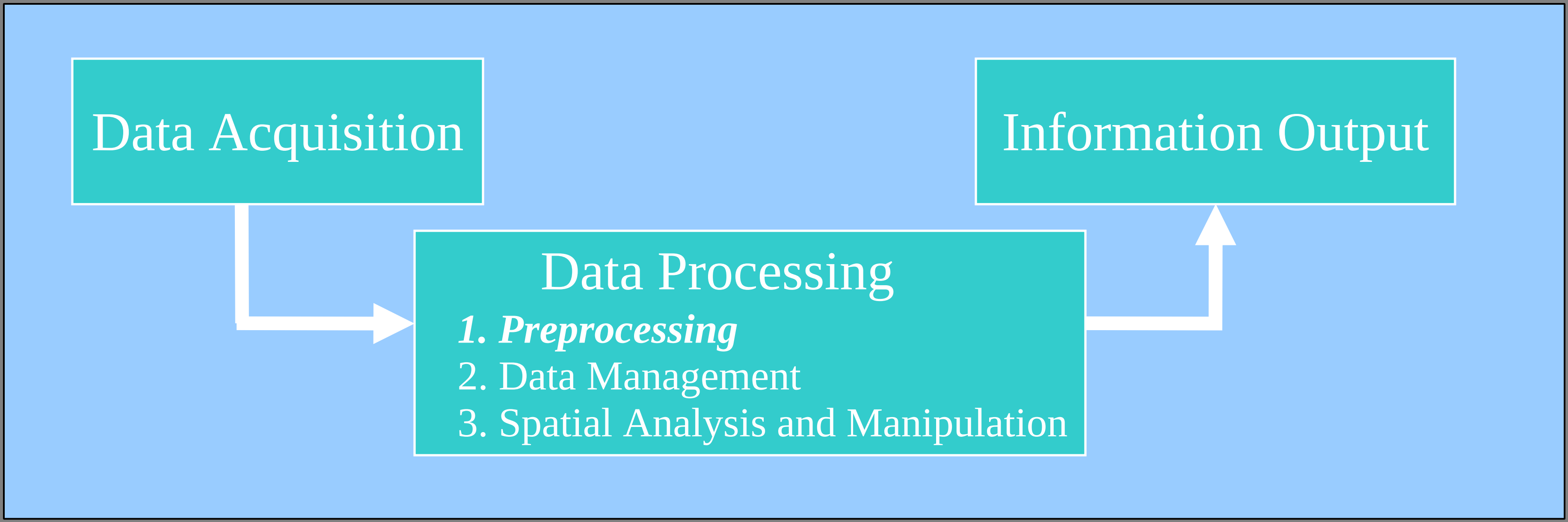
CSS 4200
Geographic Information Systems

Lecture 6:

– Data Pre-processing


Data Acquisition
Information Output
Data Processing
1. Preprocessing
2. Data Management
3. Spatial Analysis and Manipulation

Pre-processing Overview

- Preparing digital spatial data for GIS analysis
- General categories of pre-processing
  - format conversion
  - geometric transformation
  - generalization and aggregation
  - building topology
  - error detection and editing
  - edge matching
  - interpolation
  - data selection
  - data documentation
- Some of these activities also occur as part of other data processing and spatial analysis activities.

Format Conversion

- Vector to raster
- Raster to vector
- Transform formats
  - interchange with other GIS packages
  - to or from accepted or *de facto* standards (DRG, DLG, DIME, TIGER, DXF, etc.)
  - to or from more general standards (SDTS)
- Transform data types and file types within particular GIS software program

  (.e00, .shp, geodatabase)

