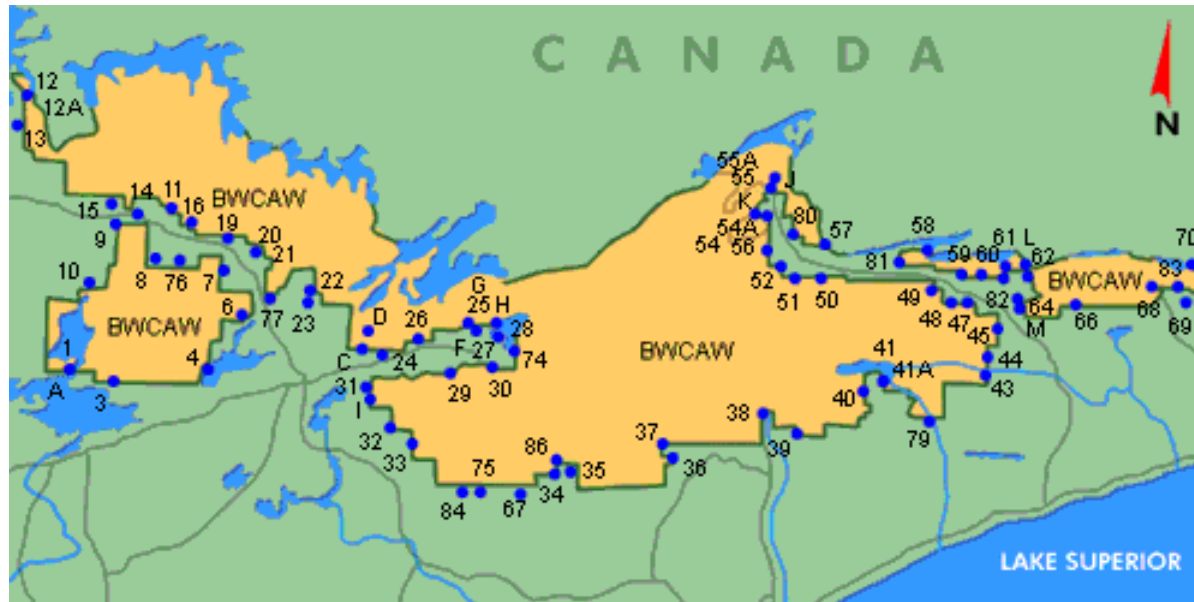


Data input and quality



Including
Georeferencing; data capture; data
sources and quality

What is georeferencing (or geocoding)?

- The process of assigning a geographic location to a geographic feature.
- This is beneficial because existing addresses can be automatically converted into a GIS database.
- The digital record for the feature must have a field which can be linked to a geographic base file with known geographic coordinates.

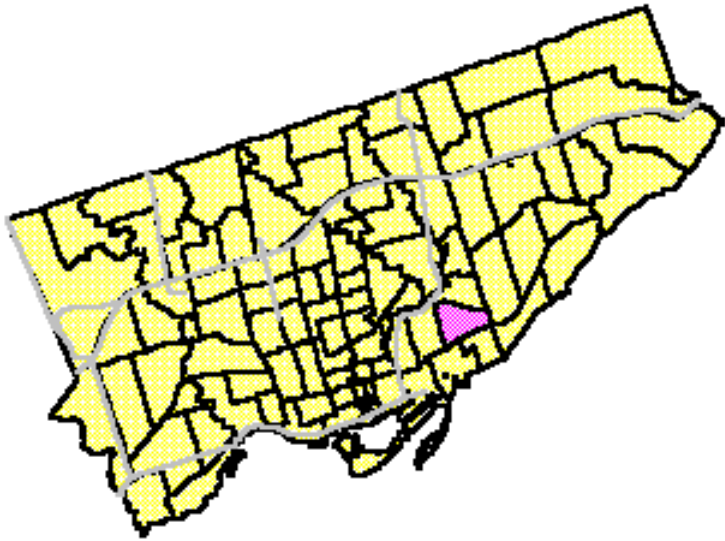
What a georeference should be.

- Geographic information requires a location.
- Any georeference should be unique (except in an OO system).
- The referencing system should be meaningful to users, and exclusive
- Georeferences should be persistent through time.

- Georeferencing is associated with a given scale.
- Georeferences are often tied to a national grid such as the National Grid of Great Britain.
- In Canada, a national grid is still being developed. See <http://www.gridcanada.ca/about.html>

