analysis (i.e., raster cells closer than the value set in RADIUS1 will not be considered for analysis).

- RADIUS2: a value used to establish the maximum distance in which to restrict a viewshed analysis (i.e., raster cells beyond the value set in RADIUS2 will not be excluded from analysis).

Generalized Raster functions (i.e., those not for specific applications such as terrain) fall into three categories:

- Local functions examine a single raster cell in isolation
- Focal, or Neighbourhood functions examine the focal cell and also its adjacent cells, or those within a specified distance
- Zonal functions examine irregular shaped sets of raster cells which share a common cell value

These functions will be described in detail in the sections which follow.

2.2.3 Local Functions and Statistics

Local functions are used to perform calculations on a single cell at a time, ignoring the value of neighbouring cells - surrounding cells have no influence on the cell in question. After performing the calculation on that cell, the function moves on to the next cell location until all cells in a raster (or within a mask) have been addressed (Figure 24).

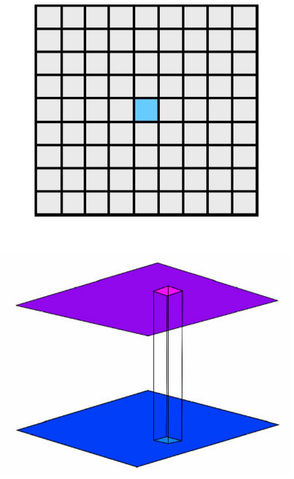
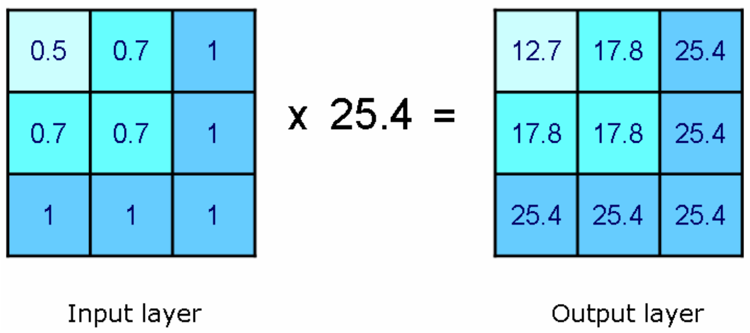


Figure 24. Local Function

Local operations can make use of either single or multiple input datasets to create a new raster.

Local Functions on a Single Raster

Local functions are able to apply any arithmetic operation on each cell in a single input layer. For example, to convert rainfall values from inches to millimetres, we could multiply each cell by 25.4 (Figure 25).

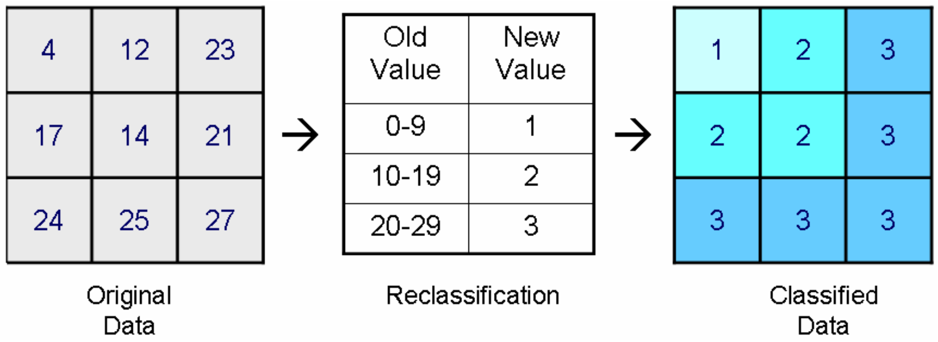


0.5	0.7	1
0.7	0.7	1
1	1	1

12.7	17.8	25.4
17.8	17.8	25.4
25.4	25.4	25.4

Figure 25. Local Function on a Single Layer

Reclassification is a local function used to re-assign values in an input raster to create a new output raster. This procedure allows you to simplify or aggregate (group data into classes) cell values within a raster dataset (Figure 26).



4	12	23
17	14	21
24	25	27

Old Value	New Value
0-9	1
10-19	2
20-29	3

1	2	3
2	2	3
3	3	3

Figure 26. Reclassification Example

Reclassification is also widely used to replace values based on new information or change cells with No Data to actual values (this operation is useful for data sets that have gaps). Section 5.3.2 in Module 1 examines reclassification in further detail.

Local Functions on Multiple Rasters

Local functions can also be applied to multiple layers. This is functionally the raster overlap operation, and may combine the attributes from each layer in many ways as discussed in the previous discussion of vector overlay. The attribute combination rules presented in Figure 5Figure may also be implemented with raster layers. For example, the dominance rule might use a *MAXIMUM* function to take the largest value from the contributing layers, the contributory rule might use arithmetic operators to combine cell values, and the interaction rule might use a series of *IF..ELSE* structures to implement decisions based on cell values encountered. Most commercial GIS applications incorporate a means of overlaying grid cell values in a variety of ways. In ArcGIS the interface is called the Raster Calculator. Figure 27Figure shows an example of two arithmetic operations used to combine raster cell values.

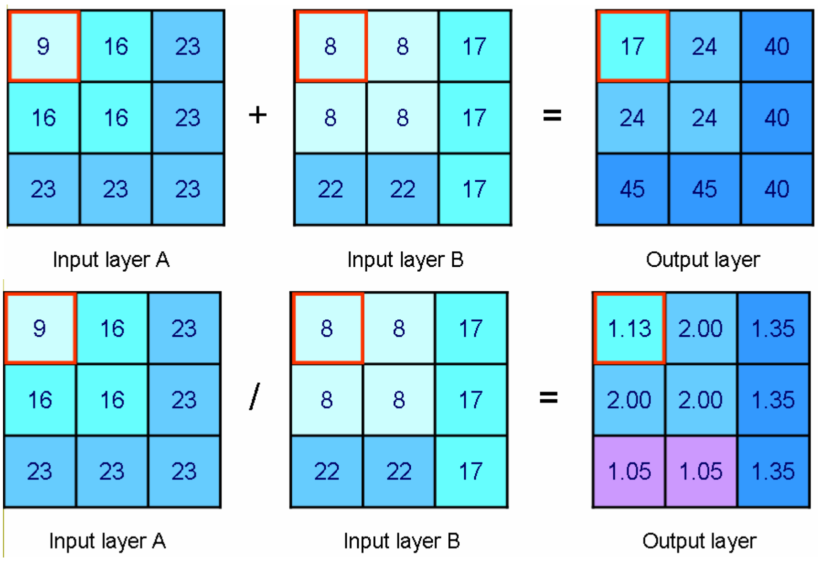


Figure 27. Local Function Applied to Multiple Layers

Local Statistics

Local (cell) statistics are another practical application of local function. This operation – often called compositing, overlaying or superimposing - involves calculating a particular summary statistic on a group of raster layers and creating an output raster containing the result. Because it is a type of local function, local statistics compares and summarizes only corresponding cells from the input rasters (i.e., analysis is performed on a cell-by-cell basis).

Below are some examples of the statistics that can be generated for a raster dataset:

- Maximum: determines the highest value among input rasters on a cell-by-cell basis
- Minimum: determines the lowest value among input rasters on a cell-by-cell basis
- Majority: determines the value that occurs most often among input rasters on a cell-by-cell basis
- Minority: determines the value that occurs least often among input rasters on a cell-by-cell basis
- Sum: calculates the sum of values from input rasters on a cell-by-cell basis
- Mean: calculates the mean (average) value from input rasters on a cell-by-cell basis
- Median: calculates the median (half of the values are above, half below) value from input rasters on a cell-by-cell basis
- Std. dev: calculates the standard deviation from input rasters on a cell-by-cell basis
- Range: determines the range in values (highest to lowest) from input rasters on a cell-by-cell basis
- Variety: determines the number of unique values from input rasters on a cell-by-cell basis

For example, you might wish to determine the maximum value between multiple input rasters; each cell in the output raster is derived from the values of the corresponding cell in each input raster. For instance, you could derive the highest rainfall for each cell location from a series of five input rasters representing five consecutive years of rainfall data.

In addition to arithmetical functions, Boolean (e.g., AND, OR, XOR, NOR) and logical operands (e.g., >, <, = etc) are often used in association with local functions.

2.2.4 Neighbourhood Functions and Statistics

Neighbourhood (or focal) functions expand on local functions in that the values of all pixels in a predetermined neighbourhood are considered when determining each output pixel's value in a new raster. Rectangles, annulus (doughnuts) and circles are the most frequently used neighbourhood configurations when performing neighbourhood analyses. In Figure 28Figure below, the blue pixel represents the focal cell while the yellow pixels constitute a 3x3 cell neighbourhood rectangle (inclusion of the focal cell in the neighbourhood is optional).

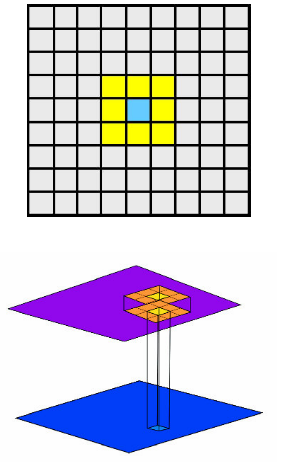
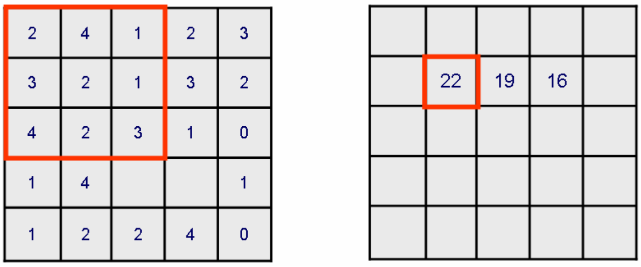


Figure 28. Neighbourhood Function

Similar to a local function, the operation moves on to the next cell location (designating it the focal cell) until all cells in a raster (or within a mask) have been addressed. As opposed to a local function using multiple input rasters, a neighbourhood function uses the values from surrounding cells to determine the values for the derivative coverage.

Neighbourhood Statistics

Like local operations, neighbourhood functions can use the same summary statistics to generate output cell values. For example, the *Sum* statistic could be employed to incorporate the data from surrounding cells into each output focal cell as shown in Figure 29 below.



2	4	1	2	3
3	2	1	3	2
4	2	3	1	0
1	4			1
1	2	2	4	0

	22	19	16	

Figure 29. Sum-Based Neighbourhood Statistic Example

Data Simplification

Data simplification (or the application of a spatial filter) is another important application of neighbourhood statistic functions. This type of operation is useful when generalizing raster data and serves to reduce the level of variation between neighbouring cells in the input layer. A typical simplification called the moving average method uses a *Mean* operator to calculate the average value on a moving 3x3 or 5x5 cell neighbourhood rectangle; the mean value derived from a neighbourhood is assigned to the focal cell of that neighbourhood in the output raster. Figure 30Figure outlines the typical statistics and their output when applying a spatial filter to a raster.

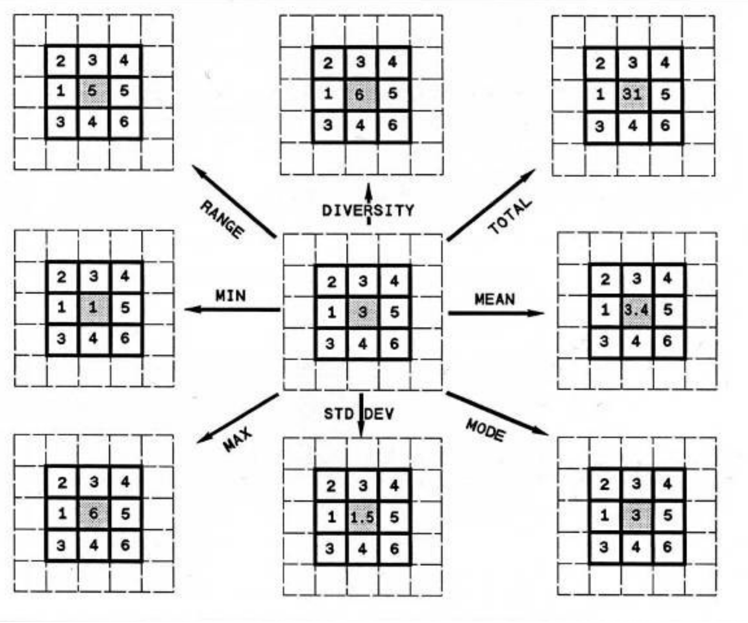
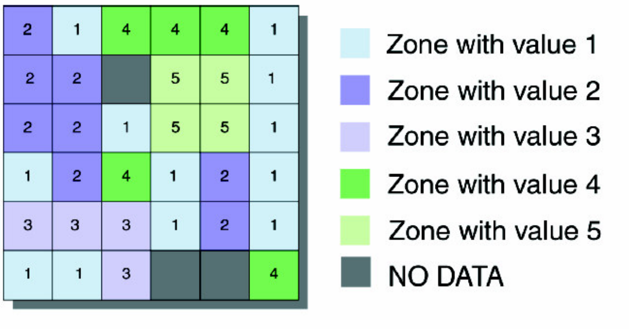


Figure 30. Typical Spatial Filters for Neighbourhood Statistics

No Data in Neighbourhood Statistics

One of the practical aspects in raster data analysis relates to gaps in cell values. Such cells can be assigned *No Data*. No Data indicates that no information or not enough information was available to assign the cell a numerical value (Figure 31).



2	1	4	4	4	1
2	2	■	5	5	1
2	2	1	5	5	1
1	2	4	1	2	1
3	3	3	1	2	1
1	1	3	■	■	4

Figure 31. No Data Values in Raster Data

Cells with No Data can be processed in one of two ways when executing a local or neighbourhood function:

1. assign No Data to output cell regardless of the combination of input cell values (i.e., as long as one input cell in the neighbourhood is No Data, then the output will be No Data); or
2. ignore the No Data cell and complete the calculation without it (i.e., calculate the maximum value in a neighbourhood, disregarding the No Data cell). Normally, No Data cells are ignored (i.e., the calculation (total sum) is executed based on neighbouring cells with values).

2.2.5 Zonal Functions and Statistics

Zonal functions perform operations on zones of common cells defined in one raster to make calculations on another layer (based on variable shaped and sized neighbourhoods). Zonal operations perform a calculation on a zone, which is a set of cells with a common value. Zones may be continuous or non-continuous (Figure 32). A continuous zone includes cells that are spatially connected, whereas a non-continuous zone includes separate regions of cells.

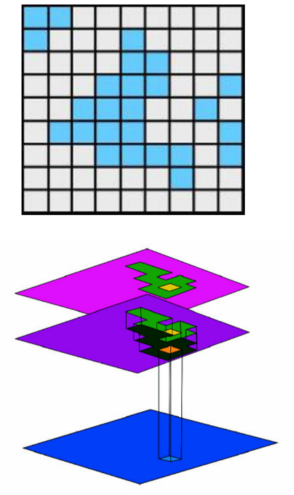


Figure 32. Zones in a Raster Dataset

Zonal operations may be performed on a single raster layer or on two raster layers (one raster contains the values to be summarized and the other defines the zones). When a single input raster is used, zonal statistical operations measure the geometry of each zone (e.g., area, perimeter, thickness, centroid, etc.). When two raster layers are used in a zonal operation the operation produces an output raster layer which processes the cell values in the input raster as per the zones defined in the zonal raster layer. A zone layer defines the zones (shape, values, and locations) and an input value raster contains the input values used in calculating the output for each zone. Figure 33Figure illustrates the results of a zonal statistics operation performed on two raster inputs to determine the mean slope of watersheds. The zones are defined by the watersheds raster and the slope raster is the other input dataset. The output summarizes the mean slope for each watershed.

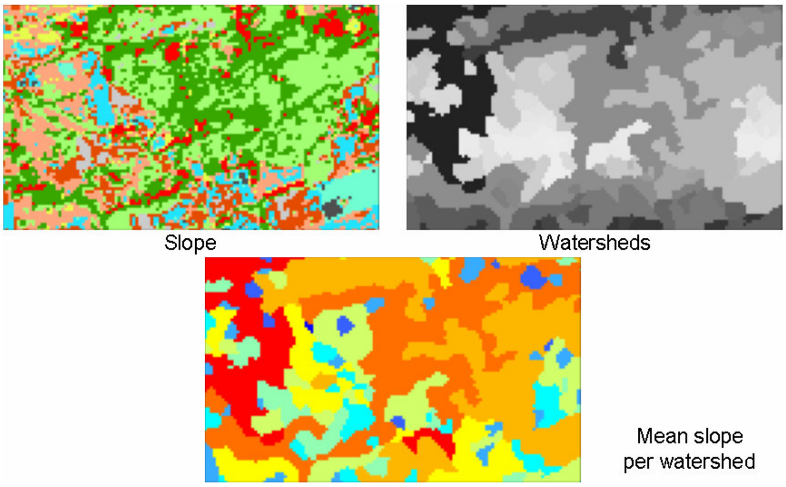


Figure 33. Illustration of Zonal Statistics

2.2.6 Distance

GIS gives us the ability to measure distances between features. At the simplest level distance statistics can be determined within the map interface through a distance query tool. The user can click on one location and then click on another and the distance between the two locations will be calculated and displayed on the screen. Typically these query tools will also allow you to draw a polyline on the screen and will summarize the distance for the length of the line.

In vector datasets, distance measurements are typically determined between point features, either located in a single dataset or in multiple datasets. For example, distances between cities can be calculated and stored as an attribute of each city location. Distances may also be determined from points in a layer to the nearest point along a line (either from a polyline or the edge of a polygon). This type of calculation would allow you to determine the distance between an offshore water quality sampling location and the nearest part of a coastline.

Distances are of interest in and of themselves; however, they can also be fed in as a variable when developing a model. For example, in a wildlife habitat model, proximity to water is critical to most wildlife species and therefore distance to the nearest fresh water source (e.g., a stream or lake) would be an important input to evaluating wildlife habitat capability. It is in these types of analyses that the functions available within raster data models are most useful. For example, when developing our wildlife habitat model we can develop a surface that specifies distance to fresh water sources. Details related to distance functions available in a raster environment are provided below.

Straight Line

Straight line distance is the physical distance between two points. It is also referred to as Euclidian distance. In a raster dataset, straight line distances are calculated between cells based on the cell centres (Figure 34).

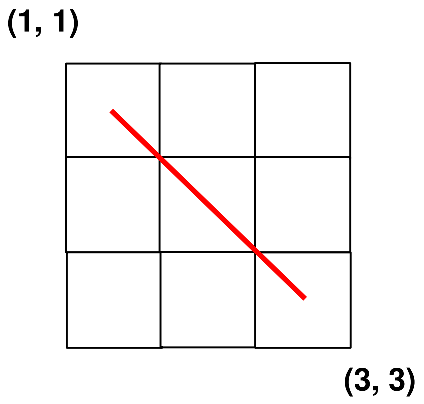


Figure 34. Calculating Distance Between Raster Cells

In Figure 34Figure we are calculating the distance between cells 1, 1 and 3, 3 (as illustrated by the red line). The calculation would be conducted using the following formula:

$$\textbf{cell size x } \surd\ (3 - 1)^2 + (3 - 1)^2$$

If the cell size was 10 metres the distance between the two points would be 28.3 metres based on the following calculation:

$$10 \text{ x } \surd\ (3 - 1)^2 + (3 - 1)^2$$
$$10 \text{ x } \surd 2^2 + 2^2$$
$$10 \text{ x } \surd 4+4$$
$$10 \text{ x } \surd 8$$
$$10 \text{ x } 2.8284 = 28.2843$$

This type of distance calculation illustrates how distance between raster cells is calculated but straight line distance functions allow us to develop output rasters that measure distance from every cell to source cells or points. This type of distance analysis allows us to determine the relationships between locations. For example, Figure 35Figure illustrates the results of a raster developed to illustrate the distances between cities (the yellow points). This type of decision support information assists regional planners when they are locating facilities (e.g., perhaps a hospital or a recreation centre can service a number of smaller communities and therefore distance can be used to help select an optimal location for a new facility).

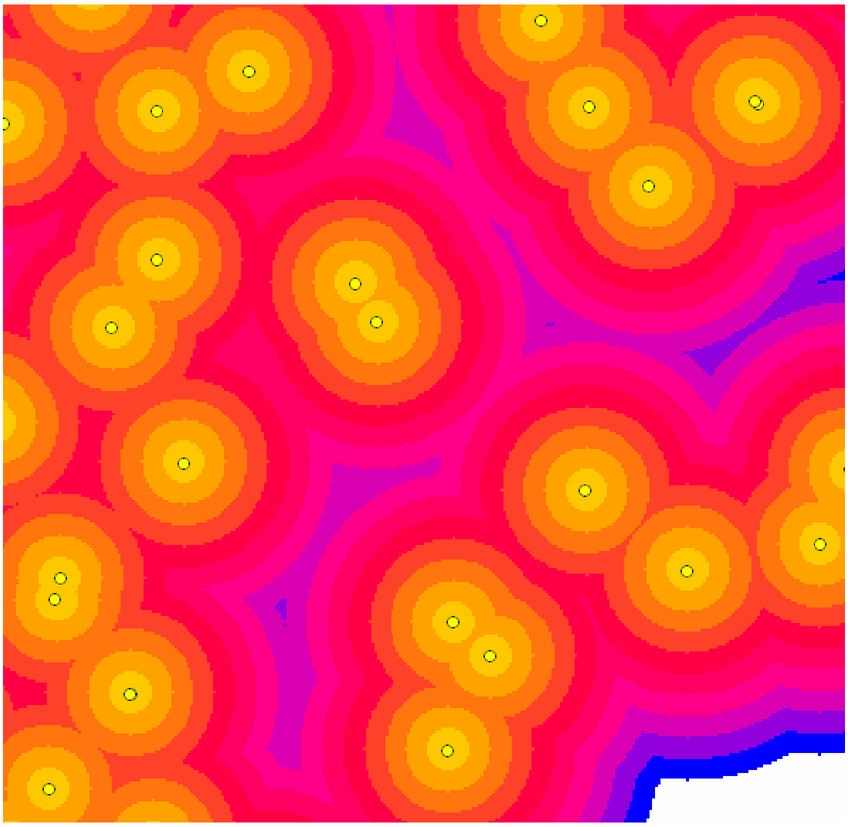


Figure 35. Straight Line Distance Raster Example

Allocation and Distance

Physical distance functions can also be used to generate allocation and direction rasters. In an allocation raster, the value of each cell is based on that of the closest source cell. An example of how an allocation raster can be applied would be an evaluation of the catchment areas for existing schools. The generation of an allocation raster would allow you to identify the school closest to each location on the map. A direction raster displays the azimuth direction from each cell to the nearest source. This function allows you to easily identify the direction of the nearest source feature (e.g., a town or a facility)

Cost Weighted

Cost weighted distance analyses calculate the cost for traveling across the landscape. This type of analysis allows you to identify the easiest route between two places based on other considerations beyond just physical distance. For example, the shortest distance between two communities may involve having to cross a mountain range and therefore, while the physical distance to avoid the mountains would increase the longer route may well have a lower cost associated with it. Typically, cost weighted distances involve the consideration of multiple factors, represented by numerous rasters. The cost represents the sum of the input raster layers. Some of the rasters may represent attribute values that lower distance costs (i.e., positive features) while others increase the cost. Costs can be represented as either actual values or as a relative numbers where factors are ranked relative to one another with higher numbers being associated with higher costs. The application of relative cost values are useful in that they allow you to standardize different cost types to a common scale.

A cost raster is generated by assembling the data associated with each of the cost factors in the analysis. A raster is then developed for each factor detailing a cost for each cell. The local statistics function may then be used to generate a derivative (output) raster that sums the total cost for each cell based on the various cost factor rasters.

Shortest Path

A cost raster allows you to determine the accumulated cost of a given route and evaluate the shortest path (the path with the lowest associated cost). The cost is determined by summing the costs associated with the links between adjacent cells (horizontal, vertical and diagonal). To determine the shortest path we must first calculate the costs between all the linked cells. We then select a source cell (the starting point) from which the cost associated with movement to that cell's adjacent cells is determined. The adjacent cell with the lowest cost then represents the next cell along the route. The cost of progressing to that cell's neighbours is then added to the total cost. The shortest path is the result of this iterative calculation.

2.3 Generalizing Data

Data can be generalized both physically (e.g., the shape or appearance of a feature can be altered) or through the classification and resultant summary of attributes (e.g., attributes can be grouped into more general classes). The following sections detail some generalization techniques that can be applied to vector and raster datasets.

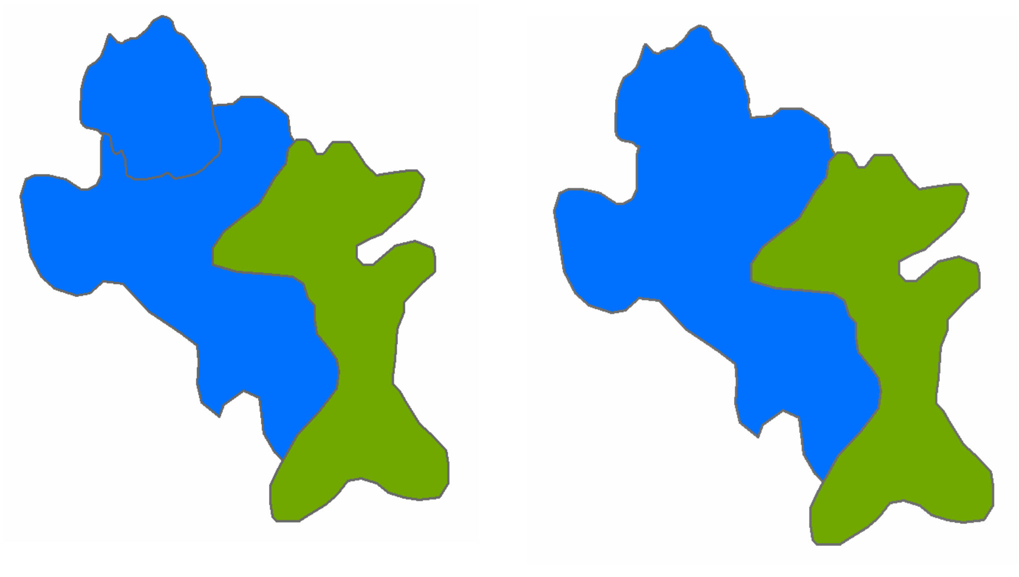
2.3.1 Vector Data

Dissolve

The dissolve function merges features having common attributes (Figure 36). It can be used for two fundamental applications:

- It allows you to simplify a classified dataset into more generalized classes. For example, land use polygons can be aggregated based on land use values (e.g., roads, buildings and paved areas could be grouped into an all-encompassing class called urban).
- It can be used to remove boundaries between polygons having identical attributes in a given field. For example, if you have merged datasets from different sources the resulting product may have a number of slivers or overlapping values. The dissolve function will allow these unnecessary boundaries to be removed from the dataset.

Dissolve can also help us summarize data because while merging the polygons we can also aggregate mathematical data. For example, using the dissolve described above to create an urban land use class we can easily determine the total area of all urban features from the resulting dissolved coverage. In addition, other related attributes (e.g., population values, annual incomes, birth rates) can be aggregated. For example, if annual income is an attribute available for a set of municipalities in a city, we can determine the total and/or average income for the entire city (or a group of municipalities within the city) during the dissolve by summing and averaging the income attribute.



Input layer **Output layer**

Figure 36. Dissolve Example

Eliminate

The eliminate function is another means of generalizing data – a new layer is created by merging selected polygons (based on a query or selection set) with neighbouring polygons. Eliminate is described in detail in Section 5.1.2 above.

Simplify Line

The process of simplifying a line involves the removal of small bends in the line. This is accomplished by removing some of the points (or vertices) within the line. The result is a generalized version of the feature (Figure 37) that maintains the general shape of the original version. Lines are simplified to improve cartographic display, for example, a detailed line digitized at a scale of 1:20,000 might have so much detail it appears fuzzy when plotted at a scale of 1:100,000. In addition, more complex lines increase the processing times associated with various analyses (e.g., buffering) and often simplifying a feature prior to a buffer analysis will not affect the accuracy of the resultant buffer polygon.

When simplifying a line it is important to specify the degree of simplification to ensure the resultant feature resembles the original feature. This is done by selecting an analysis tolerance at a scale appropriate to the source data. For example, when the goal is to improve the cartographic display of the data, the tolerance should be set equal to, or greater than, the minimum allowable spacing between graphic elements. Some trial and error may be required when selecting a tolerance appropriate for all features in a dataset.

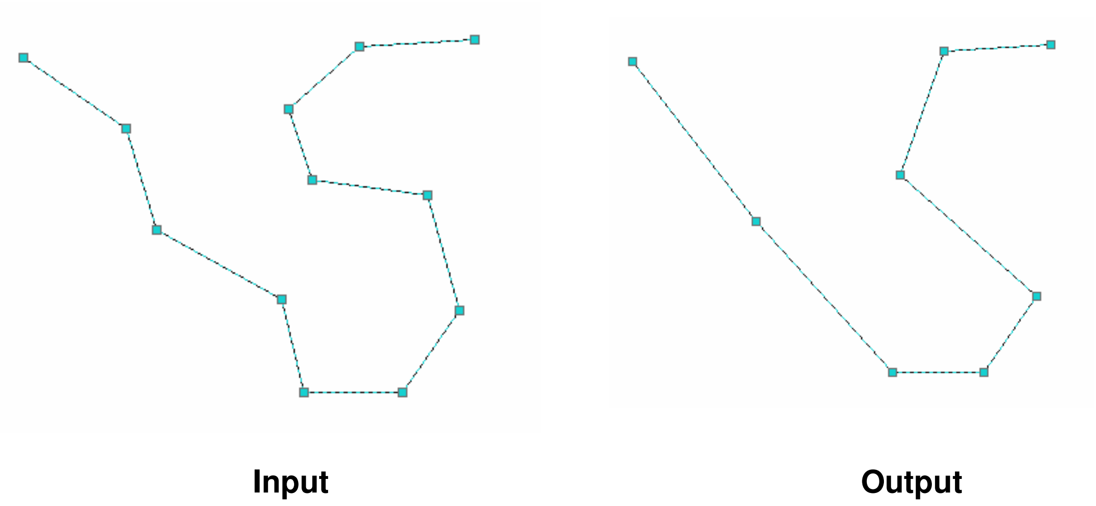


Figure 37. Simplify Line Example

Smooth Line

Lines are smoothed to improve their cartographic appearance, for example, jagged features are removed to make the line more aesthetically appealing (Figure 38). It is accomplished by reshaping the line through the application of a mathematical formula that generates new vertices (points) that are inserted into the line. There are a variety of different algorithms that can be applied including:

- Polynomial Approximation with Exponential Kernel (PAEK) calculates the smoothed version of the line using a parametric continuous averaging technique. The placement of additional points in the lines is based on a weighted average of the coordinates for all the points in the original feature. Points closer to the current coordinate are weighted more than points further away. The resulting line may not contain any of the vertices present in the original feature with the exception of the start and end points.

- Bezier Interpolation fits Bezier curves along each line segment of the original feature and then the curves are connected at the vertices through the application of the Bessel Tangent

As when simplifying a line, it is important to select an appropriate tolerance for the analysis to ensure the desired results are achieved. In line smoothing, the tolerance specifies the length of a moving path along the line used to calculate the new 'smoothed' coordinates – a higher tolerance results in a longer path (i.e., more points are factored into the result) and therefore yields a smoother line. Again a certain level of trial and error should be applied when selecting a tolerance level to ensure important characteristics (e.g., bends) are not removed.

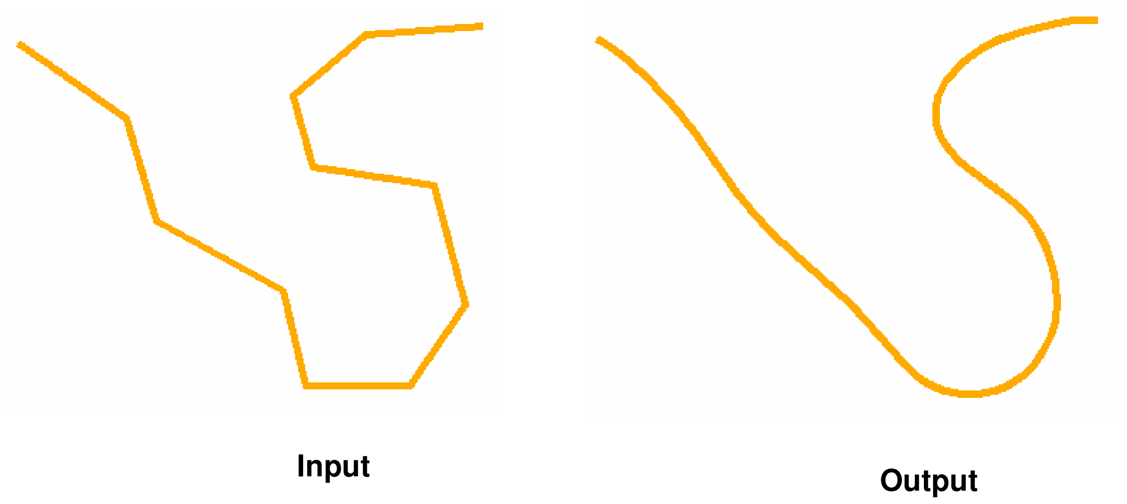


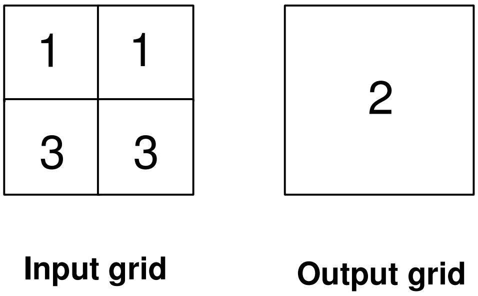
Figure 38. Smooth Line Example

2.3.2 Raster Data

Raster data can often contain data that is overly detailed or extraneous to a current analysis. For example, a land cover dataset derived from a satellite image may have numerous small groups of cells that are either misclassified or represent too small an area to be of statistical relevance. Generalization techniques, when applied to these types of examples, allow you to automate the removal of these types of regions within a raster dataset.

Aggregate

Aggregation is a resampling technique that allows you to generate a lower resolution grid (a grid with larger cell sizes) based on the attributes of an input grid. The values for each output cell can be based on the mean, median, sum, minimum or maximum of the input cells falling within the extent of the output cell. Figure 39Figure provides an illustration of an aggregation analysis where the resulting output grid is based on the mean of the input dataset.



1	1
3	3

Input grid

2

Output grid

Figure 39. Aggregate Example

Aggregate could be used to generalize a detailed grid depicting temperature values to derive a simplified mean temperature value for a large area.

Boundary Clean

Boundary clean can be used to smooth the boundaries between regions. It cleans boundaries on a relatively large scale by making a series of passes through the data – the first pass involves an examination of cells outside the region and the second considers cells inside the region. Basically all regions of less than a three by three block of cells will have their values updated based on the values of surrounding cells. During the 'outside pass', regions outside the current region (one cell in each direction) are assigned values based on the values of neighbouring higher priority zones. On the 'inside pass', cells that are not completely surrounded by cells of the same value are then evaluated and may then be replaced with the values of surrounding cells. It should be noted that thin portions of regions may be replaced, for example a region representing a linear feature (e.g., a river that is numerous cells long but only two cells wide) will be removed.

Expand

Expand allows specified regions within a raster dataset to be expanded based on a user-specified number of cells. Cells with lower priority values (determined by the user) are labelled as background cells. Cells with a higher priority (foreground cells) are allowed to expand into regions of low priority. The technique would be useful to remove no data values from a raster dataset by allowing cells with surrounding values to expand to 'fill' the no data locations.

Filtering

Spatial filtering is designed to highlight or suppress specific features in an image based on their spatial frequency. Spatial frequency is related to the concept of image texture. It refers to the frequency of the variations in tone that appear in an image. 'Rough' textured areas of an image, where the changes in tone are abrupt over a small area, have high spatial frequencies, while 'smooth' areas with little variation in tone over several pixels, have low spatial frequencies.

In practical implementation, filters are applied to the source raster by means of moving windows (kernels). A common filtering procedure involves moving a 'window' of a few pixels in dimension (e.g., 3x3, 5x5, etc.) over each pixel in the image, applying a mathematical calculation using the pixel values under that window, and replacing the central pixel with the new value. The window is moved along in both the row and column dimensions one pixel at a time and the calculation is repeated until the entire image has been filtered and a 'new' image has been generated. By varying the calculation performed and the weightings of the individual pixels in the filter window, filters can be designed to enhance or suppress different types of features. The moving window process is illustrated in Figure 40. Typically a kernel shape is square or rectangular, but circles and annuli are also used.

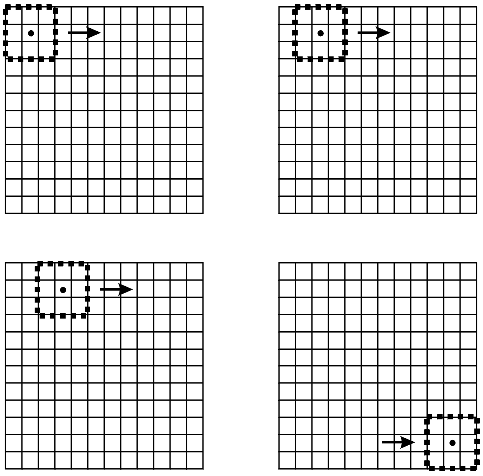


Figure 40. Moving Window Process

Filtering is widely used in many raster data analyses. Generic applications might include edge detection, blurring (smoothing), and noise removal. Noise may be erroneous data values, or spikes that can be removed from the data. For example, spikes in a digital elevation model (DEM) may be removed through 3x3 median filtering.

Thematic applications of filtering include: surface slope and aspect calculations using a DEM, calculations of weighting functions for advanced multi-criteria raster analysis and many others.

The majority filter function generalizes data by replacing cells in a raster dataset based on a values present in the majority of the cell's surrounding values. Two criteria must be met before a cell's value will be replaced:

- there must be a large enough number of surrounding cells (e.g., more than half) with a common value; and
- the cells having a common value must be spatially connected (e.g., contiguous) around the centre of the filter kernel. This criterion minimizes the potential corruption of spatial patterns in the data.

Nibble

The nibble function can be applied to edit portions of a raster dataset where the values are known to be incorrect or missing (e.g., areas of no data). A query or selection set is first applied to select the cells in the grid that are to be replaced. A mask is applied to specify the extent of the analysis – selected cells falling within the mask will be the ones replaced. The selected cells are then reassigned the values of their nearest neighbours through a Euclidean Allocation (cells are allocated based on closest proximity using a Euclidian distance (a straight-line).

Region Group

A scanning process is applied (similar to a moving window analysis) to assign a unique number to those cells falling within each region of a raster dataset. The resultant (output) dataset contains unique values for each unique region (Figure 41). The values assigned cannot be controlled by the user. The region group function allows you to examine potential spatial patterns in your data by helping you identify unique regions or zones.

1	1	1	1
3	3	4	4
3	1	1	1
2	2	2	4

Input grid

1	1	1	1
2	2	3	3
2	4	4	4
5	5	5	6

Output grid

Figure 41. Region Group Example

Shrink

Shrink allows you to change the values of spurious cells along the boundaries of regions based on the highest frequency value among the cell's surrounding cells. Shrink replaces the values for cells that are not internal cells (e.g., they are not completely surrounded by adjacent cells. It should be noted that thin portions of regions may be replaced, for example a region representing a linear feature (e.g., a river that is numerous cells long but only two cells wide) will be removed.

Thin

The thin function allows you to reduce the number of cells required to represent linear features in a raster dataset. For example, a paper map scanned at a high resolution would potentially represent linear features by a region numerous cells wide (e.g., a single-line river on the source hard copy could appear in the raster as a region 5 cells wide and 200 cells long). Thin would allow you to reclassify the cells in the raster resulting in the river in our example being represented by a region now a single cell wide by 200 cells long.

