VIEPS GIS Shortcourse

Lecture 4: Data selection and analysis

Questions we ask

- *Location: What is at . . .?*
- This question seeks to find what exists at a particular location
- A location can be described in many ways
  - place name
  - postal code
  - geographic reference
    - longitude/latitude
    - x and y.

Questions we ask

- *Condition: Where is it . . .?*
- This question is the converse of the first and requires spatial data to answer
- Instead of identifying what exists at a given location, you may want to find locations where certain conditions are satisfied
  - e.g., a packet of land that contains Ordovician turbidites and black shales, has been metamorphosed and is within 100m of a 400myo granite

Questions we ask

- *Trends: What has changed since. . .?*
- This question might involve both Location and Condition queries
- Seeks to find the differences within an area over time
  - e.g., changes in forest cover due to climate change

Questions we ask

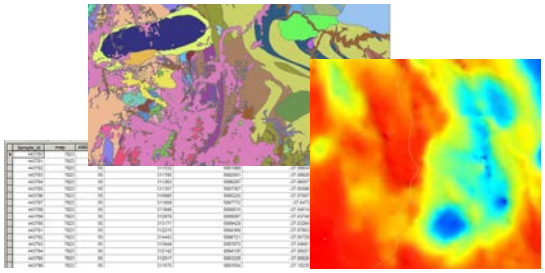
- *Modeling: What if . . .?*
- This question is posed to determine what happens if some events take place or a series of conditions are met
  - Modeling flow of toxic material in goundwater after a spill
- Answering this type of question requires both geographic and other information (as well as specific models)

Fundamental questions of GIS

- What is where?
  - Querying
- Why is it there?
  - Explaining distributions, finding new relationships
- Need information that we can query
  - Databases
    - Storage, management
  - Attributes
    - Analysis
  - Presentation Layer
    - Display

Geodatabase?

- an organised body of related geospatial information



Attributing Data

- Databases work with attributes
  - whats in "what is where?" question
- Attribute = measurement or value for the feature
- Different types (called "levels" of measurement)
  - labels (IDs)
  - qualitative categories ("good", "red", "warm")
  - quantitative categories (amounts)
  - special cases: dates, time, degrees of slope, addresses
- Features are stored in Feature Attribute Tables (e.g., AAT, PAT, etc)
  - Features are stored in rows, called records
  - Attributes are stored in columns, called fields

Attribute types in GIS

- Integers (1, -8, 357)
- Floats (3,125 = 3.125 x 103 = 3.125E3)
  - Regular floats store only 4-bit numbers (<=7 significant digits)
  - Double (floats) 8-bit is enough to store (<=15 significant digits)
- Text ("Wonga Schist")
- Date mm/dd/yyyy hh:mm:ss
  - e.g., 12/06/2004 10:20:35 am
- BLOB – binary large object can store images, pictures, any binary computer code
- GlobalID – unique number identifier for the feature

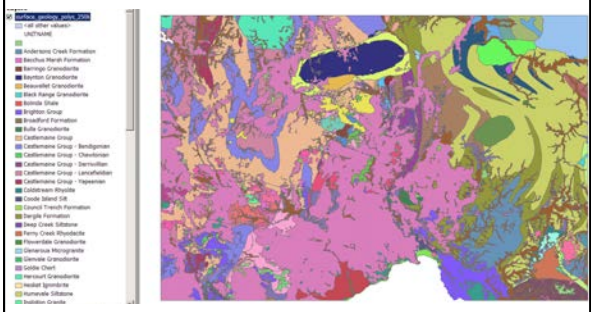
Levels of Measurement

- Not all attributes are equal
  - Stevens (1946) in Science defined "levels of measurement"
- Attempt to structure observations about reality
  - Nominal scale – named objects
  - Ordinal scale – ordered data
  - Interval scale – thermometer
  - Ratio scale – true zero
- Compare appropriate data
  - Comparing polyID with RockType makes no sense
  - Comparing alteration with gold grade does

Nominal Scale

- This is the simplest level of scaling
- Essentially, an ID (e.g. a name)
- Can be words OR numbers!
- There are no mathematical operations that can be performed between classes
- The data are purely qualitative
- Examples: "Zoning code R-1", "alfisol", "spruce forest", gender (M/F), MN 56301, SSN

Nominal Scale Map



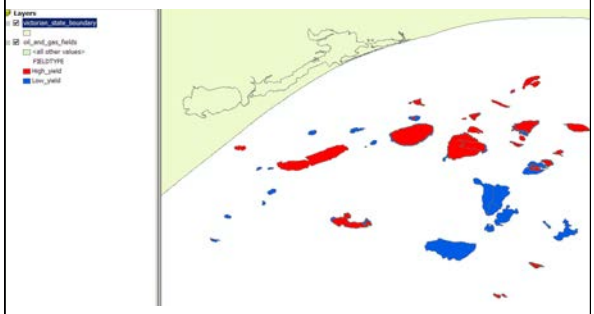
Nominal records

- You can search for:
  - exact match (DepEnv= “marine”)
  - select a group (e.g., select all drill holes labeled with a MD prefix)
    - Stawell Gold Mine – Magdala Deep Holes
- Examples of queries:
  - Find all places called “Sesame Street”
  - Find all rocks with prospectivity code of “1”
  - Find all exploration leases owned by “New Bendigo Mines”
  - Queries are written in a formal query language SQL
  - ([Rock_type]=“Granite” OR “Granodiorite”)

Ordinal Scale

- Ordered or ranked data
- Objects or events can be arranged from least to most, or vice versa
- Ordinal scaling provides no way of determining how much distance separates the items in the array
- Can search for exact match or range of values
- Examples:
  - grades (A, B, C, D, F) – if no grade point
  - qualitative categories
  - Ground cover “thick”, “thin”, “absent”
  - soils “wet”, “moderate”,”dry”
  - Zoning R-1, R-2, R-3
  - Find soils that are “wet” or “moderate”
  - Find all R-1 or R-2 parcels

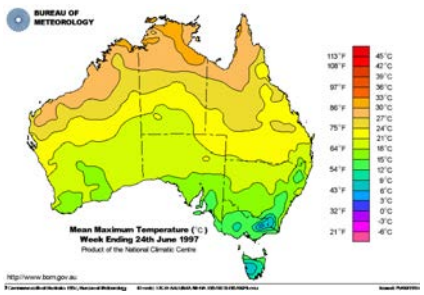
Ordinal Scale Map



Interval scale

- Objects in order of rank with a known distance between the ranks
  - Units on the interval scale are uniform
- Can subtract and add, but not multiply or delete
  - Year 2000 AD is 1000 years away from 1000 AD
  - But year 2000 is NOT twice as “big” as 1000
- Variables on the interval scale do not have zero as a starting point
  - Less common in GIS
  - Examples: temperature (ºF or ºC), pH, years BC or AD, ….

