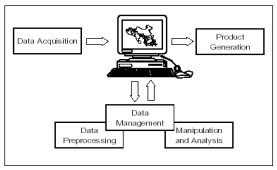
GIS basics, Data Creation, Cleaning and Editing

Rama Chandra Prasad
Lab for Spatial Informatics, IIIT Hyderabad

December, 9th 2015

GIS and Data Models

What is Geographic Information Systems?



(capture, display, manipulate, edit)

What are Digital Spatial Datasets, Data models

The contents of the **spatial database** is a model of the earth. The essential function of spatial data we store and manipulate is to subdivide the earth's surface into **entities or objects** that can be characterized.

How do we describe geographical features /entities?

Spatial data
Non-spatial data (attribute)

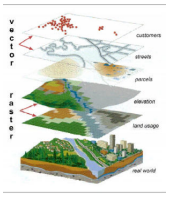
Spatial Data Types

- *continuous:* elevation, rainfall, ocean salinity
- *areas:*
  - *unbounded:* land use, market areas, soils, rock type
  - *bounded:* city/county/state boundaries, ownership parcels, zoning

  networks: roads, transmission lines, streams
- *points:*
  - *fixed:* wells, street lamps,
  - *moving:* cars, fish, deer

How do we represent these digitally in a GIS?

Data layer or data plane based on similar characteristics (e.g elevation, water lines, sewer lines)



What is the Mode of representation?

GIS DATA Models

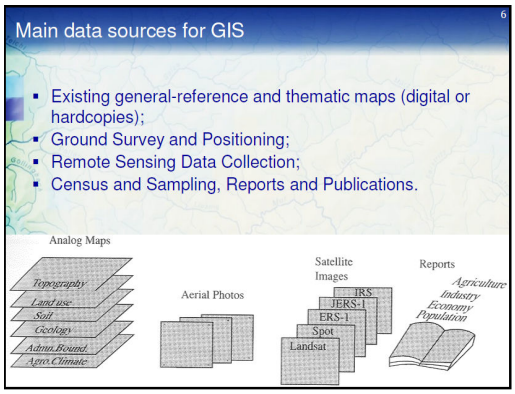
There are two fundamental approaches to the representation of the spatial component of geographic information:

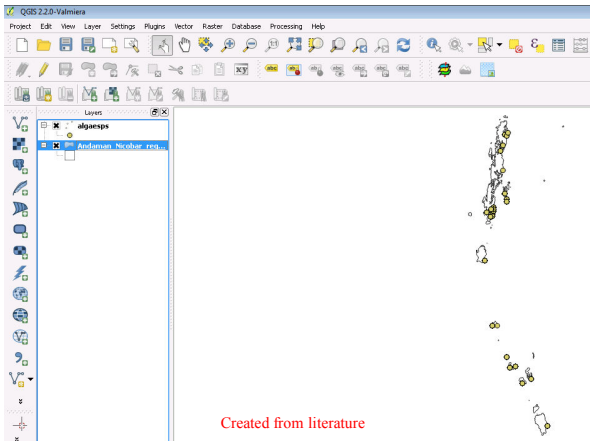
Vector Model
(features - Points, Lines, Polygons)

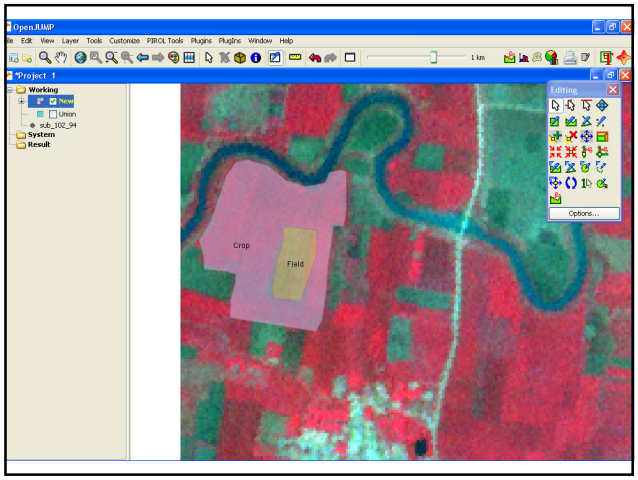
Raster Model
(Surfaces)

Data collection and Creation

- Either **digital** or **analog**.