Self-Test Questions

1. Describe the difference between the Contributory Rule and the Dominance Rule for use in combining attribute values during overlay.

2. You wish to determine the ratio of agricultural lands lying within 500m of streams to agricultural lands lying more than 500m from streams. To do this, you have two vector polygonal layers: polygons delineating agricultural land, and 500m buffer polygons generated from a streams layer. Is it appropriate to use the *Intersect* form of a vector overlay to arrive at the necessary ratio? Why or why not?

3. Describe the difference between a local, focal and zonal operation on raster layers.

4. Why might you use the Smooth function on a linear feature class?

5. Does executing a 3x3 *mean* filter on a raster grid have the effect of smoothing out extreme values, or does it enhance extreme values in the raster? Why?

References

1. Chang, K.T. *Introduction to Geographic Information Systems*. McGraw-Hill, 2006.

2. Chrisman, N. *Exploring Geographic Information Systems*. John Wiley and Sons, 1997.

3. Heywood, I., Cornelius, S. and Carver, S. An *Introduction to Geographical Information Systems*. Pearson Education, 2002,

3 Spatial Statistics

The most frequent method of reducing large quantities of data into a manageable amount of information is through the use of statistical analysis. Broadly defined as the collection, analysis, interpretation or explanation, and presentation of data, statistical analysis can provide valuable information from vast arrays of data.

Though there are many forms of statistics, two classifications of statistics are classical statistics and spatial statistics. Classical statistics is the set of methods that most people come across in government reports, sports numbers, academic research, and the media. These methods, as with all statistical methods, have restrictions on their applicability, limiting the ways in which classical statistics may be applied.

One of these restrictions is the independence of observations. This simply means that one observation is not related to another: the information from one observation gives you no information about another observation. This restriction is most often violated in the context of spatial data.

The first law of geography, usually referred to as Tobler's Law, states that "[e]verything is related to everything else, but near things are more related than distant things" (Tobler 1970: 236). This means that the information given regarding one location can be used to know information regarding another location, with that information's value decreasing as the two locations become further apart.

Because of this violated restriction a whole branch of statistics has emerged to deal with this lack of independence of observations. This module introduces some of these spatial statistical methods, with application using crime data in Lithuania measured at the municipal level.

The following six topics of spatial statistics are examined in this module:

Topic 1: Classical Correlation, Spatial Autocorrelation, and the Location Quotient
Topic 2: Modifiable Areal Unit Problem and the Ecological Fallacy
Topic 3: Pattern Analysis
Topic 4: Edge Effects
Topic 5: Density Estimation and Hot Spot Mapping
Topic 6: Local Spatial Statistics

3.1. Classical Correlation, Spatial Autocorrelation, and the Location Quotient

3.1.1. Classical Correlation

In previous modules you have covered the measurement of the mean (averages) and standard deviation (how much a variable varies). Correlation, however, is a measurement of the relationship between two variables. Essentially, when investigating correlation, you are asking, for example: what will happen to the crime rate when the level of police activity changes? These two variables are quite related, as one would expected—the more crime there is the greater the need for police officers—but a few concepts need to be covered to allow proper interpretation of the classical correlation output.

First, when analyzing data using classical correlation, you are interested in know *IF* two variables are related. Just because two variables are correlated does not mean that there is an actual relationship between two variables. In a very humorous paper, an economist showed that there is a strong relationship between the national inflation rate and dysentery, so it is important that correlations are interpreted carefully and that the two variables you choose to correlate *should* be related to each other.

Second, when two variables move in the same direction (either increase or decrease) the two variables are said to be positively correlated. It is a little confusing to think of two variables being positively correlated when they are both decreasing, but this is just the terminology. When one variable goes up and the other goes down, the two variables are said to be negatively correlated. These two types of correlation are shown in Figure 1.

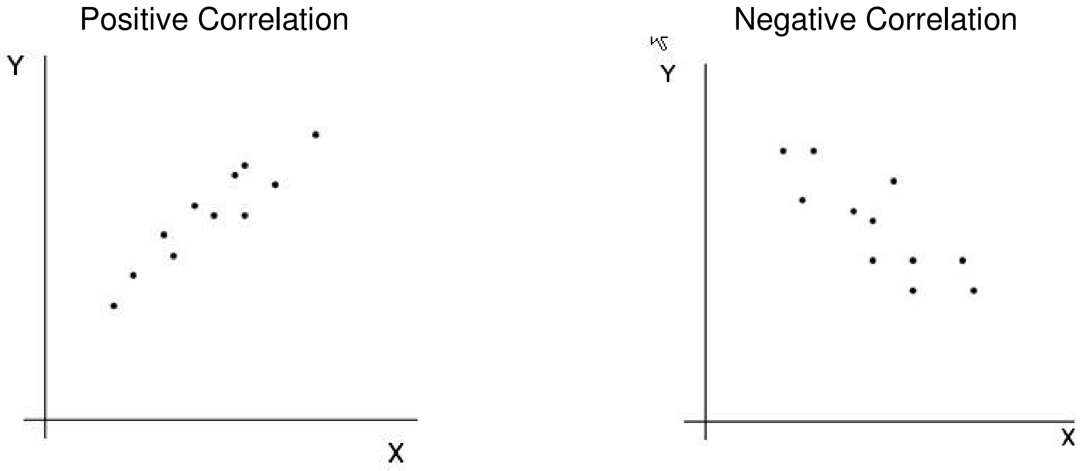


Figure 1. Classical Correlation

The calculation of classical correlation is as follows:

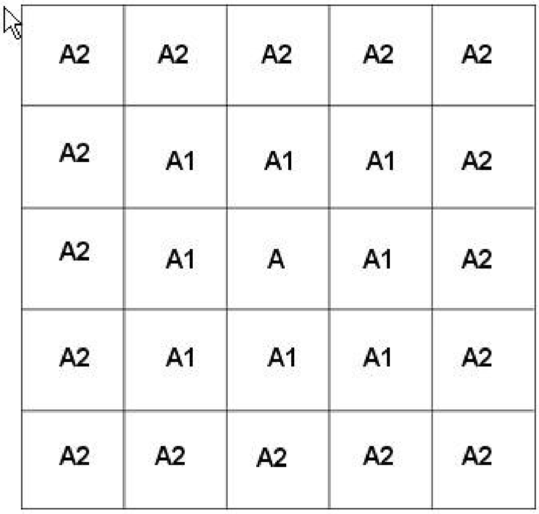
$$r = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n\right)}{S_x S_y}$$

Where *r* is called the correlation coefficient, x_i and y_i are the individual observations, *x* and *y* with the lines on top of them are the means of the two variables, *n* is the sample size, and S_x and S_y are the standard deviations of the two variables. This calculation generates a number that ranges from -1 to +1, with the positive numbers referring to positive correlation and the negative numbers referring to negative correlation. If the value is zero, there is no correlation. There are some links to further reading on the Internet, for those interested, listed in the Additional Resources section below.

3.1.2. Spatial Autocorrelation

Spatial autocorrelation is the spatial equivalent to classical correlation, with the primary difference being that when spatial autocorrelation is calculated the observations are always referring to explicitly spatial observations. Rather than asking what happens to one variable when another variables changes, spatial autocorrelation asks: how similar are neighbouring spatial units to one another with regard to some variable? For example, if one municipality has a high level of income, do its neighbouring municipalities also have a high level of income?

The biggest difference when making the calculation for spatial autocorrelation, compared to classical correlation, is that the neighbours of spatial units must be specified in order to calculate spatial autocorrelation. There are two dominant methods to determine whether or not two spatial units are neighbours: distance and contiguity. Distance, most often used in point pattern analysis, simply states that if one point is within some distance of another point they are neighbours. The distance used depends on the phenomenon being measured and the context of the study. Contiguity is simply a measure of whether or not two areal spatial units are next to each other. For example, Lithuania and Poland share a national border, so they are contiguous. What becomes important is the order of that continuity. The following example provides an explanation.



A2	A2	A2	A2	A2
A2	A1	A1	A1	A2
A2	A1	A	A1	A2
A2	A1	A1	A1	A2
A2	A2	A2	A2	A2

Figure 2. Specifying Spatial Neighbours

The spatial unit of interest is labelled as “A”. All of the other spatial units are labelled according to their level of contiguity with “A”. All those spatial units that share a boundary with “A”, even if that shared boundary is at a corner, are considered to be contiguous if order 1, labelled “A1”. This is called Queen’s contiguity, referring to the movement of the Queen in a game of chess. Those spatial units that have two boundaries between them and “A” are considered contiguous of order 2, labelled “A2”. This may be continued as distant as the researcher considers necessary, but most research uses contiguity of order 1, particularly in descriptive analyses such as spatial autocorrelation.

However, a different measurement of contiguity, Rook’s contiguity, does not consider a spatial unit with a shared boundary only at a corner to be considered contiguous. Most often, Queen’s contiguity is used with vector data because most social-economic-political spatial units are not square or rectangular. Also, using Queen’s contiguity allows for the researcher to not violate Tobler’s First Law of Geography. It is also important to note that other measures of contiguity are possible and these can affect the spatial analysis: neighbours may be based on the length of common shared borders between spatial units or bi-directional weights based on the flow of individuals into and out of adjacent spatial units. For example, to continue with the national examples, the international border shared between Lithuania and Latvia is longer than that of Lithuania and Poland, but Poland has a much larger economy. As such, the measurement of neighbours (and their respective importance) depends on the context. However, for most analyses first order Queen’s contiguity is used.

The calculation of spatial autocorrelation is most often undertaken with Moran’s *I*:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\left(\sum_{i-1}^{n}\left(y_i - \bar{y}\right)^2\right)\left(\sum_{i \neq j}\sum w_{ij}\right)},$$

where the variables representing *n* and *y* are the same as above and w_{ij} represents a matrix that defines the spatial neighbours of the spatial units. The actual formula is rather complicated and not particularly important to understand at this stage because many software programs, including ArcGIS, make these calculations for you. It is important to note that there are some similarities between the formula for Moran’s *I* and for classical autocorrelation. The primary difference is that only one variable is analyzed with spatial autocorrelation. The spatial autocorrelation occurs between the spatial units, not different variables.

Moran’s *I* has ranges from -1 to +1, just as with classical correlation. This is primarily the reason why Moran’s *I* is favoured over Geary’s *C*—this latter measure of spatial autocorrelation does not have the same numerical similarity to classical correlation. For Moran’s *I*, a value of zero means there is no spatial autocorrelation, a value greater than zero means that there is positive spatial autocorrelation, and a value lesser than zero means that there is negative spatial autocorrelation. So, when spatial autocorrelation is positive, the most often phenomenon when investigating social-economic-political spatial units, neighbouring spatial units have similar values. However, when spatial autocorrelation is negative, neighbouring spatial units have dissimilar values.

3.1.3. The Location Quotient

The location quotient is a relatively new descriptive measure in statistics. Traditionally used in economic geography to measure employment or industrial specialization since the 1940s (Isard et al. 1998; Miller et al. 1991), the location quotient has more recently been used in the spatial analysis of crime (Brantingham and Brantingham 1995, 1998). Well-known to many geographers, the location quotient measures the percentage of some activity in a spatial unit relative to the percentage of that same activity in the entire study region. This measurement of the under- or over-representation of an activity has great utility in any type of analysis that may show some regions having a high representation of a particular activity relative to other regions. Given the availability of crime data for Lithuanian municipalities, the location quotient's utility is shown in this context.

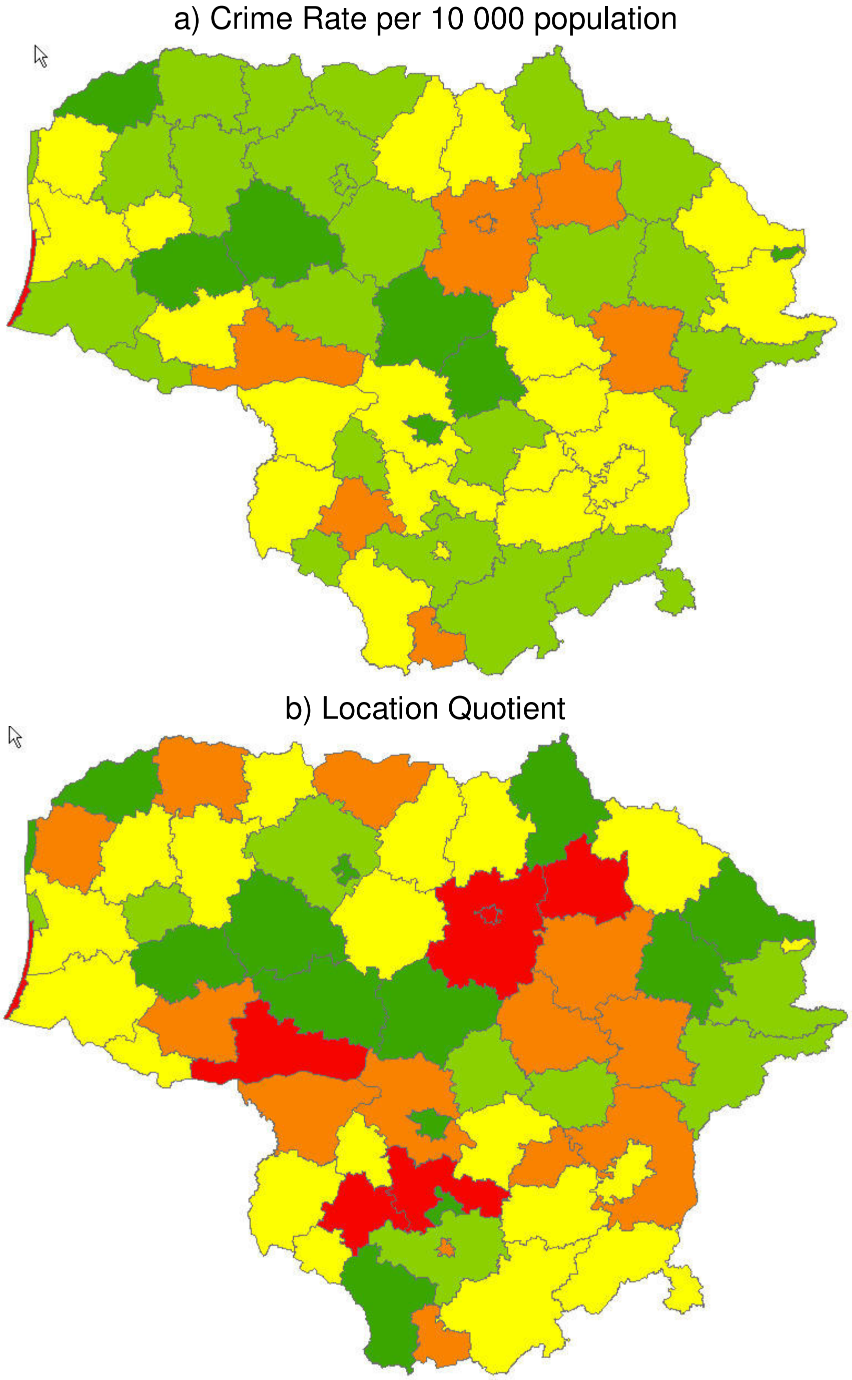
The location quotient is calculated as follows:

$$LQ = \frac{C_{in} / C_{tn}}{\sum_{n=1}^{N} C_{in} \Big/ \sum_{n=1}^{N} C_{tn}}$$

where C_{in} is the count of crime *i* in spatial unit *n*, C_{tn} is the count of all crimes in spatial unit *n*, and *N* is the total number of spatial units. In the context of this example, the location quotient is a ratio of the percentage of a particular type of crime in a municipality of Lithuania relative to the percentage of that same crime in all of Lithuania. If the location quotient is equal to one, the municipality has a proportional share of a particular crime; if the location quotient is greater than one, the municipality has a disproportionately larger share of a particular crime; and if the location quotient is less than one, the municipality has a disproportionately smaller share of a particular crime.

Consider the following example for Lithuanian municipalities. As shown in Figure 3, break and enter is not a significant problem for most Lithuanian municipalities: the legend, omitted here for brevity, goes from green (low crime) to yellow (moderate crime) to red (high crime). There are a few municipalities that have moderately high break and enter rates, but for the most part, it is not a problem in the majority of Lithuanian municipalities.

The location quotient provides a different picture of break and enters. Clearly there are some municipalities that are overrepresented (red) with regard to Lithuania as a whole. Though most of these municipalities are identified with higher break and enter crime rates, now some other municipalities appear to be overrepresented. Moderate overrepresentation (orange) is present for many municipalities that are considered to have low crime rates. This does not mean that these municipalities are not safe, crime is a phenomenon that occurs at all places and all times, but it does show that some municipalities are more popular for break and enters than others. This becomes important information to policymakers because existing police and crime prevention resources should be more focused on break and enters in these municipalities—this does not necessarily mean that more of these resources need to be allocated. The application of the location quotient is applicable to any phenomenon that may specialize in particular places, which is most of human activity.



a) Crime Rate per 10 000 population

b) Location Quotient

Figure 3. Break and Enter, Lithuanian Municipalities

3.2. Modifiable Areal Unit Problem and the Ecological Fallacy

3.2.1. The Modifiable Areal Unit Problem

The modifiable areal unit problem (MAUP) is endemic to all spatial data that has been aggregated into polygons. First outlined by Openshaw (1984), the MAUP is quite simply that the results of an analysis will be sensitive to the spatial units that are used. For example, using municipalities and counties in Lithuania for an analysis will necessarily give different results, perhaps completely opposite results—hence, the MAUP being a problem. There are two effects from the MAUP: the scale effect and the aggregation or zoning effect.

The scale effect was just mentioned above regarding the use of municipalities and counties in the same analysis. Simply put, (spatial) statistical results may change depending on the spatial resolution of the data. Generally speaking, the larger the area of a set of polygons, the more likely they are to be of similar value and, consequently, highly correlated. This is because the larger spatial unit is an average of its constituent parts. Let us assume that there is a high degree of variability in unemployment across Lithuanian municipalities and that each county contains municipalities that have both high and low levels of unemployment. Once the municipalities are aggregated into counties, it is possible that all the counties will appear to have similar levels of unemployment simply because the variability has been averaged out.

The aggregation or zoning effect occurs when statistical results change because of the methodology used to form the areal spatial unit. Again, let us consider the case of Lithuanian counties and municipalities. Each of the current ten counties has a given geographical area. Now suppose that the municipalities in each of the counties were “shuffled” such that each county now had a different set of municipalities, but the geographical area within each county is essentially unchanged—some small changes in the geographical areas may be unavoidable. By maintaining the same geographical size of each county the scale effect is avoided. However, because the municipalities in each of the counties are now different, statistical results comparing counties will also change.

In short, through the alteration of the number and arrangement of the spatial units a completely different set of results may be generated in an analysis.

3.2.2. Minimizing the Modifiable Areal Unit Problem

The MAUP occurs for a very simple reason: most of the spatial units we analyze are defined for us and statistical analyses of these spatial units were not considered. For instance, in Canada, voting districts are most often realigned to favour the governing political party when the voting districts need to be changed. Though this may serve a political purpose, it creates “false” spatial units for subsequent analysis.

The MAUP can only truly be minimized if the spatial units of analysis are specifically the appropriate units for a given analysis. In other words, the MAUP will disappear once we know the “true” spatial units. Of course, such a level of knowledge, if ever attained, will be different for every analysis: meaningful areal units for unemployment will necessarily be different from meaningful areal units for heart attack risk. The difficulty is: what do we do given that we know it is a problem and we do not currently have a solution?

The simple answer to this question is to do as many analyses as possible using different spatial units. If all of the analyses produce the same (or very similar) qualitative and quantitative results, then there is little to worry about—for example, all correlations would be positive and approximately the same magnitude. If all of the analyses produce qualitatively similar results (all have positive correlations), but the magnitude of these results have a range then err on the side of caution and do not make any overstatements regarding the relationship between two variables. If, in the worst case scenario, there is absolutely no consistency in the results (correlations are positive and negative, and of varying magnitudes) then simply do not trust the results. More analysis with different spatial units, variables, or a longer time series is necessary. Any policy based on such results will most probably be ill-informed.

Unfortunately, most analyses only have one choice of spatial units to analyze. When this situation occurs, as it most certainly will, the results must be interpreted with caution noting that they may be sensitive to the particular spatial units analyzed.

3.2.3. The Modifiable Areal Unit Problem and the Ecological Fallacy

The ecological fallacy is usually grouped together with the modifiable areal unit problem, as it is here, but it is a different issue that arises, particularly with geographical data. First noted by Robinson (1950), the ecological fallacy occurs when the results of an analysis at one spatial resolution is inferred to mean something at another spatial resolution. For example, consider a county with two municipalities. The county has an average monthly income of 1000 LTL. From this statistic, one can then infer that the average monthly income in each of the two municipalities is also 1000 LTL. However, in reality, the two average monthly incomes may be 500 and 1500 LTL.

The ecological fallacy then reduces to the fallacy of division: what is true of the whole is also true of its parts. However, one only needs to view his or her own neighbourhood to notice a wide ranging variation in monthly incomes such that any average is not representative of anyone. As such, what is true of the whole is not necessarily true of the part.

What can be done to avoid the ecological fallacy? Unlike the MAUP, the ecological fallacy is easily avoided through careful and thoughtful statements regarding inference. If your analysis only assesses counties, then any inference you make can only involve counties, similarly for municipalities, neighbourhoods and individuals. All too often academic research, government research, and the media commit the ecological fallacy.

One of the troubles with recognizing the ecological fallacy is that it occurs without the use of spatial units, the most obvious and visible form of the ecological fallacy. Groups of people, defined by any social characteristic, may have an average tendency such as high intelligence. This, however, does not mean that every member of that group of people is highly intelligent.

3.3. Pattern Analysis

3.3.1. Uniform, Random, and Clustered Distributions

The spatial patterns of both social and natural phenomena are of prime interest to people working in a variety of different work settings (government, academic, and private enterprise) as well as people working within or from a variety of disciplines—though primarily thought of as the realm of geography, pattern analysis is also prominent in the fields of ecology (or biology, more generally), medicine and public health, and statistics.

In the study of patterns, there are three general forms of spatial distributions: uniform, random, and clustered, all shown in Figure 4. It should be noted that although the patterns represented in Figure 4 are in point form and that the discussion regarding pattern analysis presented here is also in regard to points, there are methods of pattern analysis for spatial units represented as polygons. However, pattern analysis is traditionally in the realm of points so point pattern analysis is dealt with in this section of the module. Cluster analysis using areal spatial units is undertaken in Topic 5, below.

Figure 4a represents a uniform distribution of points. Sometimes referred to as a dispersed point pattern, a uniformly distributed point pattern exhibits a systematic spatial process that dictates the location of each point. In other words, if you know the location of one point and know the nature of the spatial process, you can perfectly predict where all of the other points in the pattern are going to be. In the case of Figure 4a, each point is one unit up (or down) and to the right (or left) of any point in the distribution. Such spatial point patterns rarely occur naturally, most often the result of human engineering. A prime example of such a pattern would be traffic lights at the intersections of a rectangular street network, most often referred to as a Manhattan grid. On such a street network, street are set out to be at equal intervals in both directions (north-south and east-west) so traffic lights at the intersections will produce a uniformly distributed spatial point pattern.

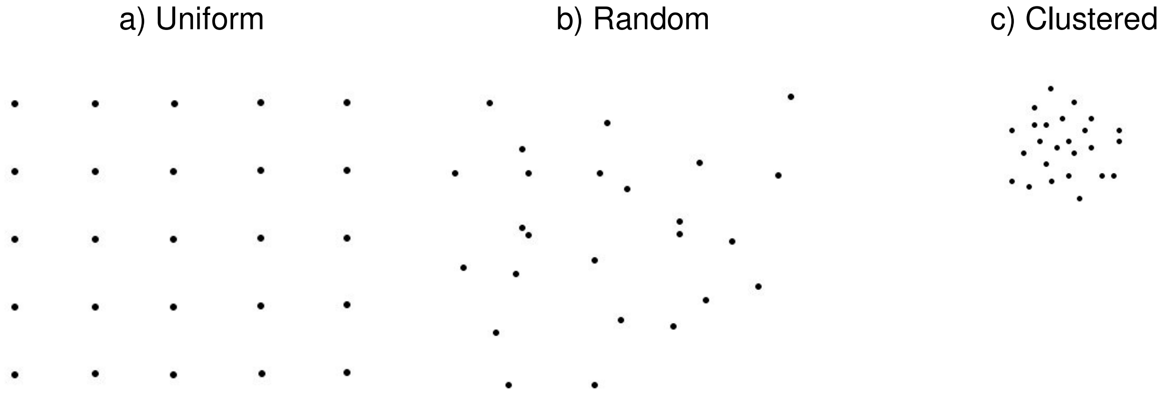


Figure 4. Uniform, Random, and Clustered Point Patterns

The randomly distributed spatial point pattern, Figure 4b, exhibits to dominant trend. With such a spatial point pattern, knowing where one point is located provides no indication of where other points are located—you are just as likely to find another point blind-folded with your finger than you are with any mathematical calculations. These spatial point pattern distributions are the products of what is called a Poisson process. This allows for easy testing of this hypothetical distribution.

Lastly, shown in Figure 4c, there is the clustered spatial point pattern distribution. With this spatial distribution, the density of points varies to a high degree because some areas have the majority of the points and other areas have none of the points. As such, the prediction of other points can take place because if you are at a location that has a point, other points are likely to be close by, and if you are at a location with no points, there are likely no other points close by. Such a clustering of points may represent, for example, the location of retail outlets in a central business district or the outbreak of disease related to an environmental toxin with a fixed source.

Two forms of pattern analysis are outlined below. The first, and most common, is nearest neighbour analysis that uses the distances between points to determine whether or not a spatial point pattern has a particular distribution. The second, quadrat analysis, places the points within areal units to assess the form of spatial distribution.

It should be noted that in much of this literature a “point” means any location on a map and an “event” is a point on the map where something has occurred and is represented by a dot: a case of disease, a crime, or a traffic accident. Though for the remainder of this topic the term “point” is used, this difference become important when discussing edge effects in the following topic.

3.3.2. Nearest Neighbour Analysis

A nearest neighbour analysis measures the distance between a point and its “nearest neighbour”, calculates the average nearest neighbour distance for all points in the data set, and then compares this average nearest neighbour distance to an expected average distance, such as a random (Poisson) probability distribution. The distances between points and their nearest neighbours are measured using Euclidean distance, or straight line distance. This measurement of distance has an advantage as well as limitations. The advantage is that the calculations are simple, as shown below, using an equal interval grid and a little geometry. However, the limitation is that Euclidean distance is not always realistic for human settlement patterns.

For example, as shown in Figure 5, there may be two dead-end streets (cul-de-sacs) that are separated by a park. For simplicity, consider that the residences on either side of the park are not even able to see each other because of trees, etc., but that the residences are actually quite close. Using Euclidean distance to measure the distance between two points may place the two residences on either side of the park as nearest neighbours when the actual network distance between these two residences (a more realistic measure of distance in this situation) may in fact be quite long. Of course, these network distances may be used in a nearest neighbour analysis. The essence of the analysis remains the same, but the calculation of distance becomes more computationally intense. The standard Euclidean distance is used in the example below.

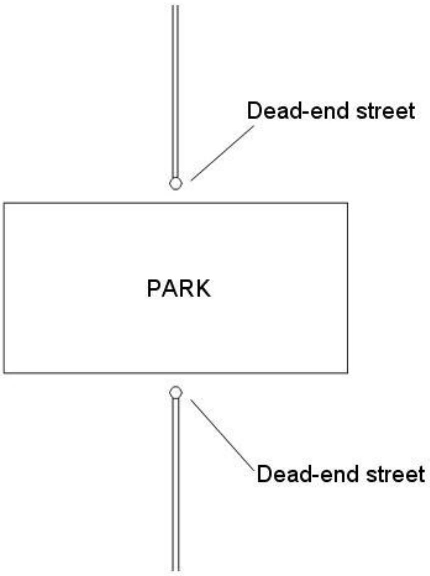


Figure 5. The Problem with Euclidean Distance

In order to ascertain which type of spatial pattern is apparent in a spatial point pattern, a number of simple calculations need to be undertaken. First, the average nearest neighbour distance:

$$NND^* = \frac{\sum_{i=1}^{n} NND_i}{n},$$

where NND_i is the nearest neighbour distance for point i an n is the number of points. Note that if the spatial point pattern is perfectly clustered, meaning that all points are in one location, then the value of NND^* is zero. Second, the expected average nearest neighbour distance for a random spatial distribution needs to be calculated:

$$NND^R = \frac{1}{2\sqrt{Density}},$$

where *Density* is the number of points divided by the area under study. These two variables are then used to calculate the standardized nearest neighbour index (*R*):

$$R = \frac{NND^*}{NND^R}$$

in order to determine the nature of the spatial point pattern. A property of the standardized nearest neighbour index, *R*, is that it has a defined range (0 – 2.149) that allows for the following classification:

Table 1. Classification Scheme for the Standardized Nearest Neighbour Index, *R*

Perfectly clustered	More clustered than random	Random	More dispersed than random	Perfectly dispersed
R = 0	*R* = 0.5	*R* = 1	*R* = 1.5	*R* = 2.149

Most often, a descriptive classification is not sufficient for an analysis, so inferential statistical testing is necessary to determine the nature of the spatial point pattern. In such a test, the null hypothesis is that the spatial point pattern is random, using the following test statistic:

$$Z_{NND} = \frac{NND^{*} - NND^{R}}{\sigma_{NND}}, \; where \; \sigma_{NND} = \frac{0.26136}{\sqrt{n(Density)}}.$$

Using a one-tailed statistical test and a probability value of 10 percent *p*-value = 0.10), the critical value for Z_{NND} is approximately 1.28. Therefore, if Z_{NND} > 1.28 the spatial point pattern is significantly more dispersed than random, and if Z_{NND} < -1.28 the spatial point pattern is significantly more clustered than random, and if -1.28 < Z_{NND} < 1.28 the spatial point pattern is insignificantly different from a random Poisson spatial process. The following example illustrates an application of nearest neighbourhood analysis and its statistical analysis.

Figure 6 shows the locations of five (5) points in a study area that appear to be in somewhat of a clustered spatial pattern. The area for this study area is 100: 10 * 10 units.

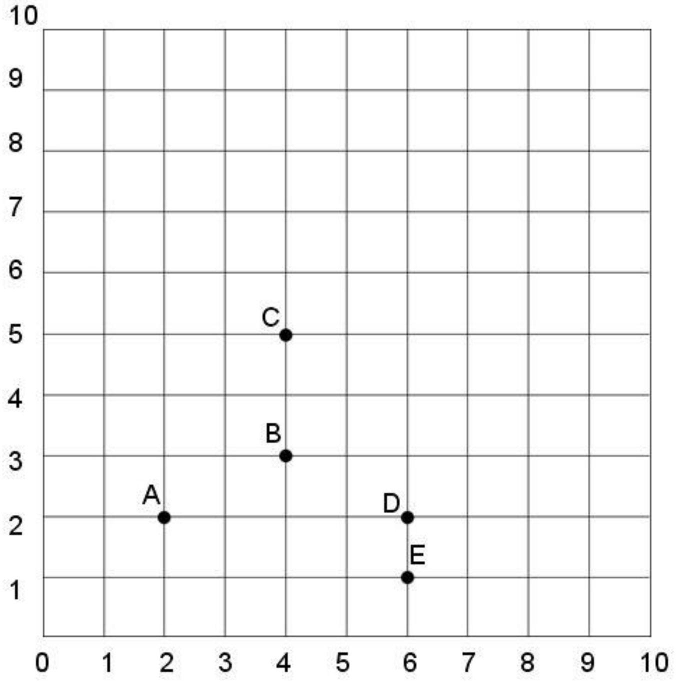


Figure 6. Location of Points for Nearest Neighbour Analysis

The points, their spatial coordinates, nearest neighbour, and nearest neighbour distance area all reported in Table 2.

Table 2. Coordinates and Nearest Neighbour (NN) Information

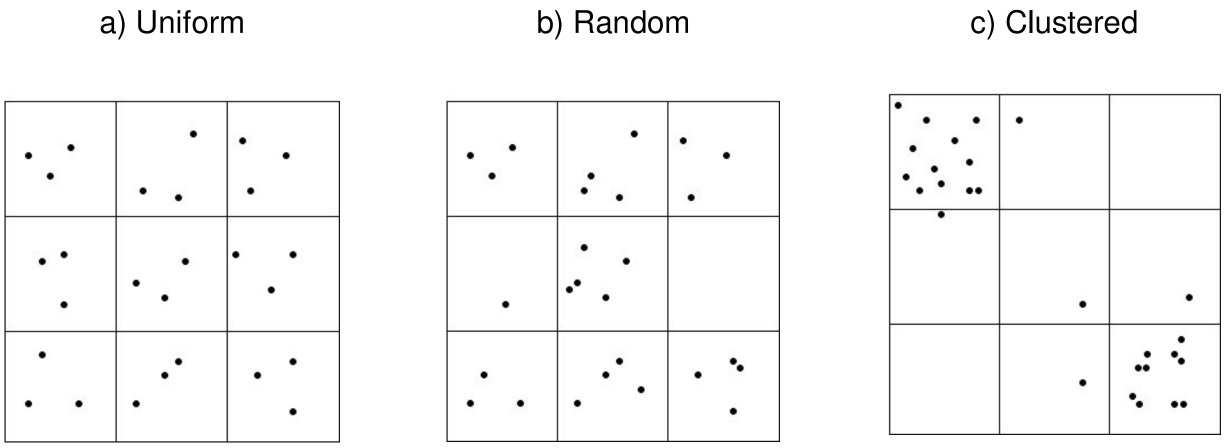
Point	X	Y	*NN*	*NND*
A	2	2	B	2.24
B	4	3	C	2
C	4	5	B	2
D	6	2	E	1
E	6	1	D	1

From these numbers, the remaining variables are easily shown to be: NND^* = (8.24)/5 = 1.648, NND^R = 1 / (2*sqrt(*Density*)) = 1/(2*sqrt(5/100)) = 1/(2*sqrt(0.05)) = 2.236, and *R* = 1.648 / 2.236 = 0.737. Consistent with the observation above, this spatial point pattern is more clustered than random. The question now is whether or not that spatial clustering is statistically significant. This question is answered through the calculation of Z_{NND} = (1.648 – 2.236) / 0.523 = -1.13. Consequently, the null hypothesis of a spatially random distribution cannot be rejected with this statistical test. Therefore, though the distribution above appears to be more clustered than random, this is not the case statistically.

3.3.3. Quadrat Analysis

Quadrat analysis is an alternative methodology used in order to determine the nature of a spatial point pattern distribution. However, rather than focussing on the distances between points, the frequency or number of points within a given area is used. In this form of pattern analysis a set of quadrats, most often square cells are superimposed over the study area and the number of points within each quadrat is counted. The nature of the spatial point pattern distribution is then determined through an analysis of the frequency of the counts in each of the square cells.

Generally speaking, if each of the cells contains the same number of points (no variability), then the spatial point pattern distribution is considered to be uniform, or dispersed; if the cells contain very different numbers of points (large variability), then the spatial point pattern distribution is considered to be clustered; and if there is a moderate amount of variability in the number of points in each of the square cells, the spatial point pattern is considered to be random. These differences are shown in Figure 7.



a) Uniform

b) Random

c) Clustered

Figure 7. Quadrat Analysis of Spatial Point Patterns

Though the different types of variability just above are quite present in Figures 7a – 7c, as with the nearest neighbour analysis discussed in the previous section the spatial point pattern distribution must be determined statistically if any inference is to be drawn from an analysis.

Though similar in form, the statistical test in quadrat analysis is simpler to execute than in nearest neighbour analysis. The only statistics needed are the mean (or average) cell frequency and the variance of the cell frequency, both classical descriptive statistics. Using these two variables, the following statistic is calculated:

$$VMR = \frac{\text{Variance of the cell frequencies}}{\text{Mean of the cell frequencies}},$$

where *VMR* is an acronym for variance-mean ratio. If the *VMR* is equal to one, the spatial point pattern distribution is considered to be random; if the *VMR* is greater than one, the spatial point pattern distribution is considered to be more clustered than random; and if the *VMR* is less than one, the spatial point pattern distribution is considered to be more dispersed than random. The test statistics is a chi-square:

$$\chi^2 = VMR(m-1),$$

where *m* is equal to the number of cells and the null hypothesis is a random spatial distribution. Given that the variance of the number of points in each cell of Figure 7a is zero, as an example, the statistical tests will be carried out using Figures 7b and 7c. The necessary variables are provided in Table 3.

Table 3. Required Information for Quadrat Analysis of Spatial Point Patterns

	Figure 7b	Figure 7c
Variance of cell frequencies	2.5	21
Mean of cell frequencies	3	3
VMR	0.833	7
χ_2 statistic	6.67	56

As shown in Table 3, Figure 7b shows definite signs of being a random spatial distribution, but does show some evidence of being more dispersed than random. The results for Figure 7c, however, show a clear degree of clustering with a chi-square statistic that easily rejects the null hypothesis of a random spatial distribution—the relevant degrees of freedom for the critical values is seven, nine observations minus the two parameters calculated (mean and variance). With a chi-square statistic of 6.67, the spatial distribution of points in Figure 7b is not significantly different from a random spatial distribution. As such, each of the examples shown in Figure 7 correspond statistically with the appearances of their spatial point pattern distribution.

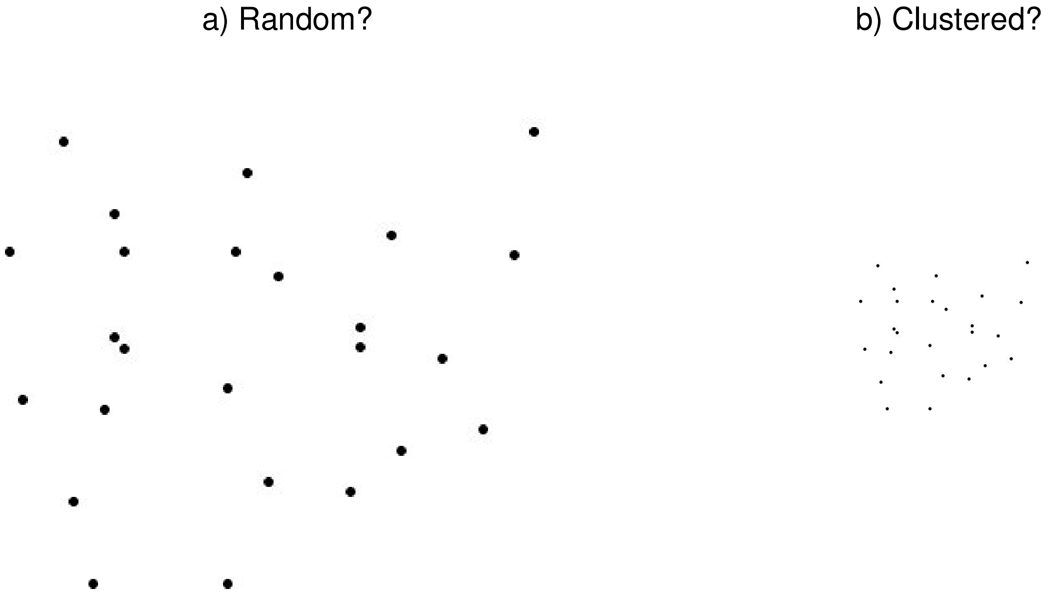
3.3.4. Limitations of Pattern Analysis

As with any form of statistical analysis, spatial or classical, pattern analysis has its limitations. The primary limitation of these analyses is that they do not tell the researcher *why* there is or is not a particular spatial distribution. Rather, these tests simply allow us to know what the particular spatial distribution is or is not.