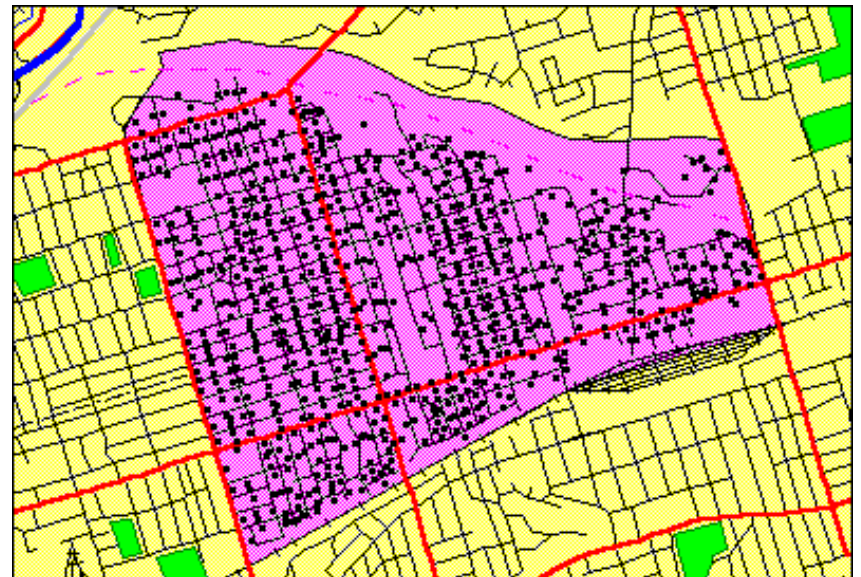
Georeferencing in Canada

- The most common form of georeferencing is based on the postal code.
- Postal Geography files allow you to analyze your data on the basis of known, established and daily functioning geographic areas..

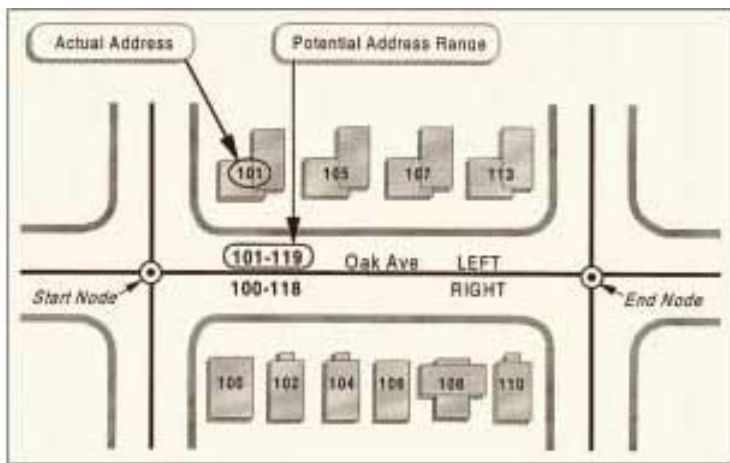


Forward Sortation Areas (FSAs). They are about 1500 in Canada. Each represents the first three digits of the postal code.

Each Postal Code in Canada is represented by points, following the referencing system of Canada Post, used to identify a geographic location. There are approximately 695 983 unique postal codes in Canada, represented by latitude and longitude coordinates. Each postal code denotes a small, defined section within an FSA



How does the computer georeference?



- How does the computer map in a GIS know where the data points should be put? It reads the x-y coordinates representing their locations.
- When locations are georeferenced, the address is represented by x-y coordinates, usually either in latitude and longitude, decimal degrees or in x-y coordinates identified by feet or meter measurements from a specific origin.
- The big headache in working with address data is that those data are often ambiguous and may be erroneously entered in field data entry situations.

Three main methods of georeferencing

1. Georeferencing by boundary. Eg. your theme table contains reference to a polygon boundary such as the province of BC, and this region can be matched to a map layer.
2. Georeferencing by postal code or zip code. Usually the theme table points to the postal code centroid (i.e. a point).
3. Georeferencing by address.

Automatic georeferencing by address

- In ArcView or other software that supports address matching, street addresses are compared against the existing street file database, and coordinates are assigned to the "hits."
- This process is sometimes called *batch matching*.

- Handling misses is done manually. The bad address is displayed with the closest possible matches the database includes. Users use these options to select the most likely match. This involves some guesswork and risks geocoding errors.
- Not all records in large data sets are likely to be successfully georeferenced.
- Surprisingly, there is no minimum standard for georeferencing. Maps can be produced and distributed based on a 25-percent *hit rate*.
- CHECK METADATA for percentage of omitted data.

Geocode Addresses

Reference Theme:

Join Field:

Using Address Style:

Address Table:

Address Field:

Zone Field:

Display Field:

Offset Distance: mi

Alias Table:

Geocoded Theme:

ArcView Help

File Edit Bookmark Options Help

Help Topics Back Glossary

Geocode Addresses (Dialog box)

This dialog box allows you to geocode a list of addresses stored in a table and add a geocoded theme to a view.


Dialog box options

Reference theme The theme in the view that is used as a reference theme for matching.

Join Field A field from the reference theme's attribute table that can be joined to the output geocoded theme. The selected field will be added to the geocoded theme's attribute table. It is useful to use a Join Field if you want to join the geocoded point with the feature in the reference theme that the point is geocoded against.

Using Address Style The Address Style that is applied to the reference theme. You can specify an address style by setting the [theme's geocoding properties](#) or pressing the Change Address Style button to change the address style.

Change Address Style Press this button to open the theme's geocoding properties dialog box for changing the address style.