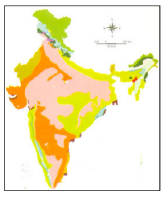




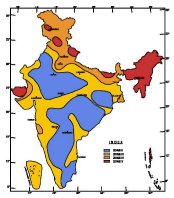
Raster data capture

Using scanner ⟶

Scanned maps and documents are used extensively in GIS as background maps and data stores.



Forest types of India

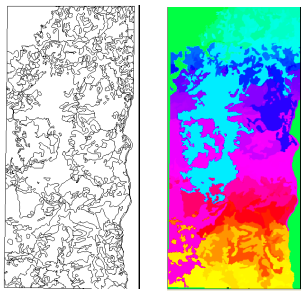


Seismic zones of India

Quality of outputs
quality of source data quality
scanning device and type of preparation

Rasterization

Rasterization refers to conversion from vector to raster data.



Forest type map of North Andaman

Error Detection and Data Cleaning

Data editing

1. Source of errors and Importance
2. Detecting and editing different types of errors
   1) Entity errors
   2) Attribute errors
3. Combining data of different sources
   1) Transformation
   2) Rubber-sheeting

Importance of data editing

- Data input involves diverse errors
- **Primary data capture**:
  - » positional accuracy of GPS receivers,
  - » device/instrument operator
  - » geometric distortion of air photo,
  - » malfunction of electronic scanners
- **Secondary data capture**:
  - » digitizing with no proper use of editing tool
  - » vectorization without post processing,
  - » georeferencing with imprecise control points
  - » data entry errors
  - » image interpretation error
- **Data transfer**: information loss caused by
  - » file conversion
  - » Resampling of data

Detecting errors in a dataset

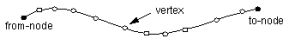
- *Data editing* is the process for detecting and eliminating errors inherent in data input, and avoiding error-prone analysis that may lead to wrongfully informed decision

  – **Entity errors**
    - Node errors: dangling nodes, pseudo nodes
    - Polygon errors: sliver, incorrect label point

  – **Attribute errors**
    - Incomplete values
    - Incorrect values

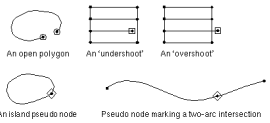
Node

- What is node?



  – A node is an endpoint of an arc. The from-node is the first vertex in the arc; the to-node is the last vertex. Arc-node relation defines connectivity

- Node types
  – Normal
  – Dangling
  – Pseudo



Node errors

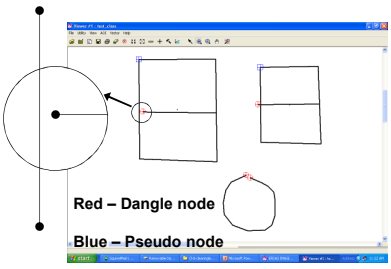
- Dangling and pseudo nodes often identify automation errors; however they can also be valid components of a feature

  – Dangling node, appear when polygon not closed
  – Pseudo node, appear where a single line connects itself

How to check node errors?

  – GIS provides functionalities for labeling potential node errors
  – But most of error detection requires visual inspection

Displaying potential node errors



How to fix node errors?

Interactive editing in a proper snapping environment
Automated data **cleaning (topology)**
Both Process - Reasonable tolerance level

Why to fix node errors?