A further difficulty that is spatial is in regard to the size of the study area. Consider the spatial point patterns displayed in Figure 8. Figure 8a shows what is perceived to be a random spatial point pattern whereas Figure 8b shows what is perceived to be a clustered spatial point pattern. The difficulty that is present is that these are precisely the same spatial point patterns shown at different scales: Figure 8a is at a cartographically larger scale (smaller area) than Figure 8b.

Suppose you were investigating the impacts of an environmental toxin that is released into the air from an industrial plant at the centre of the spatial point pattern. If one were to analyze the effect of this environmental toxin and a corresponding disease at a scale equivalent to that of Figure 8a then the industrial plant that produces the environmental toxin would not be viewed as problematic. However, if the scale of analysis was changed to that of 8b, and consider this scale appropriate for the problem at hand, then the industrial plant may be shown to be the cause of the corresponding disease.

The problem is knowing the appropriate scale of analysis. Sometimes this may be relatively easy: there may be a scientific consensus regarding the problematic concentrations of the environmental toxin in the atmosphere. The concentrations of the environmental toxin could simply be measured at various distances from the industrial plant and have the study area circumscribed based on that information. Not all such problems are resolved so “easily”, however. The most appropriate action to undertake in such a situation would be similar to that for the modifiable areal unit problem, in general: undertake the analysis at a number of scales that may be applicable and look for consistency in the results.



a) Random?

b) Clustered?

Figure 8. The Difficulty with Scale in Pattern Analysis

3.4. Edge Effects

3.4.1. What is an Edge and When Does it Have an Effect?

An edge is simply defined as the boundary of the spatial area under analysis. That is, the outer boundary of all the spatial units of analysis. For example, if an analysis was being undertaken using Lithuania municipalities as the spatial units, the national border surrounding Lithuania is the edge.

Though technically all spatial analysis, or any analysis for that matter, does have an edge, edges pose a problem in particular for point pattern analysis. This is because, as discussed above, the distances between two events or a randomly selected point and an event are used to calculate descriptive statistics and used in inferential analysis. As such, distance-based point pattern analysis methods suffer from the presence of edge effects.

In some cases, the edge does not pose a problem in the analysis. For example, a portion of the Lithuanian border is on the Baltic Sea. Consequently, we would not expect there to be any events that are outside of the edge—there may be some people, and corresponding events, that live just of the coast of Lithuania, but these events can easily be placed on a land region close to the water for an analysis.

The problems occur with edges when the study area is part of a larger geographical region that the underlying spatial process operates within. Consequently, any events occurring outside of the study area but within the larger geographical region may interact with the events within the study area (Diggle 2003). This is the edge effect. Consider the example shown in Figure 9.

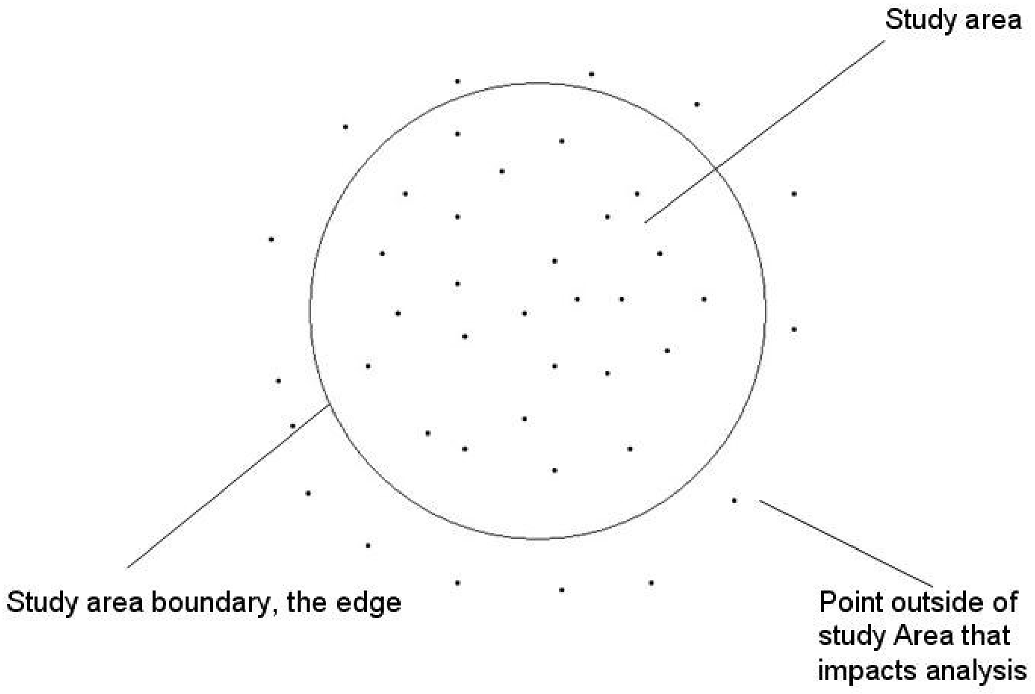

Study area
Study area boundary, the edge
Point outside of study Area that impacts analysis

Figure 9. The Presence of an Edge Effect

Figure 9 shows the study area, its boundary (edge), and some events outside of the study area that are related to the events inside of the study area. Such a situation may occur if someone is able to obtain data for one municipality on the incidence of a disease, but not for any bordering municipalities.

An excellent example of such a situation would be if there is an industrial plant located within one municipality (the study area) that creates pollution that travels through the atmosphere. The pollution from this industrial plant is not thought to impact neighbouring municipalities because of distance, or neighbouring municipalities may not cooperate with the research project because of jurisdictional issues, for example. Regardless, the events that are present outside of the study area really should be included in the analysis of the effects stemming from the pollution of the industrial plant.

3.4.2. The Impact an Edge Can Have on an Analysis

The impact that an edge may have on the analysis is a biased result. This means that the results may not be truly representative of the study. Rather, because important information was not used in the analysis, the results may be misleading.

It should be mentioned that this situation is the case in all statistical analyses. The relevant information is not always available, so important information is not included. However, as with all statistical analyses, there are methods to deal with these shortcomings stemming from the lack of available data.

3.4.3. Methods of Correcting for the Edge Effect

There are a number of methods that allow the researcher to account for the presence of edge effects. The most frequently used methods are: the adjustment method, toroidal wrapping, and the use of a buffer zone (sometimes called a guard area) (Diggle 2003).

Of these three methods for accounting for the presence of edge effects, the most common method of dealing with this problem is to create a buffer zone. A buffer zone is an area that is created a certain distance from the outer edge of the study area. This is shown in Figure 10.

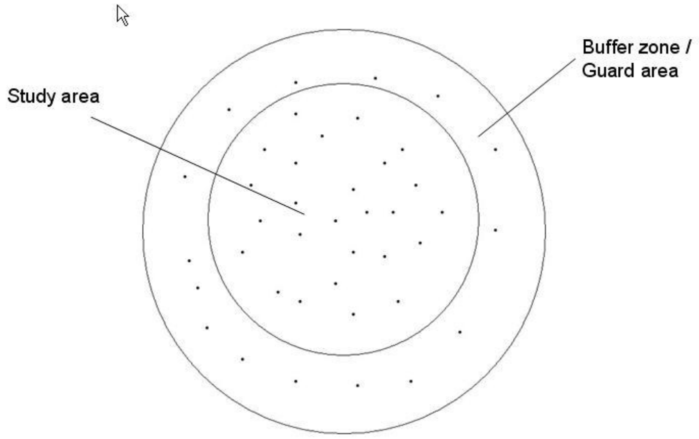


Figure 10. Buffer Zone Correction for Edge Effects

In the analysis of the point pattern, distances are not calculated for points within the buffer zone, but events in the buffer zone are allowed to be used for the calculations involving other events

(Bailey and Gatrell 1995). Effectively, what this procedure does is allow the researcher to simulate having those points outside of the study area and then undertake the analysis without concern for imposing bias on the results. As such, the use of the buffer method, as well as the adjustment method, eliminates the bias from the edge effects with the cost of increased variance from decreased data (Diggle 2003).

Regarding the size of the buffer zone, there is no specific rule used in order to determine the distance of the buffer zone from the outside of the study area. Rather, the process of creating a buffer zone for the analysis may be applied repeatedly as a sensitivity analysis.

3.5. Density Estimation and Hot Spot Mapping

3.5.1. The Concept of Density

One of the most common (and popular) methods for visualizing point data as a continuous surface is kernel density estimation. Simply put, the kernel density estimation method creates a relatively smooth surface from the point events in a data set that varies according to the density of those point events across the study area. As such, this density can be viewed as an intensity of the phenomenon under study.

One of the applications of kernel density estimation is in the identification of hot spots, usually in the case of some reported crime. Rather than trying to discern patterns on a map that potentially has tens of thousands of events—in many cases there are so many crimes (think of personal theft or automotive theft) that the entire map is covered in point events—kernel density estimation shows where the areas with the most points, the "hottest", are located. This spatial statistical method has common availability as well as visual appeal.

3.5.2. Kernel Density Estimation

There are three basic steps in kernel density estimation. Consider the study area in Figure 11: the entire study area is contained within the parallelogram; the dots within the parallelogram represent the locations within the study area that have experienced a particular event; and the entire study space, including those areas that have experienced the events, are points. In other words, as described above, a point can represent any location within the study area and an event represents a location where something has occurred.

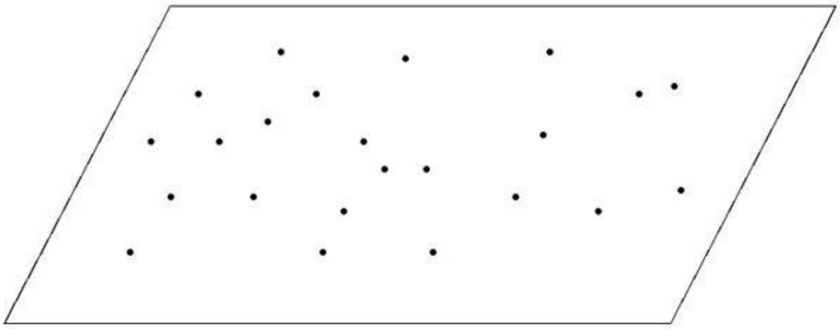


Figure 11. Kernel Density Study Area

The first step to calculate a kernel density for this study area is to superimpose a fine rectangular grid over the entire study are whether or not there are any points in all locations, shown in Figure 12. One of the good features of kernel density estimation is that the user is able to specify the grid cell size. This is a good feature because the grid cell size will likely vary from application to application and having the user to specify the grid cell size allows the user to perform a sensitivity analysis.

A sensitivity analysis may be particularly important in many applications because there may be no consensus on the grid cell size, but a range in grid cell sizes that different people have used. The user can then perform the kernel density estimation using a variety of grid cell sizes within that range in order to uncover any potential changes in the results. Ratcliffe (1999) has also proposed a method for determining the grid cell size when no other method is available: draw a rectangle around the study area that is small as it can be without entering inside of the study area, measure the shorter of the sides of that rectangle (in metres), and then divide by one hundred and fifty.

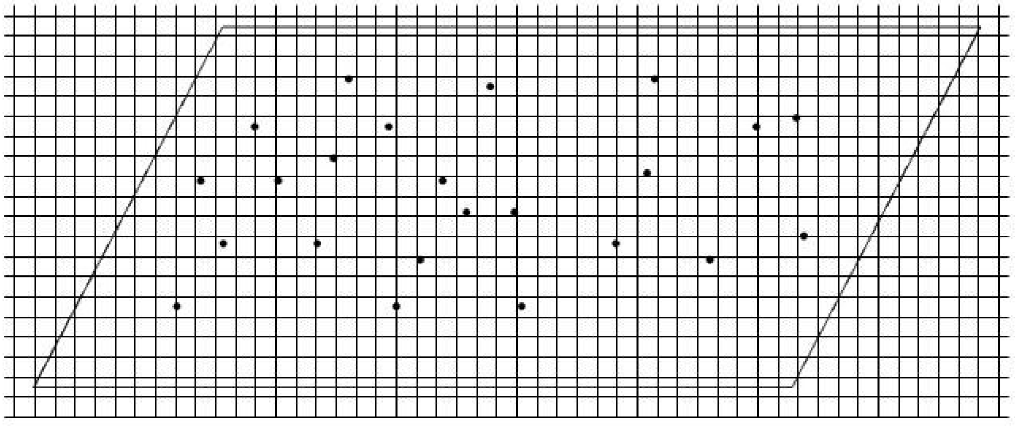


Figure 12. Kernel Density Grid

The second step is to calculate a function (the intensity function) for each of the cells in the study area. At this step, each grid cell is given a point location at its centre (the centroid) and a circle is drawn around that point with a specified radius. The radius of this circle is referred to as the bandwidth of the kernel density estimation. The more events that occur within that circle, the higher the density value for that particular grid. Additionally, all events are not given the same weight in the calculation of the density: each event is "weighted" by its distance away from the centroid, such that points closer to the centroid receive a higher weight and contribute proportionately more to the calculation of that cell's density value. This is shown in Figure 13.

The primary issue is to determine the appropriate radius, or bandwidth, and is the portion of the kernel density calculation that is most sensitive to changes. As with the grid cell size, discussed above, the appropriate bandwidth should be set according to the particular application. Other research should be consulted to determine the appropriate bandwidth, or the appropriate range to undertake a sensitivity analysis.

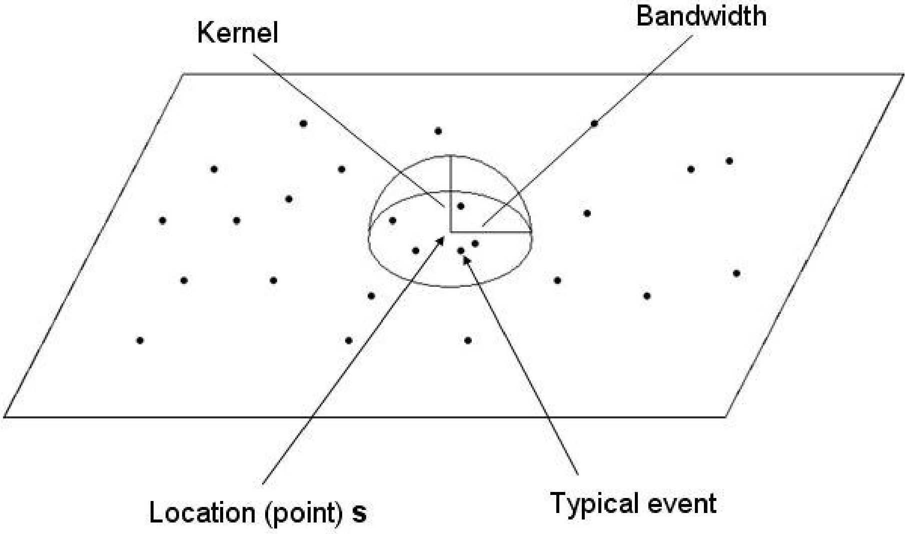


Figure 13. Kernel Density Calculation

The third step is to map the output from the kernel density estimation. Given that the estimated values represent a range of values, a graduated colour ramp is best for visualization: white or pink for low density and dark red for high density, for example.

A nice feature of the kernel density estimation is that the values generated for each of the grid cells is in a meaningful unit for describing most spatial point distribution—the number of events per square kilometre, for example. Consequently, the values in each grid cell can be compared in a meaningful way: grid cell 122 has a value of 100 and grid cell 123 has a value of 50, so grid cell 122 is twice as dense as grid cell 123.

Let’s consider a couple of simple, short examples. First, there is the equation for a kernel density. This particular equation is for a quartic kernel:

$$\lambda = \sum_{d_i \le \tau} \left(3 \big/ \pi r^2 \right) \left(1 - \frac{d_i^2}{r^2} \right)^2$$

where λ is the intensity value (kernel), π is equal to 3.14, *r* is the radius (bandwidth), and d_i is the distance between the centroid and event *i* within the bandwidth. Once the bandwidth is chosen, $3 \big/ \pi r^2$ is just a constant so it can be ignored in these examples. All that is left: $\left(1 - \frac{h_i^2}{\tau^2} \right)^2$, and let us also ignore the squared term on the outside of the brackets, again for simplicity. Set *r* = 10. Now let us consider two possible scenarios.

First there will be four points within the bandwidth: two of them will be two units from the centroid and two will be three units from the centroid. In the second scenario there will be two points within the bandwidth, both nine units from the centroid. The calculations for both examples are below.

In the first example, the value of λ is shown to be 3.74 and in the second example the value of λ is shown to be 0.38—of course, the actual values of λ would be different if the complete formula was used.

Example 1:

$$\begin{aligned}\lambda &= \left(1-\frac{4}{100}\right)+\left(1-\frac{4}{100}\right)+\left(1-\frac{9}{100}\right)+\left(1-\frac{9}{100}\right)\\ &= \frac{96}{100}+\frac{96}{100}+\frac{91}{100}+\frac{91}{100}\\ &= \frac{374}{100}=3.74\end{aligned}$$

Example 2:

$$\begin{aligned}\lambda &= \left(1-\frac{81}{100}\right)+\left(1-\frac{81}{100}\right)\\ &= \frac{19}{100}+\frac{19}{100}\\ &= \frac{38}{100}=0.38\end{aligned}$$

The first example has a much higher value of λ for 2 reasons. First, there are more events within the bandwidth, and second, the events that are within the bandwidth in example 1 are closer to the centroid. Consequently, the value of λ depends on both of these factors, so it is possible for one grid cell value to be higher than another even if the latter grid cell has more events within its bandwidth. To provide a trivial example, if there are no events within the bandwidth, $\lambda = 0$.

3.5.3. Limitations of Density Estimation

There are two primary limitations with the usage of kernel density estimation: a methodological concern and an accurate representation of events concern.

The methodological concern is probably the "worst" criticism of the use of kernel density estimates. There are two types of point data. The first type of point data represents discrete events such as a criminal occurrence, a traffic accident, and a case of disease. The second type of point data represents a measurement of a continuous surface such as temperature or pollution levels. Given that the second type of point data actually represent a continuous surface, it makes sense to transform the points into a continuous surface for analysis—for budgetary reasons, all locations cannot be measured to calculate temperatures across a country, for example.

The first type of point data, however, is not a continuous phenomenon. Consequently, there are many people who state that you cannot meaningfully transform discrete point events into a continuous surface because the underlying data themselves are not continuous. Those who map criminal occurrences are particularly guilty of turning discrete events into continuous surfaces.

However, if one considers the continuous surface to be a measure of risk, this criticism becomes less problematic. Whether or not a crime occurs at one intersection or another two intersections away may simply be random, so the risk of criminal victimization may be equal at both intersections. The difficulty arises when there is an intersection (and a small area around it), for example, that never has any crime but the kernel density estimation procedure provides that location with a risk of crime. Consequently, as with any statistical analysis, caution must be used when interpreting kernel density results.

The second limitation is one that is relatively easy to resolve. The limitations arise because many researchers who calculate kernel density estimates use only the event data for that calculation. The resulting map will indeed show where the hotpots of that activity are, but those hotspots may be misleading.

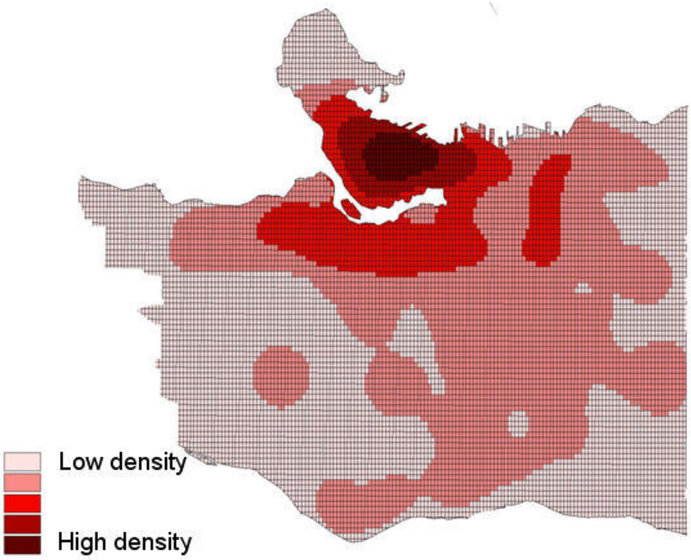


Figure 14. Single Kernel Density

Figure 14 shows a kernel density map calculated using only the event data, which makes this a single kernel density—a single set of point data are used in the calculations. Clearly there is a hotspot, high density, close to the centre of the peninsula at the top of this map—this is not a Lithuanian example. However, one must ask the question: is there something special about this location or is it a hotspot simply because there is a large population at risk in this area?

Crime rates, disease rates, traffic accident rates, etc. are calculated because if one wants to have an accurate measure of risk, one must consider the population at risk of crime, disease, and traffic accidents. Municipalities that have larger populations do not necessarily have larger volumes of crime, disease, and traffic accidents. What is more important to know is whether or not there is more crime, disease, or traffic accidents after controlling for the population at risk, usually the population of the area under study—traffic volume in the case of traffic accidents. So in Figure 14, is there really a high volume of some activity or are there simply more people in that area?

In order to address this limitation, there is a kernel density estimation technique called the dual kernel density. This dual kernel density needs two sets of point data: the event data of interest, and the population at risk of encountering that event. The difference in the resulting maps can be astounding. This is shown when considering both Figures 14 and 15.

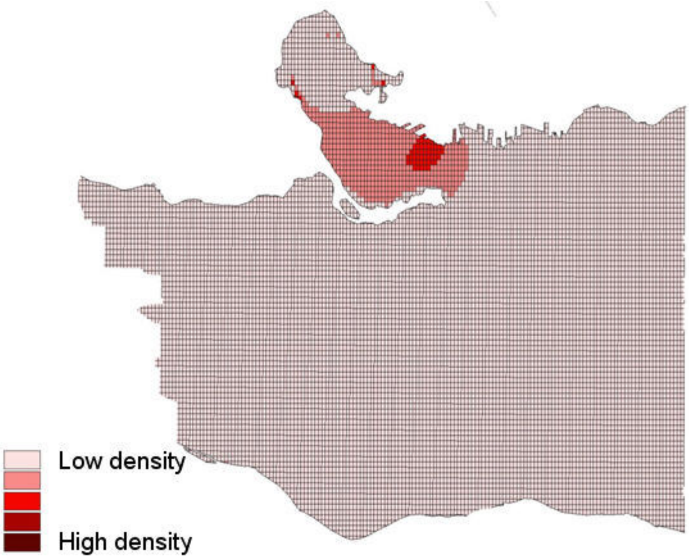


Figure 15. Dual Kernel Density

Using a dual kernel density, Figure 15 shows that there is not nearly as much high density as in Figure 14. The highest density is in the same general area on both maps, but clearly a different picture of the phenomenon has emerged. Consequently, one of the first questions one should ask when shown any type of density map is: is this a single or dual kernel density map? If the person responds with an answer of “single” or “I don’t know”, then the results of the map should be interpreted with caution because there has been no account for the population at risk.

3.6. Local Spatial Statistics

3.6.1. The Limitations of Global Spatial Statistics

Thus far, we have investigated spatial statistics that are now classified as "global". Indeed, until twenty years ago all spatial statistics could be classified in this way—it has only been approximately twelve years for areal spatial analysis. What the term global means is that one set of statistical results is generated from an entire set of data and this set of results is meant to depict the entire study area. As such, this one set of statistical results is said to represent an average set of results for all spatial units in the analysis. However, if the relationships being examined actually vary within the study area, then the global spatial statistics are actually of little value. Though a bit more extreme, a global statistic that is of little value is the average temperature in Russia. With Russia being such a large country, having one statistic to represent all of its regions provides us with little useful information. Consider the following example.

Suppose a Moran's *I* statistic was calculated using Lithuanian municipalities for automotive theft. The result would be Moran's *I* = 0.029. This result shows a very small, and statistically insignificant, degree of positive spatial autocorrelation. However, if you were to calculate a local version of Moran's *I*, described in detail below, you would find that the local Moran's *I* results range from -4.235 to 2.645, having an average of 0.134—each of the sixty municipalities in Lithuania has a local Moran's *I* statistic. Additionally, four of these local Moran's *I* statistics are significantly different from zero, contrary to the global Moran's *I* result.

Because of this known variation within most study areas there has been a relatively recent movement within spatial statistics that focuses on the differences in statistical results across space, rather than simply assuming that these differences do not exist. Indeed, this has been one of the criticisms of quantitative geography, that is searches for global generalities rather than considering the local. As such this movement in spatial statistics, or spatial analysis more generally, has been coined as local analysis, local modelling, and local spatial statistics.

3.6.2. The Utility of Local Spatial Statistics

There are a number of factors that show the utility of local spatial statistics over and above the use of global spatial statistics—it should be noted, however, that local spatial statistics are best used in conjunction with global spatial statistics rather than simply replacing them. From the example given above, local spatial statistics are multivalued, vary across space, and emphasize the differences across space within an entire study region.

Of course this is a strong case for the use of local spatial statistics, but there are more reasons why they are fruitful to an analysis. Local spatial statistics are mappable. As stated above in the Lithuanian municipality example, each areal spatial unit has a local spatial statistic associated with it. Consequently, this is simply yet another variable in the data set that can be mapped. Because this variable can be mapped, it is considered to be GIS-friendly. And lastly, local spatial statistics can be used to search for hotspots: think of two or more spatial units that have statistically significant positive local spatial autocorrelation.

The first well-known development in local spatial analysis is in regard to spatial point patterns, the geographical analysis machine of Openshaw et al. (1987). The geographical analysis machine, though extended by a number of researchers, is a computationally intense algorithm that we will not cover in any detail here. However, its general formulation is worthy of note because it is consistent across most local spatial statistical analyses.

First, there is a method of defining sub-regions within the entire study area; second, there is a method of describing the spatial point pattern within each of these sub-regions; third, there is a methodology for identifying those spatial point patterns that are significantly different from the average; and fourth, there is a method of mapping the results. Though there are, at times, substantial differences between the geographical analysis machine and more recent local spatial statistical methodologies, these characteristics are present, in some form, in all local spatial statistical methods.

3.6.3. Local Moran's *I*

Anselin (1995) introduced a local variation of the global Moran's *I* that he called local Moran's *I*. More generally, Anselin (1995) described a class of local spatial statistics called local indicators of spatial association (LISA). These local indicators of spatial association have a property that separate them from other local spatial statistical measures: a local indicator of spatial association must provide a measure of the extent to which there is a spatial clustering of similar values around each spatial unit of analysis and the sum of all the local indicators of spatial association must be proportional to its corresponding global indicator of spatial association—the global Moran's *I* in this case. That is to say that there must be some relationship between the local and global indicators of spatial association that can be defined, otherwise there should not be any similarity in the names of the local and global statistics because it would be misleading.

The local Moran's *I* is calculated as follows:

$$Local\ Moran's\ I = \frac{(y_i - \bar{y})\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_j - \bar{y})}{\left(\sum_{i-1}^{n}(y_i - \bar{y})^2\right)\Big/ n}$$

where all of the variables listed here are defined above in the section covering spatial autocorrelation. It should be noted the high degree of similarity between this local version of Moran's *I* and the global version of Moran's *I* that allows this spatial statistical measure to be classified as a local indicator of spatial association. Though this formula is complex, as with the global indicator of spatial association, the local Moran's *I* is easily computed within the ArcGIS environment using ArcToolbox.

3.6.4. Local Cluster Analysis

Local cluster analysis, as the term indicates, searches for clusters of spatial entities at a local level. Unlike Topic 3, however, the examples provided in this Topic use areal spatial units to form clusters. Specifically, it is a search for local clusters of significant local Moran's *I* values.

There are a number of steps to undertake in order to perform a local cluster analysis using the local Moran's *I* statistic. First, the local Moran's *I* statistic needs to be calculated. Using ArcGIS, this produces two new variable columns called LMiContig and LMzContig. LMiContig is the local Moran's *I* statistic value and LMzContig is the z-statistic associated with each LMi value—Contig is in reference to the use of a contiguity spatial weights matrix, discussed above.

The first variable, LMiContig, is shown in Figure 16. The degrees of green (light to dark) represent increasing negative values of the local Moran's *I* statistic and the degrees of red (light to dark) represent the increasing positive values of the local Moran's *I* statistic for Lithuanian municipalities. The white municipalities are spatial units with local Moran's *I* statistics that are close to zero. As should be rather evident, there is a large degree of variation in the local Moran's *I* statistics across Lithuanian municipalities with many of the local Moran's *I* statistics being negative, unlike the statistically insignificant positive global spatial autocorrelation. However, despite this geographical variation, it is also important to identify the significance of these local Moran's *I* statistics as well as determine the nature of any local clusters.

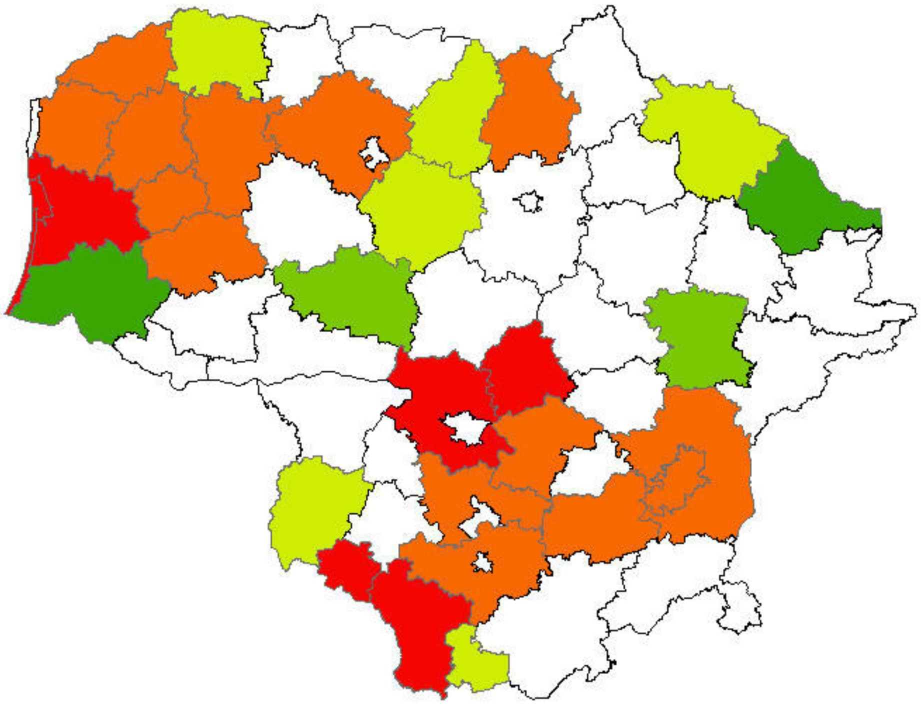


Figure 16. Local Moran's *I*, Automotive Theft Rate per 10 000 Population

The significance levels of the local Moran's *I* statistics are shown in Figure 17 (light to dark increasing in significance), with only the areas that are solid black being statistically significant. A very different picture of the spatial autocorrelation in Lithuanian municipalities should be emerging for anyone viewing both of these maps and knowing that the global spatial autocorrelation is statistically insignificant. Though there is a high degree of variation in the local Moran's *I* statistics, many of them are statistically insignificant, mirroring the global Moran's *I* statistical result. The next step is to select those Lithuanian municipalities that have statistically significant local Moran's *I* statistics, determine the nature of those Lithuanian municipalities and visualize them.

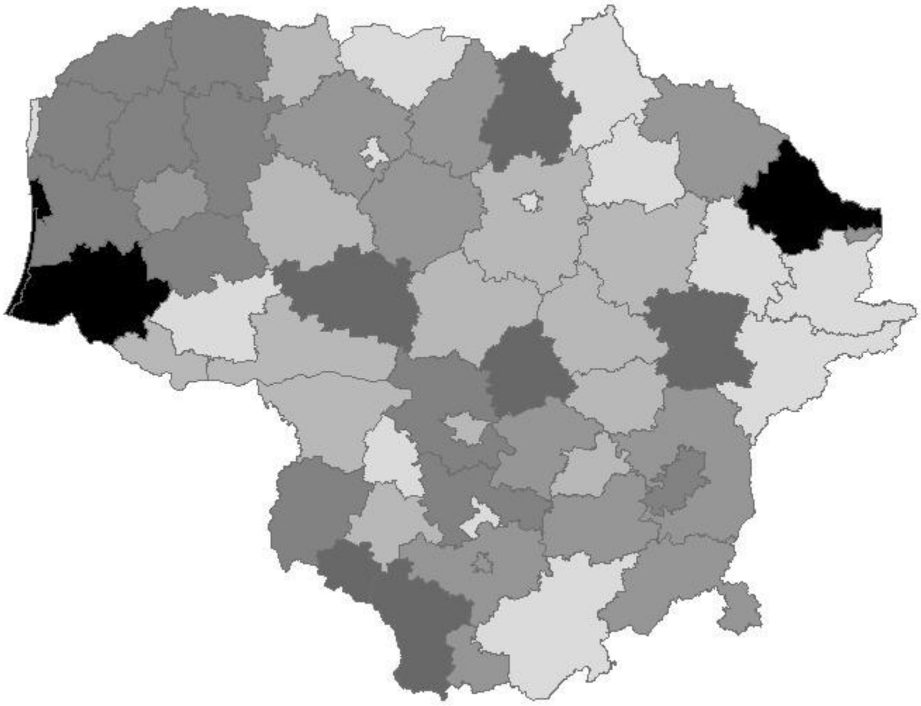


Figure 17. Local Moran's *I*, Statistical Signifiance (Z-score)

The final map for the purposes of this example is shown in Figure 18. This map represents the Lithuanian municipalities that have statistically significant local Moran's *I* statistics at the 10 percent level of significance (absolute value of the z-score > 1.28), though in this particular case there is no difference when using the 10 percent level of significance and the 5 percent level of significance (absolute value of the z-score > 1.645). Four Lithuanian municipalities have statistically significant local Moran's *I* statistics.

The categories represented are as follows: High – High (red), High – Low (pink), and Low – High (blue)—there is also the Low – Low classification that is not represented on this map. Municipalities with High – High and Low – Low categories have positive local spatial autocorrelation (similar local Moran's *I* values), and municipalities with High – Low and Low – High categories have negative local spatial autocorrelation (not similar local Moran's *I* values).

The following is a description of each of the four possible categories: municipalities that are labelled as High – High have statistically significant positive local Moran's *I* statistics, have a high automotive theft rate, and are surrounded by other municipalities that have high automotive theft rates; municipalities that are labelled as High – Low have statistically significant negative local Moran's *I* statistics, have a high automotive theft rate, and are surrounded by other municipalities that have low automotive theft rates; municipalities that are labelled as Low – High have statistically significant negative local Moran's *I* statistics, have a low automotive theft rate, and are surrounded by other municipalities that have high automotive theft rates; and lastly, municipalities that are labelled as Low – Low have statistically significant positive local Moran's *I* statistics, have a low automotive theft rate, and are surrounded by other municipalities that have low automotive theft rates.

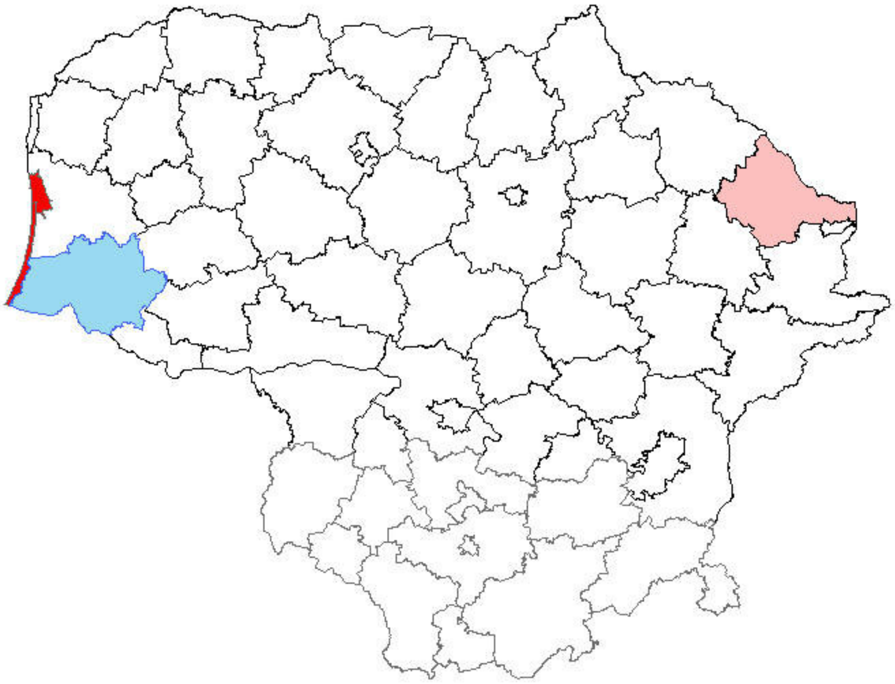


Figure 18. Significant Local Moran's *I*, By Type of Cluster

Though in this example there are only four of the sixty Lithuanian municipalities with statistically significant local Moran's *I* values, the utility of using a local spatial statistical measure should be apparent. When only considering the global measure of spatial autocorrelation, one would have thought that there was no clustering of automotive theft in Lithuanian municipalities. However, using the local measure of spatial autocorrelation there is a curious pattern that emerges, one on either side of the country.

The three municipalities that have high rates of automotive theft (Klaipeda town municipality and Palanga town municipality in the West and Kaunas town municipality in the East) are on border regions in Lithuania. This finding may have significance towards any automotive theft crime prevention. However, this may simply be a coincidence. Also, despite being a neighbour to municipalities with high automotive theft rates, Kazlu municipality has a low automotive theft rate: what would a municipality close to a small cluster of municipalities with high rates of automotive theft have such a low automotive theft rate? This question cannot be answered here, but these questions could not even have been asked if only global spatial autocorrelation was used in the analysis.

References

1. Anselin, L. (1995) Local indicators of spatial association – LISA. *Geographical Analysis* 27: 93 – 115.

2. Bailey, T.C. and A.C. Gatrell (1995) *Interactive Spatial Data Analysis*. Harlow, England: Prentice Hall.

3. Brantingham, P.L. and P.J. Brantingham (1995) Location quotients and crime hot spots in the city. In C.R. Block, M. Dabdoub, S. Fregly (eds.) *Crime Analysis through Computer Mapping*. Washington, DC: Police Executive Research Forum, pp. 129 – 149.

4. Brantingham, P.L. and P.J. Brantingham (1998) Mapping crime for analytic purposes: location quotients, counts and rates. In D. Weisburd and T. McEwen (eds.) *Crime Mapping and Crime Prevention*. Monsey, NY: Criminal Justice Press, pp. 263 – 288.

5. Diggle, P.J. (2003) *Statistical Analysis of Spatial Point Patterns, Second Edition*. London: Arnold Publishers.

6. Isard, W., I.J. Azis, M.P. Drennan, R.E. Miller, S. Saltzman, and E. Thorbecke (1998) *Methods of Interregional and Regional Analysis*. Brookfield, VT: Ashgate Publishing Limited.

7. Miller, M.M., L.J. Gibson, and N.G. Wright (1991) Location quotient: a basic tool for economic development studies. *Economic Development Review* 9: 65 – 68.

8. Openshaw, S. (1984) *The Modifiable Areal Unit Problem*. CATMOG (Concepts and
9. Techniques in Modern Geography) 38. Norwich: Geo Books.

10. Openshaw, S., M. Charlton, C. Wymer, and A. Craft (1987) Developing a mark 1 geographical analysis machine for the automated analysis of point data. *International Journal of Geographical Information Systems* 1: 335 – 358.

11. Ratcliffe, J. (1999) Hotspot Detective for MapInfo Helpfile Version 1.0.
12. http://jratcliffe.net/ware/ index.htm

13. Robinson, W.S. (1950) Ecological correlations and the behaviour of individuals. A*merican Sociological Review* 15: 351 – 357.

14. Tobler, W. R. (1970) A computer model simulation of urban growth in the Detroit region. *Economic Geography* 46: 234 – 240.