Address Table The drop-down list shows all the tables in your project. Select the table that contains address events you wish to geocode. When you select a table, ArcView reads the field names in the table to find the likely defaults for the next two options. If you want to select a new address file, you can press  to open a file. This file will not be added as a table to the project.


Address Field The field in the table that contains the addresses, street intersections or place name aliases that you want to geocode. The Batch Match and Interactive Match buttons will be disabled if no field is selected.


Zone Field (Only available if the reference theme is made matchable using an address style that contains a zone field.) The field in the table that contains the zone information of the event. The Batch Match and Interactive Match buttons will be disabled if no field is selected.

Display Field An optional field in the table that will be shown in the [Geocoding Editor](#). Displaying an additional field of information such as the name of customer or business can help you identify the address you are geocoding particularly when you are running geocoding interactively.

Note If you want to select a field in the table but you can't see the field in the drop-down list, it is possible that the field has been [hidden](#) or [renamed with an alias](#). In either case, you can open the [Table Properties](#) dialog box to reveal the field and make it visible.

Offset Distance (Only available if the reference theme is made matchable using the address style that contains information of parity and side.) The offset distance from the street segment where a point representing the location of the event is placed. A positive value places the point in the specified distance on the matched side. A negative value places the point on the opposite side.

Aliases Table The table that contains [place name aliases](#) and addresses associated with the aliases. If you want to select a new alias file, you can press  to open a file. This file will not be added as a table to the project.

Geocoded Theme ArcView provides a default theme name for the output geocoded theme. You can press  to specify a new path and file name.

Geocoding Preferences Press this button to open the [Geocoding Preferences](#) dialog box for setting the preferences such as spelling sensitivity and minimum match score.

Batch Match Press this button to invoke the Batch Match geocoding process. When it is done, it will display the [Re-match Addresses dialog box](#).

Interactive Match Press this button to invoke the Interactive Match geocoding process. ArcView will display the Geocoding Editor for you to control the geocoding process. It allows you to review the matching candidates for each address and edit the address if no matching record is found.

A few notes on geo-coding in theme tables

- You need to know the *type* of the value.
- What is the legitimate *range* of values?
- How do you *flag* missing values?

More notes on geo-coding tables data

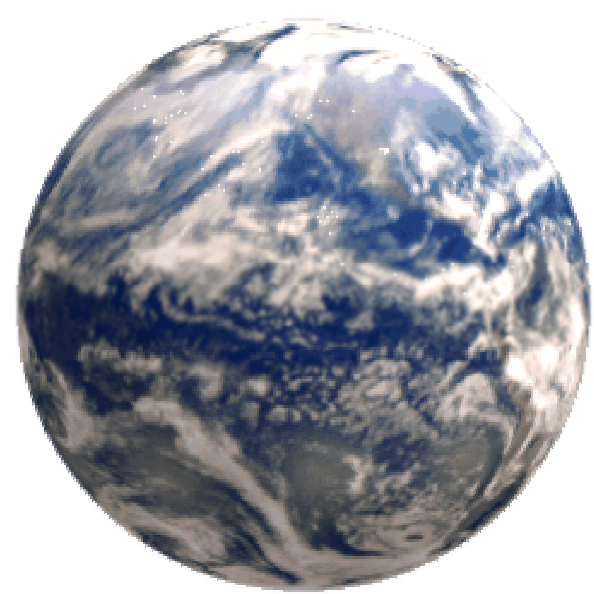
- Are duplicates allowed in the field?
- Which attribute is the *primary key*
Question: which database structure is the primary key relevant to??
- Most GIS use a data dictionary which checks values as they are entered, to ensure that they match certain criteria. You must be in a position, however, to establish those criteria.

Georeferencing absolute positions on the earth's surface

- So far, we have talked about georeferencing in relative terms that are useful for locating a location relative to administrative boundaries.
- Surveyors and physical geographers frequently want to georeference locations with respect to an absolute point on the *geoid*.