Interval Scale map



Ratio Scale

- Like interval scaling, ratio scaling involves ordering objects with known distances separating them
- The difference is that ratio magnitudes are absolute
  - True zero (absence of magnitude) as its starting point
- Very common in GIS
- Most common in analysis
- Examples: absolute temperature (deg. K), population numbers, timber yields in cu. M or in board-feet, property values or cost of land in US$, amount of rain in mm, etc.

Data structure

- GIS and DBMS
  - Retrieval:
    - Ability of the DBMS or GIS to get back on demand data that were previously stored
  - Geographic search is the secret to GIS data retrieval
  - GI systems have embedded DBMS, or link to a commercial DBMS
    - Arc/View uses dBase
    - Arc/INFO 7.x uses INFO
    - ArcGIS can link to Informix, Oracle, SQL Server, Access, etc.

Advanced Functions

- Sorting
  - e.g., sort by Sample Number in descending order
- Re-Ordering
  - e.g., put records 20-29 BEFORE 1-19
- Subsetting (also called restricting)
  - e.g., list all sample localities that contained between 2 and 10 g/t Au
- Searching by attribute
  - find all map units whose rock type includes "stone"
    - Sandstone, siltstone, greenstone, etc
- Searching by geography
  - find all lots adjacent to property A

SQL Querying

- Retrieve or manipulate data from the relational tables based on certain criteria
  - SELECT
  - INSERT
  - UPDATE
  - DELETE

SELECT

- The SELECT statement lets you select a set of values from a table in a database
- Values returned depend on the conditions specified in the SQL query
- Returned values are called "recordsets"

```
SELECT Name,Dept_ID FROM emp

SELECT * FROM emp WHERE Dept_ID='1'

SELECT * FROM dept WHERE ID>3

SELECT * FROM emp WHERE name LIKE '%aw%'

SELECT address FROM people WHERE (firstname='Tim' OR
city='Ballarat') AND age>30
```

Other SQL statements

- INSERT statement lets you insert information into a database as a new row at the bottom of the table
  - INSERT INTO emp VALUES (7, 'Uncle Bill', 2)
- DELETE statement removes records or column values from a database
  - DELETE FROM emp WHERE name = 'Jason'
- UPDATE statement updates (or replaces) specified values
  - UPDATE emp SET Dept_ID = 5 WHERE (name = 'Luke' AND ID = 3)
- DISTINCT , ORDER , MAX , MIN , AVG, SUM, COUNT(), etc.

Relational Databases

- Relational databases obtain their flexibility from being based on set theory (relational calculus) which enables sets or relations to be combined in various ways, including:
  - join/intersection
  - union (i.e. the sum of 2 sets);
  - exclusive "OR" (i.e. the difference between 2 sets)
  - and outer-join which is a combination of intersecting and exclusive "or"-ing.

Tables

Emp

Id	Name	Dept Id
1	Bill Smith	3
2	Mike Lewis	2
3	Ray Charles	3
4	Andy Mallory	4
5	Mandy Randall	6
6	Allison White	1

Dept

Id	Name
1	HR
2	IT
3	Marketing
4	Sales
5	Finance

- The ***intersection* or *join*** between 2 sets (in this case, tables) produces only those elements that exist in both sets
- Therefore, if we join *emp* and *dept* on *deptID*, we will be left with only those employees who work for a department that is in the *dept* table and only those departments which have employees who are in the *emp* table
- The ***union*** produces the sum of the tables - meaning all records in *emp* and all records in *dept* and this may be with or without duplicates

Tables

Emp

Id	Name	Dept Id
1	Bill Smith	3
2	Mike Lewis	2
3	Ray Charles	3
4	Andy Mallory	4
5	Mandy Randall	6
6	Allison White	1

Dept

Id	Name
1	HR
2	IT
3	Marketing
4	Sales
5	Finance

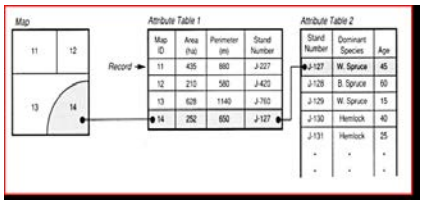
Joins

The **join** of Emp and Dept. on the department id would produce the following result:

Emp.Id	Emp.Name	Dept.Id	Dept.Name
1	Bill Smith	3	Marketing
2	Mike Lewis	2	IT
3	Ray Charles	3	Marketing
4	Andy Mallory	4	Sales
6	Allison White	1	HR

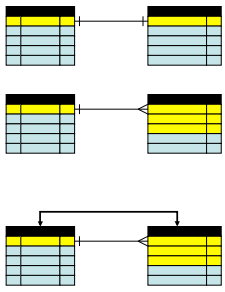
Joins

Joins in GIS may include map data as well



Joins

- One-to-one – CAN BE JOINED
  - States and their capitals
  - One state – one capital
- Many-to-one – CAN BE JOINED
  - States and Countries
  - Many states inside one country, but never one state belonging to two different countries
- One-to-many and many-to-many – CAN ONLY BE RELATED
  - Parcels and owners
  - One parcel can be owned by one or many owners
  - One owner can own more than one parcel