Additional Resources

Classical Correlation

The Statistics Homepage:

http://www.statsoft.com/textbook/stathome.html

Introductory Statistics:

http://www.psychstat.missouristate.edu/sbk00.htm

HyperStat Online:

http://davidmlane.com/hyperstat/

Modifiable Areal Unit Problem and the Ecological Fallacy

Jerry Ratcliffe's Home Page (MAUP):

http://www.jratcliffe.net/research/maup.htm

Jerry Ratcliffe's Home Page (Ecological Fallacy):

http://www.jratcliffe.net/research/ecolfallacy.htm

Density Estimation and Hot Spot Mapping

Mapping Crime: Understanding Hot Spots (U.S. National Institute of Justice):

http://www.ojp.usdoj.gov/nij/maps/ncj209393.html

CrimeStat Manual (Chapter 8):

http://www.icpsr.umich.edu/CRIMESTAT/files/CrimeStatChapter.8.pdf

or look in the Table of Contents at:

http://www.icpsr.umich.edu/CRIMESTAT/download.html

4 Geostatistics

Geostatistics is one of most challenging aspects of spatial analysis. It involves the study of the interpolation, smoothing, estimation and prediction of surface's values based on discrete measurements. This module provides an introduction to geo-statistics, including elements of exploratory spatial data analysis, structural analysis including calculation and modeling of the surface properties of nearby locations, and surface prediction and the assessment of results. Methods discussed include Inverse Distance Weighting, trend analysis with global and local polynomials, splines interpolation and techniques of kriging predictions. The notion of spatial dependency and auto-correlation is also analyzed. Categorization of geostatistical methods, recommendations for applications of these methods, and models of results validations are also discussed.

Module Outline

Topic 1: Introduction to Geostatistics:
- The First Law of Geography
- Tasks: Interpolation, Smoothing and Prediction

Topic 2: Classification of Geostatistics Methods: Global/Local, Exact/Non-Exact, Deterministic/Probabilistic

Topic 3: Interpolation and Smoothing Methods:
- IDW
- Global Polynomials
- Local Polynomials
- Splines

Topic 4: Kriging Prediction

Topic 5: Model Validation

Topic 6: Conclusion: Comparison of Geostatistical Methods

4.1. Introduction to Geostatistics

4.1.1. Motivation

A common research task is the investigation of the spatial structures of natural or social phenomena, using point observation and quantitative analysis. For example, acid precipitation, a major cause for forest decline, is usually sampled in discrete stations. The term acid precipitation refers to acid pollution of atmospheric precipitation (rain or snow). The acidity of substances dissolved in water is measured in terms of pH (defined as the negative logarithm of the concentration of hydrogen ions). According to this measurement scale, solutions with pHs less than 7 are described as being acidic, while a pH greater than 7.0 is considered alkaline. Precipitation normally has a pH from 5.0 to 5.6 because of natural atmospheric reactions involving carbon dioxide. For comparison, distilled water, pure of any other substances, would have a pH of 7.0. Precipitation is considered acidic when its pH falls below 5.6, this being 25 times more acidic than pure distilled water.

Precipitation forecasts may be based on the observations of meteorological stations with limited spatial distribution. Limitations of such an approach include the high cost of fieldwork and the inaccessibility of some regions. For example in Lithuania, a meteorological network of 16 stations was set up to monitor precipitations and other climate parameters. Annual precipitation data for 2005 ranges from $915 \geq Pr \geq 396$ mm (see Table below).

Table below shows the precipitation levels in year 2005, and the coordinates of the observation stations.

VARDAS	ID	mm	X	Y
Biržai	1	584	546484	6229692
Telšiai	13	620	389573	6206379
Utena	15	701	601070	6152592
Raseiniai	10	484	443915	6138743
Šilutė	12	835	339702	6137606
Dotnuva	2	432	492560	6136744
Nida(Neringa)	8	915	310049	6134547
Ukmergė	14	655	549330	6123740
Kybartai	4	649	420364	6056681
Varėna	16	861	535976	6012783
Lazdijai	7	722	468284	6010907
Klaipėda	5	752	323400	6174934
Šiauliai	11	396	456608	6200104
Panevėžys	9	527	522840	6177229
Kaunas	3	641	499666	6085879
Laukuva	6	642	388910	6166667
Vilnius	17	783	571223	6055189

It may be asked “What is the best way to visualize the spatial variation in the data?” One answer would be “through geographic mapping”. Mapping here can include intermediate points, with no observations but predicted values. These locations, with and without data, can be visualized in a grid covering the study area.

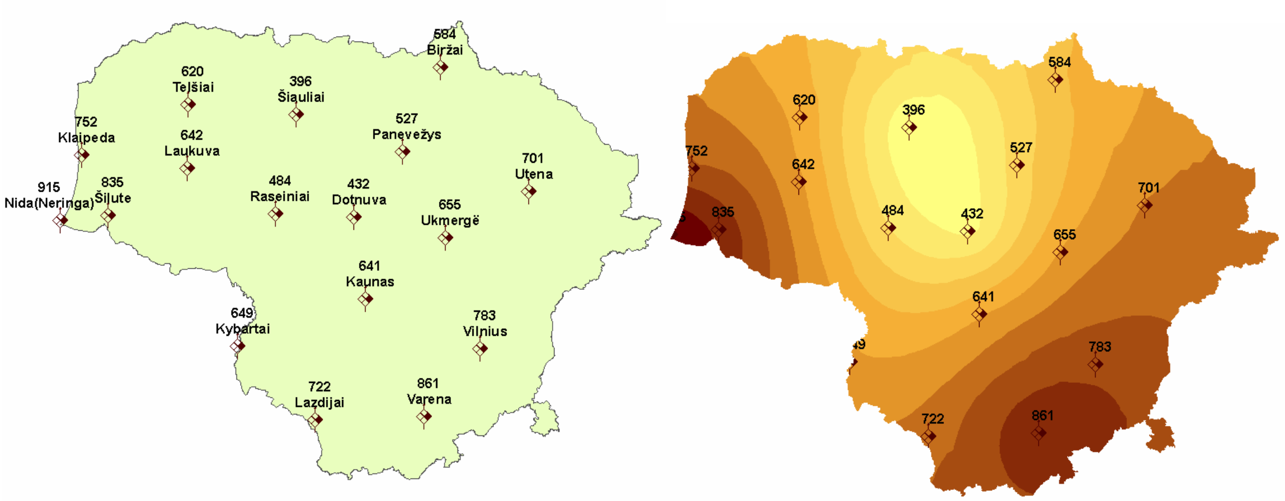


Figure 1 : The geocoded Lithuanian meteorological station network based on the table data and an interpolated surface

In the prediction of the intermediate locations, interpolation methods, such as kriging can be used. For accuracy, optimal spatial prediction and respective principles can be applied. For example, weighted averages of available observations may be used, with greater weights corresponding to proximity to the prediction point.

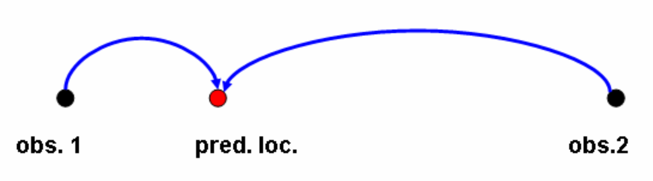


Figure 2 : Observations and unknown locations for interpolation

4.1.2. What is Geostatistics?

Geostatistics is a branch of applied statistics. The principles of Geostatistics were developed in by G. Matheron (1963, France), L.S. Gandin (1963, Soviet Union) and A.S. Goldberger (1962, USA)). The original purpose of geostatistics centered on estimating changes in ore grade within a mine. However, the principles have been applied to a variety of areas in geology and other scientific disciplines, including weather forecasting.

A unique aspect of geostatistics is the use of **regionalized variables,** which are variables that fall between random variables and completely deterministic variables. Regionalized variables describe phenomena with *geographical distribution* (e.g. elevation of ground surface or temperature). For *regional* data, the locations are known and fixed.

Although there may be spatial continuity in the observed phenomena, it is not always possible to sample every location. Therefore, unknown values must be estimated from the size, shape, orientation, and spatial arrangement of the sampled locations. If the spatial pattern changes, the estimated values also change. For **geostatistical** data, the sampling and estimating of *regionalized variables* create a pattern, which may be mapped as a raster grid or contour.

Therefore, geostatistics is about the analysis of spatially referenced phenomenon that was observed via discrete measurements. Geostatistics uses spatial coordinates to formulate continuous model of the analyzed phenomenon based on interpolation, smoothing, estimation and/or prediction techniques. These techniques use information on the spatial coordinates and their distribution of the empiric data.

Geo-statistical data may help to answer questions such as:

- Where? E.g., where are water quality measurement stations in Lithuania located?
- How many? E.g., what are the unemployment rates per municipality in Lithuania?
- How much? E.g., what are values of soil pH in Lithuania in particular locations?

4.1.3. The First Law of Geography

What is special about spatial data?

- The **location** of a sample is an intrinsic part of its definition. Spatial data often represents observations from **one** random variable. Examples include the observations of the mortality rate, soil pH, or temperature in Lithuania.

- Thus in *correlation* analysis, the linear relationship between **two** random variables is determined (e.g. between soil pH and crop production; temperature and elevation, etc.). Geostatistics is primarily concerned with the **auto**-*correlation* function, or the correlation between observations separated by a measured distances and directions. Therefore, the values are: (1) the measurement of a variable at a random point; and (2) the spatial pattern of distance and directions between the measured variables at each point. Dependence between these two sets of values may be measured with a variogram (or semi-variogram). This tool will be discussed later in this module.

- All data sets from a given area are **implicitly related** by their coordinates and are used to model spatial structure. As there may be a spatial structure to the data, values at sample points *cannot* be assumed to be **independent.** This contrasts with *classical statistics*, which assumes independence, at least within the sampling strata. These contrary assumptions (classical statistics and geostatistics) have major implications for sampling design and statistical inference.

Spatial data are characterized by spatial dependence or auto-correlation. **Spatial dependencies** can be described by "*the first law of geography*" that states "*everything is related to everything else, but near things are more related than distant things*" (Tobler, 1970). Therefore, with increasing distance, the characteristics of two locations become less similar and/or related.

The key question of geostatistics is "*How can we measure spatial dependence*?" Important measurement tools are the semivariogram and auto-covariance functions. The use of these tools will be discussed later. The variogram is an effective tool for predicting the value of non-measured points, using distance and direction with measured points, and the screening effect of clusters of such information.

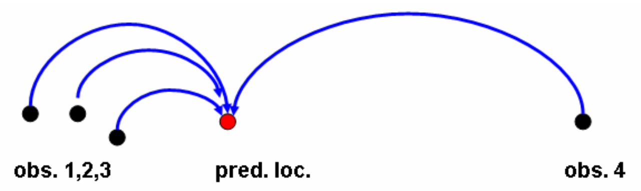


Figure 3 : Geostatistical predictors are weighted averages of available observations. Observations closer (i.e. stronger correlated – not necessary nearer) to the prediction point should be given more weight in the predictor. Screening effect: clusters of observations are down weighted.

4.1.4. Tasks: Interpolation, Smoothing, Estimation and Prediction

This module considers four of the techniques used in geostatistics to create a continuous surface from sample points across a landscape. These techniques are often called *interpolation*, but strictly speaking, that is only for points that are geographically inside the sample set (otherwise it is *extrapolation*). In context of geostatistics, these techniques can be considered as *interpolator, smoother, estimator or predictor*, depending on the particular mathematical or/and statistical technique. Interpolation methods can be adopted to allow extrapolation.

1. **Inverse Distance Weighting (IDW)** is **a** deterministic, direct interpolator.

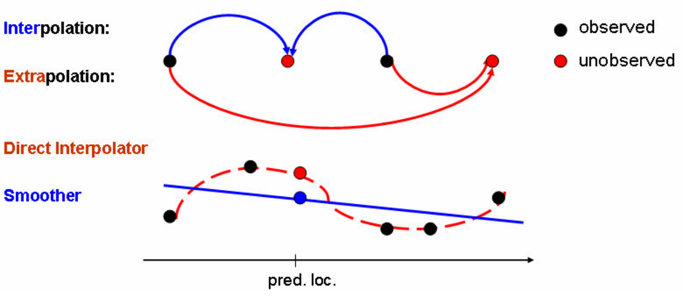


Figure 4 : Smoothing and direct Interpolation

2. **Trend surface estimation** is stochastic *smoother* based on global or local polynomials. Trend surface analysis estimates a linear trend, i.e. $\mu(s_i)$ that can be the deterministic component of the stochastic process.

3. **Splines (piecewise polynomials)** are deterministic *smoother* and based on local or global neighborhoods.

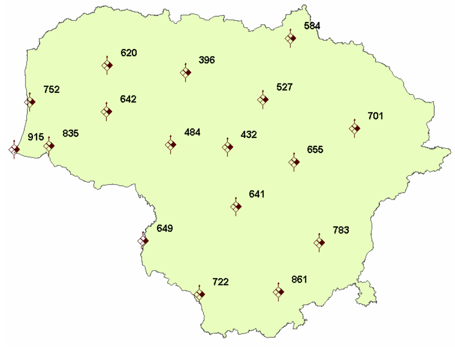
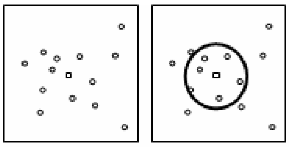
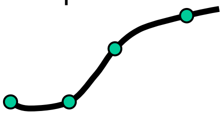
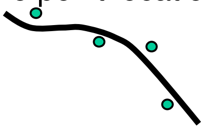
4. **Kriging** is a stochastic *predictor* and can work as smoother or direct interpolator depending on variogram or data. Data $z(s_i)$ are observations from random variables $Z(s_i)$ at locations s_i in a study area D that form a stochastic process $\{Z(s_i) : s_i \in D, i = 1,2,\ldots\}$. This process is composed of at least two components $Z(s_i) = \mu(s_i) + \eta(s_i)$ "trend plus residual" or three components $Z(s_i) =[\mu(s_i) + \eta(s_i)] + \varepsilon(s_i)$ "signal plus noise", where:

   $\mu(s_i)$ = deterministic trend,
   $\eta(s_i)$ = stochastic, correlated residual,
   $\varepsilon(s_i)$ = random noise, uncorrelated.

4.2. Classification of Geostatistics Methods

Geostatistical techniques or methods can be categorized based on the three criteria presented in Table 2.

Table 2.

Global method or **Local** method	**Global** method uses every known or sample point available to estimate an unknown value.  584 620 396 752 642 527 701 915 835 484 432 655 641 649 783 722 861 **Local** method uses a sample of known points to estimate an unknown value. Nearest N point is found and used for computation. 	
Exact or **None exact**	Is there a difference at the sample locations?	No. **Exact interpolation** predicts a value at the point location that is the same as its known value. In other words, exact interpolation generates a surface that passes through the control points.  Yes. **Inexact interpolation** or approximate interpolation predicts a value at the point location that differs from its known value. 
Deterministic or **Probabilistic**	**Deterministic** interpolators make predictions from mathematical formulas that form weighted averages of nearby known values. A deterministic interpolation method provides no assessment of errors with predicted values. Different methods use different ways to form the weighted averages. This group includes Inverse Distance Weighted, global and local polynomials, and radial basis functions or splines. **Stochastic** interpolations use weighted averages as well, but also probability models to make predictions. Stochastic interpolation methods offer assessment of prediction errors with estimated variances. This group includes kriging and all of its different sub-methods.	

The methods discussed in this module can be classified as the following:

Global		Local	
Deterministic	Stochastic	Deterministic	Stochastic
Trend Surface (inexact).	**Regression** (inexact)	**Inverse Distance Weighted** (exact) **Splines** (exact)	**Kriging** (exact)

4.3. Interpolation, Estimation and Smoothing Methods

4.3.1. Inverse Distance Weighted (IDW) Interpolation

As was stated above, an interpolation is the process of estimating unknown values that fall between known values. These points with known values are called known points, control points, sampled points, or observations. The values may describe any quantitative geographic phenomenon. With spatial interpolation, the goal is to create a surface that models the sampled phenomenon in the best possible way.

Interpolation only works where values are spatially dependant, or spatially auto-correlated, that is where nearby location tending to have similar Z values. Examples of spatially auto-correlated features are elevation, property value, crime levels, and precipitation. Non-auto-correlated example is zeppelins consumed per household. Where values across a landscape are geographically independent, interpolation does not work because value of (x,y) cannot be used to predict value of (x+1, y+1).

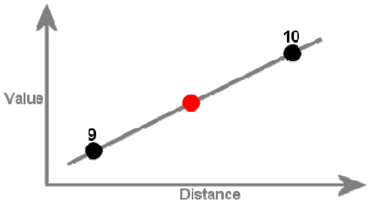


Figure 5 : One and two-dimensional interpolation: unknown value is interpolated based on values and distances to neighboring control points.

Inverse Distance Weighted method follows the principle of the First Law of Geography. IDW determines cell values using a linearly weighted combination of points. The technique estimates the Z value at an unknown point by giving increased weighting to nearer points, this creating an inverse relation between weighting and distance. This can be described mathematically by Shepard's formula of IDW below (Shepard, 1968):

$$z_j = \frac{\sum_{i=1}^{n} \frac{z_i}{d_{ij}^p}}{\sum_{i=1}^{n} \frac{1}{d_{ij}^p}}$$, where:

z_j is the estimated value at (x_j, y_j),

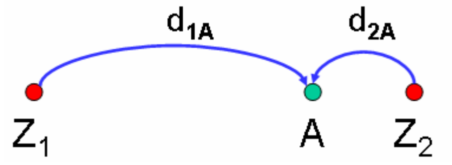
z_i is a neighboring data value at (x_i, y_i),

d_{ij} is the distance between (x_i, y_i) and (x_j, y_j),

p is the power,

$i \in [1, n]$, *n* is the number of data points in the neighborhood of z_j.

IDW linear interpolation case (power is p = 1), known values at z_1 and z_2, and correspondent distances to point A are d_{1A} and d_{2A}. Unknown value in point A will be calculated as:



$$z_A = z_1 + \frac{d_{1A}}{(d_{1A} + d_{2A})}(z_2 - z_1) = \frac{z_1 \times w_1 + z_2 \times w_2}{w_1 + w_2}$$

Where weights are $w_1 = \frac{1}{d_{1A}}$ and $w_2 = \frac{1}{d_{2A}}$, $\sum_{i=1}^{n} w_i = 1$

IDW Features: IDW can provide a good preliminary description of an interpolated surface. There are no assumptions required of the data, but there is no assessment of prediction errors. IDW works best for dense, evenly spaced sample points. It does not consider trends in the data and cannot make estimates above the maximum or below the minimum sample values.

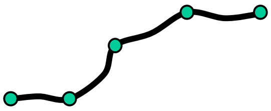


Figure 6: The IDW resulting surface will pass through the sample points, where the maximum and minimum values in the interpolated surface can only occur at known points

The **IDW model parameters:** The characteristics of the interpolated surface in IDW interpolation can be controlled by the following parameters:

- By value of the **power function**
- By the **neighborhood search strategy** that is limiting the number of input points that can be used in calculating the value of each interpolated point. Limit of control points can be defined by:
    - A **number** of closest sample points used
    - A **search radius** used for search neighborhood
    - A **shape** of search neighborhood
    - A combination of above strategies

The **Power Function:** Depending on the site conditions, the distance may be weighted in different ways. If p = 1, this is a simple linear interpolation between points. In many cases, it was found that p = 2 produce better results. In this case, close points are heavily weighted, and more distant points are lightly weighted (points are weighted by $1/d_{ij}^2$). At other sites, p has been set to other powers and yielded reasonable results. By changing the power, one can adjust the relative influence of the sample points. *Increased* power means that the output values become more localized and less averaged. *Lowering* the power that sample point values have provides a more averaged output, because sample points farther away become more and more influential until all of the sample points have the same influence.

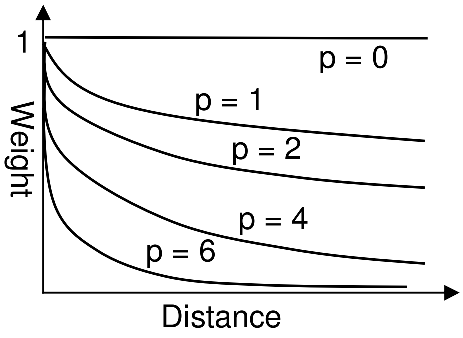


Figure 7 : Distance weighting functions

The **Neighborhood Search Strategy:** It is common practice to specify a search neighborhood to limit the number of measured values that are used to calculate the value of each interpolated point. The number of nearest neighbor points can control the sample size. In such a case, only *N* closest control points will be used for interpolation in unknown location.

A search radius, with the shape of the neighborhood defining the search boundaries, defines the sample size. Some or all of the samples that fall within a radius to calculate the unknown point value can be used.

Other parameters can be established, placing further restrictions on the selection of locations within the neighborhood. There can be a fixed search radius, which will use only the samples contained within it, regardless of number. In such a case, the *distance of the radius* of the circle used to search for points around each interpolated location and a *minimum/maximum number of points* that must be found can be specified.

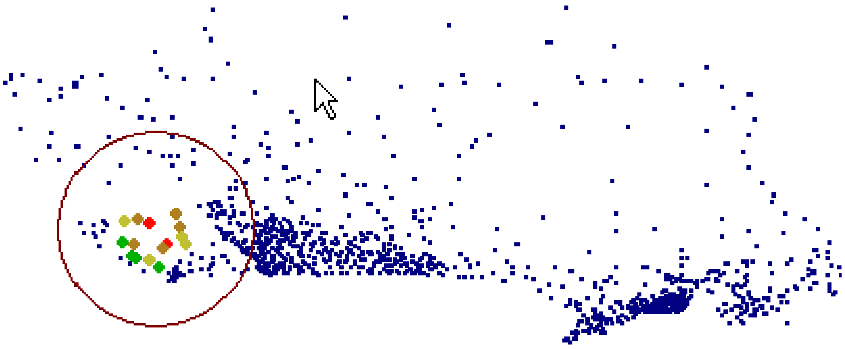


Figure 8 : A search radius strategy - only the maximum of 15 closest points that are within the searching radius are used for interpolation

A search radius also can be *variable*. A variable search radius will expand until the specified sample size is found. One can specify the *number of points* to search for when calculating a value for each interpolated cell and a *maximum distance* for the search radius.

The *shape* and *structure* of search radius also can be modified. Spatial dependencies (auto-correlation) may depend only on the distance between two locations, termed **isotropy.** If there are no directional influences in the data (*isotropy*), then a neighborhood can be a *circle*.

However, it is possible that the same autocorrelation value may occur at different distances when considering different directions. In this case, there is greater variation in some directions than in other directions. This directional influence can be seen in the semi-variogram and is called **anisotropy.** If there is directional influence in data (*anisotropy*), then a neighborhood can be an *ellipse* with the major axis running in the direction of the change. If the neighborhood is sectored, then the constraints can be applied to each *sector*.

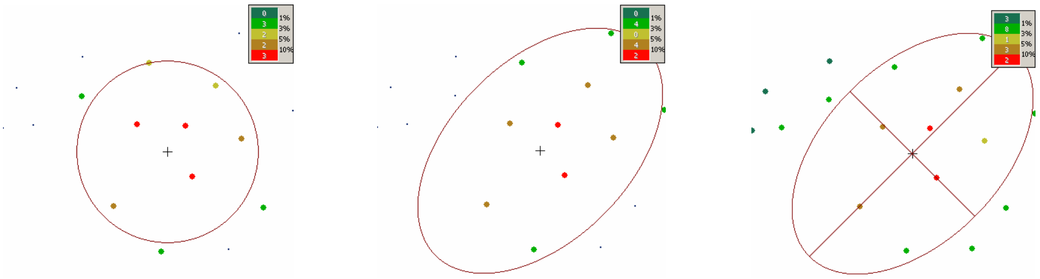


Figure 9 : Isotropy, anisotropy and structural anisotropy

It is important to explore for anisotropy and consider it within the interpolation model as a parameter (e.g. directional shape of search neighborhood). Variogram anisotropy will be discussed in more detail later.

Other restrictions on the search neighborhood are physical *barriers*, such as mountains ridges, which may prevent the interpolator from using samples points on one side of it. Modifications of IDW method can include barriers in the analysis. A barrier can be a polyline dataset used as a break that limits the search for input sample points.

IDW Summary:

- IDW is most commonly used techniques for interpolation.
- Inverse Distance Weighted method is:
  - geostatistical methods;
  - exact (in classical form);
  - local;
  - deterministic.
- There are different modifications of Shepard's IDW method, none being perfect for any application. Although methods differ in weighting (for example the following functions could be used $\frac{1}{1+cd^2}; e^{-cd}; e^{-cd^2}$, where c is a constant) and number of observations used, all IDW modifications use location and value at sampling locations to interpolate the variable of interest at unmeasured locations. Each method produces different results even with the same data.
- IDW accuracy is often judged by the root mean square (RMS) error for differences between the measured (or/and control points) and interpolated values.

4.3.2. Global Polynomials

Trend surface analysis uses global surface-fitting procedures. A surface can be approximated by a global polynomial expansion of the geographic coordinates of the control points, and the coefficients of the polynomial function are found by the method of least squares adjustment, insuring that the sum of the squared deviations from the trend surface is a minimum. An estimated trend, i.e. $\mu(s_i)$, or $Z(x,y)$ is considered as the deterministic component of the stochastic process.

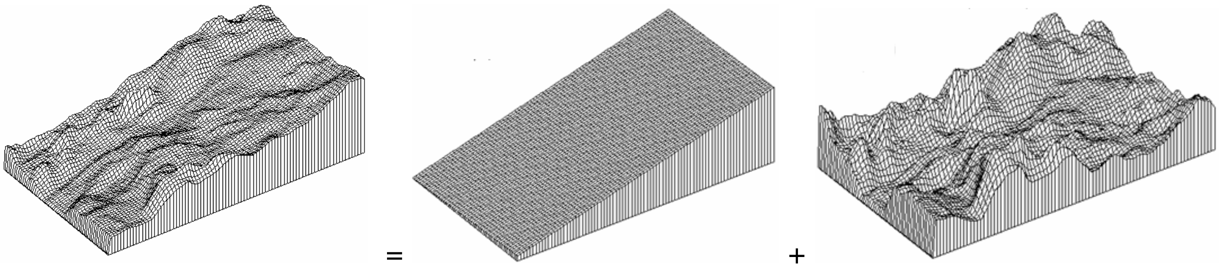


Figure 10 : Each original observation is considered to be the sum of a deterministic polynomial function of the geographic coordinates plus a random stochastic residual.

The global polynomial for the trend analysis is a linear function that has the form:

$$Z(x,y) = a_0 + a_1X + a_2Y + a_3X^2 + a_4XY + a_3Y^2...$$

Where $Z(x,y)$ is the data value at the described location, the **a**'s are coefficients, and X and Y are combinations of geographic location. Polynomial trend-surface analysis is basically a linear regression technique, but it is applied to two- and three-dimensions instead of just fitting a line.

Any desired degree of the polynomial can be chosen. Practically, for trend analysis, low degrees are used (first, second and third orders). A flat surface describes by $Z(x,y) = a_0$. A linear plane, which is first-order polynomial is $Z(x,y) = a_0 + a_1X + a_2Y$**.** Allowing for one band is a second-order quadratic polynomial is $Z(x,y) = a_0 + a_1X + a_2Y + a_3X^2 + a_4XY + a_3Y^2$ and so forth. The unknown a's coefficients are found by solving a set of simultaneous linear equations which include the sums of powers and cross products of the X, Y, and Z values.

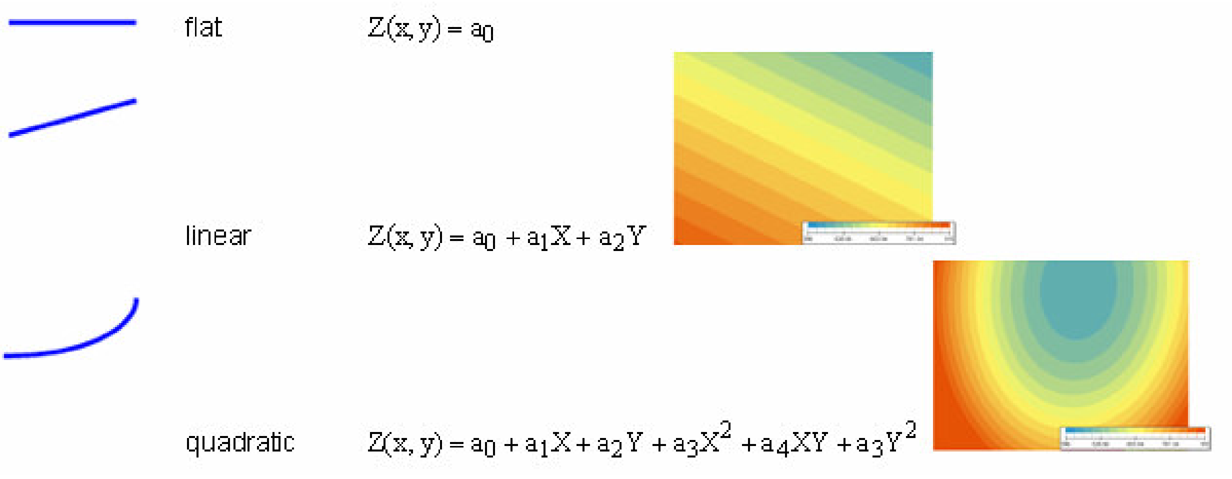


Figure 11 : Low order trend surfaces

The optimum trend, or linear function, must minimize the squared deviations from the trend. The matrix description of solution is:

$\mathbf{Z} = \mathbf{a} * \mathbf{B}$, where **Z** is a linear matrix of known values, **a** is a linear matrix of known **a**'s coefficients, $\mathbf{B}$ is quadratic matrix derived from values of X and Y of known geographic locations. The optimization solution will be $\mathbf{a} = \mathbf{B}^{-1} * \mathbf{Z}$, where $\mathbf{B}^{-1}$ is inversed matrix of $\mathbf{B}$. The example of the solution for computing first-order (linear) trend surface $Z(x, y) = a_0 + a_1X + a_2Y$ with *least square* method is below.

The sample data is presented in the figure and table.

Point	X	Y	Z Value
1	69	76	20.820
2	59	64	10.910
3	75	52	10.380
4	86	73	14.600
5	88	53	10.560
0	69	67	**?**

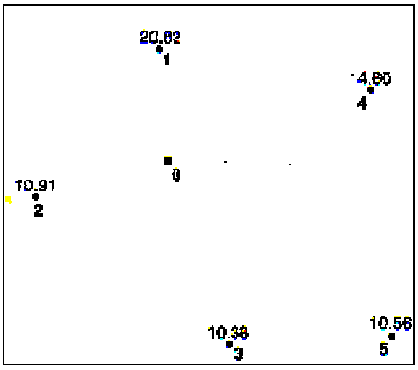


Figure 12 : Sample with known locations for interpolated surface

This surface can be represented as the following system of linear equitation:

$$
\begin{aligned}
20.820 &= a_0 + 69a_1 + 76a_2 \\
10.910 &= a_0 + 59a_1 + 64a_2 \\
10.380 &= a_0 + 75a_1 + 52a_2 \\
14.600 &= a_0 + 86a_1 + 73a_2 \\
10.560 &= a_0 + 88a_1 + 53a_2
\end{aligned}
$$

To solve for these three **a**'s unknowns for **n**th number of data points, the following three normal equations are available:

$$\sum_{i=1}^{n} Z_i(x,y) = n * a_0 + a_1 \sum_{i=1}^{n} X_i + a_2 \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n} X_i Z_i(x,y) = a_0 \sum_{i=1}^{n} X_i + a_1 \sum_{i=1}^{n} X_i^2 + a_2 \sum_{i=1}^{n} X_i Y_i$$

$$\sum_{i=1}^{n} Y_i Z_i(x,y) = a_0 \sum_{i=1}^{n} Y_i + a_1 \sum_{i=1}^{n} X_i Y_i + a_2 \sum_{i=1}^{n} Y_i^2$$

Solving these equations simultaneously will yield a "best-fit", defined by least-squares regression, for a two-dimensional, first-order (a plane) trend surface. This can be rewritten in matrix format:

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i Y_i \\ \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} X_i Y_i & \sum_{i=1}^{n} Y_i^2 \end{bmatrix} * \begin{bmatrix} a_0 \\ a_1 \\ a2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Z_i \\ \sum_{i=1}^{n} X_i Z_i \\ \sum_{i=1}^{n} Y_i Z_i \end{bmatrix}$$

With initial data values:

$$\begin{bmatrix} 5 & 377 & 318 \\ 377 & 29.007 & 23.862 \\ 318 & 23.862 & 20.414 \end{bmatrix} * \begin{bmatrix} a_0 \\ a_1 \\ a2 \end{bmatrix} = \begin{bmatrix} 67.270 \\ 5043.650 \\ 4445.800 \end{bmatrix}$$

After the adjustments:

$$\begin{bmatrix} 23.2102 & -0.1631 & -0.1684 \\ -0.1631 & 0.0018 & 0.0004 \\ -0.1684 & 0.0004 & 0,0021 \end{bmatrix} * \begin{bmatrix} 67.270 \\ 5043.650 \\ 4445.800 \end{bmatrix} = \begin{bmatrix} -10.094 \\ 0.020 \\ 0.3470 \end{bmatrix}$$

The linear function is defined as $Z(x,y) = -10.094 + 0.020X + 0.347Y$. Once the coefficients have been estimated, the polynomial function can be evaluated at any point within the study area (e.g. for $Z(69,67) = -10.094 + 0.020 * 69 + 0.347 * 67 = 14.535$). It is a simple matter to create a grid matrix of values by substituting the coordinates of the grid nodes into the polynomial and calculating an estimate of the surface for each node. Because of the least squares fitting procedure, no other polynomial equation of the same degree can provide a better approximation of the data.

The global polynomial surface changes gradually and captures coarse-scale patterns in the data. Global polynomial interpolation creates a slowly varying surface using low-order polynomials that possibly describe some physical process (e.g., pollution and wind direction). The calculated surfaces are highly susceptible to outliers (extremely high and low values, especially at the edges).

Global polynomial estimation can be used for fitting a surface to the sample points when the surface varies slowly over area of interest. It is also applied to examining and/or removing the effects of long-range or global trends. The *trend* polynomial analysis can help identify global trends in the input dataset. Global polynomial approximation also can be use for *regression analysis* between two or more variable.

Global polynomial Summary:

- Trend-surface analysis is a mathematical method used to separate "regional" from "local" fluctuations
- Global polynomial is a quick deterministic interpolator that is an inexact smoother.
- There are very few decisions to make regarding model parameters (only polynomial order).
- There is no assessment of prediction errors.
- Locations at the edge of the data can have a large effect on the surface.
- There are no assumptions required of the data.

4.3.3. Local Polynomials

Local polynomial interpolation is similar to global polynomial interpolation, except that it uses data within localized “windows” rather than complete datasets. It fits local trends within a “window” and it uses weights. The window can be moved around, and the surface value $\mu_0(x_i, y_i)$ at the center of the window is estimated at each point. Weighted least squares can be represented as below:

$$\sum_{i=1}^{n} w_i (Z(x_i, y_i) - \mu_0(x_i, y_i))^2,$$

$$w_i = \exp(-3d_{i0}/a),$$

Where $\mu_0(x_i, y_i)$ is the value of the polynomial, d_{i0} is the distance between the point and the center of the window and **a** is a parameter that can be used to control how fast weights decay with distance.

Local polynomial interpolation fits the specified order (e.g., zero, first, second, and third) polynomial using all points only within the defined neighborhood. The neighborhoods overlap, and the value used for each prediction is the value of the fitted polynomial at the center of the neighborhood. A first-order polynomial looks like as $\mu_0(x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i$, for second-order polynomial is $\mu_0(x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i^2 + \beta_4 y_i^2 + \beta_5 x_i y_i$, and so on. The minimization occurs for the parameters *$\{\beta_j\}$.*

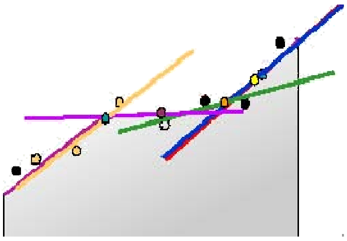


Figure 13 : *Local polynomial* interpolation fits many polynomials, each within specified overlapping neighborhoods. Thus, local polynomial interpolation produces surfaces that account for more local variation.

Global polynomial interpolation is good for creating smooth surfaces and for identifying long-range trends in the dataset. However, in the earth sciences the variable of interest usually has short-range variation and a long-range trend. When the dataset exhibits short-range variation, local polynomial interpolation maps can capture the short-range variation.

Local polynomial interpolation is sensitive to the neighborhood distance. An operator can interactively choose this distance. As with IDW, it is possible to define a model that accounts for anisotropy by choosing an appropriate shape of the neighborhood search.

Local Polynomial Summary:

- Local polynomial is a quick deterministic interpolator that is smooth and inexact.
- There are more parameter decisions – the polynomial order, size and shape of neighborhood and overlapping.
- There is no assessment of prediction errors.
- The method provides prediction surfaces that are comparable to kriging with measurement errors.
- There are no assumptions required of the data.

4.3.4. Splines or Radial Basis Functions (RBF)

Spline methods are a series of exact interpolation techniques that fits the surface through each measured sample value. There are few different spline functions (http://en.wikipedia.org/wiki/Spline_(mathematics)). Each spline function has a different shape and results in a different interpolation surface. Spline is conceptually similar to fitting a rubber membrane through the measured sample values while minimizing the total curvature of the surface. Spline methods are a form of artificial neural networks.

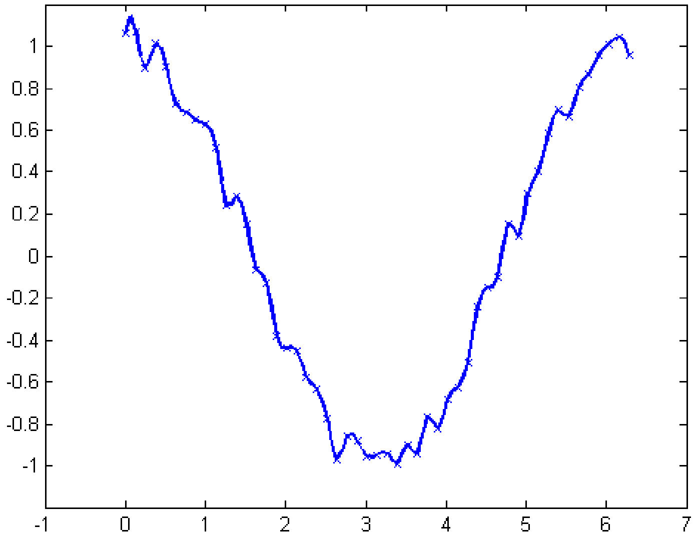


Figure 14 : Interpolation with a spline function

Splines are used for calculating smooth surfaces from a large number of data points. The functions produce good results for gently varying surfaces such as elevation. The techniques are inappropriate when there are abrupt changes in the surface values within a short horizontal distance.

Splines Summary:

- Radial basis functions are deterministic interpolators that are exact.
- There are more parameter decisions than IDW; therefore, it is more flexible than IDW.
- There is no assessment of prediction errors. Radial basis functions do not allow investigating the autocorrelation of the data.
- The method provides prediction surfaces that are comparable to the exact form of kriging.
- Radial basis functions make no assumptions about the data.

4.4. *Kriging Prediction*

The South African engineer D. G. Krige was the first to formalize a method that uses a mathematical model of the *semivariogram* for estimating a surface at grid nodes. The kriging method, named after its founder, predicts the best linear unbiased estimates of a surface at specified locations, based on the assumptions that the surface is *stationary* and the correct form of the semivariogram has been chosen. The kriging procedures incorporate measures of error and uncertainty in determining estimates. The calculation of unknown values is similar to IDW and based on weights assigned to known values. However, these weights are only optimal weights and the semivariogram is used for calculation weights. Semivariogram weights depend on the known sample distribution – distance and direction.