Geoid

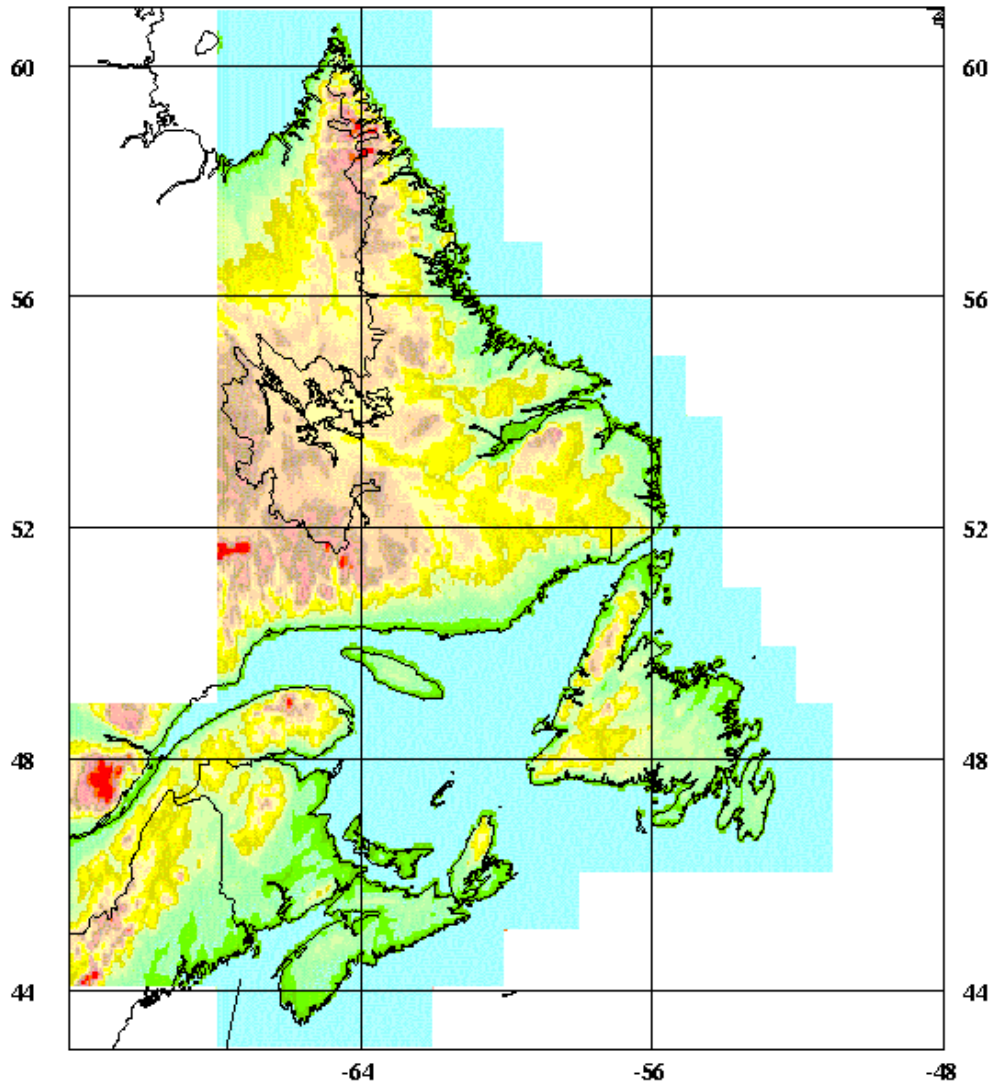
- Geoid refers to the shape of the earth if the oceans flowed under continents, and allowed a single sea level to emerge.
- The geoid rises over continents (where the crust is thicker) and lowers over oceans.
- The geoid also manifest various local bumps.



Keeping track of the geoid

- The geoid is measured in terms of vertical and horizontal position using *datums*.
- A datum tells you how the reference (map surface) is related to the earth. It uses vertical and horizontal measurements.
- Such *geodetic* measurements are usually collected and controlled by national governments.

- Vertical datum for Newfoundland and Labrador



DTED Maritimes / Newfoundland and Labrador

Data Collection/Input

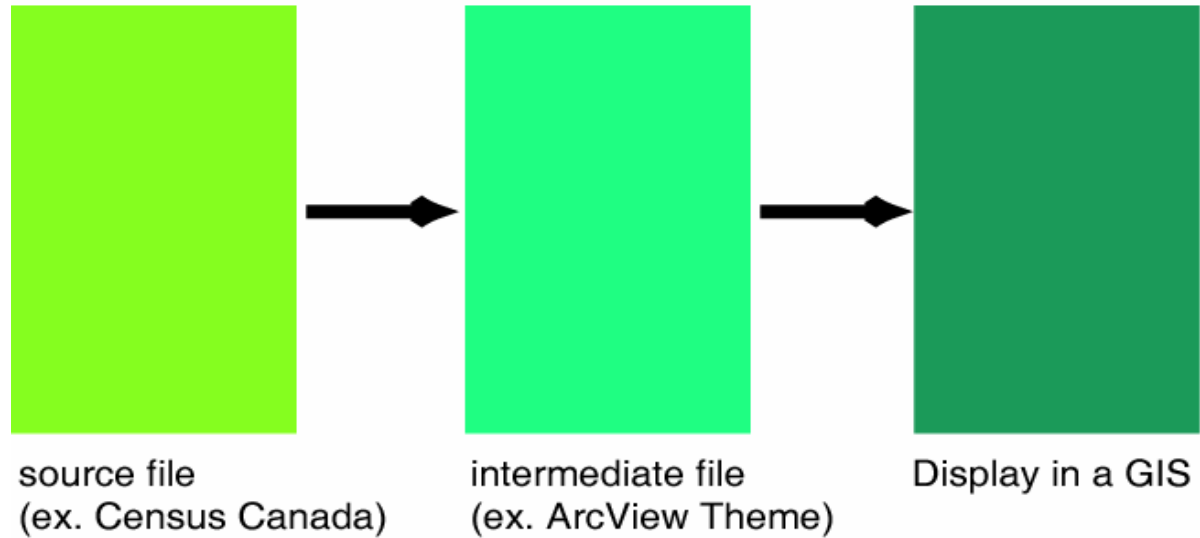
- There are many sources of GIS data, but because they all end up in some kind of data model, we tend to talk about raster or vector data (with object data usually geometrically described in vector format).
- Primary data is that captured using direct measurement specifically for use in GIS (ex. SPOT data; some GPS data)