Vector to Raster Conversion

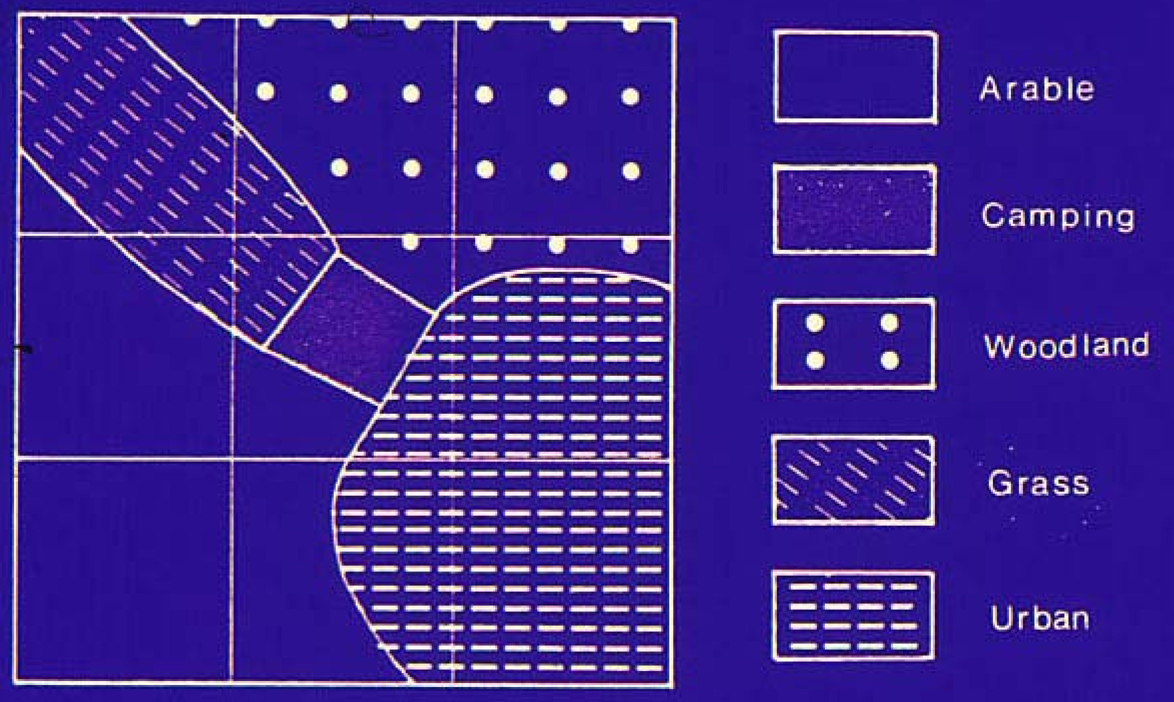


Fig. 6.3 Grid cells that are too large to resolve spatial details present coding problems. How should the central cell be coded? (from Burrough and McDonnell 1998)

Cell size (m) = [(msd, m^2)/(#grid cells)]$^{0.5}$

msd = minimum size delineation

Minimum Size Delineation

(also, Minimum Mapping Unit)

- Cartographic limit
- Generally, ~1 cm^2 (or ¼" x ¼") delineation
- For a given map scale (e.g., 1:24,000):

msd: 1/sf = md/cgd

1/24,000 = 0.01m/*x*

x = 240 m

surface area = $(240m)^2$

= 57,600 m^2 / 10,000m^2/ha

= 5.7 ha

therefore,

cell size ~ $[57{,}600\ m^2 / 30\ \text{cells*}]^{0.5}$

~ 44 m x 44 m

* Number of cells depends on how many grid cells (rasters) you want to represent the smallest delineation on map, or to comply with grid-cell resolution of other raster datasets in your geodatabase.

Geometric Transformation

- Transform the geometric characteristics of data layers
- Rectification vs. registration
  - rectification: refers to **<u>correcting</u>** geometric relationships
  - registration: refers to **<u>matching</u>** data layers
  - both follow a common conceptual process
- Usually performed in preprocessing and output

Rectification

- Rectification
  - processing to produce geometric relationships that are more correct (defining & removing distortions)
  - general rectification process
    1. select a general model of the geometric relationship between the data and the desired base
    2. calibrate model using homologous points or features (e.g. Ground Control Points, GCPs) to produce a specific model
    3. check specific model using additional (independent) homologous points or features (Check Points)
    4. if acceptable, apply specific model to the entire data set

Geometric Transformation

- Registration
  - processing to match positions between two data sets
  - often in concert with rectification but emphasize matching layers rather than removing distortions
  - registration also occurs when digitizing

- Transformation between coordinate systems
  - correctly done using at least two steps
  - must be judicious when performing coordinate system transformations