Union

The **union** of Emp and Dept. would produce the following results

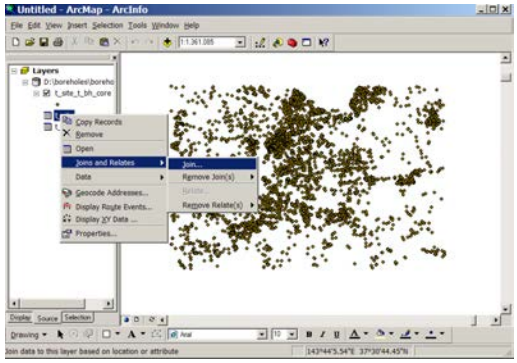
Id	Name
1	Bill Smith
2	Mike Lewis
3	Ray Charles
4	Andy Mallory
5	Mandy Randall
1	HR
2	IT
3	Marketing
4	Sales
5	Finance

SQL Join syntax

SELECT Emp.Name, Dept.Name
FROM Emp, Dept WHERE
Emp. Dept_ID = Dept.ID

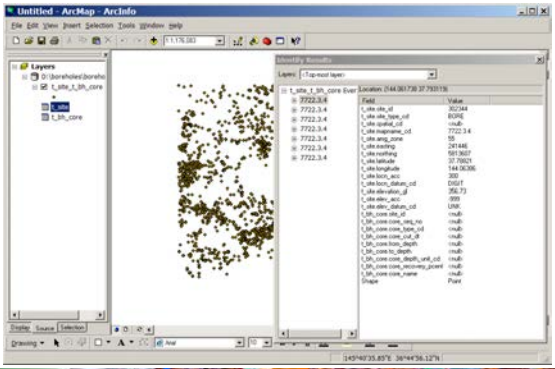
but the good news is…

Joins in ArcGIS

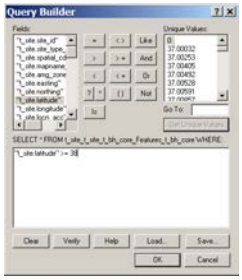


Joins in ArcGIS

Joins in ArcGIS



Select queries in GIS

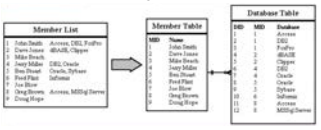


Normalisation

- The process of structuring data to minimize duplication and inconsistencies
- Break down single tables into two or more tables and defining relationships between those tables
- Stages
  - Normal Forms
  - Each more stringent
- Higher levels of normalisation increases the efficiency of a particular database
  - Also increases complexity
- Usually use the Third Normal Form
- Generally, there is no real need to know the terminology, applying the principles of normalization are generally enough unless you get caught at a dinner party with a database designer

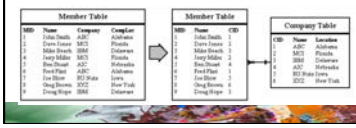
Normalisation

Eliminate Repeating Groups



Eliminate Redundant Data

Eliminate Columns Not Dependent On Key



Metadata

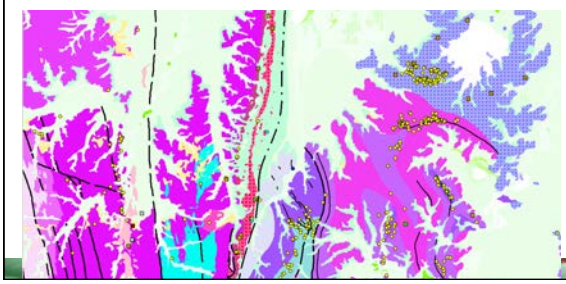
- Metadata is simply data about data
- Information about:
  - Content
  - Projection and coordinate system
  - Source
  - Quality
  - Condition
  - Other relevant characteristics of the data
- e.g. Geological data, MGA 94, from Geoscience Victoria, locations accurate to 20m, extrapolated beneath cover

Analysis: not why is it there, but why is it *there*?

- Spatial analysis is a study of the pattern and relationships between points, lines, areas, and surfaces
- It is about discovering/creating new relationships! (as opposed to merely searching for existing stuff)
- Creates new maps (shapefiles, coverages)

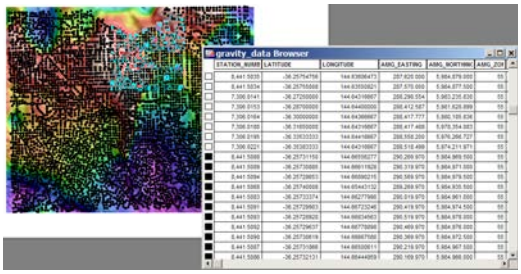
Types of analysis

- Showing the geographic distribution of data
  - Pins on a wall map
  - Powerful pattern recognition

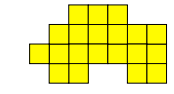


Types of analysis

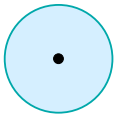
- Querying GIS data
  - Text based queries based on attributes
  - Spatial queries based on location



Types of analysis

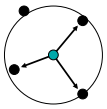


Area Calculation



Buffers

Overlays



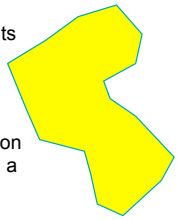
Nearest Neighbor

Spatial Analysis Operations

- Possibilities are limitless!
- Measurement and Classification
  - *Isolated features on a single coverage*
  - How far? How big? Find all parcels of class X…
  - Common statistics (mean, median, min, max, SD
- Overlay Operations
  - *Requires two or more coverages*
  - A + B , A AND B
- Neighborhood Operations
  - *Works with immediate neighbors on one coverage*
  - Finding neighbors, topography functions
- Connectivity Operations
  - *Works with distant neighbors and across the whole coverage*
  - Buffers, network analysis, barriers

Measurement and Classification

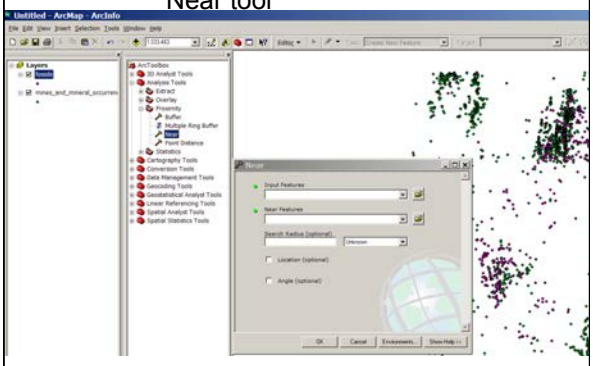
- Distance measurement
- Measure distances between points
  - careful with projections!
- Measure perimeters of polygons
- Areas of polygons
- Can calculate common statistics on those, e.g., mean polygon size in a coverage



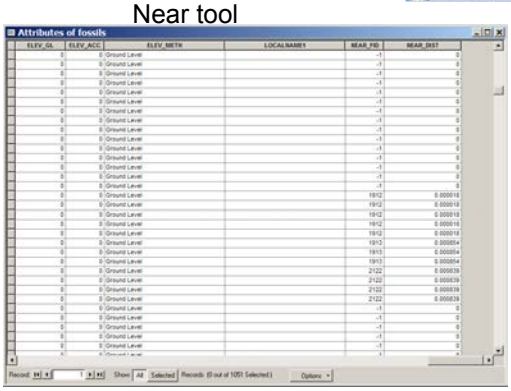
What are you measuring?