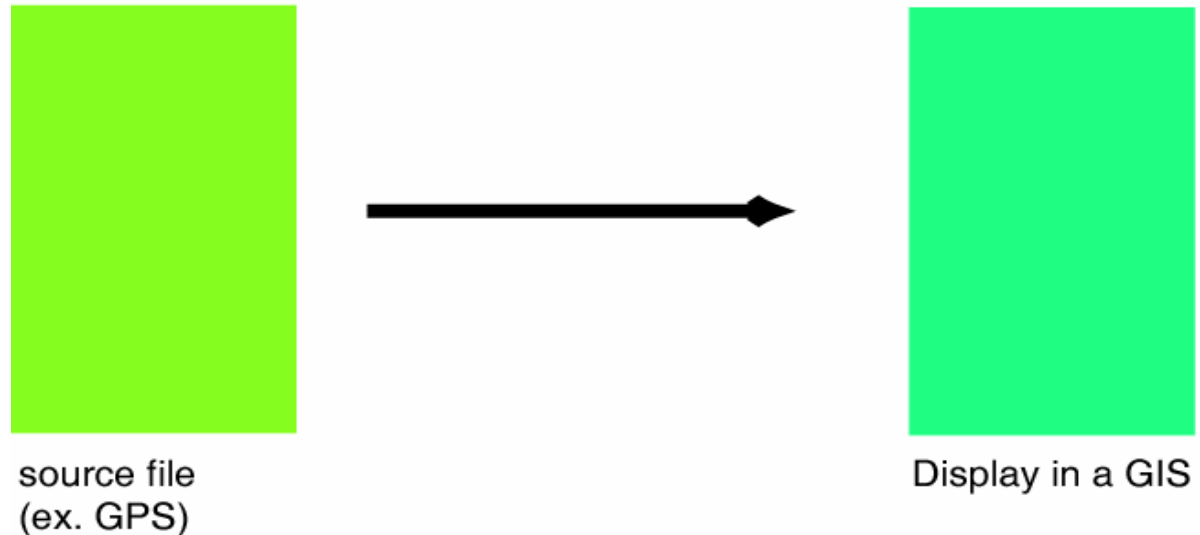
- Secondary data are collected for another purpose, and need to be converted for use in a GIS.
- Census data, scanned aerial photos, and digitized maps are secondary data sources.
- In the early days of GIS, data capture often accounted for 85%+ of the cost of a GIS project. Today, it is between 15% and 50%.

There is always compromise involved in using data-especially secondary data.