Observed values are only one of many possible realizations of a random “stochastic” process. At each point $\vec{s}$, an observed value Z is one possibility of a random variable $z(\vec{s})$. There is only one value that is sampled and it is only one realization of a process that can produce different values. Each point has its own random process, with the same form. However, there may be spatial dependence among points. In this case, points are *not* independent. As a whole, they make up a stochastic process over the whole field R (i.e. the observed values are assumed to result from some random process but one that respects certain restrictions, in particular spatial dependence). The set of observed values $Z = \{Z(\vec{s}), \forall \vec{s} \in R\}$ is called a regionalized variable. This variable is doubly infinite by 1) number of points and 2) possible values at each point.

Regionalized variable theory uses a related property called the ***semivariance*** to express the degree of relationship (or autocorrelation) between points on a surface. The semivariance is simply half of the variance in the values between each point pair that is separated by a known distance. The ***variogram*** is a representation which plots *semivariances* against its separated distance. This type of representation is discussed in detail below.

The idea of stationarity is used to obtain the necessary replication. Stationarity is an assumption that is often reasonable for spatial data. There are two types of stationarity. One is called the first-order or ***mean stationarity***. In geostatistics, it is assumed that the mean is constant between samples and is independent of location.

The second type of stationarity is called ***intrinsic stationarity*** for semivariograms and *second-order stationarity* for covariance (http://en.wikipedia.org/wiki/Covariance). The intrinsic stationarity for semivariograms is based on the assumption that the variance of the value’s difference is the same between any two points that are the same distance and direction apart no matter which two points you choose. Second-order stationarity is the assumption that the spatial auto-covariance is the same between any two points that are at the same distance and direction apart no matter which two points you choose. The auto-covariance is dependent on the distance between any two values and not on their locations.

However, in reality, first-order stationarity is often not verisimilar. The observed mean value is often different in several regions or has an obvious trend. Second-order stationarity is also often not plausible, thus, it is observed that covariance often increases without bound as the area increases. Solutions can be to use the *differences* between values, not the *values* themselves, and in a “small” region. The *differences* between values are the same over the whole area. In addition, if the trend is *subtracted*, the residuals may comply with first-order stationary. These and other solutions (e.g. to use empirical variogram mean differences), as you will see below, are used for kriging prediction.

4.4.1. Variography

Each pair of observation points has a variance. The variance is the difference between two variables at two locations, raised to the second power. The semi-variance is variance divided by two. The ***semivariance*** is defined as:

$$\gamma(\vec{s}_\mathrm{i},\vec{s}_\mathrm{j})=\frac{1}{2}\left[\mathrm{Z}(\vec{s}_\mathrm{i})-\mathrm{Z}(\vec{s}_\mathrm{j})\right]^2,$$

Where $\gamma(\vec{s}_\mathrm{i},\vec{s}_\mathrm{j})$ - semivariance and $\mathrm{Z}(\vec{s}_\mathrm{i})$ and $\mathrm{Z}(\vec{s}_\mathrm{j})$ are values in two locations $\vec{s}_\mathrm{i}$ and $\vec{s}_\mathrm{j}$, which is separated by a known distance. The formula calculates half the difference squared between the values of the paired locations. Semivariances can be used as a measure of spatial dependences or autocorrelation.

The semivariances can be summarized in a ***variogram***. A variogram is obtained from the data. The variogram is a point *"cloud"* that plots the variance between two values of the same variable at two locations for $n(n-1)/2$ points, where n is the number of observed points. The semivariances are plotted against distance in a *variogram "cloud"*. Along the ordinate *x*-axis, *variogram* plots the distance separating two locations; along the abscissa *y*-axis, *variogram* plots the semivariance that is used to quantify autocorrelation.

The semivariance generally increases with distance and variograms are described by ***nugget, sill***, and ***range*** parameters. *Sill* is maximum semivariance or the height that the semivariogram reaches, and represents variability in the absence of spatial dependence. It is often composed of two parts: a discontinuity at the origin, called the *nugget* effect, and the *partial sill*, which added produce the *sill*. *Nugget* is semi-variance as the separation approaches zero and represents variability at a point that cannot be explained by spatial structure. The nugget can be divided into measurement error and micro-scale variation and since either component can be zero, the nugget effect can be comprised wholly of one or the other. *Range* is separation between point-pairs at which the sill is reached or distance at which there is no evidence of spatial dependence.

As $\mathrm{Z}(\vec{s}_\mathrm{i})$ and $\mathrm{Z}(\vec{s}_\mathrm{j})$get farther apart, they become less similar, and so the difference in their values, $\mathrm{Z}(\vec{s}_\mathrm{i})-\mathrm{Z}(\vec{s}_\mathrm{j})$, will become larger. The *anatomy* of a typical semivariogram is represented in the following figure:

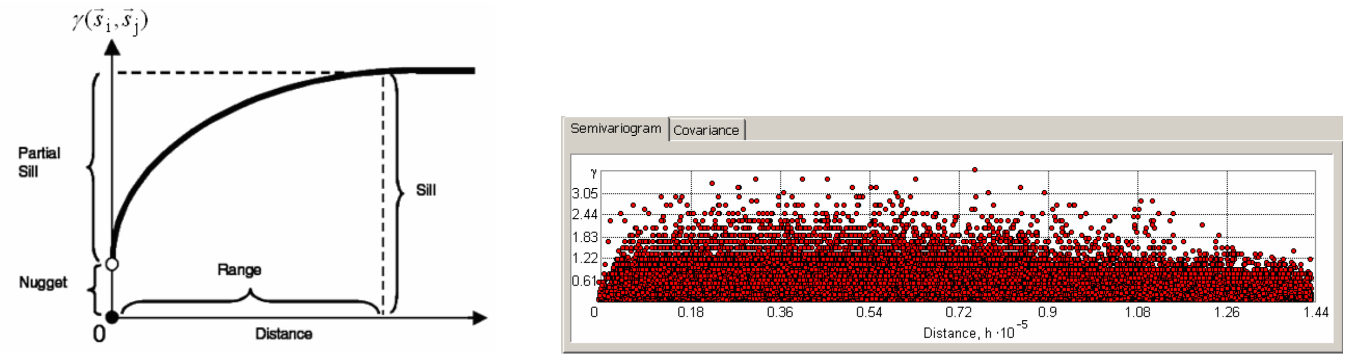


Figure 15 : The anatomy of a typical semivariogram and the semivariogram view in ArcGIS Geostatistical Analyst

There are $n(n-1)/2$ point pairs that are used to calculate and build the variogram. These involve large numbers and can become unmanageable to plot. For example, with 200 points, there are 19,900-point pairs. To reduce the number of points in the ***empirical semivariogram***, the pairs of locations are grouped based on their distance from one another into ***lag bins*** or by *separations* $\vec{h}$. For example, compute the average semivariance for all pairs of points that are greater than 100 meters but less than 200 meters apart. This is repeated for all samples that are h distance apart and the average squared difference obtained. Therefore, the empirical semivariogram is a graph of the averaged semivariogram values on the *y*-axis, and $\vec{h}$ distance (or lag) on the *x*-axis.

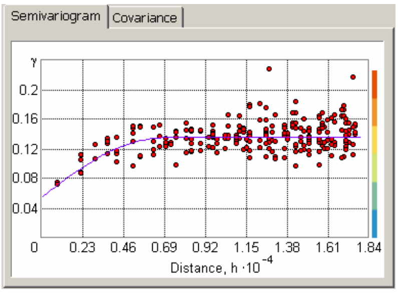


Figure 16 : The *empirical* semivariogram view in ArcGIS Geostatistical Analyst. For each bin, only the average distance and semivariance for all the pairs in that bin are plotted as a single point on the empirical semivariogram cloud graph.

Note that binning is the intrinsic stationarity assumption that allows replication. Mean values are replaced with mean differences, which are the same over the whole field, at least within some 'small' lag separation $\vec{h}$. Thus, it uses "averaging" in the semivariogram formula above and *the empirical semivariogram* can be estimated for distances that are multiple of $\vec{h}$ as:

$\bar{\gamma}(\vec{h}) = \frac{1}{2m(\vec{h})}\sum_{\mathrm{k}=1}^{m(\vec{h})}\left[\mathrm{Z}(\vec{s}_{\mathrm{i}}) - \mathrm{Z}(\vec{s}_{\mathrm{j}})\right]^2$ - the semivariance is equal to the average of the squared differences between pairs of points within a bin spaced at distance $\vec{h}$.

Where, $\vec{h}$ is regularly spaced points distance (separation or lag distance); $\mathrm{Z}(\vec{s}_{\mathrm{i}})$ and $\mathrm{Z}(\vec{s}_{\mathrm{j}})$ are values in two locations $\vec{s}_{\mathrm{i}}$ and $\vec{s}_{\mathrm{j}}$; $m(\vec{h})$ is the number of point pairs separated by vector $\vec{h}$ or number of points within a bin. In practice, there have to be enough known points in order to define the set of vectors in each “bin”.

Bins are commonly formed by dividing the sample area into a grid of cells or sectors that are used to calculate the empirical semivariogram. The size of the cell is called *lag size* and the number of cells is called *number of lags*. Narrow intervals mean more resolution, but fewer point pairs for each sample.

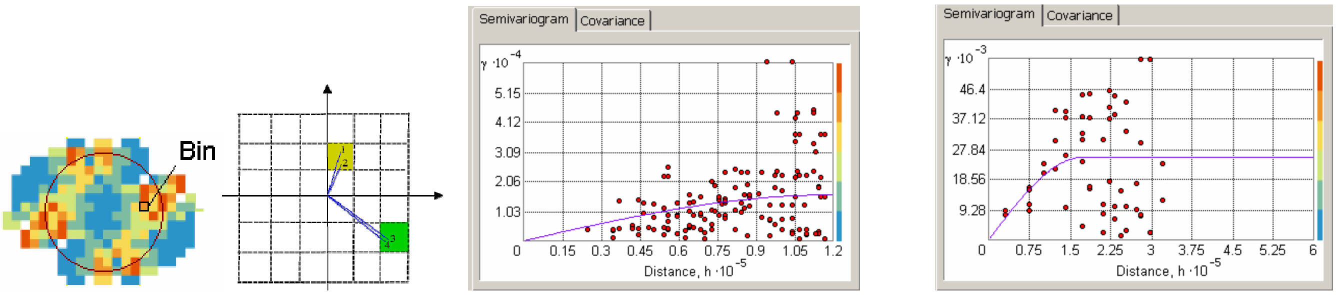


Figure 17 : A measure of the similarity between a variable’s values (the co-variance between the two values) for distance h apart is obtained. Variograms with the respective 10000 and 50000 meters lags (bin sizes) are produced different outputs (in the right-hand side of the figure)

Consider the following example for binning the empirical semivariogram:

Locations $\vec{s}(x_i, y_i)$	Value in Location $Z(\vec{s}_i)$
(1,3)	105
(1,5)	100
(4,3)	110
(5,1)	115
(4,5)	100

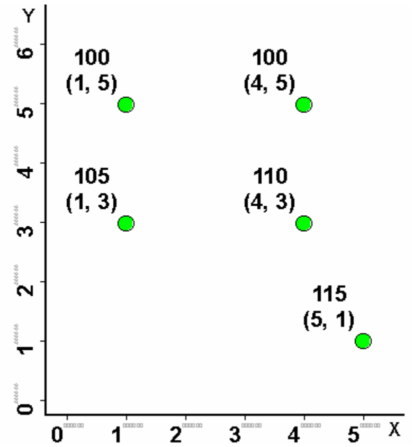


Table 3 : Calculation of semivariances $\gamma(\vec{s}_i,\vec{s}_j)=\frac{1}{2}\left[Z(\vec{s}_i)-Z(\vec{s}_j)\right]^2$ **:**

Locations $(\vec{s}_i,\vec{s}_j)$	Euclidian Distance Calculations $\gamma(\vec{s}_i,\vec{s}_j)$	Distances $d(\vec{s}_i,\vec{s}_j)=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$	Differences^2 $\left[Z(\vec{s}_i)-Z(\vec{s}_j)\right]^2$	Semivariances $\gamma(\vec{s}_i,\vec{s}_j)$
(1,5),(4,3)	sqrt[(1-4)^2 + (5-3)^2]	3.606	100	50
(1,5),(1,3)	sqrt[0^2 + 2^2]	2	25	12.5
(1,5),(4,5)	sqrt[3^2 + 0^2]	3	0	0
(1,5),(5,1)	sqrt[4^2 + 4^2]	5.657	225	112.5
(4,3),(1,3)	sqrt[3^2 + 0^2]	3	25	12.5
(4,3),(4,5)	sqrt[0^2 + 2^2]	2	100	50
(4,3),(5,1)	sqrt[1^2 + 2^2]	2.236	25	12.5
(1,3),(4,5)	sqrt[3^2 + 2^2]	3.606	25	12.5
(1,3),(5,1)	sqrt[4^2 + 2^2]	4.472	100	50
(4,5),(5,1)	sqrt[1^2 + 4^2]	4.123	225	112.5

Table 4 : Binning the empirical semivariogram $\bar{\gamma}(\vec{h})=\frac{1}{2m(\vec{h})}\sum_{k=1}^{m(\vec{h})}\left[Z(\vec{s}_i)-Z(\vec{s}_j)\right]^2$ **with lag** $\vec{h}=1$ **meter:**

Lag distances in meters	Distance Pairs	Average Distance	Semivariance $\gamma(\vec{s}_i,\vec{s}_j)$	Average $\gamma(\vec{s}_i,\vec{s}_j)$
From 1 to 2	2, 2	2	12.5, 50	31.25
From 2 to 3	2.236, 3, 3	2.745	12.5, 12.5, 0	8.33
From 3 to 4	3.606, 3.606	3.606	50, 12.5	31.25
From 4 to 5	4.472, 4.123	4.298	50, 112.5	81.25
More then 5	5.657	5.657	112.5	112.5

Question: How does one find a lag size? A rule of thumb is to multiply the lag size by the number of lags; the product of these numbers should be about half the largest distance among all points.

So far, the values in the semivariogram cloud are put into bins based only on the *distance.* This variogram is called ***omnidirectional*** and ***isotropic*** (Greek *"iso"* + *"tropic"* = English “same” + “trend”). So far, the direction between a pair of locations of lag *h* was not specified in order to constrict the variogram, but variation may depend on *direction*, not just distance.

There are two types of directional components that can affect the predictions in output surface:

- A ***global trend*** is an overriding process that affects all measurements in a deterministic manner. The global trend can be represented by a mathematical formula (e.g., a polynomial) and removed from the analysis of the measured points but added back before predictions are made. This process is referred to as de-trending.
- ***Anisotropy*** (Greek *“an”* + *“tropic”* = English “not-” + “trend”) is a characteristic of a random process that shows higher autocorrelation in one direction than another.

Anisotropy arises due to directionality of a process, for example, sand content in a narrow flood plain has much greater spatial dependence along the axis parallel to the river; secondary mineralization is changing near an intrusive dyke; population density is different in a hilly terrain with long, linear valleys.

So here again, the notion of the phenomenon, such as *anisotropy* and respective ***directional*** *empirical* ***variogram,*** can be brought back. A directional variogram defines the spatial variation among points separated by space lag $\vec{h}$. The difference from the omnidirectional variogram is that $\vec{h}$ - a vector rather than a scalar. The number of directions may be different. A directional variogram is estimated using the same equation as the omnidirectional. Note that *nugget* must logically be *iso*tropic (it is variation *at* a point).

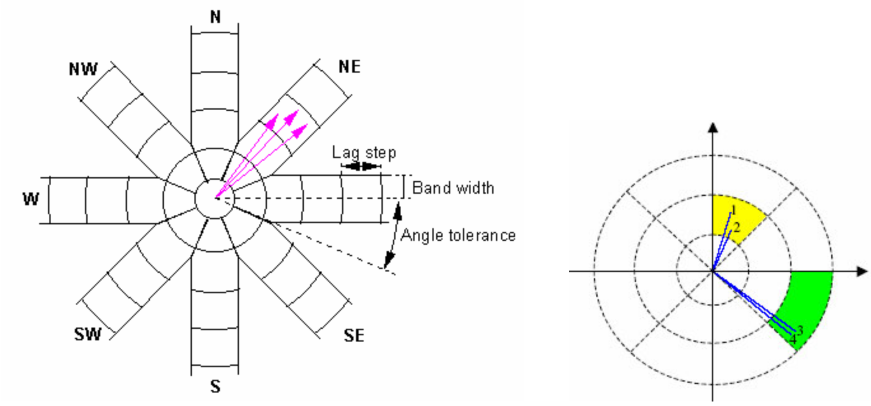
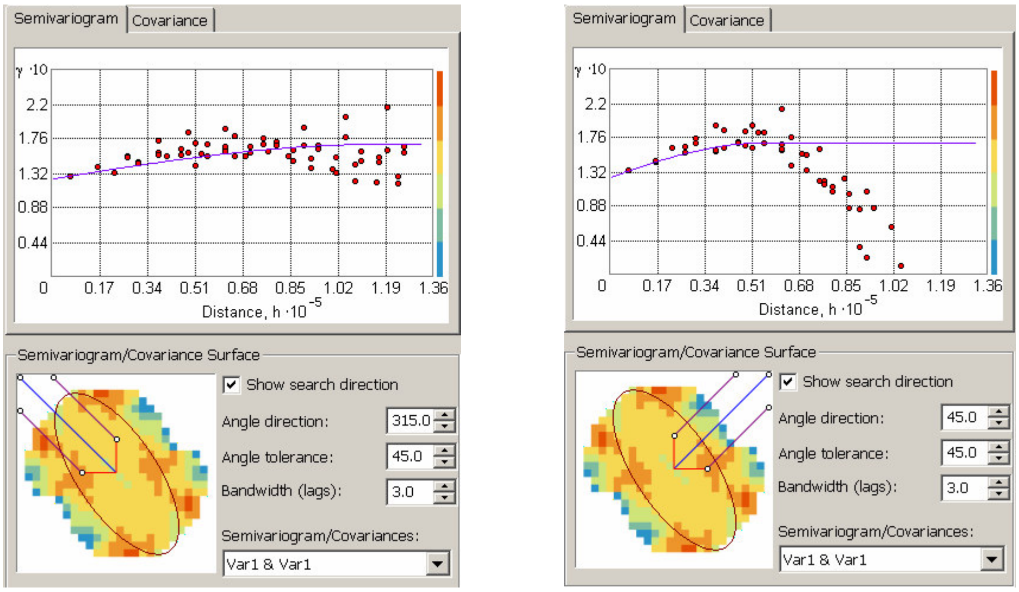


Figure 18 : Point pairs are grouped based on common *separation* distances and directions (bandwidth sectors)

The indicators of anisotropy are that the semivariogram is changed notably; ranges and/or sills are considerably different for the respective directions. There are two types of anisotropy indicators:

1. Sill is the same, but different ranges in different directions; this is *geometric*, or also called *affine*, anisotropy.
2. Range is the same, but sill varies with direction, this is *zonal* anisotropy.

See more about types of anisotropy at http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Accounting_for_directional_influences.



☑ Show search direction

Angle direction: 315.0

Angle tolerance: 45.0

Bandwidth (lags): 3.0

Semivariogram/Covariances: Var1 & Var1

☑ Show search direction

Angle direction: 45.0

Angle tolerance: 45.0

Bandwidth (lags): 3.0

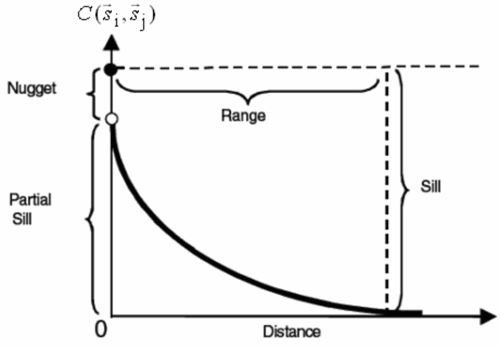
Semivariogram/Covariances: Var1 & Var1

Figure 19 : *Zonal* anisotropy: variogram (range) changes notable and depends on search directions (respectively north-west and north–east search directions)

Therefore, if directional differences in the spatial dependences are detected, it can be accounted for in the semivariogram or covariance models. This, in turn, has an effect on the geostatistical prediction method.

Another measure that is used to estimate the strength of a spatial correlation as a function of distance is ***auto-covariance*** and ***autocorrelation***. The spatial auto-covariance is computed within the same variable, using pairs of observations. Each pair of observations $(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}})$ has a $(\mathrm{Z}(\vec{s}_{\mathrm{i}}) - \overline{\mathrm{Z}})(\mathrm{Z}(\vec{s}_{\mathrm{j}}) - \overline{\mathrm{Z}})$ auto-covariance, showing how they jointly differ from the variable's $\overline{\mathrm{Z}}$ mean. Spatial autocorrelation can also be analyzed using covariance functions and correlograms. The *auto-covariance* function is $C(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}}) = \mathrm{cov}(\mathrm{Z}(\vec{s}_{\mathrm{i}}), \mathrm{Z}(\vec{s}_{\mathrm{j}}))$, where cov is the covariance. See more about the measures of spread at the http://www.spatialanalysisonline.com/output/html/Measuresofspread.html.

The covariance depends on the separation between points. The individual covariance has to be summarized as a *auto-covariance function* of spatial separation. Once this is done, then the covariance between any two locations in space can be predicted.



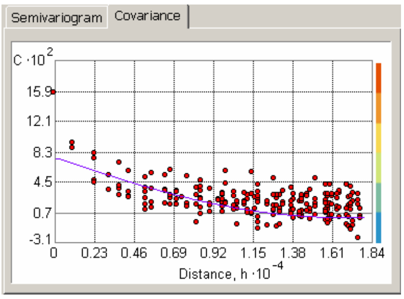


Figure 20 : The anatomy of auto-covariance function and the auto-covariance function's view in ArcGIS Geostatistical Analyst

The idea of correlation to *one* variable is called ***auto-correlation*** (the prefix *auto-* means "self" and refers to a single variable). In such cases, a correlation is controlled by some other dimensions, such as *space,* if the variable is collected at points in space or *time,* if the variable is collected as a time-series. A measure of how much the variable is *correlated to itself*, considering the other factor (time or space), is auto-correlation.

Auto-covariance is just a scaled version of auto-correlation. Auto-correlation $r(\vec{s}_i,\vec{s}_j) = C(\vec{s}_i,\vec{s}_j)/\sigma^2 = C(\vec{h})/C(0)$ is just auto-covariance normalized by total variance ($C(0)$ - variance with distance equals 0), when two locations, $\vec{s}_i$ and $\vec{s}_j$ are close to each other, then they are expected to be similar, so their auto-covariance (auto-correlation) will be large. The population standard deviation is $\sigma = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(Z(\vec{s}_k) - \overline{Z})^2}$. As $\vec{s}_i$ and $\vec{s}_j$get farther apart, they become less similar and so their covariance becomes zero.

The following expression is known as the *autocorrelation* coefficient for lag of $\vec{h}$.

$$r(\vec{h}) = \frac{\sum_{k=1}^{m(\vec{h})} (Z(\vec{s}_i) - \overline{Z})(Z(\vec{s}_j) - \overline{Z})}{\sum_{k=1}^{n} (Z(\vec{s}_k) - \overline{Z})^2}$$

The top part of this expression is like the covariance, but at a lag of $\vec{h}$, and the bottom is like the covariance at a lag of 0. These two components are the *autocovariance* at $\vec{h}$ and 0 lags. The set of values $\{r(\vec{h})\}$ can then be plotted against the lag $\vec{h}$, to see how the pattern of correlation varies with lag. This plot is known as a *correlogram*, and provides a valuable insight into the behavior of the autocorrelation at different lags or "distances".

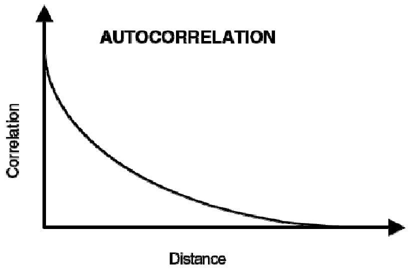


Figure 21: The anatomy of a *correlogram*

Summary:

- The covariance function decreases with distance, so it can be considered a similarity function.
- The semivariance of the difference increases with distance on semivariogram, so it can be considered a dissimilarity function.

There are mathematical relationships between the semivariogram and the covariance functions that appear as $\gamma(\vec{s}_\mathrm{i},\vec{s}_\mathrm{j}) = \mathrm{sill} - C(\vec{s}_\mathrm{i},\vec{s}_\mathrm{j})$. If the regionalized variable is stationary, the semivariance for a distance d is equal to the difference between the variance and the auto-covariance for the same distance:

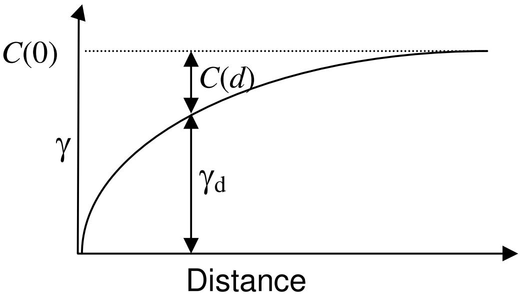


Figure 22 : Relationship between semivariance γ and autocovariance $C(d)$ for a stationary regionalized variable. Where $C(0)$ is the variance of the observation, or the auto-covariance at $d = 0$.

If the regionalized variable is not only stationary, but also is standardized to have a mean of zero and variance of 1, the semivariogram is a mirror image of the autocorrelation function:

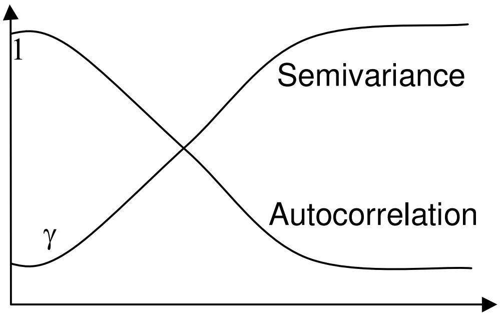


Figure 23: Relationship between semivariance and autocorrelation for a stationary regionalized variable

So far, the models of empirical semivariogram and covariance clouds, which provide information on the spatial autocorrelation of datasets, have been discussed. However, empirical semivariogram and covariance models do not provide information for all possible directions and distances. The main application of geostatistics is the prediction (*optimal interpolation*) of attribute values at unsampled locations (kriging).

For this interpolation reason, it is necessary to ***fit*** or approximate the empirical semivariogram/covariance cloud by a continuous function or curve. This function is described the *theoretical variogram* model, which expresses semivariance as a *function of separation vector.* The theoretical variogram model is a function that will characterize the dependence existing between variables at different points in space. This dependence is assumed to be a function of the distance and direction that separates values of variables.

This fitted model is used in the kriging equations. Abstractly, this is similar to regression analysis, where a continuous line or a curve of various types is fitted. However, only some functions can be used (*authorized models*) to fit semivariogram and covariance clouds. Any variogram function must be able to model the following: monotonically increasing values**,** possibly with a fluctuation (hole); constant or asymptotic maximum (sill), non-negative intercept (nugget) and anisotropy. Theoretical variograms must obey mathematical constraints so that the resulting kriging equations are solvable (e.g., positive definite between-sample covariance matrices).

The following functions are authorized to model the empirical semivariogram: circular, spherical, tetraspherical, pentaspherical, exponential, Gaussian, rational quadratic, hole effect, K-Bessel, J-Bessel, etc (see the descriptions of some mathematical models at http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?topicname=how_kriging_works). Any linear combination of authorized models is also authorized.

The selected model influences the prediction of the unknown values, particularly when the shape of the curve near the origin differs significantly. The steeper curve near the origin will bring the more influence of the closest neighbors to the prediction. As a result, the output surface will be less smooth.

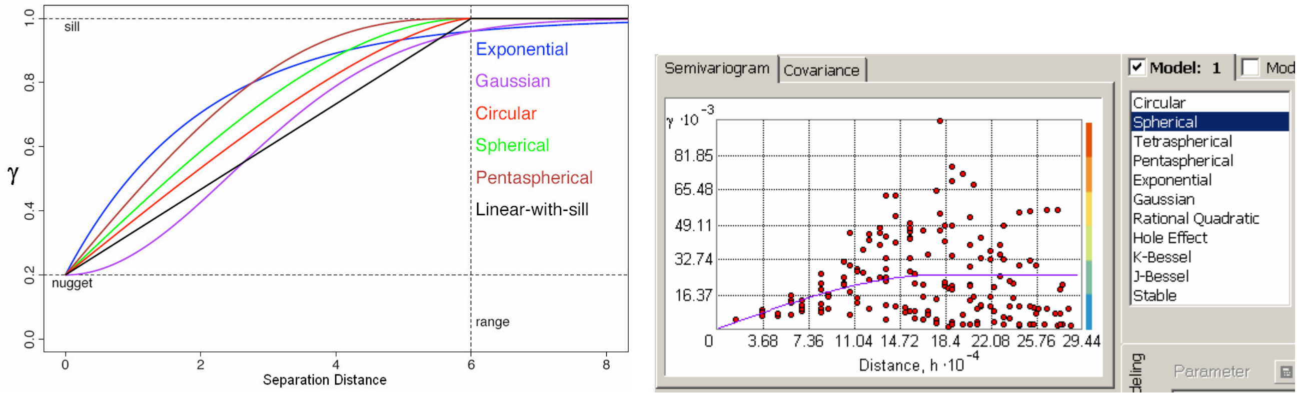


Figure 24 : Comparison of variogram models and spherical model (blue curve), which is derived from the intersection of random spheres of a given size, is used to fit the *empirical* variogram.

There are no exact rules on choosing the “best” variogram model or function. Each model is designed to fit different types of phenomenon more accurately. For example, a Gaussian model might be expected for a phenomenon that physically must be very continuous, such as the surface of a ground-water table. A model that looks appropriate could be picked up based on an expert's examination of empirical semivariogram or covariance functions, and validation and cross-validation statistics as a guide can be used (will be discussed later in the module).

Let's look on an example of a theoretical variogram. The theoretical variogram is needed because the empirical semivariogram values cannot be used directly in the matrix calculations in the kriging system (discussed in next section). Thus, values from empirical variograms can introduce negative standard errors for the predictions. Instead, the authorized fitted model has to be used when determining semivariogram values for various distances. These authorized models are designed in such way that they cannot introduce negative standard errors for the predictions.

The *empirical* variogram in the previous example was calculated as the following:

Lag distances in meters $(\vec{h})$	**Empirical semivariances** $\gamma(\vec{h})$
From 1 to 2	31.25
From 2 to 3	8.33
From 3 to 4	31.25
From 4 to 5	81.25
More then 5	112.5

For example, a simple linear authorized model may use $\gamma(\vec{h}) = c * \vec{h}$, where c is constant and defines the slope of the theoretical variogram line. Based on the regression analysis adjustment for the averaged $\gamma(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}})$ from the table above, c is calculated to be 13.5. So, the theoretical model will be $\gamma(\vec{h}) = 13.5\vec{h}$.

Based on this model, the theoretical semivariances between known points $\gamma(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}})$ are calculated and presented in the following table:

Locations $(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}})$	**Distances** $d(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}}) = \sqrt{(x_{\mathrm{i}} - x_{\mathrm{j}})^2 + (y_{\mathrm{i}} - y_{\mathrm{j}})^2}$	**Theoretical semivariances** $\gamma(\vec{s}_{\mathrm{i}}, \vec{s}_{\mathrm{j}}) = 13.5\vec{h}$
(1,5),(4,3)	3.606	48.681
(1,5),(1,3)	2	27
(1,5),(4,5)	3	40.5
(1,5),(5,1)	5.657	76.3695
(4,3),(1,3)	3	40.5
(4,3),(4,5)	2	27
(4,3),(5,1)	2.236	30.186
(1,3),(4,5)	3.606	48.681
(1,3),(5,1)	4.472	60.372
(4,5),(5,1)	4.123	55.6605

The following is a conceptual **summarization** of the variography technique:

- A quantitative measure of the data value confidence is a statistical parameter called variance. Variance is a measure of the uncertainty of a value. In the kriging method, every known data value and every missing data value has an associated variance. For a constant or exact known value, the variance is zero. In worst case scenarios, when there is no trust on a data value, the variance of such a value is one on a normalized scale.

- Conceptually, as it is shown later on, the variance in kriging plays the role of a weighting function. For example, there is a single data value with zero variance. On can be relatively assured that the missing values physically close to the known location will be well approximated by the known value. With the points further away, there will be less certainly about unknown values; the uncertainty increases with distance from the known value.
- Variance for each known data value can be set to the uncertainty of that value. The estimation for each of the unknown values is more concerned with the relative changes of uncertainties rather than with their absolute values.
- Each known data value has a variance function associated with it that is used to determine variance of data values around the known location. The shape of the function is empirical and can be one of four forms: linear, spherical, exponential, or Gaussian. These curves have their minimum value (usually 0) at the known data location and their maximum value (usually 1) at some specified distance (range) away from that point. All locations outside the range are considered to be unaffected by the known data value. Just as one value for the uncertainty of all known data values can be set, the same, one range can be used throughout the calculations.
- The variance and range allows for a variance discontinuity at the known data value, commonly referred to as the nugget. This causes a step increase in variance just away from the known data value. In the best case scenario, this value set would be 0, producing no nugget effect.
- The variogram is *estimated* from the data in two steps:
  1. At first, the empirical variogram is estimated with a particular lag size and anisotropy parameters;
  2. In addition, a model to theoretical variogram is fitted with a particular authorized function.

4.4.2. Kriging: Regionalized Variable Theory

A unique aspect of geostatistics is the use of *regionalized variables,* which are variables that fall between random variables and completely deterministic variables – a concept that assumes the spatial variation of *regionalized variables* is sum of:

- Structural deterministic component, having constant mean or trend - $\mu(\vec{s}_\mathrm{i})$
- Random, but spatially correlated component - $\eta(\vec{s}_\mathrm{i})$
- Spatially uncorrelated random noise - $\varepsilon(\vec{s}_\mathrm{i})$

Thus, random regionalized variables $\mathrm{z}(\vec{s}_\mathrm{i})$ at locations $\vec{s}_\mathrm{i}$ is described as $\mathrm{z}(\vec{s}_\mathrm{i}) = \mu(\vec{s}_\mathrm{i}) + \eta(\vec{s}_\mathrm{i}) + \varepsilon(\vec{s}_\mathrm{i})$. In this model, the nugget effect can be composed of the variance of $\eta(\vec{s}_\mathrm{i})$ that is called micro-scale variation, plus the variance of $\varepsilon(\vec{s}_\mathrm{i})$ that is also called measurement error.

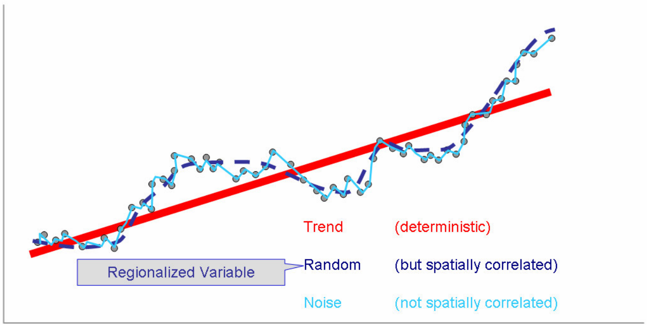


Figure 25 : Kriging methods depend on mathematical and statistical models. Kriging methods rely on the notion of autocorrelation.

The general *formula* for Kriging interpolators is formed as a weighted sum of the data:

$\mathrm{z}(\vec{s}_0) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \lambda_{\mathrm{i}} Z(\vec{s}_{\mathrm{i}})$, where $Z(\vec{s}_{\mathrm{i}})$ is the measured value at the i-th location, λ_{i} is an unknown weight for the measured value at the i-th location, $\mathrm{z}(\vec{s}_0)$ is the prediction in location $\vec{s}_0$ and n is the number of measured values. Each point $\mathrm{z}(\vec{s}_0)$ is predicted as the *weighted average* of the values at *all* sample points $Z(\vec{s}_{\mathrm{i}})$.

Kriging is similar to IDW in that it weights the surrounding measured values to derive a prediction for an unmeasured location. The general formula for both interpolators is formed as a weighted sum of the data. However, in IDW, the weight depends solely on the distance to the prediction location. Kriging weights come from a semivariogram that was developed by looking at the spatial nature of the data - the spatial autocorrelation is quantified by using semivariances. Thus, kriging is based on the theory of random processes, with covariances depending only on separation (i.e. a variogram model). Nevertheless, in Kriging the weights are based not only on the distance between the measured points and the prediction location, but also on the overall spatial arrangement among the measured points.

In addition, the weights used in kriging involve not only the semivariances between the points to be established and the known points (IDW method uses distances between the points to be established and the known points), but also those between the known points.

Various kriging techniques are based on certain assumptions. Thus *simple* kriging is linear with a known trend; *ordinary* kriging is linear with an unknown flat trend; *universal* kriging is linear with an unknown polynomial trend; *co-kriging* is linear and multivariate and can have different types of trends; *Trans-Gaussian* kriging is linear after transformation with a flat trend; *indicator* and *disjunctive* kriging is non-linear and works with threshold or binary data; *block kriging* is linear and works with average value over some small area (block) rather than at a point.

First, let us examine **ordinary** kriging. In *ordinary* kriging, the λ_{i} weight depends on a fitted model to the measured points, the distance to the prediction location, and the spatial relationships among the measured values around the prediction location. Predictions are made as linear combinations of known data values (a weighted average). The *ordinary* kriging prediction is ***unbiased.*** The known points are predicted *exactly*; they are assumed to be without error, even if there is a nugget effect in the variogram model.

In ordinary kriging it is also true what points closer to the point are predicted to have larger weights. Clusters of points “reduce to” single equivalent points (i.e., over-sampling in a small area cannot bias result). Closer sample points “mask” further ones in the same direction. Error estimates are based only on the sample configurations, not the data values.

The theory of regionalized variables leads to an “optimal” interpolation method, in the sense that the prediction variance is minimized. In ordinary kriging, prediction error should be as small as possible. Ordinary kriging, as a “best linear unbiased predictor”, has to satisfy certain criteria for optimality. However, it is only “optimal” with respect to the *chosen fitted model*!

In ordinary kriging models the value in location $z(\vec{s}_i) = \mu(\vec{s}_i) + \eta(\vec{s}_i)$ is the sum of a regional mean $\mu(\vec{s}_i)$ and a spatially-correlated random component $\eta(\vec{s}_i)$. The regional mean $\mu(\vec{s}_i)$ is estimated from the sample, but not as the simple average, because there is spatial dependence. It is *implicit* in the ordinary kriging system. Therefore, ordinary kriging predicts at points, with ***unknown*** mean (must be estimated) and there is no trend (or flat trend).

When making predictions for several locations with ordinary kriging, it is expected what some of the prediction values will be above the actual values and some below. Nevertheless, on average, the difference between the predictions and the actual values should be zero (*first order* stationary condition). This is referred to as “making the prediction ***unbiased”*** and this is the main constraint of ordinary kriging. Formally, this constraint can be used to satisfy the sum of the weight λ_i assigned to each sample point sum to one. The unbiased condition is:

$$\sum_{i=1}^{n} \lambda_i = 1$$

The variance at any point is finite and the same at all locations in the field; and the covariance structure depends only on the separation between point pairs (*second order* stationary condition).