- Euclidean Distance - $\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$
- Network bound distance
- Long distance
  - Projection

Near tool

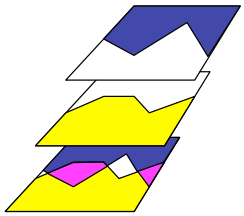


Near tool



Overlay Operations

- More complex (and powerful) than just viewing several layers of data at once
- You control how the GIS overlays the data
- Creates "new" datasets

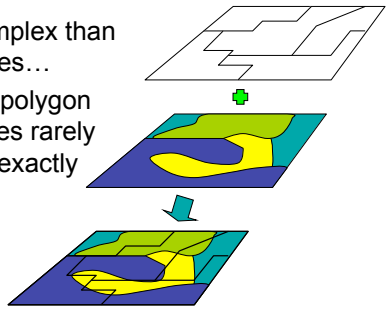


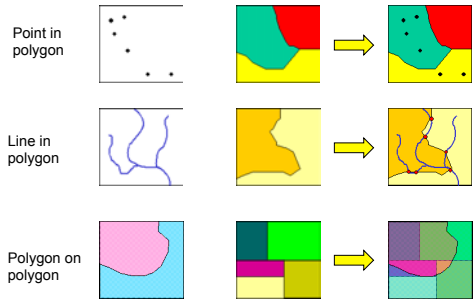
Overlay Operations

- Arithmetic
  - Includes addition, subtraction, multiplication, square roots, etc.
  - Works only with ratio and sometimes with interval scale data
- Logical (same as Boolean)
  - Finding areas where a specified set of conditions occurs or does not occur (True or False)
  - Works with any data, most frequently with nominal data

Vector Overlays

- More complex than raster ones…
- because polygon boundaries rarely coincide exactly





Applying weighting

9	9	1	
9	9	9	Prospectivity (model)
5	1	5	

9 – excellent
5 – average
1 - poor

Weighting Factor: x1

18	18	18	
10	18	18	Assay results
10	2	2	

9 – excellent
5 – average
1 - poor

Weighting Factor: x2

9	9	9	
5	9	9	Access
1	5	9	

9 – excellent
5 – average
1 - poor

Weighting Factor: x1

36	36	28
24	36	36
16	8	16

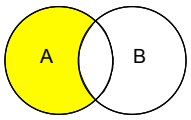
Result

Suitability analysis

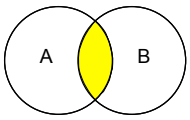
- Boolean logic is common in suitability analysis
- Consider a logical expression
  - FIND City population > 500,000 AND State pop.>5,000,000
- Operators are <, >, =, <=, >=
- "City population" and "state population" are operands
- AND, OR, NOT and XOR are called Boolean connectors
  - Find all boys in the classroom roster
    - NOT girls
  - Find all conifer forest stands that are not tamarack (tamarack is not evergreen)
    - Conifer AND Evergreen
  - Find all soil orders suitable for farming in Minnesota
    - Mollisols OR alfisols
  - I can afford a DVD player XOR a digital camera (but not both)

Logical expressions

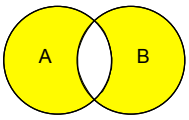
A ***NOT*** B



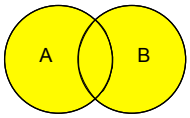
A ***AND*** B



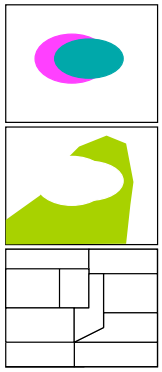
A ***XOR*** B



A ***OR*** B



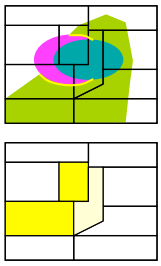
Overlay Example



Granite and alteration aureole

Fe rich black shales

Government exploration lease map

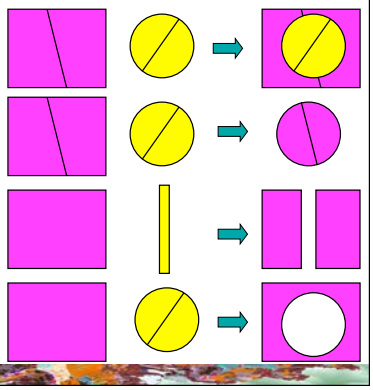


Overlay with interpretation

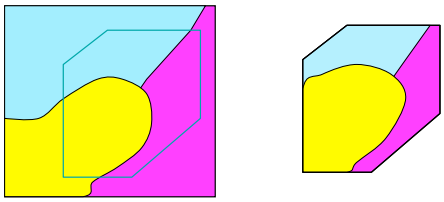
Acquisition targets

Map manipulation

- **Update** – cut and paste
- **Clip** – cookie cutter
- **Split** – cuts into fragments
- **Erase** – removes shape

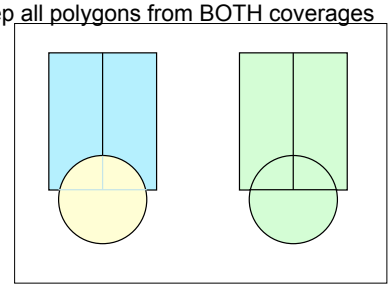


Clip a coverage



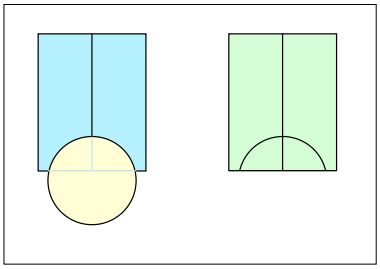
Overlays in ArcGIS

- UNION - overlay polygons from 2 coverages, keep all polygons from BOTH coverages