Translation OF DATA



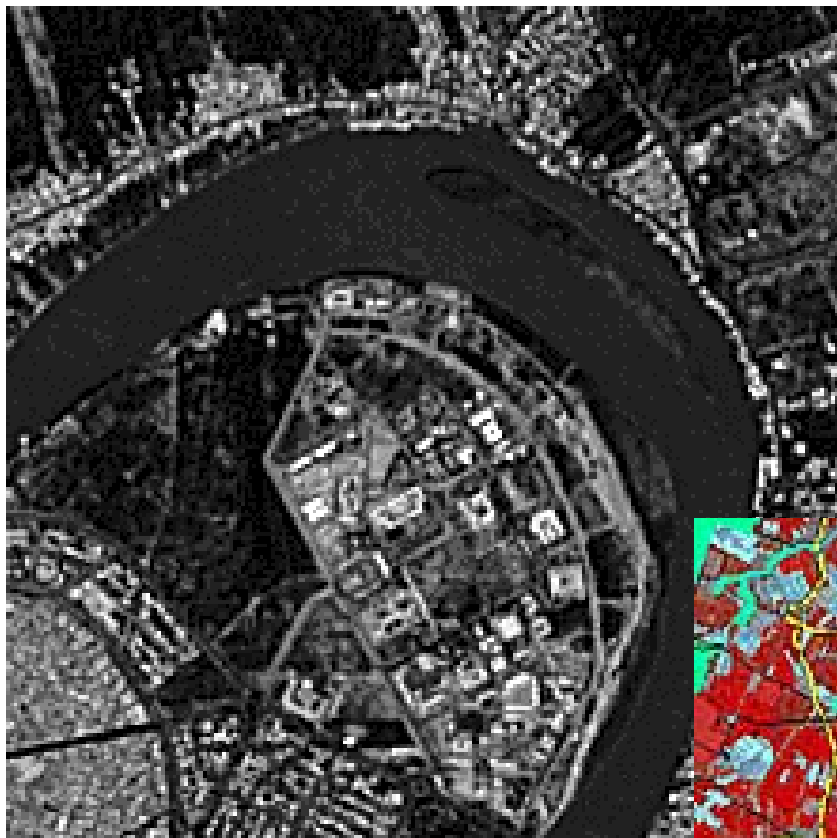
Direct input of DATA



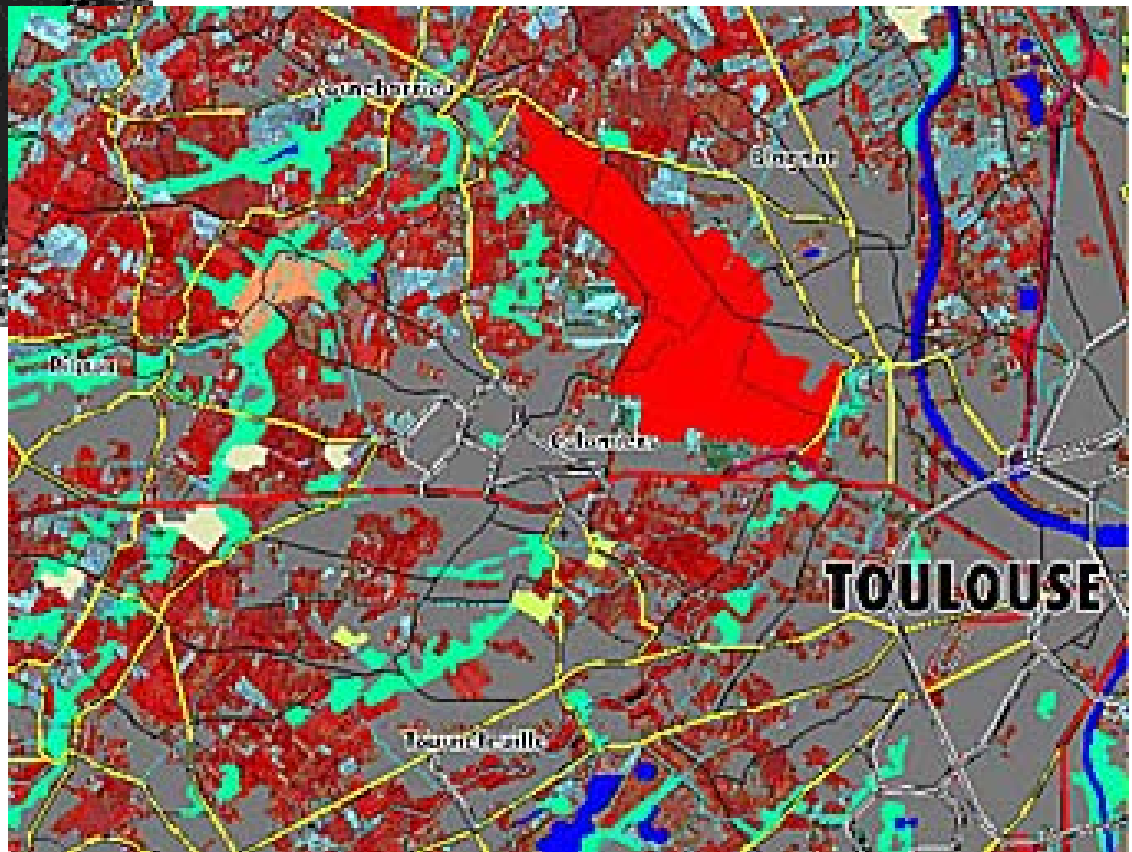


Primary Data Capture

- Most raster data now comes from remotely sensed imagery.
- RS imagery transmits information about the physical, chemical, and biological properties of the world without taking measurements.
- It is based on the translation of spectral signatures that are derived from sensors that take measurements throughout the visible spectrum from visible light to microwave frequency (2.4 ghz)



The image on the left is from SPOT (Systeme Probatoire D'Observation de la Terre)

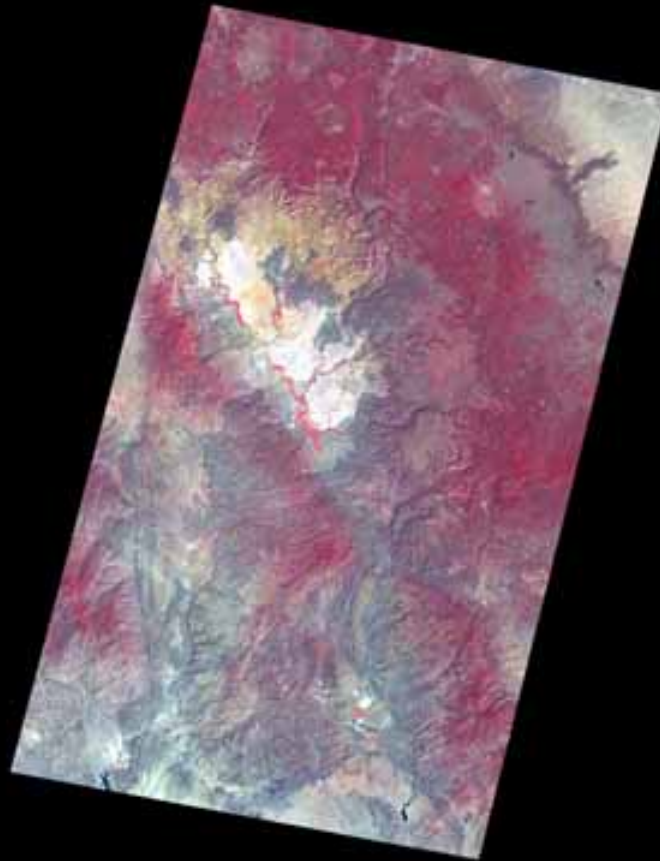


The image on the right is a raster GIS product produced from a SPOT image.