Registration

- Attempt to match the geometry of a data layer to another
- Making a location in one layer correspond the appropriate location in another layer
  - the other layer may be less correct geometrically but have preferable characteristics
  - registering a layer to a more correct standard will provide a degree of rectification
- Rectified data layers do not necessarily register (unless rectification also performed registration)

Models of Geometry

- Interpolative (approximate) Models
  - approximate the geometric relationship using relatively simple mathematical expressions that do not attempt to describe the geometric or physical properties of the relationship
  - often use general polynomials of some selected degree
  - many also use a "finite element" approach that interpolates between nearest control points
  - these models provide basis for procedures known as "**rubber-sheeting**"

Models of Geometry (con't)

- Geometric Models
  - attempt to accurately describe the geometric and physical properties of the relationship between the data and the base
  - sometimes termed exact, mathematical, or parametric models
  - might be theoretically superior, but
    - may not be able to come up with a model that is reasonable
    - model may be much more complex
    - in many situations, results are little better than when interpolative models are used

Models of Geometry (con't)

- Hybrid Models
  - include properties of both interpolative and geometric models
  - include simplified geometric properties but do not attempt to be rigorous in geometric or physical descriptions
  - accomplished by considering simple mathematical approaches to scaling, rotation, translation, etc.
  - these models remain simple, usually low order polynomials
  - example: four parameter (conformal) transformation

Models of Geometry

- General Idea:

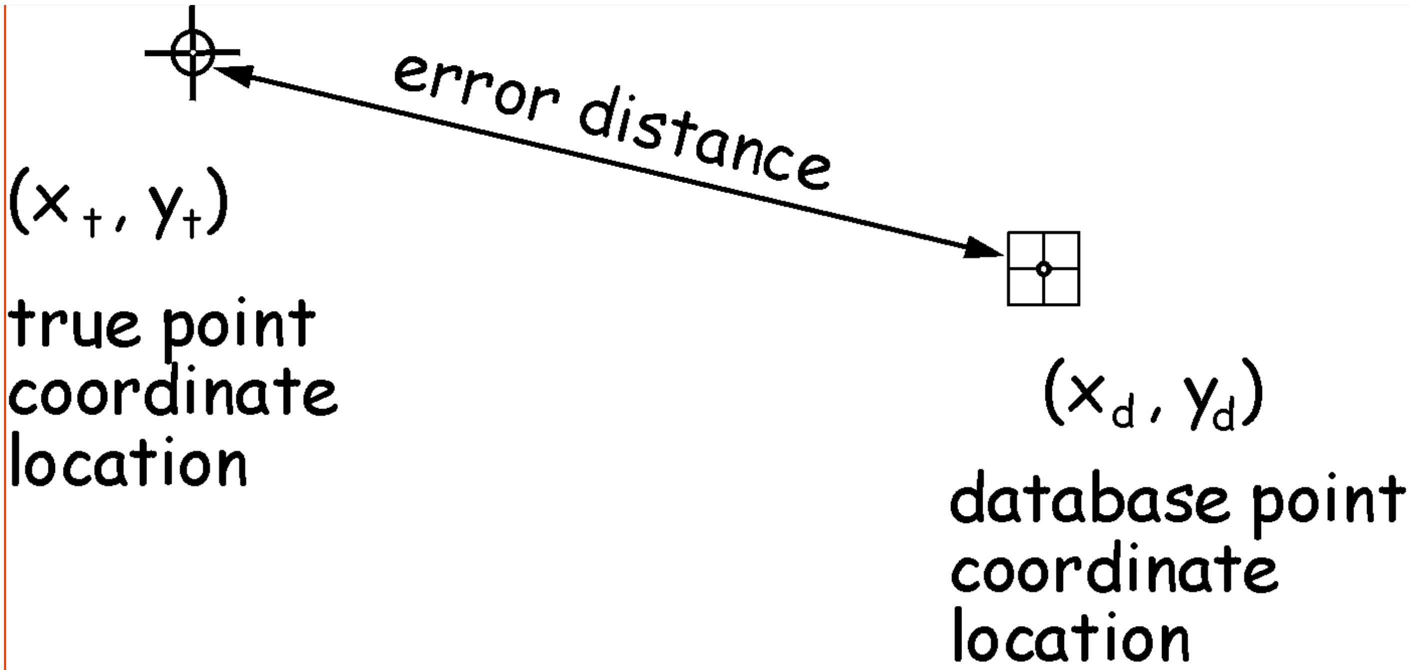
$$X' = f_X(X, Y)$$
$$Y' = f_Y(X, Y)$$

- Example of a General Model

$$X' = aX + bY + c$$

- Example of a Specific Model

$$X' = 5.3X - 28.3Y - 450{,}543.2$$



*For planar (Cartesian) grid coordinates only:

$$\text{error distance} = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2}$$

Root Mean Square Error ($RMSE_{x,y}$) = $[e_1^2 + e_2^2 \ldots + e_n^2 / n]^{0.5}$

Generalization and Aggregation

- Basic Ideas
  - often more data or detail than needed to adequately perform the GIS analysis
  - attempt to reduce the amount of data while maintaining the information needed by the projects
    - foolish to create and maintain a spatial database that is too complicated, detailed, and large
    - must be careful if the database is designed for long-term use
    - applies to attribute as well as graphic aspects of spatial data

Generalization and Aggregation

- Generalization (thinning, smoothing)
  - remove extra points or cells along linear features,

    OR
  - remove isolated features or excessive detail
- Aggregation
  - combine many detailed spatial objects into fewer, more general spatial objects
  - aggregation is also considered a method in Spatial Analysis and Manipulation
- Often need capabilities for merging features (line removal, “eliminate, “dissolve”)

Land Use and Land Cover Classification System For Use with Remote Sensor Data

Level I	Level II
1 Urban or built-up land	11 Residential 12 Commercial and services 13 Industrial 14 Transportation, communications, and utilities 15 Industrial and commercial complexes 16 Mixed urban or built-up land 17 Other urban or built-up land
2 Agricultural land	21 Cropland and pasture 22 Orchards, groves, vineyards, nurseries, and ornamental horticultural areas 23 Confined feeding operations 24 Other agricultural land
3 Rangeland	31 Herbaceous rangeland 32 Shrub and brush rangeland 33 Mixed rangeland
4 Forest land	41 Deciduous forest land 42 Evergreen forest land 43 Mixed forest land
5 Water	51 Streams and canals 52 Lakes 53 Reservoirs 54 Bays and estuaries
6 Wetland	61 Forest wetland 62 Nonforested wetland
7 Barren land	71 Dry salt flats 72 Beaches 73 Sandy areas other than beaches 74 Bare exposed rock 75 Strip mines, quarries, and gravel pits 76 Transitional areas 77 Mixed barren land
8 Tundra	81 Shrub and brush tundra 82 Herbaceous tundra 83 Bare ground tundra 84 Wet tundra 85 Mixed tundra
9 Perennial snow or ice	91 Perennial snowfields 92 Glaciers

Building Topology

- Process of relating spatial objects (e.g. points, lines, areas) to each other
- Encoding higher order spatial objects from more primitive objects (e.g. from points to lines, polygons)
  - Raster topology is systematic and (usually) simple
  - Vector topology involves filling in tables for each type of spatial object
- Object-oriented analysis continues this idea to higher levels
- Conclude with error detection and editing

Error Detection and Editing

- Functions that add, delete, or change values or positions of spatial objects
- Quality or utility of editing functions is important in judging the quality of a GIS software program
- Two standards:
  - initial data processing
    - verification (detection of problems)
    - editing (correcting or accepting problems)
  - updating

Edge Matching

- Goal is a “seamless” data base
  - features that cross the artificial boundaries of the tiles will often not match (for various reasons)
  - need ways to ensure cartographic features and labels (taxonomy) at edges of tiles of map areas match throughout the spatial database
  - For example,

    Detailed soil survey map unit boundaries match cartographically and taxonomically between survey areas (e.g., counties, counties and states, counties/states and countries…)

GIS Data Quality

- Accuracy and precision
- Timeliness
- Completeness
- Logical consistency
- Spatial variability
- Appropriateness
  - scale or resolution
  - classification (intent)
  - format