Data Manipulation and Analysis

Dr.Weerakaset Suanpaga
(D.ENG RS&GIS)

Department of Civil Engineering
Faculty of Engineering , Kasetsart University
Bangkok, Thailand

http://pirun.ku.ac.th/~fengwks/gis/lecture/3datamanagement.pdf

Data Manipulation

- Data Manipulation deals with handling spatial data for a particular purpose.

Data Analysis

Data Analysis deals with the discovery of general principles underlying the total phenomenon.

Example 7 Operations in data manipulation and analysis

1. Reclassification and Aggregation
2. Geometric Operations
   - Rotation, Translation and Scaling
   - Rectification and Rotation

3.Centroid Determination

4.Data Structure Conversion

5.Spatial Operations

– Connectivity and Neighbourhood Operations

6.Measurement

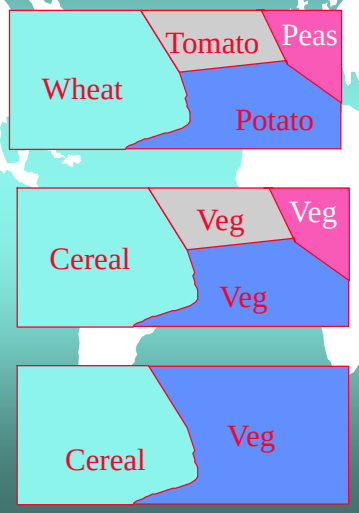
– Distance and Direction
– Statistical Analysis
– Descriptive Statistics
– Regression, Correlation and Cross-Tabulation

7.Modeling

1.Reclassification *and* Aggregation

- Data may not be compatible with the user need or for further analysis
- Data may be at different resolution than needed by the user

2.Attribute Aggregation



CROP TYPE MAP

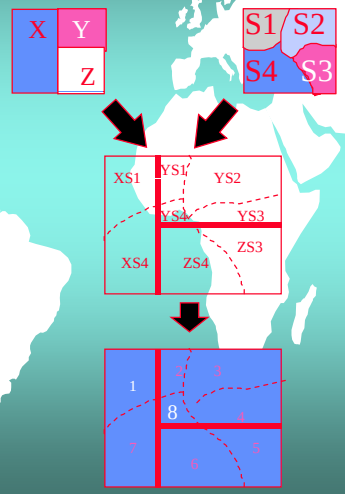
RECODED MAP OF VEGETABLE AND CEREAL CROP

REDUNDANT BOUNDARIES REMOVED

Overlay

- Polygon overlay or dissolve techniques involve the compositioning or extracting multiple maps in order to create a new dataset
- Mathematical overlay : for the purpose of area and measurement and multiple attribute modeling

Polygon Overlay



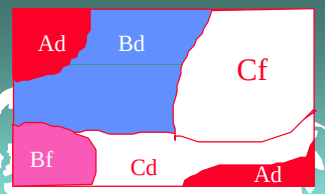
LAND HOLDING

ID TYPE	ID	SOIL
1	X	S1
2	Y	S1
3	Y	S2
4	Y	S3
5	Z	S3
6	Z	S4
7	X	S4
8	Y	S4

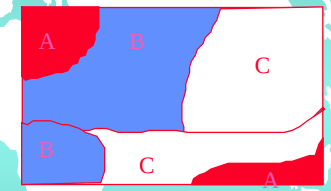
Map Dissolve

- To extract a single attribute from a multiple attribute polygon.
- Steps
  - Reclassifying soil areas by soil type only
  - Dissolve boundaries between areas of same soil type
  - Merge polygons into large objects

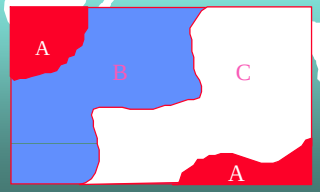
SOIL TYPES A, B, C WITH GROWTH POTENTIALS