The most important thing about RS imagery is resolution

- For GIS users, resolution is the defining characteristic of imagery.
- There are three fundamental aspects of resolution: spatial, spectral and temporal.
- Spatial = size of smallest object that can be discerned. Measured in pixel size (eg. 1 m x 1m.)
- Spectral = the part of the e/m spectrum that is being measured.
- Temporal = frequency with which imagery is collected for the same area. This can vary from every 26 days (SPOT) to only once (airborne).

EPA & USGS NALC PATHFINDER WRS2 Path 037 Row 036
08/23/72



Low resolution over Prescott, AZ

Advantages to raster data capture

- Can use overlapping imagery to create stereo images with elevation.
- Consistency of data is useful for GIS.
- Can acquire temporal sequences.
- Great for small scale/large area.
- BUT, often commercially available imagery has low resolution (see above).
- Plus cost of data acquisition can be high.

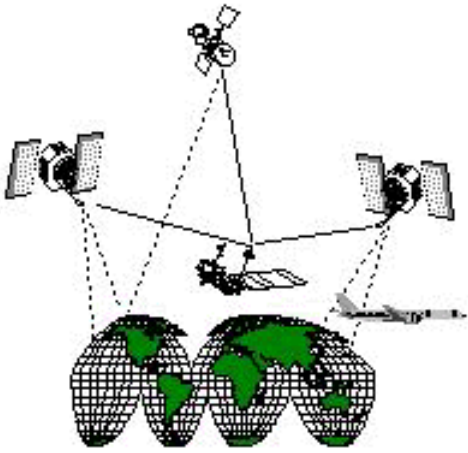
Vector Data Capture

- Two main sources of primary vector data are ground surveying and GPS. In the past, most vector data was digitized from existing NMA maps.
- Field surveying is based on the principle of triangulation.

- Field surveying have traditionally been based on measurements taken using transits and theodolites.
- More recently, electro optical devices have been used. These are called Total Stations as they measure both angles and distances to within a mm.
- Field surveying is expensive, but remains the most accurate mode of mapping location.

Trimble Total Station 5700 includes a GPS

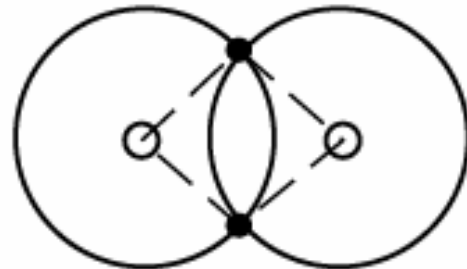
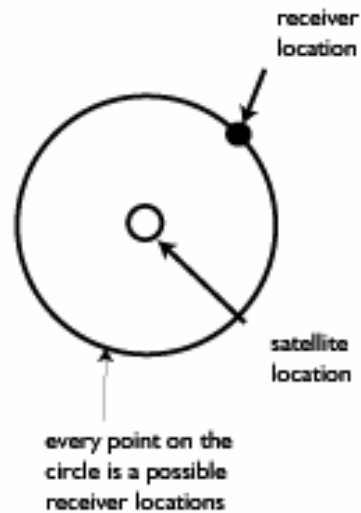