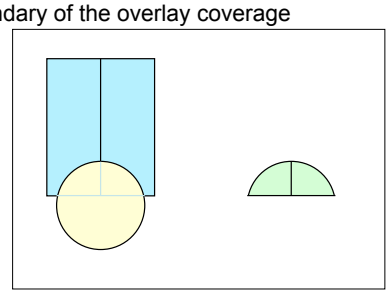
Overlays in ArcGIS

- IDENTITY - overlay points, lines, or polygons and keep ALL from ONLY the INPUT coverage



Overlays in ArcGIS

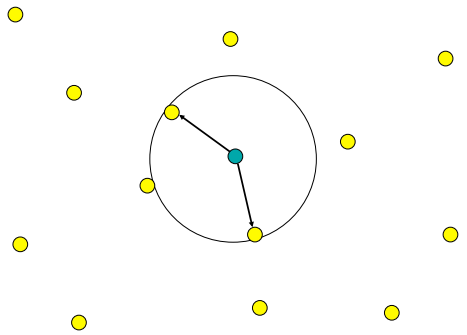
- INTERSECT - overlay points, lines, or polygons, but keep ONLY those from INPUT that fall within the boundary of the overlay coverage

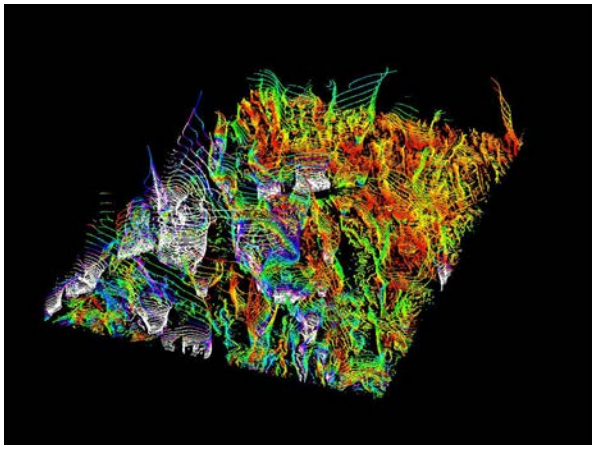


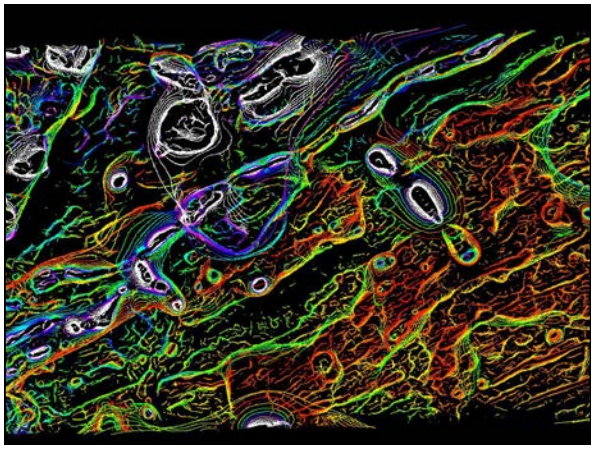
Neighborhood Operations

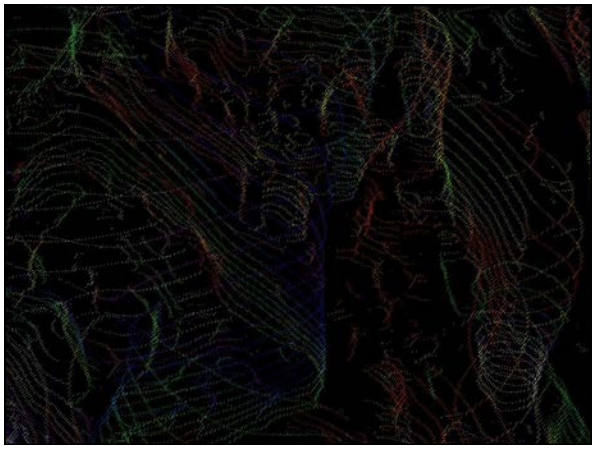
- Search for nearest neighbor
- Topographic Functions
- Contour Generation (from DEM)
- Interpolation - adding missing values