- Because it ensures correct topology that forms the basis for further analysis
  - **It validates connectivity**
    - If the data is used for network operations such as hydrological model or routing
  - **It validates contiguity**
    - If the data is used for spatial overlay such as point-in-polygon, intersection between line and polygon

Polygon errors

- Most common polygon errors
  - Missing or multiple label points
  - Sliver polygon
- How to identify polygon errors?
  - Label error
  - Sliver polygon: select areas where its area is less than reasonably small value
- How to fix polygon errors?
  - Label errors can be checked
  - Sliver polygon: merge silver polygon to neighboring polygon (use eliminate command)

Attribute errors

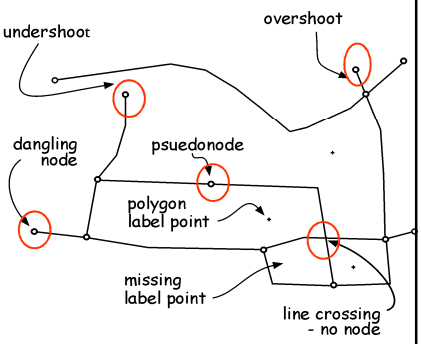
- Missing attribute values
  - Compare the unique list of values to master list
- Incorrect attribute values
  - Typing errors: spelling check

Errors: *detection and removal*

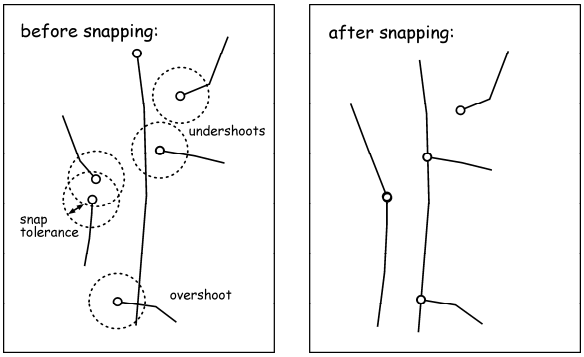
- GIS packages commonly use topological structure checking to detect errors
- Editing based on ***node snapping*** used to correct errors: snapping conducted based on *tolerances* -- snap if within 1 foot, for example
- ***Care must always be taken to assure that topological "cleaning" does not itself introduce errors (e.g. snapping nodes and lines together which shouldn't be)***

Manual Digitizing Process -

- Digitizing errors-
  - Positional errors are inevitable
  - Undershoots
  - Overshoots
- Node and line snapping-
  - Snap tolerance / snap distance



Manual Digitizing Process - cont.



- Snapping - defined by snap tolerance

Editing the data

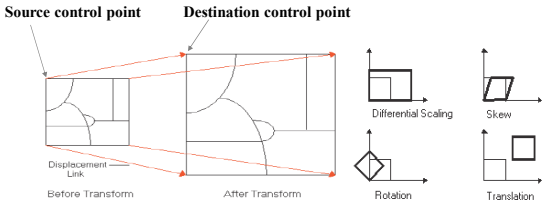
1. Add missing arc
2. Correct overshoot / under shoot
3. Fix an open polygon
4. Split & unsplit
5. Move & add vertex
6. Labels – Add, delete or change

Combining data of different sources

- When you work on data of different sources and they are not well aligned spatially even after projection change, it is necessary to make spatial adjustments of a layer to the layer with a higher accuracy
- Two kinds of spatial adjustments: **transformation** and **rubber-sheeting**
- Spatial adjustment allows you to improve data quality

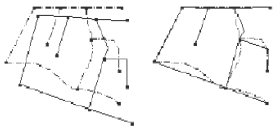
Transformation

Transformation converts data from one coordinate system to another; can be used to shift your data within a coordinate system (shift, rotate, enlargement and so on)

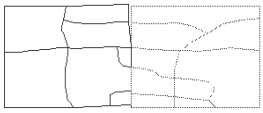


→ **Displacement links**

Rubber-sheeting



Edge-matching



Rubber-sheeting is usually used to correct for geometric distortions; they may be introduced by imperfect registration in map compilation, lack of geodetic control in source data, and so on.

- Rubber-sheeting adjusts source layer (dashed line) to target layer (solid line) based on displacement links
- Source layer is adjusted to target layer by setting snapping properties

Warping