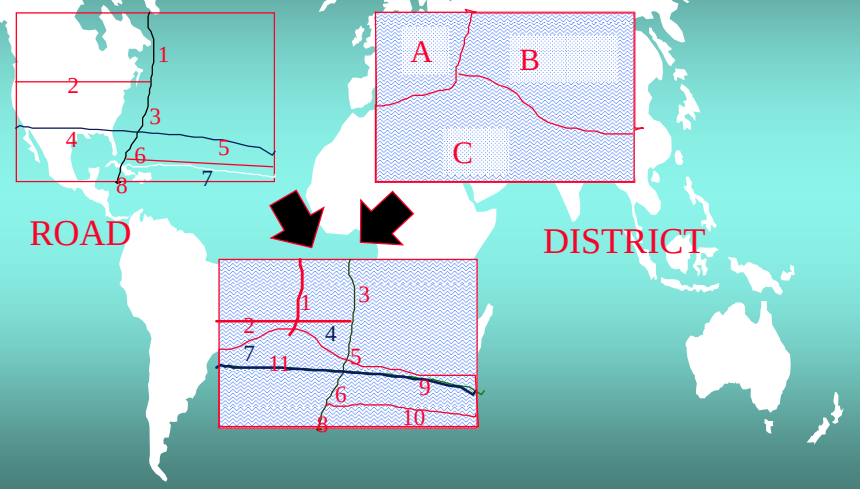
SOIL TYPES **A, B, C**



SOIL TYPES **A, B, C**



Overlay Line on Polygon

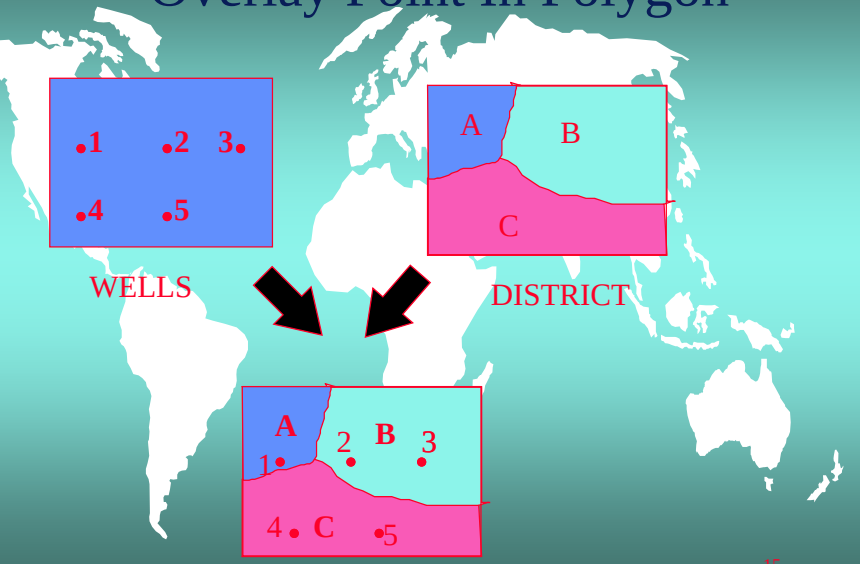


ID	ROAD
1	35
2	22
3	35
4	60
5	60
6	35
7	82
8	35

ID	DISTRICT
A	Kanpur
B	Fatehpur
C	Banda

ID	ORIGINAL	ROAD	DISTRICT
1	2	22	Fatehpur
2	2	22	Kanpur
3	1	35	Fatehpur
4	3	35	Fatehpur
5	4	60	Banda
6	4	60	Banda
7	5	60	Banda
8	6	35	Banda
9	6	35	Fatehpur
10	7	82	Banda
11	8	35	Banda

Overlay Point In Polygon



ID	BLOCK
1	Rampur
2	Mandhana
3	Nankari
4	Bithur
5	Bilhaur

ID	DISTRICT
A	Kanpur
B	Fatehpur
C	Banda

ID	DISTRICT	LOCATION
1	Kanpur	Rampur
2	Fatehpur	Mandhana
3	Fatehpur	Nankari
4	Banda	Bithur
5	Banda	Bilhaur