Using the *unbiased* constraint together with an “optimal” interpolation assumption that the prediction variance is as small as possible - $(z(\vec{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\vec{s}_i))^2 = \min$, that is the difference between the true value $z(\vec{s}_0)$ and the predictor $\lambda_i Z(\vec{s}_i)$ in unknown location $\vec{s}_0$. The solution to the minimization, constrained by unbiasedness, gives the ordinary kriging equations:

$$\begin{bmatrix} \gamma(\vec{s}_1,\vec{s}_1) & \dots & \gamma(\vec{s}_1,\vec{s}_n) & 1 \\ \dots & \dots & \dots & \dots \\ \gamma(\vec{s}_n,\vec{s}_1) & \dots & \gamma(\vec{s}_n,\vec{s}_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \dots \\ \lambda_n \\ \Psi \end{bmatrix} = \begin{bmatrix} \gamma(\vec{s}_1,\vec{s}_0) \\ \dots \\ \gamma(\vec{s}_n,\vec{s}_0) \\ 1 \end{bmatrix}$$

or in matrix notation: $\mathrm{A} * \boldsymbol{\lambda} = \mathrm{b}$, where $\mathrm{A} = \begin{bmatrix} \Gamma & 1 \\ 1^T & 0 \end{bmatrix}$, $\boldsymbol{\lambda} = \begin{bmatrix} \Lambda \\ \Psi \end{bmatrix}$, $\mathrm{b} = \begin{bmatrix} \Gamma_0 \\ 1 \end{bmatrix}$. See more about matrix expressions at the http://www.spatialanalysisonline.com/output/html/Matrixexpressions.html.

The γ_{ij} semivariance values in matrix A and b are taken from the mathematical expression of the semivariogram (*fitted model*). Also, a fourth variable is introduced called the LaGrange multiplier Ψ, to assure that the minimum possible estimation error is obtained. The Ψ depends on covariance structure of the sample points.

This is a system of $n+1$ equations in $n+1$ unknowns, so can be adjusted optimally, as long as A is a *positive definite* matrix and this is ***guaranteed*** by using *authorized* fitted models! This system has the following solution in matrix notation:

$\lambda = A^{-1}b$, where A^{-1} is inversed matrix of A.

The *weights* for each predicted point, based on the *point configuration* and the *modelled variogram*, are computed by an *optimization* criterion, which in ordinary kriging is *minimizing the prediction variance*.

The term “ordinary” infers there is no trend or strata; the regional mean must be estimated from the sample. One of the main issues concerning ordinary kriging is whether the assumption of a constant mean is reasonable. Sometimes there are good scientific reasons to reject this assumption.

Ordinary kriging can use either semivariograms or covariances to express autocorrelation, it can use transformations and remove trends, and it can allow for *measurement error*. The ordinary kriging *prediction error* or kriging variance $\hat{\sigma}$ at a point depends on the semivariances between the prediction point and the sample points b and the weights (including LaGrange multiplier) λ computed in the ordinary kriging system. The ordinary kriging variance at the point is computed from $\hat{\sigma}^2(\vec{s}_0) = b^T\lambda$.

The variance measure $\hat{\sigma}$ is the important difference between kriging method and other interpolation methods e.g. IDW. The $\hat{\sigma}$ can be used for each predicted point to estimate the reliability of interpolation.

Ordinary kriging can be summarized by engaging in the following computational steps when predicting the value for *each* unsampled point:

1. Compute distances between *all pairs of sample points* and after compute the respective semivariances using the *fitted variogram model* and these distances for matrix A.

2. Compute distances between the *prediction unknown point and each sample known point* and after compute the respective semivariances using the *theoretical variogram* model and these distances for the vector b.

Steps 1 and 2 of the kriging system use modeled (or fitted, or theoretical) semivariances. Different fitted models could give different kriging weights to the sample points and these will give different prediction results.

3. Complete A and b with 1’s and a 0.

4. Solve the kriging system of linear equations for the n weights λ_i and the LaGrange multipler Ψ. The kriging equations are solved separately for each point $\vec{s}_0$, using the semivariances around that point, in a *local neighborhood*; this gives a different set of weights λ_i for each point to be predicted.

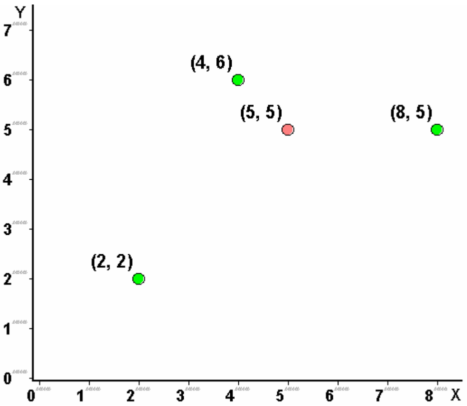
5. Predict the point as the weighted average by λ_i of the sample points from $z(\vec{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\vec{s}_i)$.

6. Compute the prediction error $\hat{\sigma}$ as the scalar product of b and the λ vector.

Note: The variogram model $\gamma(\vec{h})$ used in these equations is estimated only once, using information about the spatial structure over the *whole study area*, and so, the semivariances between sample points $\gamma(\vec{s}_i, \vec{s}_j)$ are computed only once for any point configuration. However, the semivariances at a sample point $\gamma(\vec{s}_i, \vec{s}_0)$ must be computed separately for each point to be predicted.

Let's consider a sample from the following locations:

Locations $\vec{s}(x_i, y_i)$	
(2,2)	known
(4,6)	known
(8,5)	known
(5,5)	unknown



Exponential fitted model $\gamma(\vec{h}) = c(1 - e^{(-\frac{\vec{h}}{a})})$ is used with the following variogram parameters: sill $c = 10$, effective range $a = 1.5$, and nugget $c_0 = 0$ to fit theoretic variogram. Based on this model, semivariances between known points $\gamma(\vec{s}_i, \vec{s}_j)$ and known and predicted point $\gamma(\vec{s}_i, \vec{s}_0)$ are calculated and presented in the following table, as well as distances:

Location pairs	**Distances between sample points** d_{ij}	**Distances between sample points and prediction point** d_{i0}	**Semivariances between sample points** $\gamma(\vec{s}_i, \vec{s}_j)$	**Semivariances between sample points and prediction point** $\gamma(\vec{s}_i, \vec{s}_0)$
(1,2)	4.472		9.493	
(1,3)	6.708		9.886	
(1,4)		4.243		9.409
(2,3)	4.123		9.360	
(2,4)		1.414		6.105
(3,4)		3		8.647

Ordinary kriging equations, which based on *Exponential* fitted model, can be represented as the following matrix:

$$\begin{bmatrix} 0 & 9.492 & 9.886 & 1 \\ 9.492 & 0 & 9.360 & 1 \\ 9.886 & 9.360 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \Psi \end{bmatrix} = \begin{bmatrix} 9.409 \\ 6.105 \\ 8.647 \\ 1 \end{bmatrix}$$

Solution of these equations will be the weights λ_i:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \Psi \end{bmatrix} = \begin{bmatrix} 0.1989525 \\ 0.5321374 \\ 0.2689101 \\ 1.6990883 \end{bmatrix}$$

The unbiasedness constraint is satisfied - $0.1989525 + 0.5321374 + 0.2689101 = 1$. The prediction error $\hat{\sigma}$ can be computed as:

$$\hat{\sigma}^2(\vec{s}_0) = \begin{bmatrix} 9.409 & 6.105 & 8.647 & 1 \end{bmatrix} * \begin{bmatrix} 0.1989525 \\ 0.5321374 \\ 0.2689101 \\ 1.6990883 \end{bmatrix} = 9.145$$

Now let us briefly at other kriging techniques. There may be situations where *the regional mean is known*. Sometimes it makes sense to assume a physically-based model that gives a known trend. In the **simple kriging** model, the value in location $z(\vec{s}_i) = \mu + \eta(\vec{s}_i)$ as the sum of a *known constant* μ and a spatially-correlated random component $\eta(\vec{s}_i)$. For simple kriging, because μ is known exactly, $\eta(\vec{s}_i)$ is also known exactly at the known locations. If $\eta(\vec{s}_i)$ is known, then it is easier to estimate the autocorrelation than if $\eta(\vec{s}_i)$ had been estimated. Simple kriging uses the residuals (the difference between the model and the observations), assuming that the trend in the residuals is known to be zero.

There is *no need for a LaGrange multipler* in the simple kriging system. The simple kriging estimate *without the constraint* that weights sum to 1, that is $\sum_{i=1}^{n} \lambda_i \neq 1$. However, any bias from the weights must be compensated with respect to the (known) mean when predicting at a point. The equations of the simple kriging system are:

$$\begin{bmatrix} \gamma(\vec{s}_1,\vec{s}_1) & \ldots & \gamma(\vec{s}_1,\vec{s}_n) \\ \ldots & \ldots & \ldots \\ \gamma(\vec{s}_n,\vec{s}_1) & \ldots & \gamma(\vec{s}_n,\vec{s}_n) \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \ldots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \gamma(\vec{s}_1,\vec{s}_0) \\ \ldots \\ \gamma(\vec{s}_n,\vec{s}_0) \end{bmatrix}$$

Simple kriging also allows for measurement error.

The theory of regionalized variables can incorporate cases of *first-order non-stationarity* (i.e. where a significant *trend surface* to the geographic coordinates exists or *strata* have significantly different means). The *intrinsic* hypothesis only needs *local* first-order stationarity, so ordinary kriging can be applied in local neighborhoods and would work in these cases. However, even then, useful information about spatial structure is discarded. Accounting for a *global* trend would improve predictions and allow one a better understanding of the processes that form the spatial variation.

Specially designed kriging methods could model both a *global* trend or stratification, and a *local* spatial-dependence structure at the same time. One such method is **universal kriging** – a procedure that includes a global trend as a function of the geographic coordinates within the kriging system. Another one is **regression kriging**, also called “kriging after de-trending”, that models the *trend* (geographic or feature space) and its *residuals* separately.

Universal kriging is a mixed interpolator that models a *global trend* as a function of the *geographic coordinates* in the kriging system. In universal kriging, the value of variable z at location $\vec{s}_\mathrm{i}$ is modeled as the sum of a regional *non-stationary trend* $\mu(\vec{s}_\mathrm{i})$ and a *spatially-correlated random component* $\eta(\vec{s}_\mathrm{i})$:

$\mathrm{z}(\vec{s}_\mathrm{i}) = \mu(\vec{s}_\mathrm{i}) + \eta(\vec{s}_\mathrm{i})$, where $\mu(\vec{s}_\mathrm{i})$ is not a constant as in ordinary kriging, but a *deterministic function* of position (in geographic space) (i.e. the global *trend*). This trend is modeled as a linear function of p-order known *base functions* $\mathrm{f}_\mathrm{j}(\vec{s}_\mathrm{i})$ (e.g. global polynomials) and p unknown constant coefficients (or model parameters) β_j as $\mathrm{z}(\vec{s}_\mathrm{i}) = \sum_{\mathrm{j}=0}^{\mathrm{p}} \beta_\mathrm{j} \mathrm{f}_\mathrm{j}(\vec{s}_\mathrm{i}) + \eta(\vec{s}_\mathrm{i})$. The $\eta(\vec{s}_\mathrm{i})$ is the *spatially-correlated error*, which is modeled as before, with a variogram, but now only considering the *residuals*, *after the global trend is removed.* A universal kriging point is *predicted* the same as in ordinary kriging - $\mathrm{z}(\vec{s}_0) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \lambda_\mathrm{i} Z(\vec{s}_\mathrm{i})$, only the weights λ_i for each sample point take into account both the global trend and local effects.

The sample *base functions* for *linear* drift (or for first order polynomial) are $\mathrm{f}_0(\vec{s}_\mathrm{i}) = 1$, $\mathrm{f}_1(\vec{s}_\mathrm{i}) = x$, $\mathrm{f}_2(\vec{s}_\mathrm{i}) = y$, where x is abscissa and y the ordinate of a point. For *quadratic* drift, also second-order terms will be included that are $\mathrm{f}_3(\vec{s}_\mathrm{i}) = x^2, \mathrm{f}_4(\vec{s}_\mathrm{i}) = xy$ and $\mathrm{f}_5(\vec{s}_\mathrm{i}) = y^2$. Note that $\mathrm{f}_0(\vec{s}_\mathrm{i}) = 1$ estimates the global mean as it is in ordinary kriging.

The *unbiasedness* condition for universal kriging is expressed with respect to the *trend* as well as the overall mean (as in ordinary kriging): $\sum_{\mathrm{i}=0}^{\mathrm{p}} \lambda_\mathrm{i} \mathrm{f}_\mathrm{j}(\vec{s}_\mathrm{i}) = \mathrm{f}_\mathrm{j}(\vec{s}_0)$ and $\sum_{\mathrm{i}=1}^{\mathrm{n}} \lambda_\mathrm{i} = 1$. The expected value at each point of all the *functions* must be that predicted by that function. The first of these is the overall mean as in ordinary kriging.

The semivariances $\gamma(\vec{s}_\mathrm{i}, \vec{s}_\mathrm{j})$ for universal kriging are based on the *residuals*, not the original data, because the random part of the spatial structure applies only to these residuals. The fitted model of variogram is obtained in three steps:

1. Calculate the best-fit trend surface that will be used in universal kriging

2. Subtract trend from the sampled values to get residuals
3. Model the variogram of the *residuals*

The model *parameters* for the residuals will usually be very different from the original variogram model. It often has a lower sill and shorter range because the global trend has been taken out of some of the variation and the long-range structure.

The universal kriging system solves the following equations:

$$\begin{bmatrix} \gamma(\vec{s}_1,\vec{s}_1) & \ldots & \gamma(\vec{s}_1,\vec{s}_n) & 1 & \mathrm{f}_j(\vec{s}_i) & \ldots & \mathrm{f}_j(\vec{s}_i) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \gamma(\vec{s}_n,\vec{s}_1) & \ldots & \gamma(\vec{s}_n,\vec{s}_n) & 1 & \mathrm{f}_j(\vec{s}_i) & \ldots & \mathrm{f}_j(\vec{s}_i) \\ 1 & \ldots & 1 & 0 & 0 & 0 & 0 \\ \mathrm{f}_1(\vec{s}_1) & \ldots & \mathrm{f}_1(\vec{s}_n) & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathrm{f}_p(\vec{s}_1) & \ldots & \mathrm{f}_p(\vec{s}_n) & 0 & 0 & \ldots & 0 \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \ldots \\ \lambda_n \\ \Psi_0 \\ \Psi_1 \\ \ldots \\ \Psi_p \end{bmatrix} = \begin{bmatrix} \gamma(\vec{s}_1,\vec{s}_0) \\ \ldots \\ \gamma(\vec{s}_n,\vec{s}_0) \\ 1 \\ \mathrm{f}_1(\vec{s}_0) \\ \ldots \\ \mathrm{f}_p(\vec{s}_0) \end{bmatrix}$$

Global universal kriging can be using *all* sample points when predicting each point. This gives the same results as **regression kriging.** Such global variation is appropriate if there is a regional trend. Universal kriging can also be *local* when the number sample points are restricted to *neighbours* around the prediction point. *Local* variation of universal kriging allows the trend surface to vary over the study area, since it is re-computed at each point. Similar to *simple kriging*, if *the trend is known, a "**simple**"* variant can be used to universal kriging.

In **regression kriging**, $\eta(\vec{s}_i)$ can be obtained by subtracting the p-order polynomial from the original data, when $\eta(\vec{s}_i)$ are assumed to be random. The mean of all $\eta(\vec{s}_i)$ is 0. Conceptually, the autocorrelation is now modeled from the random errors $\eta(\vec{s}_i)$. Regression kriging can be considered as a type of a polynomial regression. However, instead of assuming the errors $\eta(\vec{s}_i)$ are independent, they are modeled as if autocorrelated. Universal and regression kriging allow for measurement error as well.

Regression kriging is accomplished by taking the following steps:

1. Calculate *trend* and get its prediction error of linear model
2. Subtract trend to get *residuals* for known points
3. Model the kriging residuals – this can be done for simple kriging when the known mean of the residuals is 0 and the prediction error of residuals can be determined
4. Add trend back to modeled residuals in order to get estimates in unknown points
5. Add the two prediction variances (prediction error of linear model and prediction error of residuals) at each point to get the overall error

It may be the case that the observed data is *binary* (with values of 0 or 1) or a variable that is continuous may be reclassified into a binary variable by choosing some threshold. For example, if values are above the threshold, they become a 1, and if they are below the threshold, they become a 0. For example, surface can be classified as land as 1 and water body as 0. The **indicator kriging** model is $\mathrm{z}(\vec{s}_i) = \mu(\vec{s}_i) + I(\vec{s}_i)$, where $\mu(\vec{s}_i)$ is an unknown constant and $I(\vec{s}_i)$ is a *binary* *variable.* Using binary variables, indicator kriging calculates exactly the same as ordinary kriging. The interpolations will be between 0 and 1 and predictions from indicator kriging can be interpreted as probabilities of the variable being a 1 or of being in the class that is indicated by a 1.

The extension of the theory of kriging of regionalized variables to several variables, which have a *multivariate spatial cross-correlation* as well as the individual *univariate spatial auto-correlation*, is called **co-regionalization**. **Co-kriging** is a method of using supplementary information on co-regionalized variables (*co-variables*) to improve the *prediction* of a *target* variable.

The idea of co-regionalization is that the process that drives one variable is the same, or at least related to, the process that describes the other variables. For example, distribution of heavy metals in soil can relate pollution or distribution of water pH level can relate to pollution and elevation. All the variables involved have to be *regionalized* variables and, in addition, if they are related both in geographic space, they are *co-regionalized*.

For example, a *target variable* has relatively few observations and can involve expensive additional measurement. However, when more values of a *second variable* are available and a second variable is *co-related* with the target variable, then this becomes the *co-variable*. This second variable is, for example, easy and cheap to measure, so there are many observations of it. Typically, there are *more observations of the co-variable* (i.e. the target variable was not measured at some points).

Therefore, for theory of co-regionalization involves two types of variograms: the first variogram is *direct* - that is a single variogram for each regionalized variable; the second is a *cross* variogram – a pair of regionalized variables. The *cross empirical* variogram will be estimated for distances that are multiples of $\vec{h}$ as $\gamma_{1,2}(\vec{h}) = \frac{1}{2m(\vec{h})}\sum_{\mathrm{k}=1}^{m(\vec{h})}\left[\mathrm{Z}_1(\vec{s}_\mathrm{i}) - \mathrm{Z}_1(\vec{s}_\mathrm{j})\right] * \left[\mathrm{Z}_2(\vec{s}_\mathrm{i}) - \mathrm{Z}_2(\vec{s}_\mathrm{j})\right]$, where are $\mathrm{Z}_1(\vec{s}_\mathrm{i})$ is a *target* or main variable of interest and $\mathrm{Z}_2(\vec{s}_\mathrm{i})$ is a *co-variable* or co-regionalized variable. ***Cross-variograms*** can depict either a positive or a negative spatial correlation.

The direct and cross-variograms must be modeled together, with some restrictions to ensure that the resulting co-kriging system can be solved. Therefore, the co-kriging method uses information on *several variable types.* The general models of ***ordinary*** **co-kriging** are:

$$\mathrm{z}_1(\vec{s}_\mathrm{i}) = \mu_1(\vec{s}_\mathrm{i}) + \eta_1(\vec{s}_\mathrm{i})$$
$$\mathrm{z}_2(\vec{s}_\mathrm{i}) = \mu_2(\vec{s}_\mathrm{i}) + \eta_2(\vec{s}_\mathrm{i})$$

Where $\mu_1(\vec{s}_\mathrm{i})$ and $\mu_2(\vec{s}_\mathrm{i})$ are *unknown constants*. There are two types of random errors, $\eta_1(\vec{s}_\mathrm{i})$ and $\eta_2(\vec{s}_\mathrm{i})$*,* so there is autocorrelation for each of them and cross-correlation between them. Ordinary co-kriging attempts to predict $\mathrm{z}_1(\vec{s}_\mathrm{i})$, just like ordinary kriging, but it uses information in the covariate $\{\mathrm{z}_2(\vec{s}_\mathrm{i})\}$ in an attempt to do a better prediction.

The target variable is $\mathrm{z}_1(\vec{s}_\mathrm{i})$, and both autocorrelation for $\mathrm{z}_1(\vec{s}_\mathrm{i})$ and cross-correlations between $\mathrm{z}_1(\vec{s}_\mathrm{i})$ and the other variable type $\mathrm{z}_2(\vec{s}_\mathrm{i})$ are used to make better predictions. It uses information from co-variable $\mathrm{z}_2(\vec{s}_\mathrm{i})$ to help make predictions, but it requires much more estimation, which includes estimating the autocorrelation for each variable, as well as all cross-correlations.

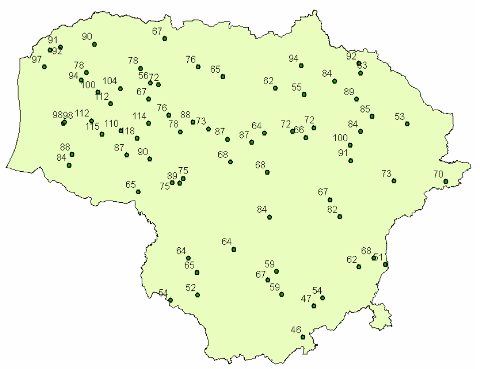
Theoretically, co-kriging can do no worse than kriging because, if there is no cross-correlation, it backs on to a just autocorrelation for $z_1(\vec{s}_i)$. However, practically, the estimation of unknown autocorrelation parameters can introduce more variability and decrease the calculation precision of the predictions.

The *formula* for ordinary kriging interpolators is formed as a weighted sums of the data and will be $z(\vec{s}_0)=\sum_{i=1}^{n}\lambda_i Z_1(\vec{s}_i)+\sum_{i=1}^{n}\omega_i Z_2(\vec{s}_i)$, where $Z_1(\vec{s}_i)$ is the measured value of target variable, λ_i is an unknown weight for the measured value for the target variable; $Z_2(\vec{s}_i)$ is the measured value of co-variable, ϖ_i is an unknown weight for the measured value for the co-variable; $z(\vec{s}_0)$ is the prediction in location $\vec{s}_0$ and n is the number of measured bin's values.

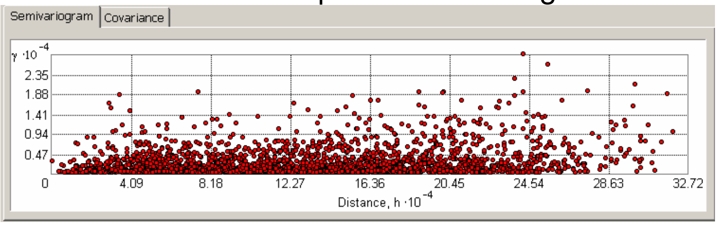
The other co-kriging methods, including universal co-kriging, simple co-kriging, indicator co-kriging, are all generalizations of the foregoing methods to the case where multiple datasets are used. Co-kriging can allow for measurement error in the same situation as for the various kriging methods (ordinary kriging, simple kriging, and universal kriging).

The kriging methods also can be summarized as steps in optimal spatial prediction modeling:

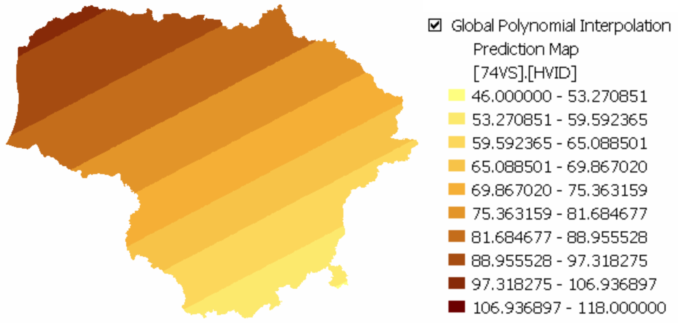
1. Sample phenomena (e.g. long-term average values of water levels), preferably at different resolutions:



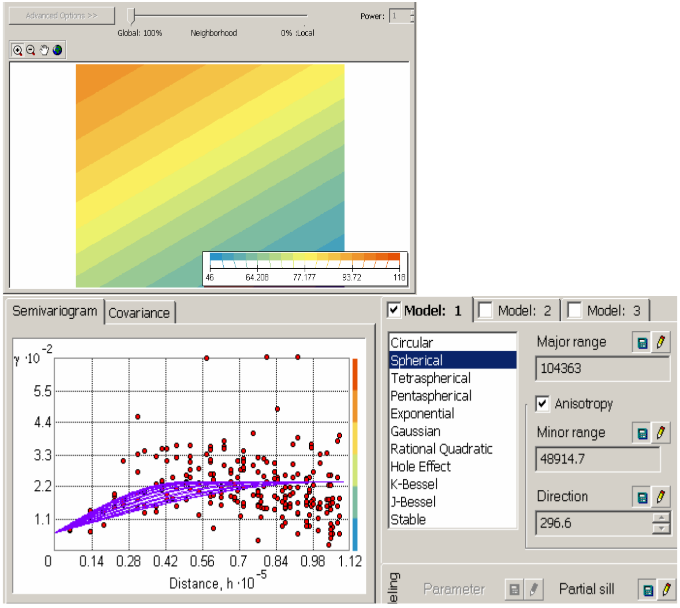
2. Calculate the experimental variogram:



3. Check for trend (e.g. visualize by global polynomial). If it exists, fit with a trend surface model (here is linear in north-west direction):



4. Model or fit the variogram with one or more authorized functions to the residuals $\eta(\vec{s}_{\mathrm{i}}) = \mathrm{z}(\vec{s}_{\mathrm{i}}) - \mu(\vec{s}_{\mathrm{i}})$. The $\eta(\vec{s}_{\mathrm{i}})$ is obtained by subtracting the first-order polynomial (top-left-hand figure) from the original data. Each blue curve on the left-hand variogram represents the fitted spherical model for the particular search direction (e.g. north, north-west, etc.) (anisotropy models). In the right-hand variogram, the north-west fitted model (blue curve) follows the major range's direction.



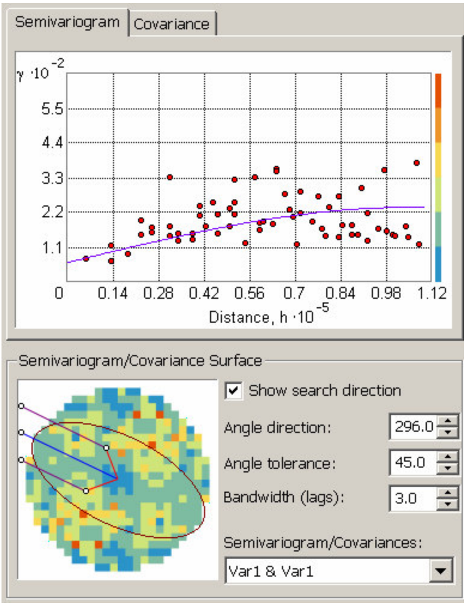
☑ Model: 1 □ Model: 2 □ Model: 3

Major range: 104363

☑ Anisotropy

Minor range: 48914.7

Direction: 296.6



☑ Show search direction

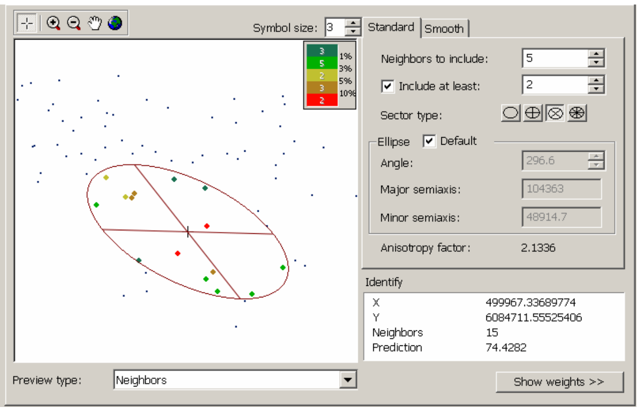
Angle direction: 296.0

Angle tolerance: 45.0

Bandwidth (lags): 3.0

Semivariogram/Covariances: Var1 & Var1

5. Some kriging systems, with the variogram models of spatial dependence, can be applied to produce kriging predictions based on the variogram and trend model at each predicted point. Predictions are often done at each point on a regular grid (e.g. a raster map). For a local approach, the kriging equations are solved separately for each point $\vec{s}_0$, using the semivariances around that point, in a *local neighborhood*; this gives a different set of weights λ_{i} for each point to be predicted (left-hand figure). The predicted plot of the fitted line through the scattered known points is given in blue with the equation given just below the plot (shown on the right-hand figure). If the data were not autocorrelated, all predictions would be the same or every prediction would be the mean of the known data, in this case the blue line would be horizontal. With autocorrelation and a good kriging model, the blue line should be closer to the black dashed line.



Symbol size: 3

Neighbors to include: 5

☑ Include at least: 2

Sector type:

Ellipse ☑ Default

Angle: 296.6

Major semiaxis: 104363

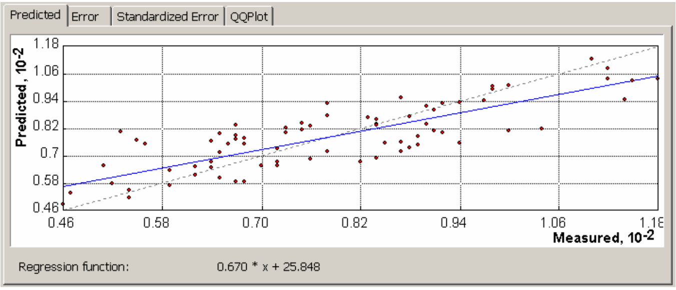
Minor semiaxis: 48914.7

Anisotropy factor: 2.1336

Identify

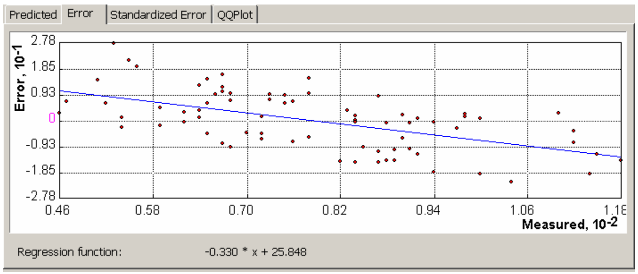
X	499967.33689774
Y	6084711.55525406
Neighbors	15
Prediction	74.4282

Preview type: Neighbors

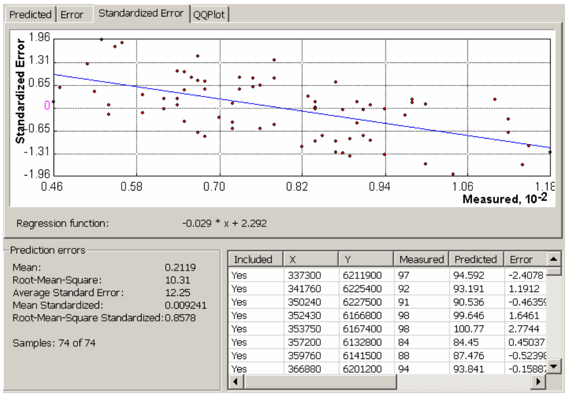


Regression function: 0.670 * x + 25.848

6. Calculate the error of each prediction; this is based only on the sample point locations, *not* their data values. The error plot is the same as the prediction plot, except that the known values are subtracted from the predicted values (left-hand figure). For the standardized error plot, the known values are subtracted from the predicted values, and then divided by the estimated kriging prediction errors $\hat{\sigma}$ and other derivative standardized errors (right-hand figure).



Regression function: -0.330 * x + 25.848



Regression function: -0.029 * x + 2.292

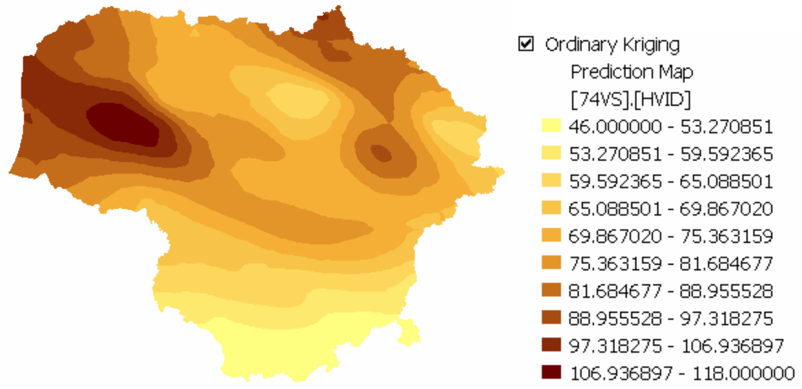
Prediction errors

Mean: 0.2119
Root-Mean-Square: 10.31
Average Standard Error: 12.25
Mean Standardized: 0.009241
Root-Mean-Square Standardized: 0.8578

Samples: 74 of 74

Included	X	Y	Measured	Predicted	Error
Yes	337300	6211900	97	94.592	-2.4078
Yes	341760	6225400	92	93.191	1.1912
Yes	350240	6227500	91	90.536	-0.46359
Yes	352430	6166800	98	99.646	1.6461
Yes	353750	6167400	98	100.77	2.7744
Yes	357200	6132800	84	84.45	0.45037
Yes	359760	6141500	88	87.476	-0.52398
Yes	366880	6201200	94	93.841	-0.15887

7. Map the *predicted values*:



4.5. Model Validation

With any interpolation method, one would like to know how good the results will be. The model, therefore, needs justification. This involves a **model validation** approach and methodology. Model validation needs to be applied during the building of the model, including *calibration* of the model.

The main approaches used to compare model predictions with reality are:

1. Use of independent ***measures*** of validity for *data* and *models*; some of the measures can be represented as a surface.
2. Separate ***validation*** dataset by dividing it into two samples - the training dataset and the test dataset - and compare the predicted values used from training datasets for specified locations with the values from the test datasets.
3. ***Cross-validation*** using *calibration* datasets.
4. Create many interpolation surfaces and use measures when ***comparing*** models.

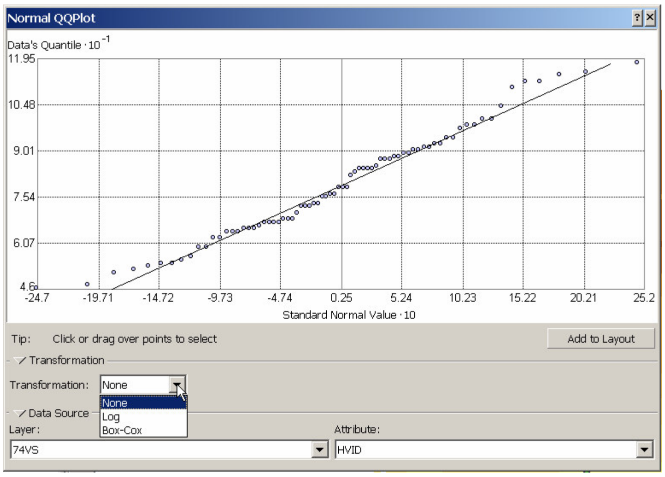
4.5.1. Data Validity

Kriging is based on data modeling. The data have to be tested before they are used in spatial modeling in order to allow for proper interpretation.

At the *data exploration* stage, usually preceded by interpolation, the data have to be examined for normality, outlets, global trend, stationarity, spatial autocorrelation, etc. Some of the kriging and co-kriging methods (e.g. ordinary kriging) require that the data come from normal distributions, therefore, data has to be explored for normality. If data did not normally distributed, normal score transformations may be necessary to apply to the data.

The indicators (measures) that data is normally distributed can be obtained from a histogram. For a normal distribution, a histogram has the unimodal bell shape, the mean and median values are very close, and the kurtosis is close to 3. Normality also can be verified with numeric tests (e.g. Shapiro-Wilks or Anderson).

In addition, the Normal QQPlot of the quantiles of the input dataset versus quantiles of the standard normal distribution can be used. QQ plots are graphs on which quantiles from two distributions are plotted relative to each other. For two identical distributions, the Normal QQPlot will be a straight line. Therefore, comparing this line with the distribution of sampled points on the Normal QQPlot provides an indication of univariate normality. If the distribution of sampled data is asymmetric (i.e., far from normal), the points will deviate from the Normal QQPlot line.



Tip: Click or drag over points to select

Transformation: None

Layer: 74VS

Attribute: HVID

Figure 26 : The Normal QQPlot against the long-term average values of water levels, and the transformation options in the ESRI Geostatistical Analyst. See more about transformation at the http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Box-Cox%2C_Arcsine%2C_and_Log_transformations and http://www.spatialanalysisonline.com/output/html/Datatransformsandbacktransforms.html.

One can use the global polynomial interpolation methods for visual inspection of linear or other trends in data. Cross-section plots of data values along ordinate and abscissa coordinate axes can also give an idea about global data trends.

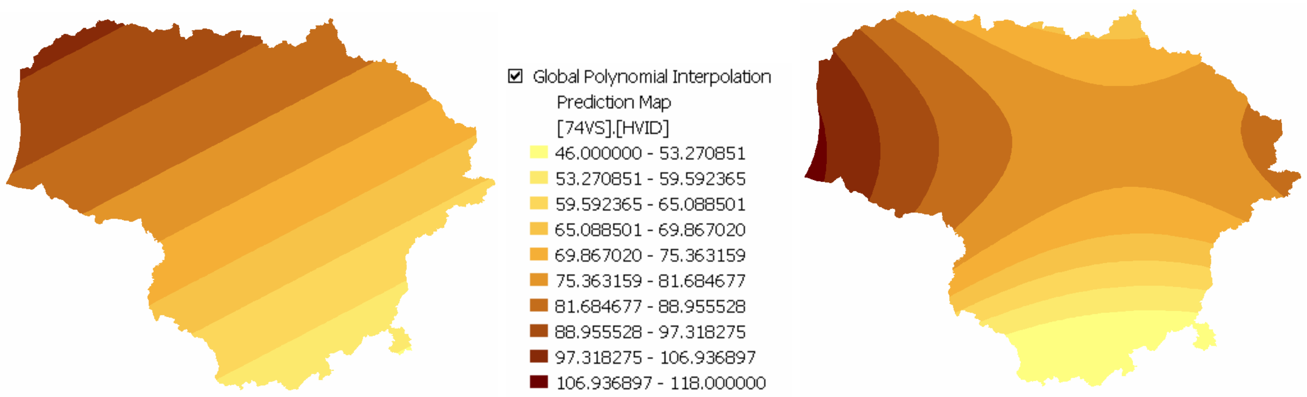
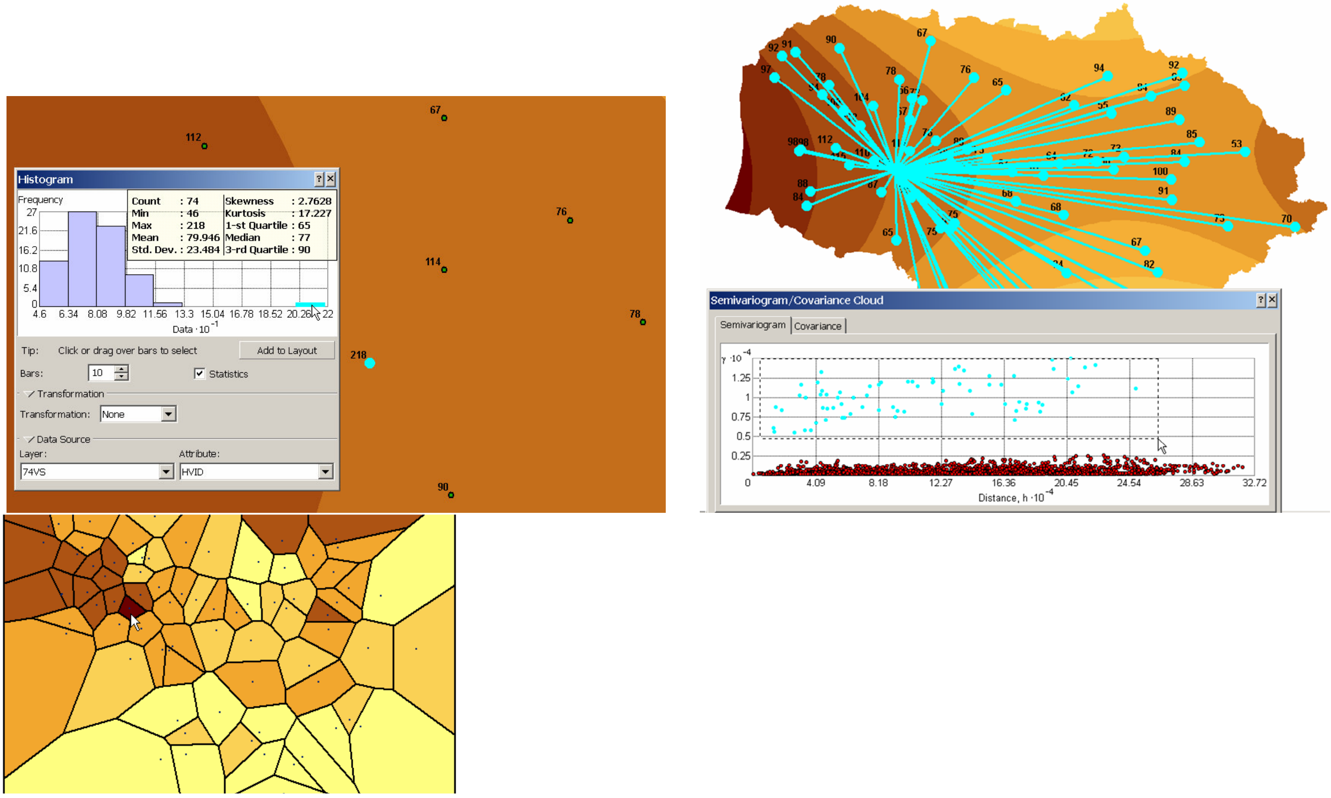


Figure 27 : The long-term average values of water levels has linear (first order) trend (the map on left- hand side). There is no explicit second order trend (the map on right hand side)

The data outliers or abrupt changes in data values, which can be caused by real abnormalities in the phenomenon or measured errors, can be investigated by using the histogram (points on the tail of the distribution), semivariogram (pairs of points with high values in the semivariogram cloud, regardless of distance) and Voronoi diagram (high dissimilarity between neighbors). See about the Voronoi diagram at the http://en.wikipedia.org/wiki/Voronoi_diagram.