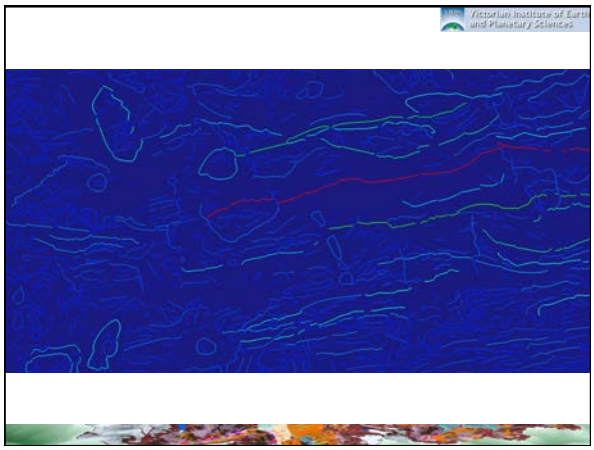
Nearest neighbour

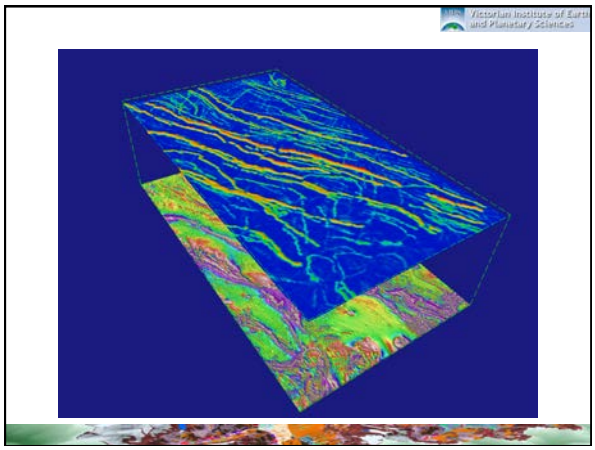




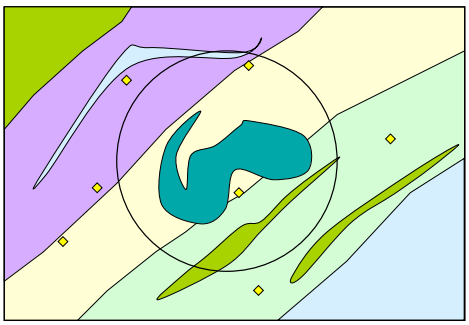






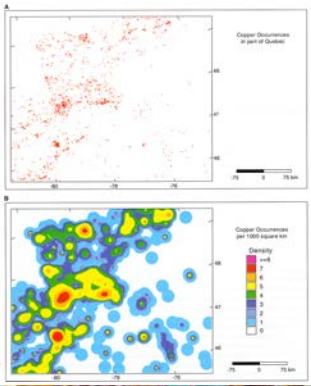


Nearest neighbour



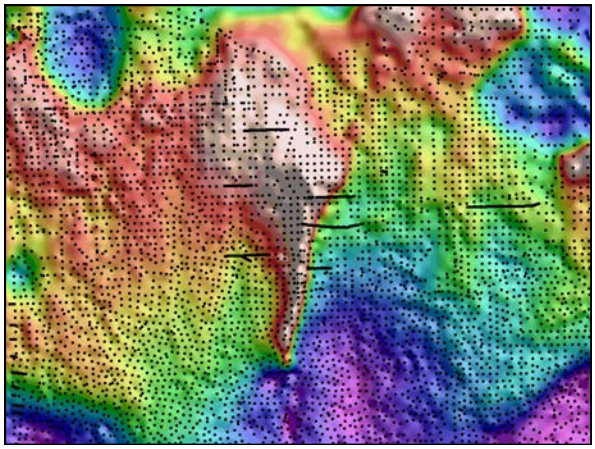
Point to area – density mapping

- Create a raster from point data
- Visualise a surface for continuous data
- Visualise relationships from sample density surface



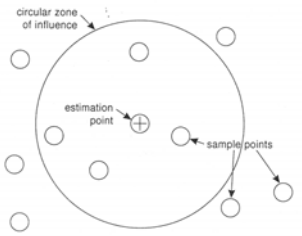
Interpolation

- Calculating missing values
  - 1,2,X,X,5,6,7,X,9,10,X,X
  - 1,2,3,4,5,6,7,8,9,10,11,12
- A few different algorithms
  - Linear
  - Non-linear
- Generates surface from a network of points
  - Errors will be greater where points are scarce
  - Errors will be smaller where points are numerous



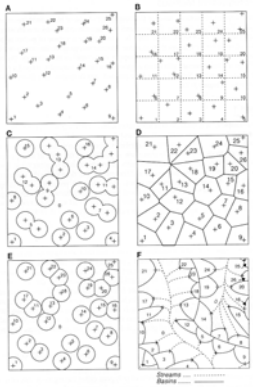
InInverse distance weighting

$$\hat{z}_0 = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}$$



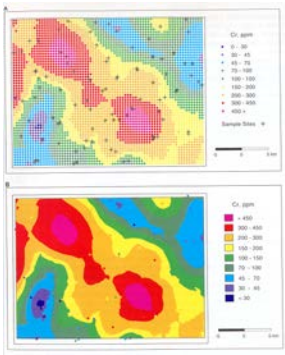
Point to area

- A – sample data
- B – cellular grid
  - Multiples aggregated
  - Nulls null or averaged
- C – circles
  - Overlap aggregated
- D – Thiessen polygons
  - 1 poly per point
- E – Thiessen circles
- F – points to catchments



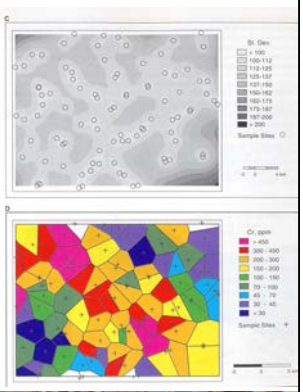
Points to area

- Cr distribution
  - A – Kriging estimates at points on a grid
  - B – Grid conveted to a raster
  - Could have been threaded to create contours



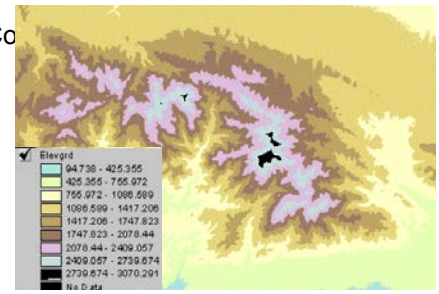
Points to area

- C – kriging standard deviation
- D – Thiessen polygons of same data



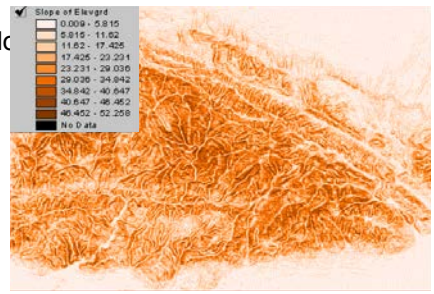
Topographic Functions

- Co



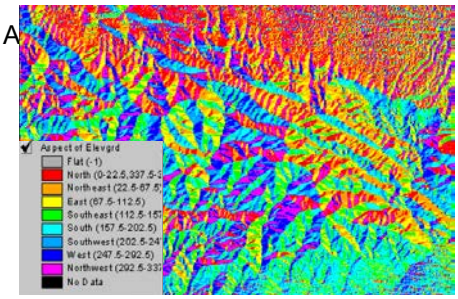
Topographic Functions

- Slo
  e



Topographic Functions

- A



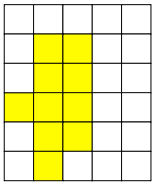
Connectivity Operations

- Contiguity
  - finding areas of same value
- Buffers
- Network Analysis
- Barriers
- Intervisibility (Viewsheds)