Spatial Aggregation

- It involves increasing the size of the elemental unit in the database
- For Raster datasets only
- Regions of less than a specified size is ignored for a particular application

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	2	2	2
1	1	1	1	1	1	2	2	2	2
1	1	1	1	1	2	2	2	2	2
1	1	1	2	2	2	2	2	2	2
1	1	1	2	2	2	2	2	2	2
2	2	2	2	2	2	2	1	2	2
1	1	2	2	2	2	2	2	2	2

1 - SUBURBAN

2 - URBAN

1	1	1	1	2
1	1	1	2	2
1	2	2	2	2
2	2	2	2	2

1	1	1	3	3
1	1	3	2	2
1	3	2	2	2
3	2	2	3	3

1 - SUBURBAN

2 - URBAN

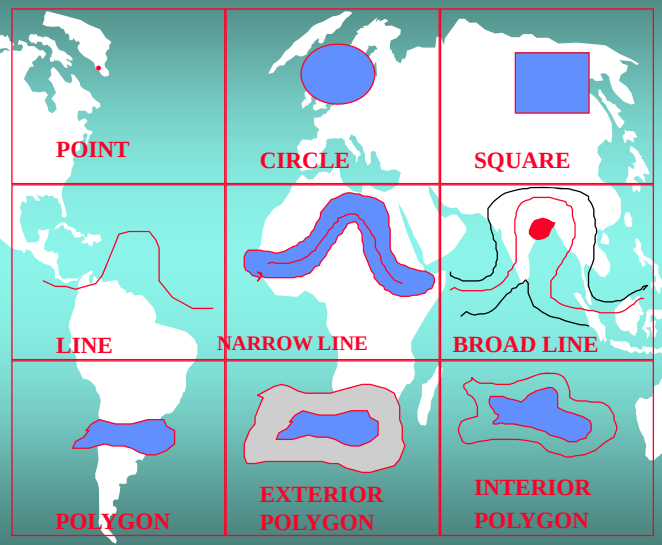
3 - MIXED

- In Vector dataset : Merging of adjoining polygons based on their attributes
- These processes of changing the mean resolution of the data change the effective size of the MINIMUM MAPPING UNIT
- Decision Rule for mapping

Buffer Generation

- Generation of new polygon from points, lines and polygon features within the database
- Circular or square buffer can be calculated

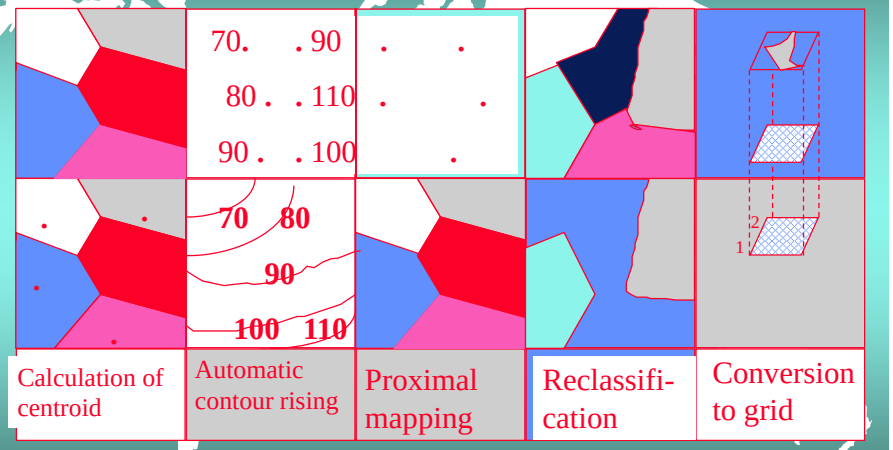
Buffer Generation



Map Abstraction

- Map Abstraction consists of
  - Calculation of Centroid
  - Automatic Contouring
  - Proximal Mapping
  - Reclassification
  - Conversion to Grid