Histogram

Count : 74
Min : 46
Max : 218
Mean : 79.946
Std. Dev. : 23.484
Skewness : 2.7628
Kurtosis : 17.227
1-st Quartile : 65
Median : 77
3-rd Quartile : 90

Tip: Click or drag over bars to select

Bars: 10

☑ Statistics

Transformation: None

Layer: 74VS

Attribute: HVID

Figure 28 : View on the outlier – the point with 218 value

The directional semivariogram can be used to investigate isotropic or anisotropic surfaces, as discussed in section 1.4.

4.5.2. Independent Model Validity Measures

Different independent measures of validity can be used to estimate an interpolation result.

Along with these measures are:
Root-mean-squared error (***RMSE***) of the values or residuals - the actual value against estimated value from the model in the *validation* dataset - is $RMSE = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{z}(\vec{s}_{\mathrm{i}}) - \mathrm{Z}(\vec{s}_{\mathrm{i}}))^2}$, there $\mathrm{z}(\vec{s}_{\mathrm{i}})$ is the predicted value and $\mathrm{Z}(\vec{s}_{\mathrm{i}})$ is the observed value.

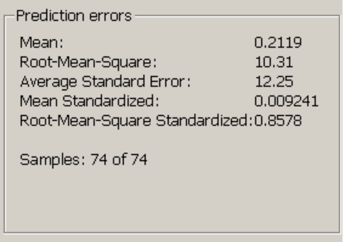
Bias or *mean error* (***ME***) of estimated value against actual mean of the *validation* dataset is $ME = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{z}(\vec{s}_{\mathrm{i}}) - \mathrm{Z}(\vec{s}_{\mathrm{i}}))$

The *kriging prediction error* $\hat{\sigma}$ at a point or kriging variance at the point is computed from $\hat{\sigma}^2(\vec{s}_0) = \mathrm{b}^T\lambda$.

The *average kriging prediction error* is $\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\hat{\sigma}(\vec{s}_{\mathrm{i}})}$.

The *mean standardized prediction error* is $\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{z}(\vec{s}_{\mathrm{i}}) - \mathrm{Z}(\vec{s}_{\mathrm{i}}))/\hat{\sigma}(\vec{s}_{\mathrm{i}})$.

The *root-mean-square standardized prediction error* is $\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left((\mathrm{z}(\vec{s}_{\mathrm{i}}) - \mathrm{Z}(\vec{s}_{\mathrm{i}}))/\hat{\sigma}(\vec{s}_{\mathrm{i}})\right)^2}$.



```
Prediction errors
Mean:                            0.2119
Root-Mean-Square:                10.31
Average Standard Error:          12.25
Mean Standardized:               0.009241
Root-Mean-Square Standardized:0.8578

Samples: 74 of 74
```

Figure 29 : Prediction errors for the resulting surface of the long-term average values of water levels

When one compares these models, one should look for a model that satisfies the following conditions:

1. The lower root-mean-squared error

2. Mean error is near zero
3. The mean standardized prediction error should be nearest to zero
4. The average prediction error should be nearest to the root-mean-squared error
5. The standardized root-mean-squared prediction error should be nearest to one

Thus, the kriging *prediction error* $\hat{\sigma}$ can be estimated for each predicted point. Therefore, if a stochastic method, such as kriging is used, it allows the quantification of prediction errors and representation of these errors as an error surface or map. Deterministic interpolation methods (IDW, splines, etc) do not consider errors! Therefore, only RMSE could be used for comparisons between a deterministic method and another interpolation method.

However, the question is “How to calculate RMSE for exact interpolation?” The RMSE for an exact interpolation can be calculated from *cross-validation*. In cross-validation, each sample point is removed; interpolation surface is created without this point, and this surface is compared to the predicted value for removed location. Such process is repeated for each sample point. The cross-validation RMSE is a summary statistic quantifying the error of the prediction surface. See more about the cross-validation below in the text.

So far, maps of predictions, called ***prediction*** maps or interpolated maps, were created. Prediction maps are produced from the interpolated values. An ***error*** map (e.g. the kriging *prediction error* map) shows error is simply the square root of the variance of a kriging prediction or estimate $\hat{\sigma}^2$. An error map quantifies the uncertainty of the prediction. If the data comes from a normal distribution, the true value will be within ± 2 times the prediction error about 95 percent of the time.

Besides making predictions, the variability of the predictions from the true values is estimated. If the *average kriging prediction* error is greater than the *root-mean-squared* error, the variability of the predictions is overestimated; if the average *kriging prediction* error is less than the *root-mean-squared* errors, the variability in the predictions is underestimated.

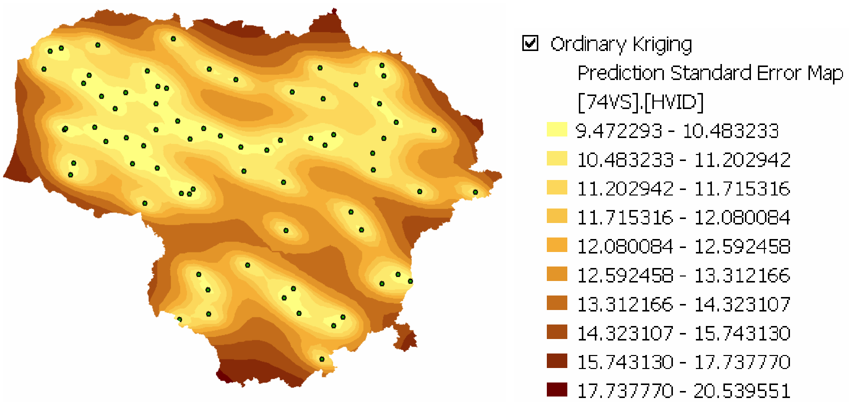


Figure 30 : The error map of long-term average values of water level estimates. Errors are larger in sparsely sampled areas

It is possible to derive two other error estimation representations from the error map. These are ***quantile*** and ***probability*** maps. The values of a ***quantile*** map reflect the upper or lower limits of the true values. Quantile maps represent surfaces of values where the predictions exceed (or do not exceed) the values at the *specified* probability. For example, if the quantile probability is set up to 0.5, an output map will be produced the predicted median values at each known location. If the quantile probability is set up to 0.75, an output map will be produced where there is a 75% chance that an unknown value is below the surface value, and a 25% chance that the unknown value is above the surface value.

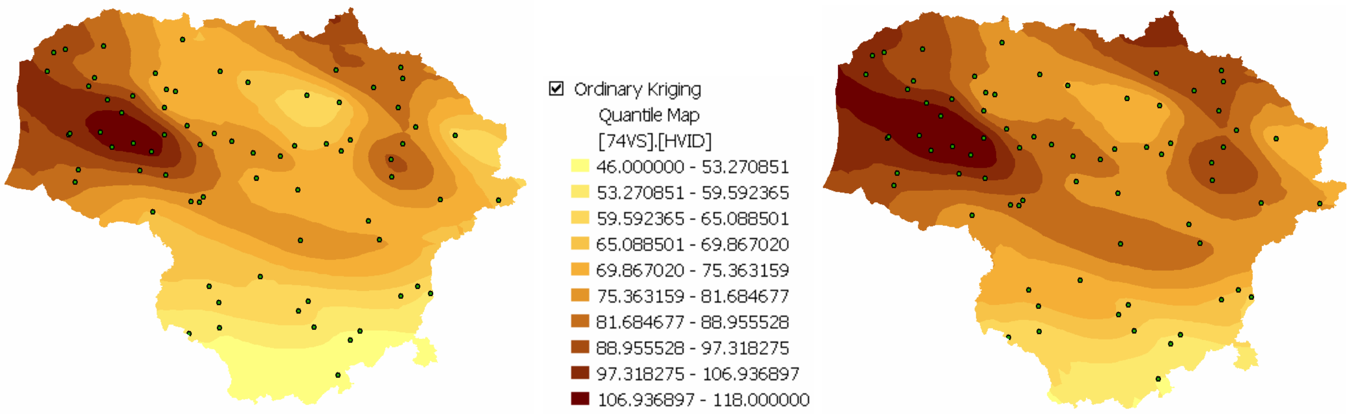


Figure 31: The quantile maps of long-term average values of water levels estimates respectively for the specified 0.5 and 0.75 quantile probabilities

Probability maps show the probability that the true value is above (or below) some specified threshold value. For example, if the threshold value is set up to “*exceed 100*”, the map will show the surface of probabilities when values may exceed the 100-measurement mark.

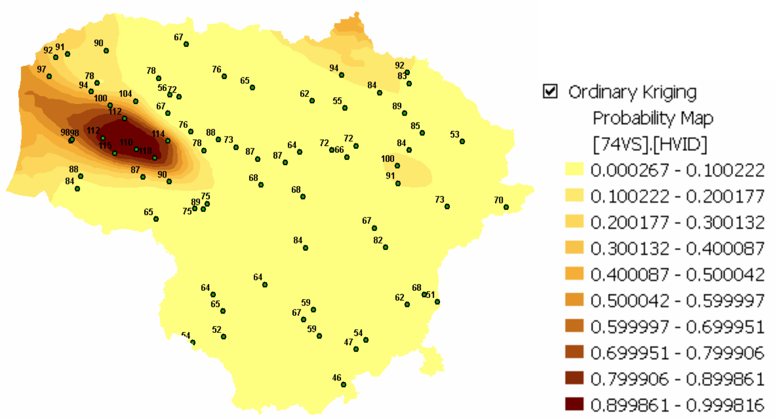
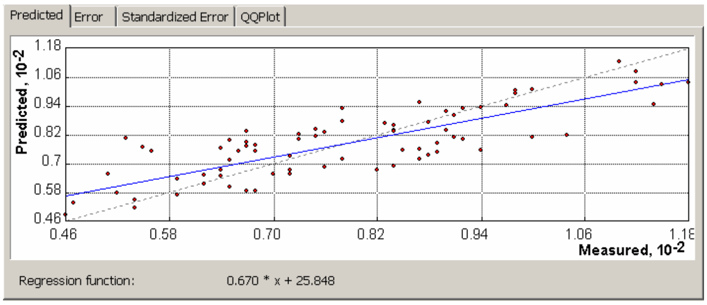


Figure 32 : The probability map of long-term average values of water levels estimates when the specified threshold value exceeds 100

In order to use probability and quantile maps confidently, the data have to come from a full multivariate normal distribution.

Few graphical plots can be used for validation of results of kriging prediction. A scatter plot of predicted values (blue fitted line in the plot given below) versus sample values (dots) is one such plot. It might be expected that the fitted line will be a diagonal line (the black dashed line). However, the slope is usually less than one. It is a property of kriging that tends to under-predict large values and over-predict small values (that is result in surface smoothing), as shown in the following figure.



Regression function: 0.670 * x + 25.848

Figure 33 : A scatter plot of predicted and sample values. The fitted line through the scatter of points is shown in blue with the equation given just below the plot.

If the sample data were not autocorrelated or independent in space, every prediction would be the same and equal to the mean of the measured data. In such a case, the blue line would be horizontal. With autocorrelation and a right kriging model, the blue line should be closer to the diagonal black dashed line. The tighter the scatter about the diagonal line, the better.

In addition, the *error* QQ-plot can be used to show the distribution of the prediction error against the corresponding standard normal distribution. The error plot is the same as the prediction plot, except the measured values are subtracted from the predicted values.

For the *standardized error* QQ-plot, the sample values are subtracted from the predicted values and divided by the estimated kriging errors.

If the standardized errors of the predictions from their measured values are normally distributed, the points should be close to the dashed line that represents the normal distribution (in the plot given below). If the errors are normally distributed, it confirms the appropriateness of using the methods that rely on normality (e.g., ordinary kriging).

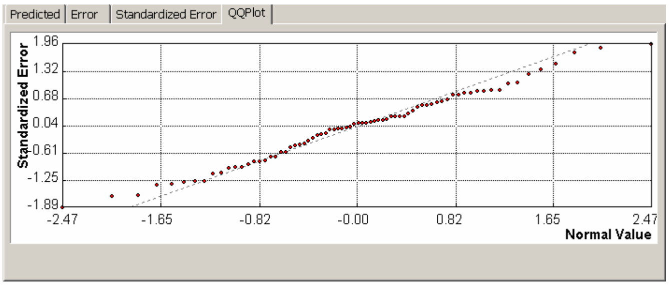


Figure 34 : QQ-plot of the prediction *standardized error*. The long-term average values of water levels are close to the normal distribution

4.5.3. Model Validation via Subsetting

The most rigorous way to assess the quality of an output surface is to compare the predicted values for specified locations with those that are independently measured and used as control points. Two independent datasets are used for the validation. The validation first creates a model for only a subset of the data, which is called the *training* dataset, and afterwards uses the *test* dataset of control points to check the model. For example, the *Root-mean-squared error* between predicted *training* values and *test* values in control locations can be used for the assessment.

In such a case, validation creates a model for only a subset of the data - the *training* dataset; therefore, it does not directly check a final model, which should include the entire available dataset. Rather, validation checks whether a *training* dataset model is valid, for example, choice of semivariogram model, lag size, and search neighborhood. The *training* dataset model is used for the whole dataset.

4.5.4. Model Cross-validation

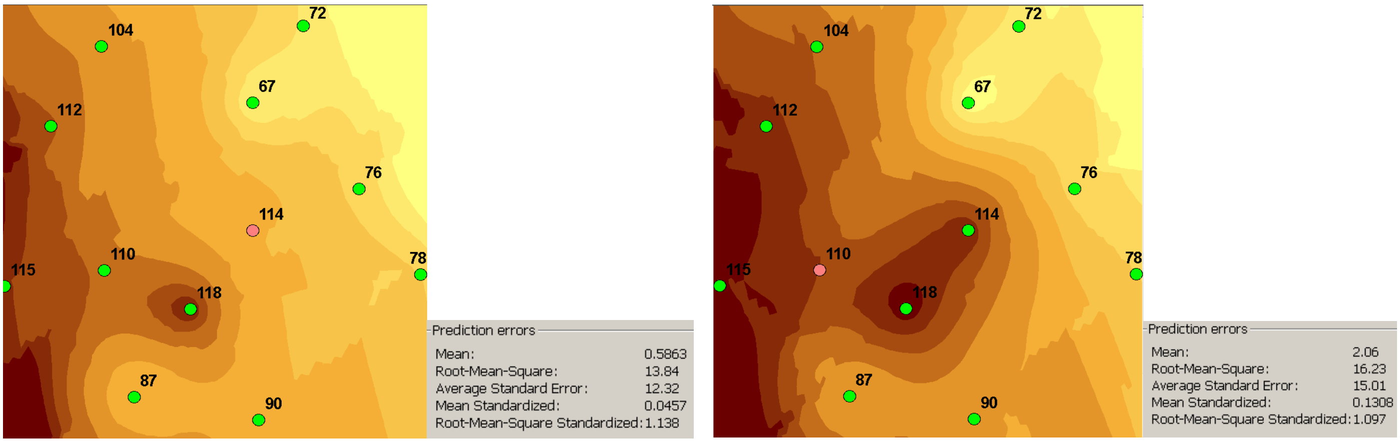
If an independent data set is not available to evaluate a model, the *same sample points,* which were used to estimate the model, are used to validate that same model. Thus, cross-validation uses all of the data to estimate the trend and autocorrelation models. It removes each sample point, one at a time, and predicts the associated data value and prediction errors. By judging errors, outlets could be found and after completing cross-validation, some data locations may be set aside as unusual, and refine the trend and autocorrelation models.

Cross- validation executes the following steps:

1. Compute experimental variogram with all sample points and its theoretical model
2. For each point:
    a. Remove the point from the sample set
    b. Predict *at that point* using the *other points* and the modeled variogram
3. Summarize the deviations of the model from the actual point. Models can be compared by their summary error statistics, by also looking at individual predictions of interest. The following error statistics for cross-validation can be used:
    a. *Root-mean-square error* that is lower is better; it is computed for independent validation

b. The *mean standardized prediction error* of residuals with kriging variance; computed for independent validation

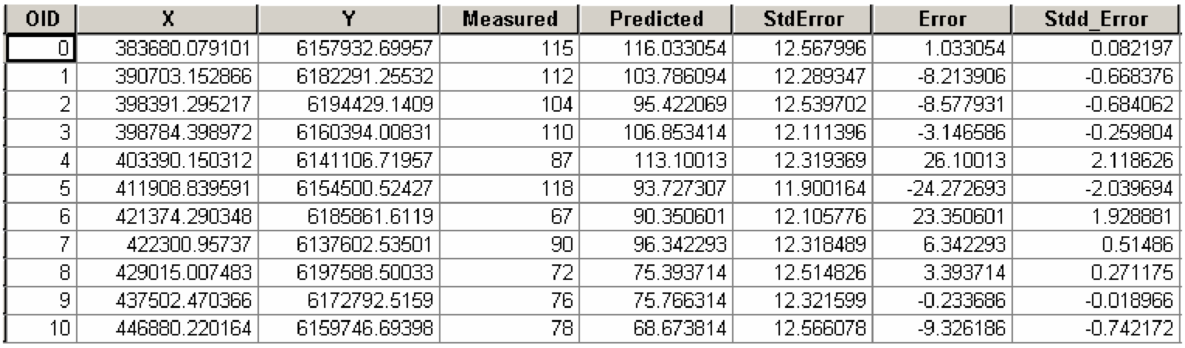
Other than that, the types of graphs and summary statistics used to compare predictions to true values are similar for both validation and cross-validation.



Prediction errors
Mean: 0.5863
Root-Mean-Square: 13.84
Average Standard Error: 12.32
Mean Standardized: 0.0457
Root-Mean-Square Standardized: 1.138

Prediction errors
Mean: 2.06
Root-Mean-Square: 16.23
Average Standard Error: 15.01
Mean Standardized: 0.1308
Root-Mean-Square Standardized: 1.097

Figure 35 : Cross-validation removes each sample data location (red dot) one at a time and produces a model



OID	X	Y	Measured	Predicted	StdError	Error	Stdd_Error
0	383680.079101	6157932.69957	115	116.033054	12.567996	1.033054	0.082197
1	390703.152866	6182291.25532	112	103.786094	12.289347	-8.213906	-0.668376
2	398391.295217	6194429.1409	104	95.422069	12.539702	-8.577931	-0.684062
3	398784.398972	6160394.00831	110	106.853414	12.111396	-3.146586	-0.259804
4	403390.150312	6141106.71957	87	113.10013	12.319369	26.10013	2.118626
5	411908.839591	6154500.52427	118	93.727307	11.900164	-24.272693	-2.039694
6	421374.290348	6185861.6119	67	90.350601	12.105776	23.350601	1.928881
7	422300.95737	6137602.53501	90	96.342293	12.318489	6.342293	0.51486
8	429015.007483	6197588.50033	72	75.393714	12.514826	3.393714	0.271175
9	437502.470366	6172792.5159	76	75.766314	12.321599	-0.233686	-0.018966
10	446880.220164	6159746.69398	78	68.673814	12.566078	-9.326186	-0.742172

Figure 36 : Cross-validation results: cross-validation compares the measured and predicted values for all points

With enough points, the effect of the removed point on the model, which was estimated using that point, is likely to be minor.

4.5.5. Model Comparisons

It is common practice to create many surfaces before one is selected as "best". Each surface systematically can be compared with another by using prediction errors. When comparing models, the “best” ones will be with the *mean standardized prediction error* nearest to zero, the smallest *root-mean-squared error*, the *average prediction error* nearest the *root-mean-squared error*, and the *standardized root-mean-squared prediction error nearest* to one.

There are two issues to consider when comparing the results from different methods and/or models: one is optimality and the other is validity (or the correct variability). It is important to get the correct variability. In kriging, the predictions depend on the kriging prediction errors $\hat{\sigma}_i$.

For example, the root-mean-squared error may be smaller for a particular model. Therefore, this model may be considered as the "optimal" model. However, when comparing to another model, the root-mean-squared error may be closer to the average estimated prediction error. If the *average kriging prediction error* is close to the root-mean-squared prediction error, this is a more valid model, because only the estimated kriging prediction error assesses uncertainty of the prediction independently of the actual data values (only variogram model is required). If the *average kriging error* is greater than the root-mean-squared prediction error, the variability of predictions is over-estimated; if the *average kriging error* is less than the root-mean-squared prediction error, the variability in predictions is under-estimated.

Therefore, the variability in prediction can be assessed correctly by the *root-mean-squared standardized prediction error*. Thus, the root-mean-squared standardized error should be close to 1 if the *prediction kriging errors* are valid.

4.5.6. Data Sample Considerations

Sample size and distribution can influence the accuracy of prediction. In general, more sample points are better - an increased sampling rate (samples taken closer together) and the local variation will be more accurately captured and appropriate for large-scale (small-area) studies. However, it will introduce a higher data gathering cost. With low density of sample points, the sensitivity of local variation will be lost and only the regional variation will be captured; this would be more appropriate for small-scale (large-area) studies.

The purpose of sampling design is to establish the structure of spatial dependence (e.g. semivariogram) with a minimal number of samples that will produce optimal sample spacing on a map. Moreover, kriging methods include solutions to accomplish the task of sampling cost minimization.

As mentioned above, in kriging the estimation error is based *only* on the sample *configuration* and the chosen *model* of spatial dependence, not the actual data values. Therefore, *if* the spatial structure (variogram model) is known, maximum or average prediction errors can be determined *before* sampling is computed. Based on the prediction errors, sampling decisions can be made based on a *cost-benefit* framework before fieldwork is undertaken.

Optimal point configuration can be established based on the following *optimization criteria* of known kriging systems as some numerical measure of the quality of the sampling design:

1. Minimize the *maximum* kriging *prediction error* in the study area
2. Minimize the *average* kriging *prediction error* over the entire area
3. Maximize the *information in a sample variogram* in order to allow reliable variogram estimation.

4.6. Conclusion and Comparison of Geostatistical Methods

The key concepts that were introduced in this module relate to *spatial dependence* or *spatial correlation* that is described in general as “the value of a variable at a point in space or time is related to its value at nearby points”. Knowing the values of sample points allows one to predict (with some degree of certainty) the value at any given chosen point.

Here the concept of correlation between variables is applied to correlation *within* or *to one* variable, using *distance* or *time* to model the spatial relation. The *auto-correlation* term is used to describe such correlation. Main question was “How to describe the auto-correlation?” Spatial *structure* is the nature of the spatial relation: how far, and in what directions, is there spatial dependence? How does the dependence vary with distance and direction between points?

A function of the distance and direction separating two locations is used to quantify autocorrelation. Spatial structure can be described by range, direction, and strength. The type of interpolation method that can be used will depend on many factors. A common approach is to try different interpolation methods and compare the results to determine the best interpolation method for a given situation. Still, real-world knowledge of the subject matter can affect what interpolation method to use.

The quality of a sample point set can affect the choice of an interpolation method as well. Support of a sample or the physical dimensions it represents will define coarser or finer resolutions of the prediction result.

Some features of module interpolation methods are summarized in the following table:

Method	Type of Interpolator	Output Map Type	Advantages	Disadvantages	Assumptions
Inverse Distance Weighted	Deterministic interpolator	Prediction	Few parameter decisions	No assessment of precision errors; produces "bulls eyes" a round data locations	Not required
Global polynomial	Deterministic estimator	Prediction	Few parameter decisions	No assessment of precision errors; may be too smooth: edge points have large influence	Not required
Local polynomial	Deterministic estimator	Prediction	More parameter decisions	No assessment of precision errors; may be too automatic	Not required
Splines	Deterministic smoother	Prediction	Flexible and automatic with some parameter decisions	No assessment of precision errors; may be too automatic	Not required
Kriging	Stochastic predictor	Prediction, Errors, Probability, Quantile	Very flexible: allows assessment of spatial autocorrelation: can obtain prediction errors: many parameter decisions	Need to make many decisions on transformations trends, models, parameters, and neighborhoods	Some methods require that the data comes from a normal distribution
Co-kriging	Stochastic predictor	Prediction, Errors, Probability, Quantile	Very flexible: can use information in multiple datasets; allows assessment of spatial cross-correlation; many parameter decisions	Need to make many decisions on transformations trends, models, parameters, and neighborhoods	Some methods require that the data comes from normal distribution

Module self-study questions:

1. Define autocorrelation.
2. Explain the difference between global and local interpolation methods.
3. What is an exact interpolation method?
4. Explain semivariance as a measure of spatial dependency.
5. Define the elements of nugget, range, and sill in a semivariogram.
6. What is the purpose of fitting a semivariogram with a mathematical model?
7. How does universal kriging differ from ordinary kriging?
8. Describe how a cross-validation analysis is performed.

Recommended Readings:

[1] Geostatistical Analyst, ESRI ArcGIS 9.2 Desktop Help, http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=An_overview_of_Geostatistical_Analyst.

[2] Deterministic and Geostatistical Interpolation methods sections, Geospatial Analysis: Web site, M. J. de Smith, M. F. Goodchild, P. A. Longley, http://www.spatialanalysisonline.com/output/ .

References

1. Johnston K., Ver Hoef J. M., Krivoruchko K., and Lucas N. (2003) *ArcGIS® 9 Using ArcGIS® Geostatistical Analyst,* ESRI.
2. Berke, O. (1999) Estimation and Prediction in the Spatial Linear Model. *Water, Air, and Soil Pollution* 110, 215-237.
3. Kang-Tsung Chang (2006) Introduction to Geographic Information Systems, Third Edition, The McGraw Hill.
4. Bailey, T. & Gatrell, A. (1995) *Interactive Spatial Data Analysis.* Addison Wesley Longman, Harlow.
5. Cressie, N. (1993) *Statistics for Spatial Data* (Revised Edition). Wiley, New York.

Terms used

- Geostatistics
- Regionalized variables
- Auto-correlation
- Spatial dependence
- Inverse distance weighting
- Trend
- Global and local polynomial
- Splines
- Kriging
- Regression
- Power function
- Neighborhood search strategy
- Omnidirectional and isotropy
- Anisotropy
- Semivariance
- Variogram
- Empirical semivariogram
- Lag bins
- Auto-covariance
- Theoretical, authorized or fitted variogram
- Ordinary, simple, universal, regression kriging and co-kriging
- Unbiased
- Validation
- Cross-validation
- RMSE
- ME
- Kriging prediction error
- Average kriging prediction error
- Root-mean-square standardized prediction error
- Error, quantile and probability maps

5 Spatial Modelling

In its simplest application, a Geographic Information System (GIS) can be used to organize and edit datasets and display the results in the form of a map which can be viewed in digital format or output to hardcopy. However, the ability of a GIS to generate models simulating real-world events is perhaps its most significant feature. In reality any map is a spatial model as it provides a representation of the real world, however, GIS can be used to analyse multiple variables and then interpret these variables to simulate or predict potential events. This capability allows us to interpolate information for large areas of the landscape. For example, it would be impractical (i.e., due to budgetary constraints) to conduct a wildlife habitat field survey for an entire country, however, a GIS can be used to model wildlife habitat suitability for large areas.

The following module details some of the basic spatial modelling functions available:

Topic 1: Modelling spatial problems
Topic 2: Classification of spatial models
Topic 3: Model types
Topic 4: Model builder
Topic 5: Model examples

5.1 Modelling Spatial Problems

5.1.1 Overview

Spatial modeling is the most sophisticated level of spatial analysis. Models enable us to represent, in a simplified form, some aspect of reality. As an abstract or partial representation, models help explain or predict how a certain human or natural phenomenon or system works (or could work) in the real world. They can be used to represent past, present, or future conditions. Models generally consist of a set of spatial variables tied together by an arithmetic operation or set of commands.

5.1.2 Developing a Model

Developing a model involves working through a series of conceptual steps, from identifying the problem through to implementing the results. It is critical to the development of a model to understand the problem at a conceptual level prior to developing the model and therefore following the steps detailed below prior to beginning work on the computer is highly recommended.

Step 1: Identify the Problem – In order to establish a model that addresses a specific problem, it is necessary to clearly define the problem and goal(s) of the model. Inherent in this step is establishing parameters by asking the following types of questions:

- What phenomena are being modelled?
- Why is the model necessary?
- What is the spatial scale and extent of the model (i.e., has the study area been defined)?
- What time period is pertinent – are we assessing single or multiple time periods?

For example, as part of an environmental impact assessment, a wildlife biologist might wish to quantify the amount of wildlife habitat potentially impacted within the zone of influence of a proposed development site. A model, or potentially a series of models, integrating variables related to wildlife habitat could be used to solve the problem. When identifying the problem the first step in this example would be to identify the species that are potentially present in the project area and then select the species that habitat models will be generated for. For example, it may be determined that half-a-dozen indicator species can be used to represent the various types of wildlife present (e.g., a predator species, a prey species, a bird, a raptor, a furbearer and an ungulate). These species then become the phenomena being modelled. In wildlife habitat models the spatial extent of the study area is typically represented by the maximum home range extent of the species or by the watersheds or ecosystems surrounding the proposed project area. These types of boundaries can be used to help define the spatial extents of the model. For the purpose of quantifying the amount of habitat that might be lost it would be effective to assess the wildlife habitat available in multiple time periods. This would be done by modelling the present conditions (baseline) and then comparing those results against a second set of model results depicting the habitat available after the project was developed.

Step 2: Break the Problem Down – After you have established the problem and the goals, breaking the problem down into its constituent parts helps create more manageable steps. This involves identifying the objectives required to reach your goal, the phenomena involved, and the interactions between these phenomena. Establishing the phenomena allows the modeller to identify and assemble the datasets required for processing. Often, a flowchart can be useful in visualizing and understanding the spatial and attribute relationships between the constituents.

In our wildlife habitat assessment example, the modeller would assess the habitat requirements associated with each species being modelled (e.g., elevation, slope, proximity to water, vegetation) and then compile a catalogue of relevant datasets and define how their associated attribute and spatial relationships contribute to creating potential habitat for each species. For example, if we know a species exists below 500 metres elevation, on slopes less than 10%, within 250 metres to a source of freshwater and prefers coniferous forest habitat, we can use the data from a digital elevation model (DEM) to identify areas meeting the elevation and slope requirements. A hydrology layer would be required to model proximity to water and a vegetation or land cover dataset could be used to identify coniferous forest habitats. The intersection of these four data layers allows us to identify potential habitats for our species of interest.

Step 3: Develop and Calibrate the Model – Identifying the tools and mathematical operations required for analysis is addressed at this stage. Using the data examined in Step 2, the modeller builds the model from these tools and operations. Subsequent repeated running of the model allows the modeller to calibrate the model. Calibration involves comparing the results of the model with its input data and subsequently adjusting the parameters or mathematical operations to arrive at more accurate results.

Using a GIS, the modeller would calibrate the model by repeatedly running the model and comparing the results to the data. Adjustments to parameters and operations would be used to refine the model results.

In our wildlife habitat map example, the initial results would illustrate potential habitat based on the variables considered; we may then want to refine the results based on other criteria. For example, perhaps ideal habitat (with a high rating) is within 100 metres of a source of freshwater and habitat falling between 100 and 250 metres of freshwater would be assigned a moderate habitat rating.

Step 4: Validate the Model Results – Validation is an evaluation of the model's capacity for accurately predicting the real world phenomena. This involves comparing the results to field data and/or running the model using a different set of data representing conditions that are unlike those used in the calibration phase. If another set of data are not available, a suitable alternative consists of splitting the one dataset into two subsets: one for developing and calibrating the model, the other for the validation process.

Error Detection

Visual Inspection – one method for identifying errors is simply by looking at the output to see if the results seem logical and consistent and that the output data reflects the input.

Documentation – metadata for input datasets can be used to ensure you are using the data appropriately, based on scale, accuracy, attribute values.

Validation Rules – topological and attribute domain ranges are two useful validation rules that allow you to check the data against the design.

Consistency of Results – collecting data again and repeating the conversion and processing steps can be used to compare the results of a model.

Ground Truthing – verifying the results of a model in the field is a dependable method for validating data.

Statistics – correlating the model results with a closely associated variable is a statistical approach to data verification.

In our example, the wildlife biologist could validate the habitat suitability model results by comparing areas having high predicted wildlife habitat ratings with known wildlife habitat locations (e.g., denning sites).

Step 5: Implement the Model Results – Once the model has been validated, the modeller can then implement the model results.

By using the validated results of the model, our wildlife biologist can determine how much habitat is currently present and then overlay the proposed development footprint to determine the amount of habitat lost after the proposed development is operation. In addition, the model results could be used to identify high priority areas for subsequent field surveying efforts.

Limitations of Modelling

It is critical to understand that all models have limitations. Factors that may limit the use or implementation of a model include:

- insufficient data – for example data at the required level of detail may not be available for all, or part of, the study area
- lack of user understanding of the data – for example, many models represent a probability analysis rather than an absolute interpretation of the data
- inappropriate modelling – in this case the model would be generating incorrect results either based on a faulty assumption in the model design or as a result of an error in executing the model

5.2 Classification of Spatial Models

Models may be classified by purpose, methodology, or logic, with some classifications falling within more than one category.

5.2.1 Purpose

Descriptive Model

A descriptive model describes the current conditions of a real world environment (natural or anthropogenic). A simple thematic map showing land use can be considered a descriptive model in that it represents an environmental characteristic that actually occurs on the ground. Other examples include digital elevation models, land use coverage, meteorological maps, and vegetation cover maps.

Explanatory Model

As indicated by their name, explanatory models attempt to explain or account for the occurrence of existing phenomena by identifying the factors involved and assessing their relative influence. Examples of explanatory models include soil erosion models, ozone depletion, and algae blooms.

Predictive Model

A predictive (or prescriptive) model predicts (by using the factors identified in an explanatory model) where you might find occurrences of a particular environmental phenomena. For example, a habitat capability map, derived from vegetation and slope raster datasets, might be used to predict the capacity of an area to support the habitation by a certain animal species. Forecasting sea level rise based on changes in climate (e.g., mean temperature values) would be another example of a predictive model.

Normative Model

Normative models attempt to affirm how phenomena (especially network) ought to operate in the real world; they recommend optimal solutions for given situations. Examples include food aid distribution, traffic volume levels, and route planning (e.g., travelling salesperson problems).

5.2.2 Methodology

Stochastic

A stochastic (or probabilistic) model is represented by a mathematical equation where at least one of its variables or parameters is assumed to contain some level of randomness. Because of this randomness, some degree of error or uncertainty is accepted and expressed as a measure of probability along with the predictions of the model. An example might be an evaluation of the probability of the occurrence of landslides due to forest harvest areas. Another example of a stochastic model would be the application of a kriging analysis to a series of water quality sampling measurements. This would result in an interpolated surface depicting the potential concentration of a given contaminant (a predictive map) and the generation of a standard level of error for each predicted value.

Deterministic

Contrary to the stochastic method, a deterministic model does not consider randomness to be a part of any of the variables or parameters present in the mathematical equation.

5.2.3 Static or Dynamic

A static model deals with a phenomenon occurring at a specific point in time, whereas a dynamic model is used to consider and highlight the changes of a phenomenon over time and the interactions between multiple variables. For example, a static model might summarize the density of road features in analysis units for a particular year. A dynamic model would build upon that model by summarizing the change in road density between successive years.

5.2.4 Based on Logic

Inductive
An inductive model moves from the specific to the general, basing its conclusions on evidence observed in previous studies. Typically, inductive models attempt to identify general conditions or rules when important themes and relationships are not well understood. An archaeological potential model that relies strictly on the distribution of existing sites is an example of inductive reasoning.

Deductive
Deductive models derive specific conclusions by using general premises established in scientific theory or physical laws - where the variables and their interactions are well understood. An archaeological potential model that uses physical constraints (e.g., slope, aspect, proximity to water) to predict the location of sites is an example of deductive logic.

5.3 Model Types

5.3.1 Representation Models

Binary

A binary model selects features from layers or multiple raster datasets using logical expressions, much like a data query yields a selection set. The output is a raster or vector layer with binary values: spatial features or raster cells that meet the criteria are coded 1 (true) while those features that do not agree with the criteria are coded 0 (false). Binary model analysis can be either vector- or raster-based:

> Vector-Based Binary Model – a vector-based binary model uses overlay operations (e.g., intersect, union) to combine the shapes and attributes of the contributing data.
>
> Raster-Based Binary Model – a raster-based binary model extracts desired criteria by directly querying multiple raster layers.

The most common application of binary raster data is site selection analysis, where multiple criteria are evaluated to determine the most appropriate location for a certain facility or specific land use. Two major methods are used to model siting analysis: select one site based on the evaluation of a set of pre-selected sites using stringent selection criteria; or the evaluation of all potential sites.

Supposed a mining company wanted to select potential gold mining sites in a valley using the following criteria:

- a minimum ore concentration per cubic metre
- minimum open pit size of 10 hectares
- minimum flood potential
- sites must be less than 30 kilometres from an existing haul road
- sites must be on land having less than 10 percent slope

Steps for building the model might include:

- gather all contributing datasets required and pre-process in preparation for analysis (e.g., from a DEM, create a binary grid depicting areas having a slope value less than 10 percent and then the conversion of this grid to a vector polygon layer)
- create a 30 kilometre buffer zone of all roads
- use overlay function (e.g., union) to combine the other layers with the road buffer
- query the derivative merged layer to locate which areas satisfy the criteria listed above

Ranking

Also referred to as an index model, a ranking model calculates an index value specific to each unit area, with the end result being a ranked map. Similar to binary models, ranking models evaluate multiple criteria through the use of overlay operations and raster arithmetic. The difference lies in the fact that while binary models are given a yes/no value (1 or 0), ranking models apply an index value to each unit area.

Commonly, a weighted linear combination method is used to determine the index value for each unit area. This method is completed in three steps:

1. Relative Importance – all criteria are evaluated against each other to determine the relative importance of each criterion weighted on a scale of 0.0 to 1.0 (0 – 100%).
2. Standardize Criteria – the relative importance values are standardized between criteria. This step facilitates the comparison of multiple variables.
3. Calculate Index Value – an index value is calculated for each unit area by adding the weighted criterion values and then dividing by the total of the weights.

Ranking models are used predominantly in creating suitability or vulnerability maps which are discussed further in the section on model examples (Section 1.5).