Contiguity

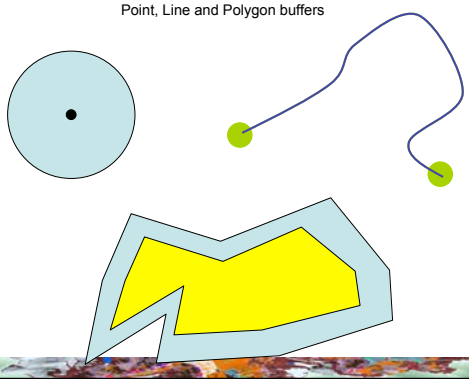
- The topological identification of adjacent polygons
- Contiguity is not concerned with the exact locations of polygons, only their relative positions

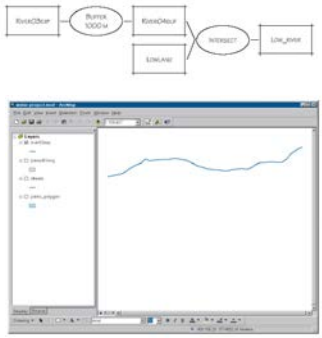
2	5	3	7	1
3	1	1	3	4
2	1	1	6	4
1	1	1	6	1
7	1	1	4	2
7	1	7	5	4

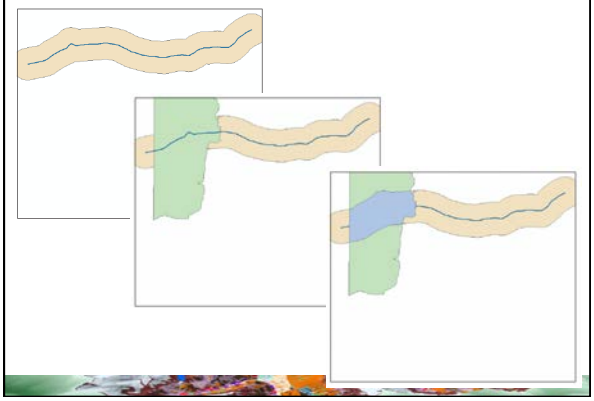


Buffer objects

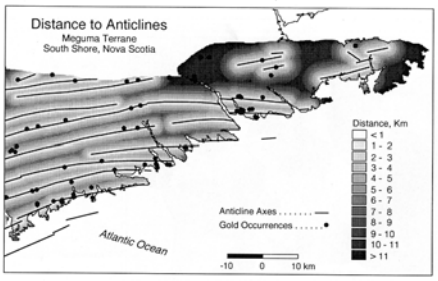
Point, Line and Polygon buffers



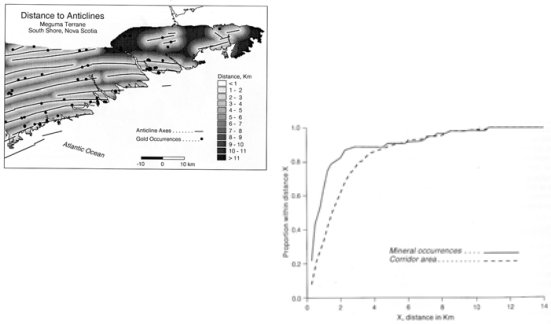




Object Dilation



Object Dilation



Network

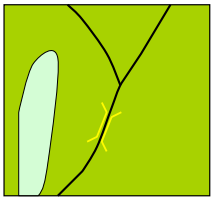
- Connected linear features
- Each segment is called a link
- Impedance is the speed (cost) of traversing the link
- Some links may be one-way only
- There can be forbidden turns
- The goal is usually to find shortest and/or fastest route connecting two or multiple points

Routing

- Navigation
- Motion analysis
- Logistics

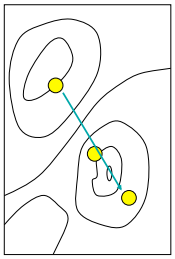
Barriers

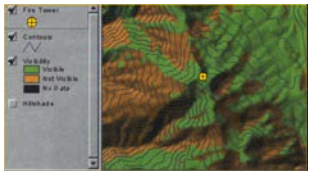
- Relative barriers
- Absolute barriers
- Access points
- Impedance values



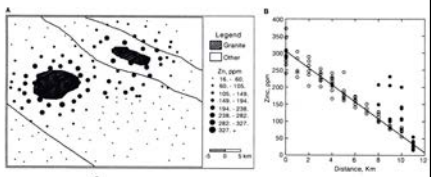
Topographic analysis

- Viewsheds

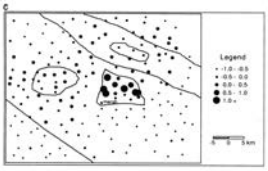




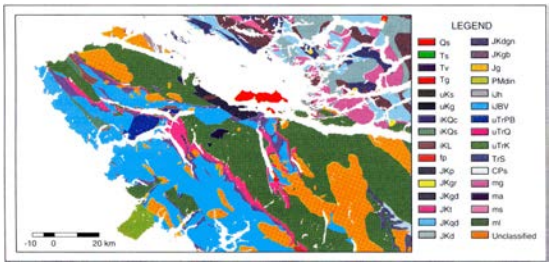
Analysis – simple modelling



- Develop model from analysis
- Model does not fit all data
- Expand model
  - Chemistry of marsh
  - organics



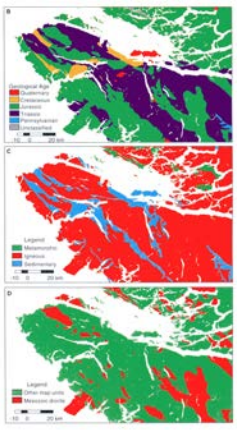
Map reclassification



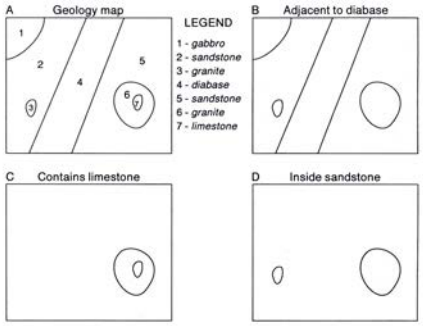
Geology map with 32 classes in attribute table (stratigraphy)