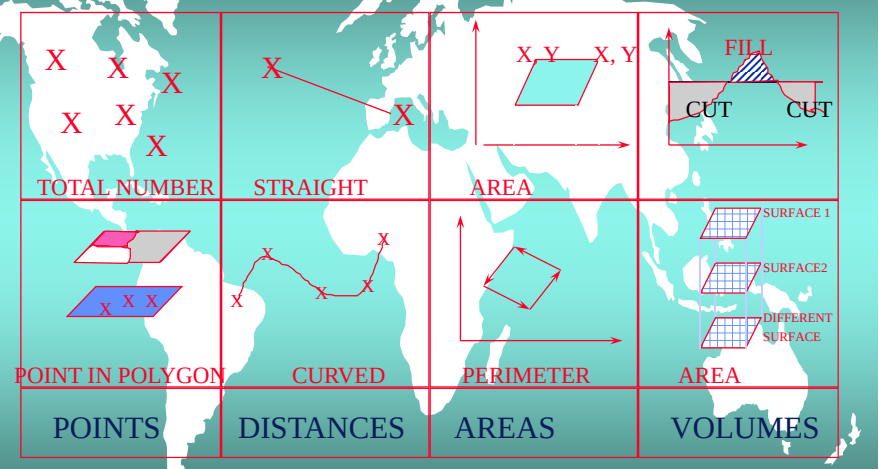
Map Abstraction



Measurement

- Measurement tasks are
  - Points : Inclusion of a point in polygon and enumeration of points inside polygon
  - Distance : Linear and Curvilinear
  - Area and Perimeter
  - Volume : Cutting and Filling

Measurements



Centroid Determination

- Centroid
  - Average location of a line or polygon
  - Center of mass of a two-or-three-dimensional object

- For Vector dataset:
  - Average the location of all the infinitesimal area elements within the polygon and finally determining the coordinate location of the area's centroid
- For Raster dataset:
  - Average the coordinates of all Raster elements that combine an implicitly defined POLYGON and finally providing centroid

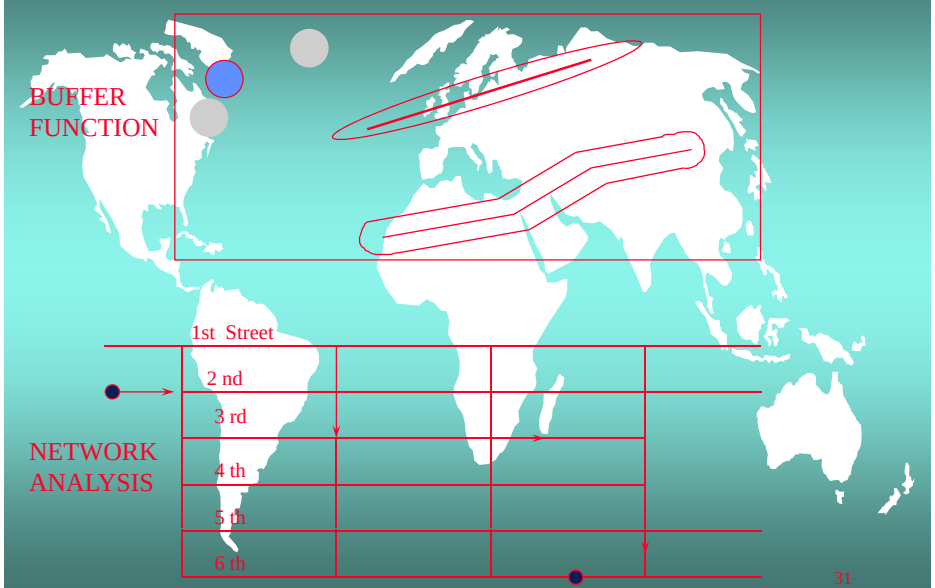
Data Structure Conversion

- Conversion from one format of data structure to another for :
  - Portability into different systems
  - Processing for external modeling and porting back to same system
- Generally done as a preprocessing
- It is must in a system