5.3.2 Process Models

Process models integrate existing environmental information to simulate a real-world process. Because they are used to interpret the interaction of multiple variables and predict what may occur in the real world, process models are classified as both dynamic and predictive. Typically, process model analyses are examined in a raster-based GIS environment. The raster world allows for more complex arithmetic computations and standardization between many disparate data sources. Soil erosion models, discussed in detail in Section 1.5.4, are good examples of process models.

5.4 Model Builder

5.4.1 Overview

Model Builder is an interface within ESRI's ArcGIS product that allows multiple processes to be combined facilitating the development of models. It enables you to visualize work flow (in the form of flow chart diagrams) and author and automate geoprocessing tasks that would normally be executed in single steps in ArcMap. It also has the resultant advantage of allowing you to document the steps involved in the development of a model. While the development of the initial version of a model might take a little more time than conducting the steps manually it is extremely useful when conducting multiple runs of a model – the model can be run on different data or small changes in the model can be made and the model rerun to examine model alternatives and assumptions. Processes in Model Builder can be either ArcGIS system tools (ArcTools tool) or custom tools created by the modeller.

5.4.2 How to Build a Model

The Model Builder interface consists of a window displaying the diagram of the model (the process and model elements), a main menu bar containing the functions, and a toolbar that holds the tools and functions used to interact with the model elements in the display window (Figure 1).

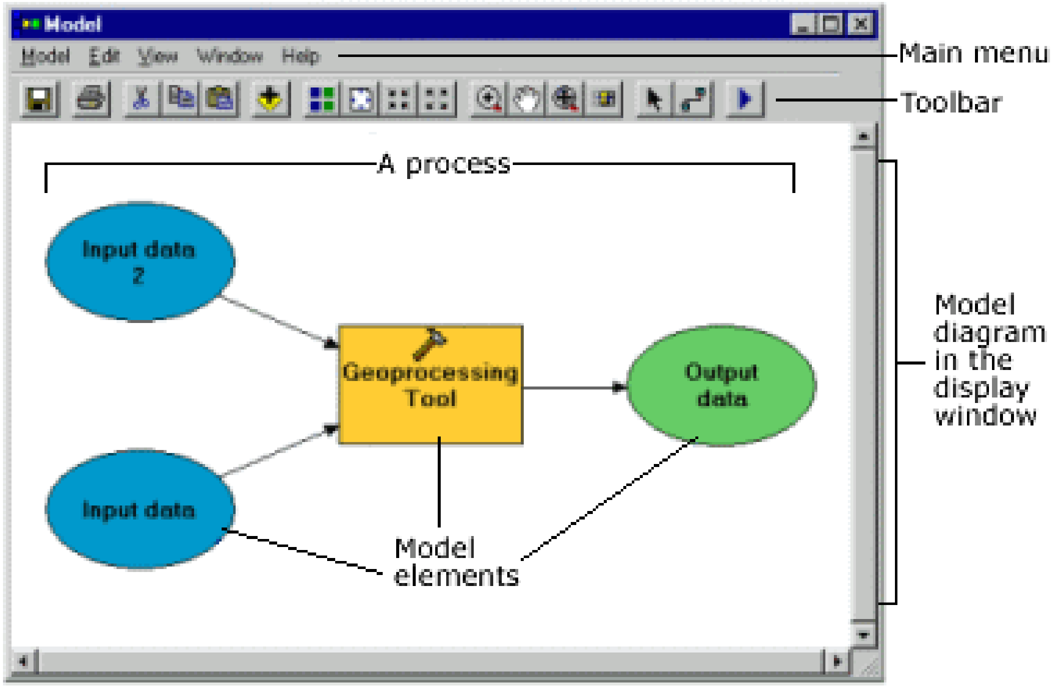


Figure 1. Model Builder Interface.

Model elements may take the form of input data, geoprocessing tools, or output data layers. Processes are constructed by dragging and dropping tools and data into the display window, then entering parameters for each tool. When connections are made between an input data variable and a geoprocessing tool, the input data value (i.e. the name of the data layer) is automatically entered as the input parameter in the tool's dialog box. Running the model executes the process(es), based on the layer and tool layout and order of steps as diagrammed by the modeller in the display window.

After the desired data and tools have been added and suitably connected in the display window, the model can be saved. The saved model configuration can serve as a template and used with different model parameters/data inputs.

The steps used in the gold mine site selection example above could be inserted into Model Builder and run as an automated and comprehensive process. The slope layer and road buffer processes could be designed in Model Builder and then connected to a union tool, ultimately yielding a layer with appropriate locations for the mine.

It is extremely important when implementing a model using the Model Builder interface to double check the processing steps associated with each model element and to review the output data at each step to ensure the model is creating the expected results. As with any automated process it can often be difficult to determine if a mistake exists within the model and therefore the quality assurance and quality control procedures are of great importance.

5.5 Model Examples

5.5.1 Wildlife Habitat

Habitat Suitability

The ability to evaluate the past, present and future condition of landscape-scale variables, relationships and dependencies help provide an insight into ecosystem health and function. Wildlife habitat maps provide a useful planning and management tool as they can serve as an indicator of the overall condition of an area while also providing information specific to an individual species. For example, they can be used to provide a regional perspective (e.g., as a variable in a constraints map) or for a specific purpose (e.g., to help select sampling locations for field survey). The maps show the location and extent (availability) of habitat and indicate the relative suitability of the habitat through the assignment of a habitat suitability index or class (Figure 2). Terrain, elevation, vegetation, hydrological features and human disturbance layers can be integrated and, based on habitat requirements, used to map wildlife habitat suitability. For example, a species might prefer a specific elevation range, vegetation type and require a 200 metre proximity to a freshwater source. All of these variables can be quantified within a GIS and a model developed and applied to available data.

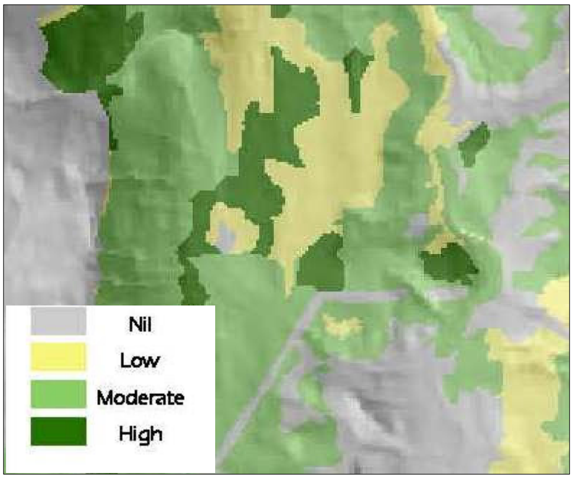


Figure 2. Example of a Habitat Map.

The data needed to perform this type of analysis is as follows:

- Ecological Land Classification (ELC) mapping and/or vegetation data
- Terrain
    - Slope
    - Aspect
    - Elevation
- Base map information
    - Hydrology
    - Human disturbance

- Discipline knowledge
  - Wildlife experts
  - GIS and data integration

Habitat Fragmentation

The mapping of habitat fragmentation is another tool to help manage environmental change. Habitat fragmentation maps integrate wildlife habitat map layers with maps depicting the extent of human disturbance to illustrate the size of available habitat patches and areas of continuous habitat. The aggregate maps help us quantify the amount of viable habitat and the potential loss resulting from existing and/or proposed development activities. They help us identify: the total area and average patch size of habitat; potential increases in the amount of edge habitat (this can be a positive or negative factor depending on the species of interest); the potential decrease in the amount of interior habitat; patches of habitat that have the potential to become isolated; and the potential increase in smaller habitat patches.

The results can be used to help identify the locations of major habitat reservoirs and habitat refuges that are essential to the continued success of the species. A habitat reservoir is a large area (the size of which is dependent on the species of interest) of habitat that has sufficient size and ecological integrity to support a range of native species including species that need interior habitats. A habitat refuge is a small patch of habitat that provides food, shelter and other needs for wildlife. It may include human-modified ecosystems. Refuges generally are not large enough to maintain the genetic diversity of a population but may act as important ‘stepping stones’ to habitat reservoirs for species and for maintaining ecological functions.

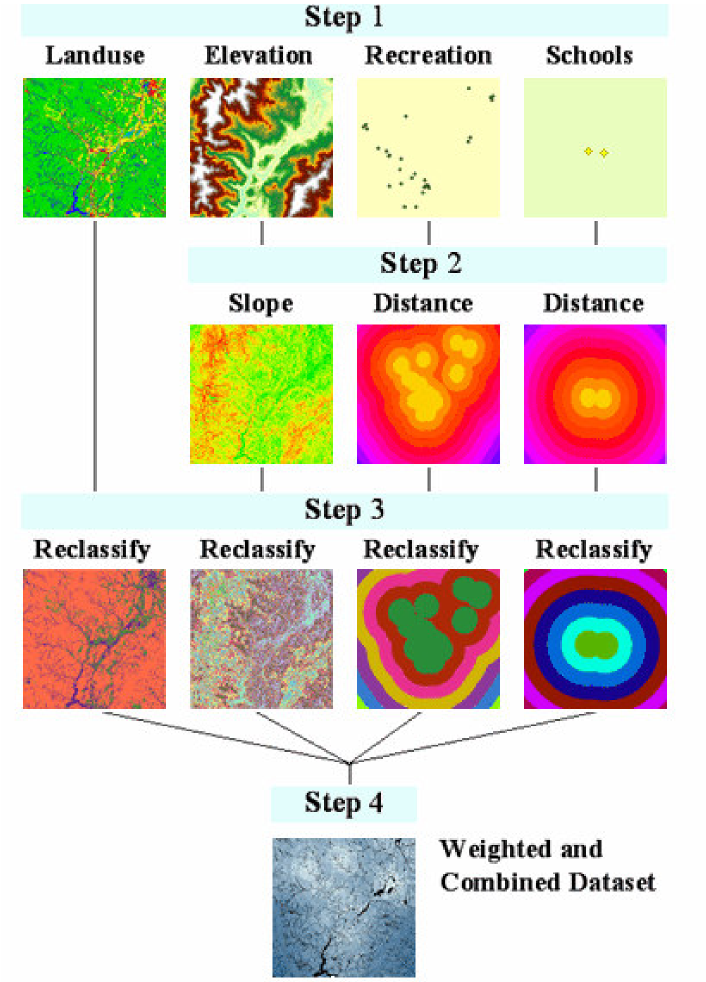
The data needed to perform this type of analysis is as follows:

- Ecological Land Classification mapping and/or vegetation data
- Wildlife habitat maps
- Terrain (optional)
  - Slope
  - Aspect
  - Elevation
- Base map information
  - Human disturbance
- Discipline knowledge
  - Wildlife experts
  - GIS and data integration

5.5.2 Locational Suitability Models

Suitability Maps

Raster reclassification (changing cell values) is sometimes done with the aim of assigning rank or weighting to cells to convey a sense of importance, sensitivity, or hierarchy. This process is often used in the creation of a suitability map (e.g., a habitat suitability model based on multiple input raster layers with values ranging from 1-6 [low to high]) (Figure 3).



Step 1 - Input Datasets: determine what datasets are needed as input layers to the model

Step 2 - Create Derivative Datasets: generate offshoot rasters from original data (e.g., slope and aspect datasets can be derived from an elevation raster)

Step 3 - Reclassify: reclassify input datasets to a common value range so that data can be compared and combined on an even scale (playing field)

Step 4 - Assign Weights and Combine Rasters: give added weight to more influential datasets and combine to create a suitability raster

Figure 3. Reclassification - Creating a Suitability Map.

Inherent in the reclassification process is the output of a new raster as illustrated in the example above. Reclassification is predominantly used to reduce the number of output categories in preparation for combining data (overlay analysis).

Constraints Mapping

Constraints mapping is a type of suitability map. The analysis identifies opportunities and restrictions (constraints) to project construction. The process involves assembling a variety of different spatial data layers (e.g., terrain and topography, wildlife habitat, slope, protected areas, soils, heritage resource information), assigning the data in each layer a sensitivity rating (based on scientific and local knowledge), and combining the layers to develop a single map that integrates all of the source data layers. The resultant derivative map product identifies and synthesizes the complex relationships between different environmental datasets, while also considering location and operational limitations posed by project design. The constraints map serves as a visual decision-support tool that delineates areas determined to be of environmental and cultural importance based on the occurrence of sensitive landscape features. The constraints designation can be either avoidance of the site (e.g., a 'no-go' area) or a graduated level of concern. Ideally the constraints map helps guide the placement of surface facilities into the least constraining locations, thereby lowering the environmental and cultural impact of industrial activities.

The type of derivative surface that is created depends on the values and weights applied to the input data layers. For example, corridor delineations depend on slope and ecological land classification (ELC) type for defining where ideal animal corridors are present within the landscape, while accessibility mapping will depend on slope and distance from communities.

Overview

Constraints mapping identifies opportunities and restrictions to project siting by integrating a series of spatial layers into a single map. Figure 4 below provides an overview of the constraints mapping process.

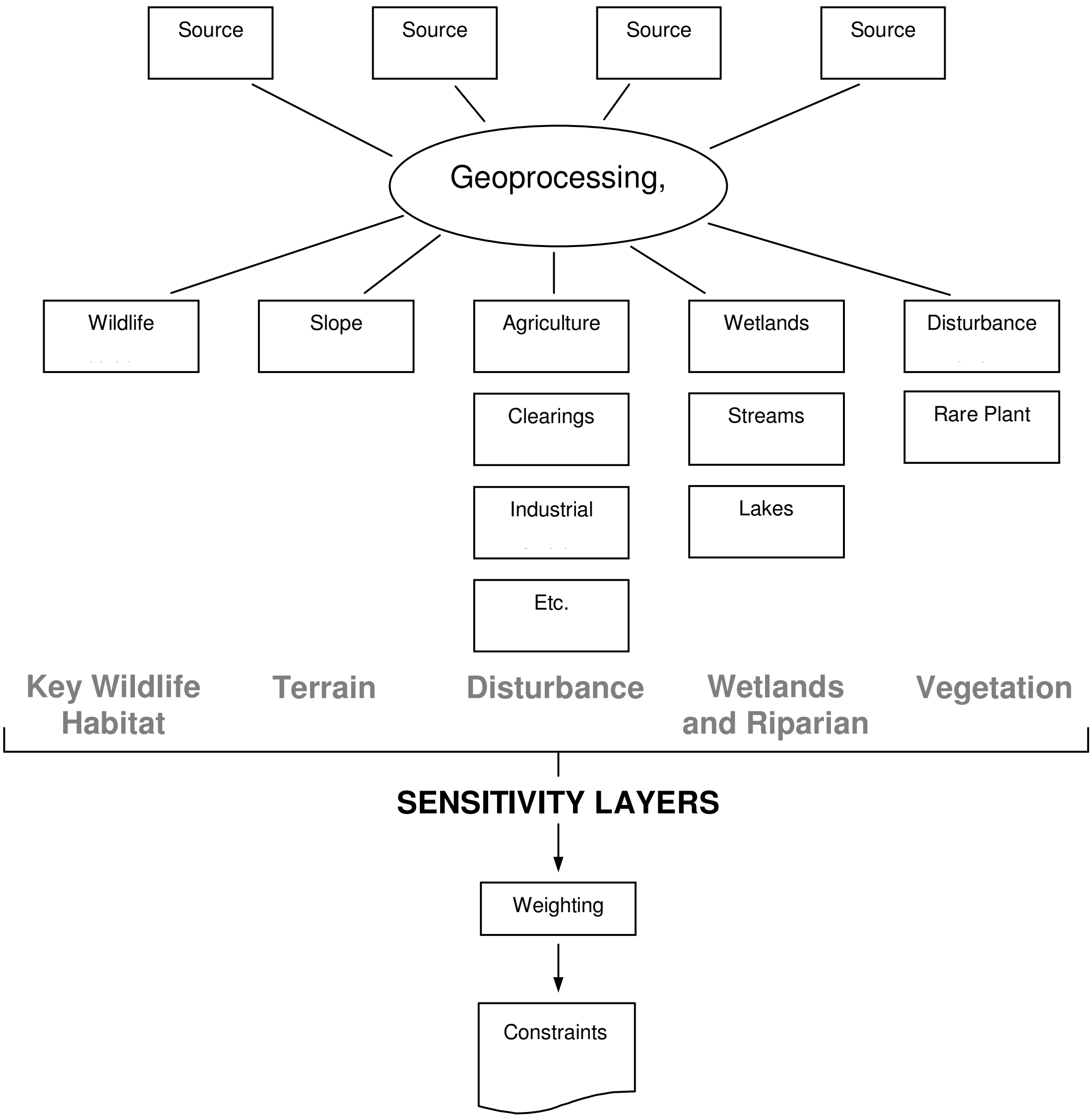


Figure 4. Example Sensitivity Layer Structure.

The implementation of constraints mapping is tiered and the definitions discussed here are directly related to the modeling methodology used and described below.

- The constraints map represents the visual result or output of the constraint model. The constraints map is created from a mathematical combination of any number of sensitivity layers.
- A sensitivity layer represents the sensitivity of a given discipline or subject of interest (e.g., wildlife, traditional knowledge or vegetation) relative to the objective of minimizing environmental and cultural impacts while project siting. Each sensitivity layer is the result of combining input data from potentially many GIS datasets that depict features on the landscape such as roads, forest cut blocks or wildlife habitat.

The constraints mapping process encompasses the following general steps:

1. Data preprocessing or modeling (e.g., creation of a slope surface or wildlife models);
2. Assigning constraint characteristics (zone of influence buffers, sensitivity values and combination rules);
3. Aggregating input data to create sensitivity layers;
4. Assigning weighting factors to sensitivity layers; and
5. Combine sensitivity layers to derive the constraints map.

GIS Data Preparation

GIS datasets are grouped according to their constraints characteristics. For instance, a polygonal wetland dataset would likely contain polygons identifying the spatial location of different types of areal wetland features. These may include unique types of wetlands such as; marshes, bogs, fens and swamps if these features are to be rated differently. Alternatively, the unique features could be grouped in a single class (e.g., all wetlands) if the rating values are consistent between the wetland types.

To be incorporated into the model, each feature in a GIS dataset must be assigned a sensitivity value - a constraint coefficient. The value assigned represents the sensitivity of a particular feature to the constraints objective. For instance, being interested in minimizing the social and cultural impacts of siting a project, we are aware of the importance of wetlands to wildlife and ecosystem health – particularly marshes and swamps. These wetlands are therefore assigned a relatively high sensitivity value (which recognizes that development may be more constrained where these features exist). In the constraints model, the possible sensitivity values ranged from zero (0.0) to one (1.0). A value of zero represents no sensitivity and a value of one represents the highest sensitivity.

Building Sensitivity Layers

Each sensitivity layer is produced by combining one or more GIS datasets. In some instances it is found that features from different GIS datasets overlap each other and have different sensitivity values (Figure 5). In the creation of a ‘Wetlands and Riparian’ sensitivity layer for example, it is common for a wetland buffer with a sensitivity value of 1.0 to overlap an intermittent stream buffer with a sensitivity value of 0.5. Where GIS features contributing to the same sensitivity layer overlap, the maximum sensitivity is assigned to the calculated sensitivity layer. In this case, the value of 1.0 was assigned.

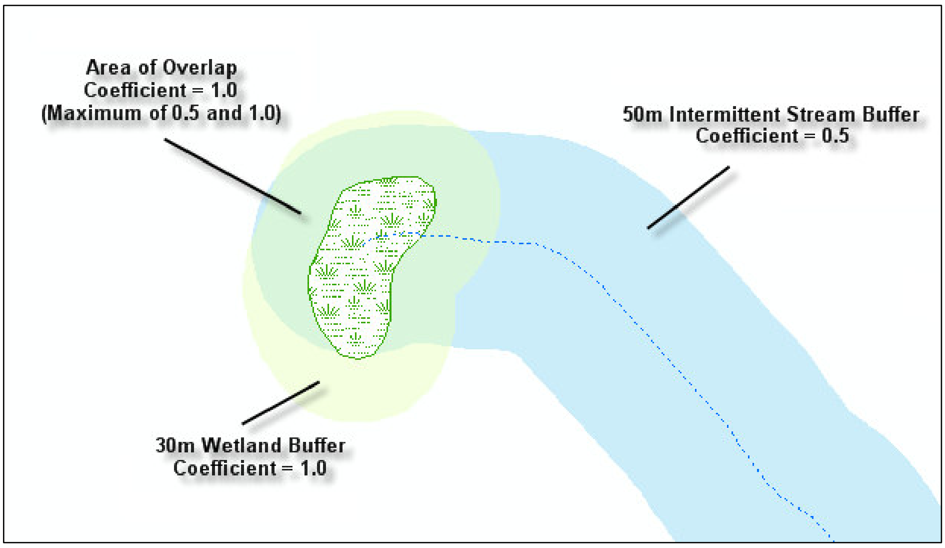


Figure 5 Example of Overlapping Coefficients.

All GIS datasets are combined using this 'conservative' rule where the maximum sensitivity of all inputs to a sensitivity layer value was used.

Building the Constraints Map

The final constraints map is produced by combining the sensitivity layers. The layers are combined using linear weighted summation. Each sensitivity layer is assigned a weight that specifies its relative importance with respect to the other sensitivity layers. The weights are used to specify how much each sensitivity layer should contribute to the model. The total of all weights must sum to 1.0 (or 100%). Table 1 presents an example of how a number of sensitivity layers might be weighted in the constraints mapping process.

Table 1. Sensitivity Layer Weighting

Sensitivity Layer	Weight
Wetlands and Riparian	0.492
Key Wildlife Habitat	0.268
Disturbance	0.140
Vegetation	0.053
Slope	0.047
Total	1.000

For a given point on the constraints map, the resultant constraint value is calculated as follows:

> Constraint Map Unit Value = (wetland and riparian sensitivity value x 0.492) + (key wildlife habitat sensitivity value x 0.268) + (disturbance sensitivity value x 0.140) + (vegetation sensitivity value x 0.053) + (slope x 0.047).

The following is an example calculation for a map unit (e.g., a grid cell) where the final constraint value was determined to be 0.707. The example Sensitivity Value is the value for that sensitivity layer at a given pixel location. A Weighted Sensitivity is calculated for each map unit in the study area. Weighted Sensitivity values range from 0.0 (no constraints) to 1.0 (fully constrained). The constraint value represents the sum of the weighted sensitivity values for a given map unit.

Table 2. Constraints Calculation Example

Sensitivity Layer	Sensitivity Value	Weighting Value	Weighted Sensitivity
Wetlands and Riparian	1	0.492	0.492
Key Wildlife Habitat	0	0.268	0.000
Disturbance	1	0.140	0.140
Vegetation	0.8	0.053	0.042
Terrain	0.7	0.047	0.033
		TOTAL (Constraint Value)	0.707

Figure 6 depicts an example of a constraints map. Brown areas are the most constrained while dark green areas are the least constrained. In this example, it is apparent that the wetland and riparian features contributed heavily to the final constraints map, since the buffered hydrological features are clearly visible in the resultant surface.

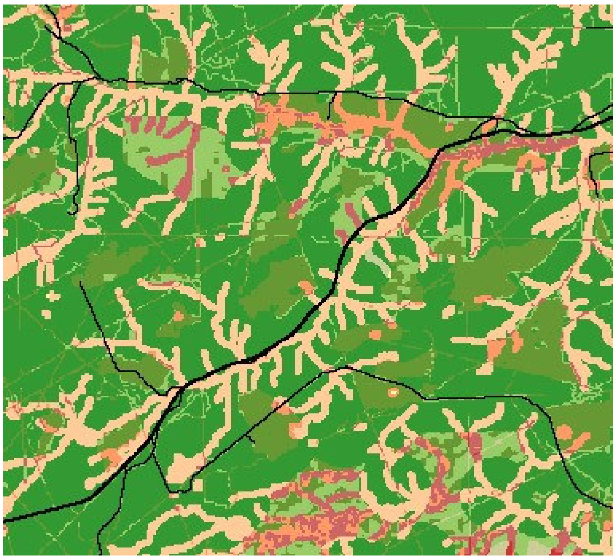


Figure 6 Example Constraints Map.

Assumptions and Limitations

Constraints maps should be interpreted with care due to limitations of the model. The most obvious limitation is that individual sensitivity values (for individual features on the landscape) are lost in the final output. Because inputs from many GIS data sources are included in the output, there is a dilution of sensitivity of individual features. This happens in two ways:

1. For a sensitivity layer, the sensitivity at a given location on the landscape is determined by choosing the highest of all sensitivities from the input GIS data. That means that if there were three features in the same area all with a sensitivity of 1.0, they would be considered no more important than an area were there was only one feature with a sensitivity of 1.0. Both areas would be assigned a sensitivity of 1.0.
2. The weighting mechanism used in the final combination of sensitivity layers can dilute the final constraint value. If, for example, the Wetlands and Riparian sensitivity layer had been

weighted much lower (e.g., 0.10), the other layers would have contributed much more to the final result and the constraints map would be much different. The relative importance (or lack thereof) of the wetland and riparian features would not be evident in the final output.

Application

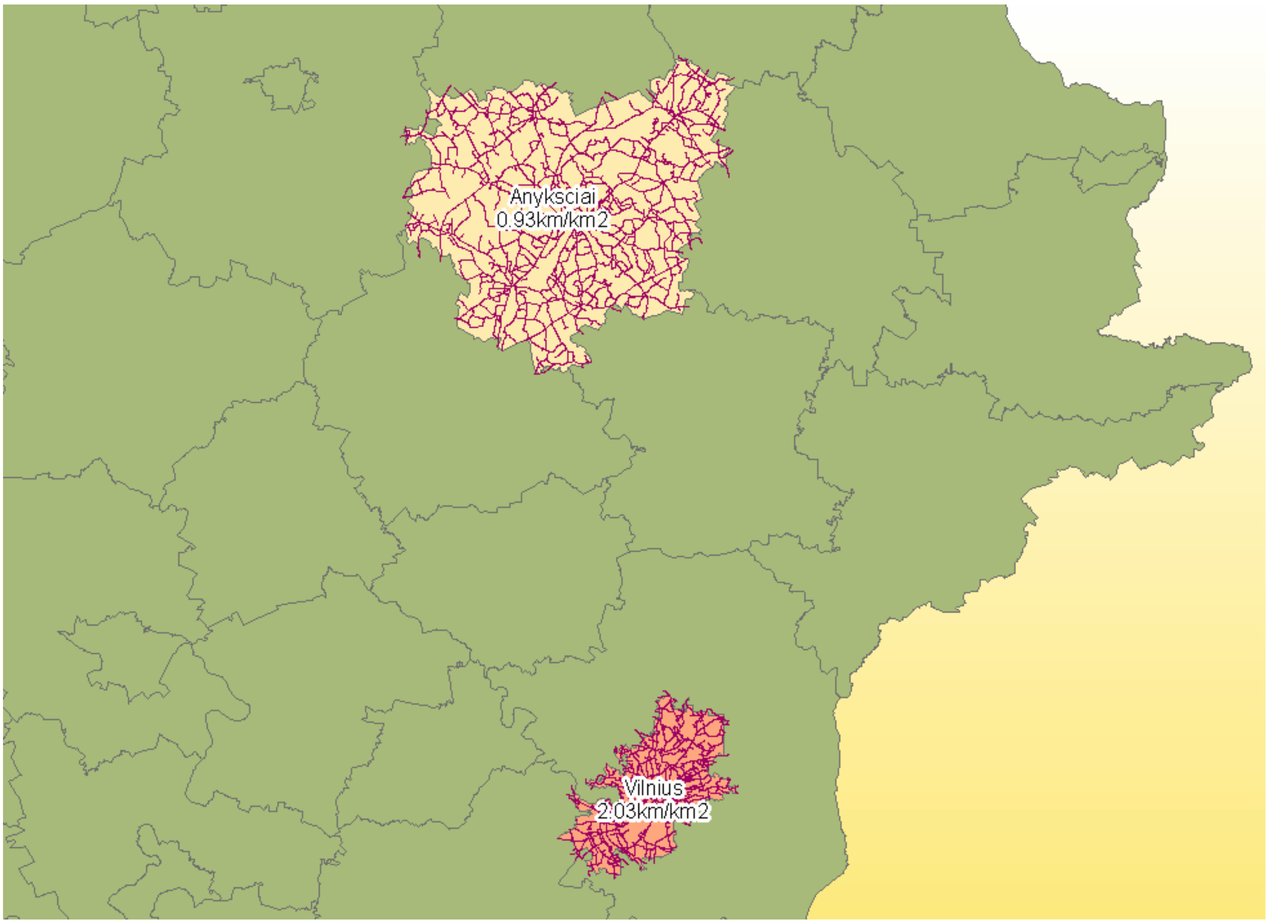
Constraints mapping can provide information on environmental constraints as per the example detailed, however, it can also be used to identify opportunities for siting a project. Ideally two separate analyses would be run in parallel: the constraints analysis examining environmental sensitivities; and an opportunities analysis examining existing development and terrain conditions to identify the 'positive' factors associated with a given location (e.g., close proximity to existing access rights-of-way, low slope, minimum number of road-stream crossings). Potentially the results of the two approaches can either be used independently or integrated into a single map that identifies optimal locations based on both engineering-based siting requirements and minimizing environmental impact. The primary goal of the constraints process is to provide information for more informed decision making, thereby minimizing the effects of industrial development. The constraints map helps effectively identify potential, low impact development sites and, while field verification of these sites is still a requirement, the constraints map allow the effort associated with field programs to be focussed on a much smaller subset of the landscape.

5.5.3 Road Density

Road density provides a useful environmental indicator to help assess the existing and potential impacts of human disturbance on wildlife and fish habitat. It helps us measure the amount of activity in an area and the level of habitat fragmentation. Typically, road density values are summarized as kilometres of road per square kilometre or alternatively the road density in summarized for an area (e.g., by watershed or by a jurisdictional boundary) to facilitate comparison between different regions (Figure 7). When conducted for multiple points in time (e.g., the 1980s, 1990s and 2000s) the results allow us to compare development trends over time and potentially identify those areas under stress.

A road density analysis can be conducted with all roads contributing equally to the final density statistic or, alternatively, a weighting can be applied to roads with a greater width or a higher traffic volume resulting in a weighted road density measurement.

Figure 7. Road Density in Anyksciai and Vilnius



5.5.4 Universal Soil Loss Equation

The Universal Soil Loss Equation (USLE) is a conventional model of soil erosion that takes into account climate characteristics, soil properties, topography, surface conditions, and human activities. It predicts the average soil loss attributed to the runoff from various slopes associated with agriculture, rangeland, and other managed land systems (e.g., construction sites).

The equation is A = R K L S C P
where:

A = average soil loss (e.g., tons/acre/year)
R = rainfall runoff erosivity factor (derived from the energy in an average rainfall)
K = soil erodibility factor (average soil loss in tons/acre/year at a standard slope length and steepness)
L = slope length factor
S = slope steepness factor
C = crop management factor (effect of crop management factors on soil erosion)
P = support practice factor (determined by contouring, strip cropping, terracing, and subsurface drainage)

Each of the above factors contributes to a simulation of conditions that affect the severity of soil erosion at a particular location.

GIS enables the model to incorporate the spatial portion of the equation:

- precipitation data to serve as the rainfall runoff erosivity factor (R)
- soil property map generates the soil erodibility factor (K) (Figure 8).

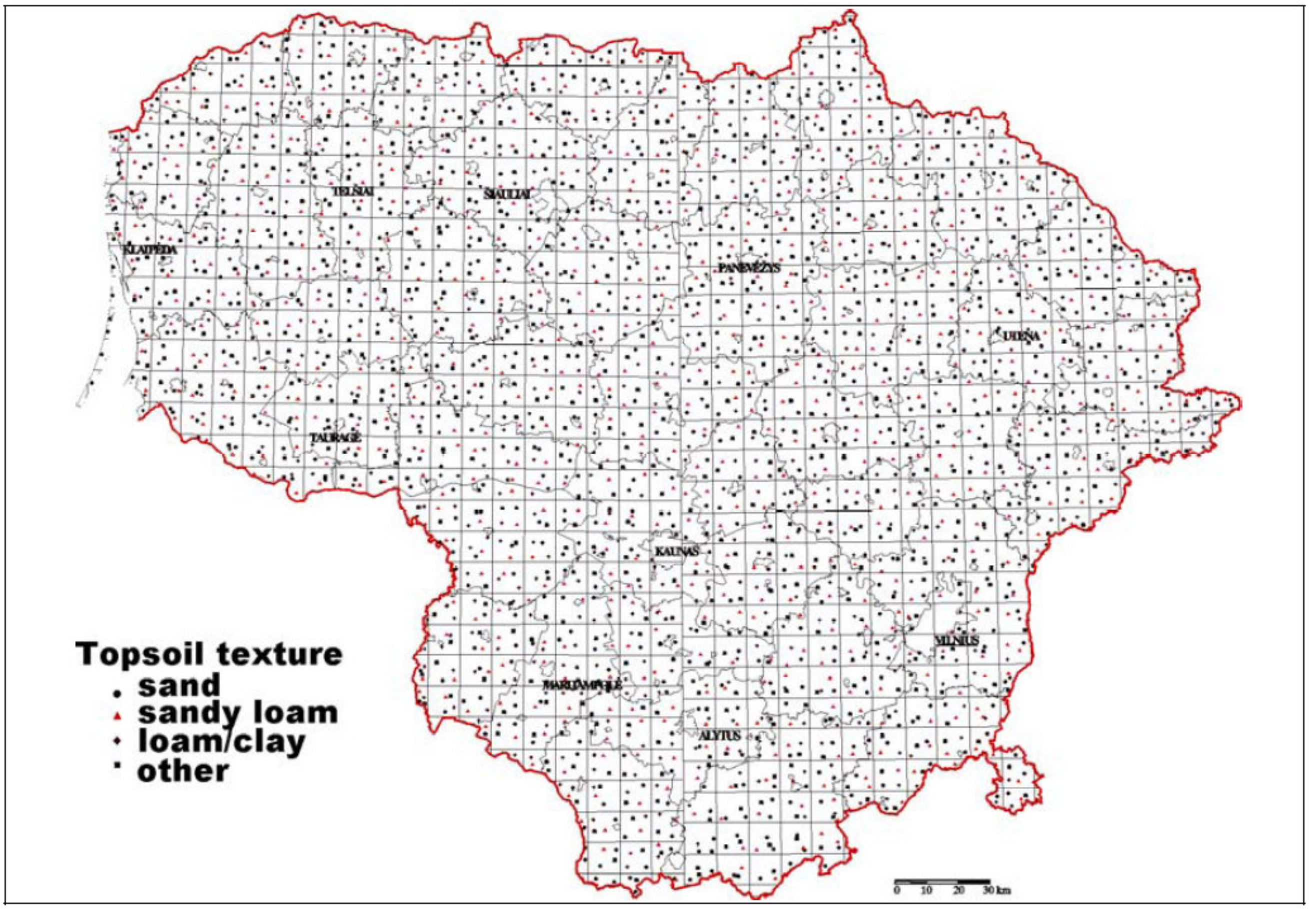


Figure 8. Lithuanian Soil Property Map

- Digital Elevation Models are used to calculate the slope length (L) and steepness (S) for each cell
- a land cover/land use maps may be used as a source for crop management practices (C) and support practices (P) (Figure 9).

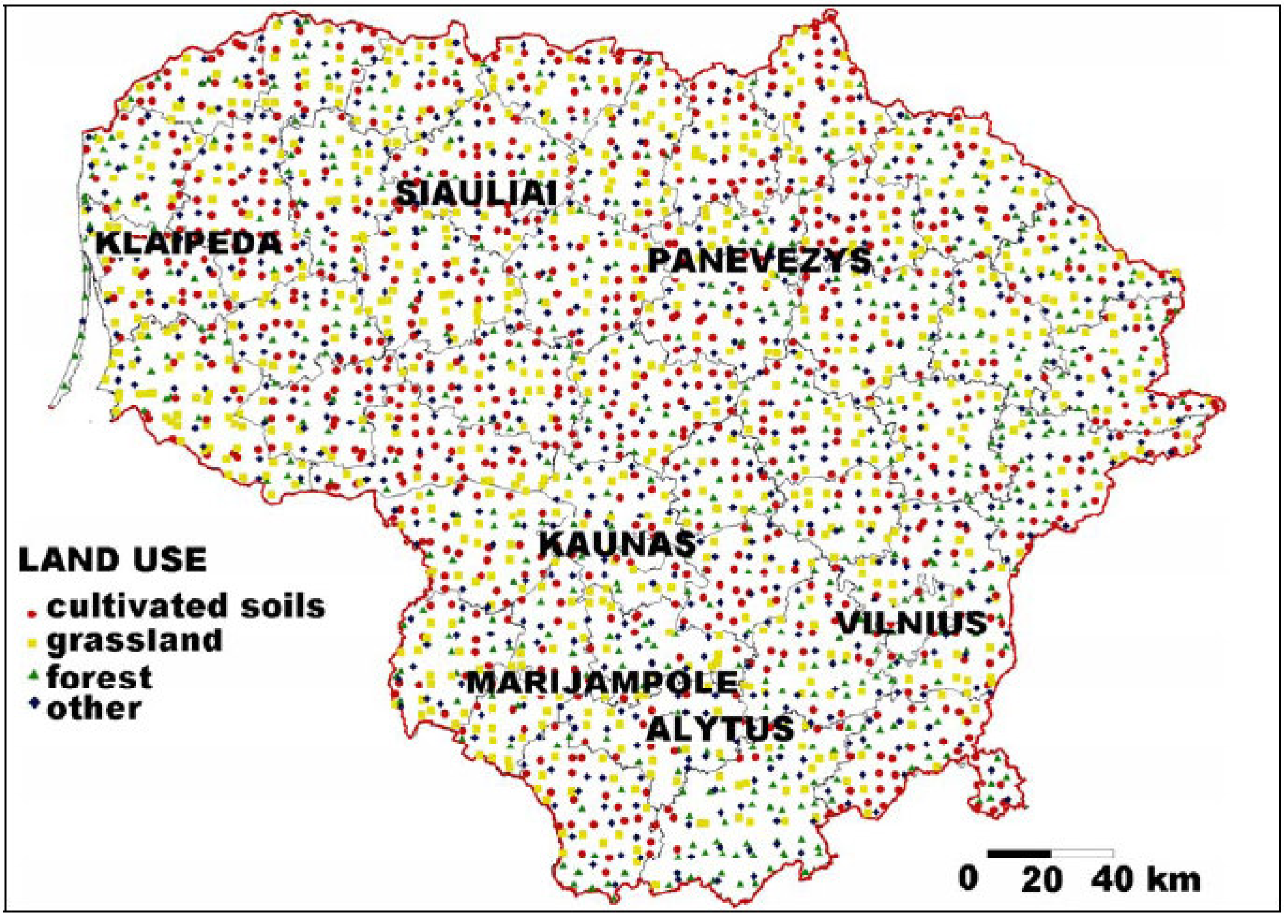


Figure 9. Lithuanian Land Use Map.

After importing and pre-processing, an overlay operation is used to reclassify the data into equivalent units, combine the input data using the Universal Soil Loss Equation, and generate a derivative soil erosion potential layer.

Module Self-Study Questions:

1. What is a model, and what are the limitations of models?
2. What are the ways in which model results may be validated?
3. Describe the nature of a binary model. What types of problems might one use a binary model to solve?
4. Is the Road Density model presented in section 5.5.3 an example of a Descriptive, Explanatory, Predictive or Normative model? Why?
5. Is the Universal Soil Loss Equation presented in section 5.5.4 an example of a Representation or a Process model? Why?

References

1. Blyth, C. Ann, and David Cake, *Constraints Mapping as a Decision Support Tool for Project Siting.* GeoTec 2005 Conference. Vancouver, BC, February 13-15, 2005.

2. Chang, K.T. *Introduction to Geographic Information Systems*. McGraw-Hill, 2006.

3. Chrisman, N. *Exploring Geographic Information Systems*. John Wiley and Sons, 1997.

4. Heywood, I., Cornelius, S. and Carver, S. An *Introduction to Geographical Information Systems*. Pearson Education, 2002,