Map reclassification

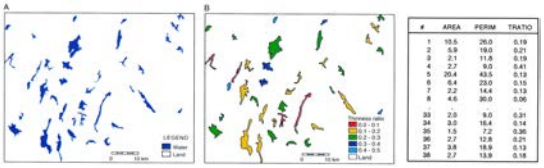
- New maps
  - Generalised to 7 classes
    - Age
  - Generalised to 3 classes
    - Rock type
  - Generalised to 2 classes
    - Mesozoic diorite



Topological analysis



Geometrical analysis



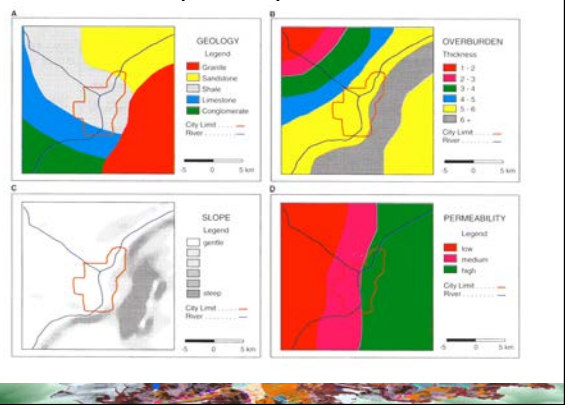
Geometrical analysis of lake shapes

• Area, perimeter and thinness derived from polygon attributes

Multiple map analysis

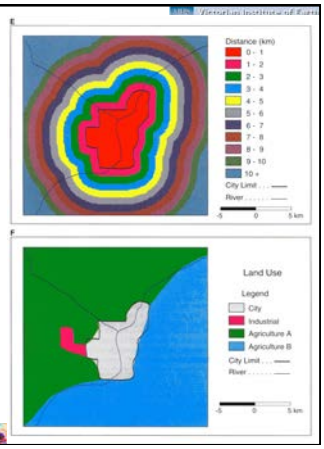
- Modelling multiple maps
  - Boolean
  - Index overlay
    - Weighted combination
  - Fuzzy Logic
  - Weights of evidence
    - Probability model

Landfill example – input data

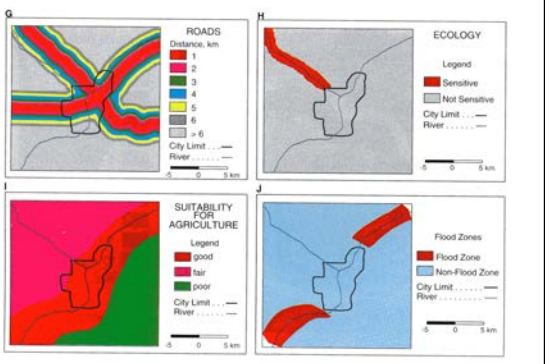


Example - landfill

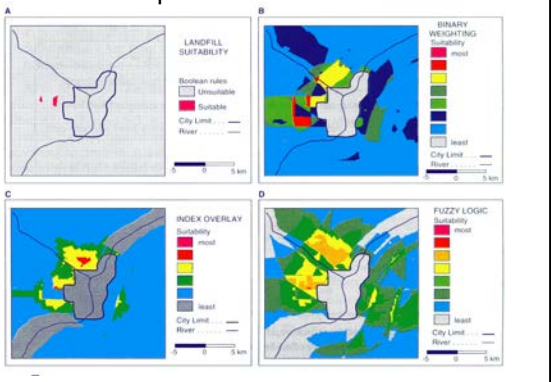
- Input data
  - Buffered distances to city
  - Land use map



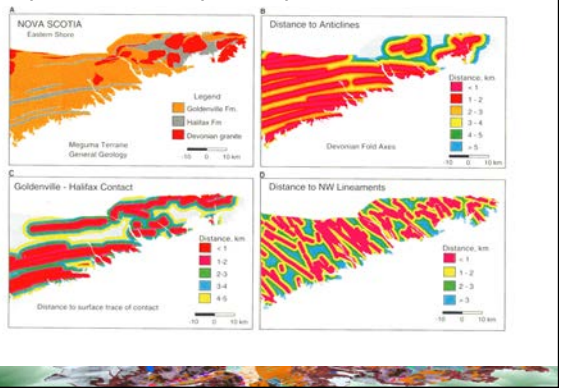
Landfill example – input data



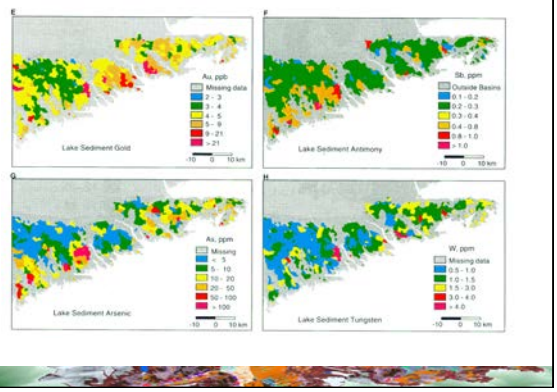
Landfill example – results data



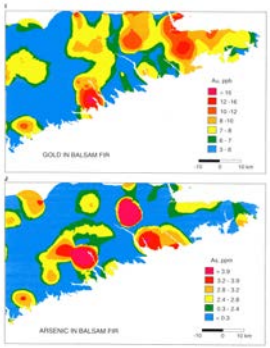
Exploration example – input data



Exploration example – input data

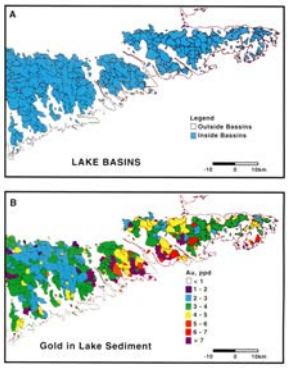


Exploration example – input data



Breakpoint classification

- Catchments regions as separate objects (polygons)
  - Each has geochem attributes
- Catchments reclassified using gold levels and breakpoints



Exploration example – analysis