Connectivity Operations

- Network Analysis:
  - Optimum Corridor or Travel Selection
  - Hydrology and Discharge Estimation
  - A complex but useful function, found in some system, it is to be able to identify the separate watersheds in an area, through run-off direction calculations that are based on terrain descriptors

SPATIAL ANALYSIS



- Alternate Route for emergency vehicles
  - Combination of total length of the route and congestion on surface streets
  - Time of the day

- It is a complex problem in system analysis, and is not a part of general purpose GIS
- Separate software modules are required to solve these problems

Statistical Analysis

- Why ?
- Quality Assurance during preprocessing
- Summarizing a dataset as a data management report
- Deriving new data for analysis

- These have importance for information generation
- It forms a common feature in modern GIS

Essential Tools/Operations of Statistical Analysis

- Utilized for overall information flow in GIS
- The popular tools are:
  - Descriptive Statistics
  - Histogram or Frequency Count
  - Extreme Values
  - Correlation and Cross-Tabulation

Descriptive Statistics

- Mean, Median and Variance value in a data layer
- Higher order statistical moments such as the coefficient of skewness and Kurtosis are rarely used

Histogram or Frequency Counts

- Histogram displays the distribution of attribute value in a layer / region
- The calculation is straight forward in Raster Layer

- In Vector Database, it is carried out using the area of each polygon to appropriately weigh the attribute or base the histogram on a *per polygon* analysis
- Useful as data screening tools and can help us to formulate hypotheses during analysis

Extreme Values

- Locating maximum or minimum values in a specified area

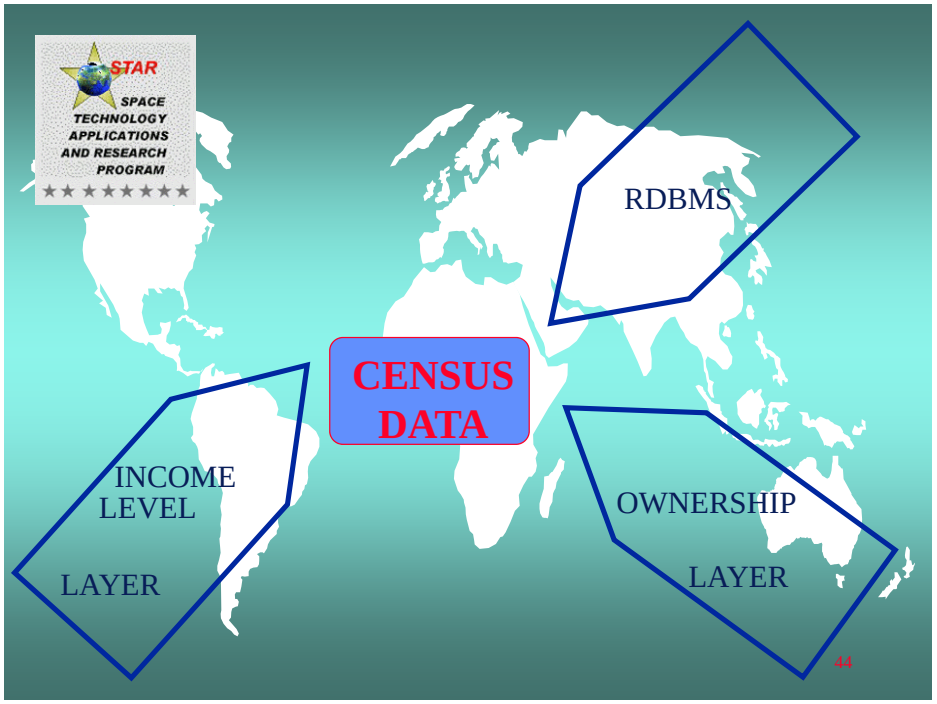
Correlation and Regression

- Comparison of spatial distribution of attributes in two or more data layer
  - Correlation Coefficient
  - Linear Regression Equation

- Cross-Tabulation is used to compare the attributes in two datalayers by determining the joint distribution of attribute
- When working simultaneously in both categorical and continuous variables, the appropriate statistical model is an analysis of variance (or covariance)

Average per capita income

	<$5,000	<$12,500	<$22,500	>$22,500
OWNER	154	354	673	982
RENTER	269	627	513	451



Specific Analysis

- If there is any relationship between the level of income and the probability of home ownership
- For this kind of analysis there are standard statistical tests that may be applied to determine whether the arrangement of data in the cells of the table might have arisen by chance

- The table is based on categorical data
  - Household Ownership : Nominal Variable
  - Per capita Income : Ordinal Variable
- In this table, one continuous ratio variable plus a nominal variable is termed into an integer-valued ratio variable

Frequently, we realize that statistical capability of a GIS is inadequate for an analysis problem

↓

In this case, intermediate output file for data to be transferred or analyzed is used in supporting powerful statistical analysis packages like SPSS, MINITAB, and BIOMED etc.

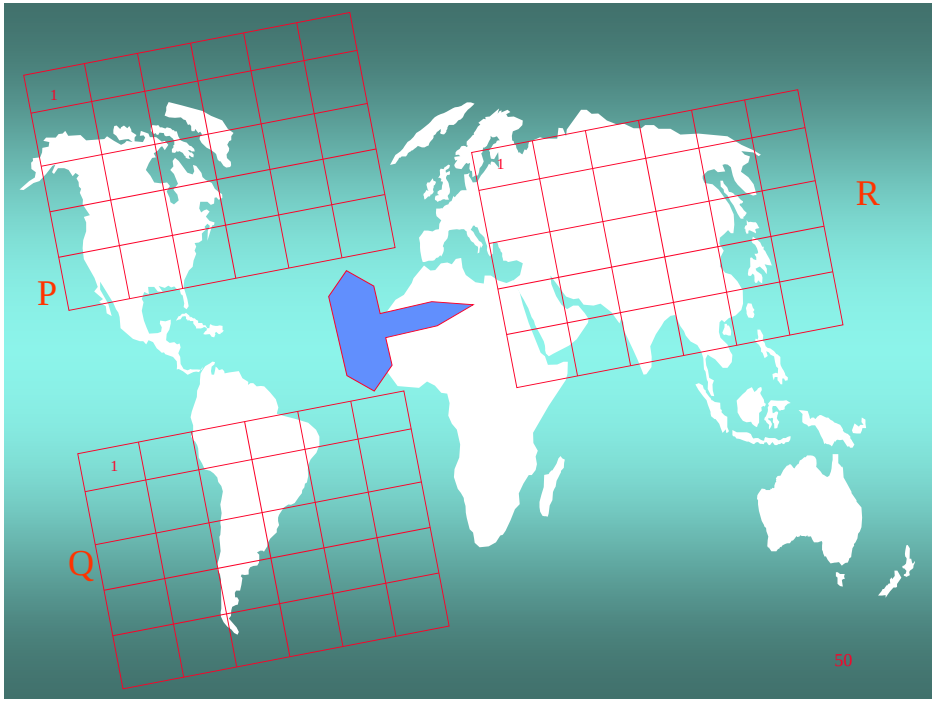
↓

New value Added or derived data/information may be incorporated again in GIS database for further analysis or presentation in map form

Raster Data Overlay

- Raster layers can be overlaid
- Raster overlay much more efficient than vector overlay
- There is cell-to-cell comparison or analysis in different layers
- Operational time increases with more cells

- The arithmetic operations on two thematic layers, P and Q, produce a new thematic layer R,
  - R = P + Q
  - R = P - Q
  - R = 2P -3Q
  - R = P/Q
  - R = P*Q
  - R = (P*P-Q*Q)**2
  - Logical Operations
    - if P > 30, R = 1; else R = 0
    - R = Max (or Min) of P or Q
    - And many more…..



Procedure for integrated data analysis

- State the problem
- Adapt the data for geometric operation
- Perform the geometric operation
- Adapt attribute for analysis
- Perform attribute analysis
- Evaluate Result
- Redefinition and new analysis, if needed

Question?

Thank you for your attention