GPS

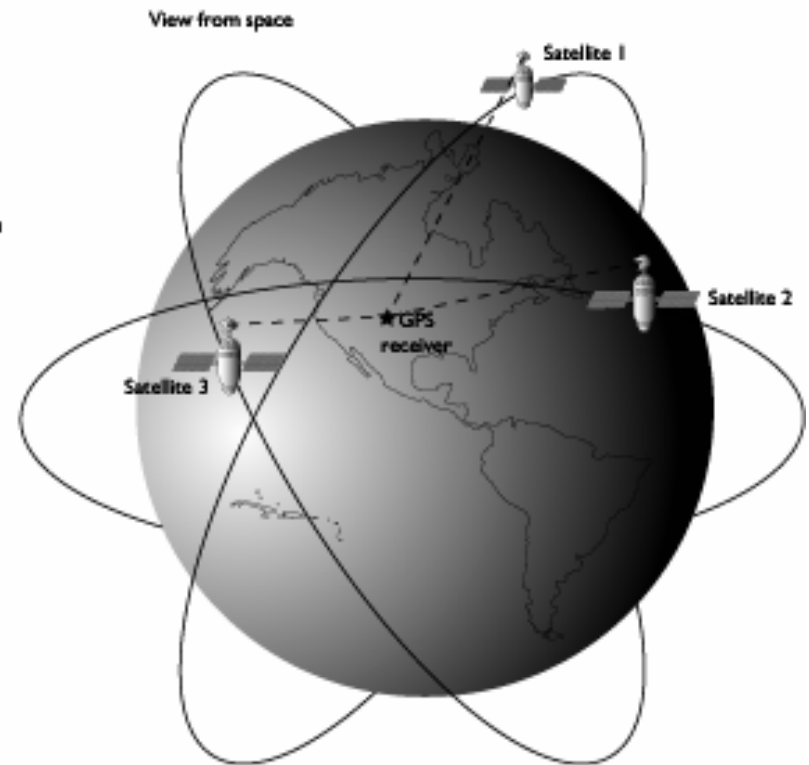
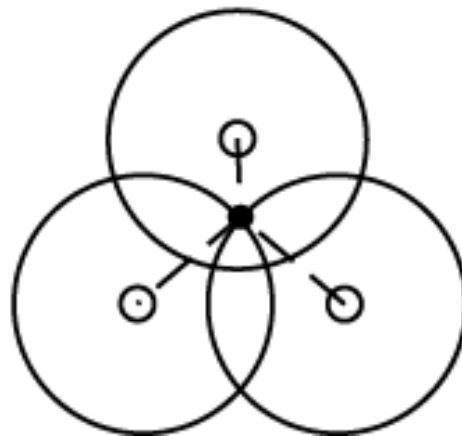
- GPS (global positioning system) uses 27 NAVSTAR satellites that orbit the earth at 12,500 miles!
- Developed by the US military.
- Selective availability (dithering of the signal) was removed in May, 2000.
- Most GPS record to within 10 meters.

1. A signal is sent out from the satellite to the GPS receiver which measures how long it took for the signal to reach it. It uses the length of travel time for the signal to calculate a circular range of possible locations.



2. Using the signal from a 2nd satellite, possible locations of the receiver on the ground are narrowed to the two points where the arc intersect.

3. When a third satellite locates the receiver, the actual location can be determined. Most GPS receiver give a location within 100 meters. Using additional satellites will increase locational accuracy.



Summary of principles of GPS

- Based on the length of time that a signal takes to travel the 12,500 miles from a satellite to a receiver on the ground.
- Uses a coded radio signal that includes exact position of the satellite in time and space.
- By measuring the distance of three+ satellites, the location of the receiver is located through triangulation.

Sources of error in GPS

1. Signal degradation as a result of atmospheric conditions.
2. Minor variations in the location of satellites.
3. Inaccuracy in the timing clocks.
4. Receiver error.
5. Variation in the reflection of signals from different objects on the ground.

- Differential GPS is the most accurate form of GPS, compensating for human and radio error by using two GPS: one roving, and another, stationary reference unit to monitor timing errors.
- If the fixed receiver has a known position, deviations from the measurement of the roving receiver (other than those based on distance from the differential receiver) are compensated for (i.e. eliminated).
- In many Canadian cities, especially on the coastlines, government agencies broadcast a differential signal that can be used to correct GPS signals..

An end to US domination of GPS?

- Up until the present, the USA has run the 24 Navstar satellites that provides us with global GPS coverage.
- The EU has planned a fleet of 30 satellites dedicated to the broadcast of positioning data.
- Galileo will supplement and improve the existing GPS satellite coverage.
- The EU will pay for Galileo, but everyone will benefit.

Galileo's services

- Most GPS services offered by Galileo will be free, but consumers can elect to pay for real-time monitoring of the system's accuracy.
- This will buy encrypted "integrity" messages" that warn of glitches and errors in position (used for precision navigation). This service will be free for search and rescue and operations where lives are at stake.

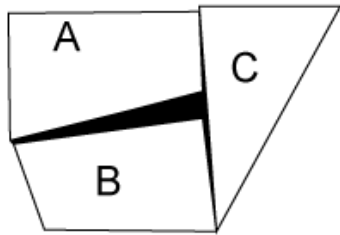
Secondary data capture

- Digitizing was historically a major form of data input but has almost been abandoned today.
- Digital surveys and already existing maps available on the www are of a much better quality.
- Very few GIS projects today start with no data, mostly because most maps of the world have already been digitized.
- Digitizing – converts physical representations, from traditional maps and diagrams, into digital data.

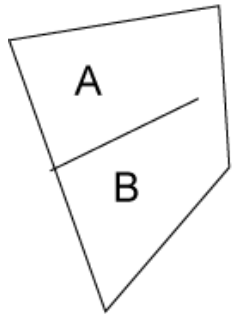
- Traditionally, digitizing consumed huge amounts of resources . It was (i) time consuming; (ii) reinforced error and introduced new error; (iii) was of low quality by survey standards.
- Digitizing must be geo-referenced from the local coordinates of the digitizing apparatus to a permanent, know location on the earth's surface, in a reference system.

- Once digitized, data must be checked for error including omissions (the polygon you forgot to digitize) and inaccuracies. There are a number of errors associated with digitizing and they include
 - gaps between lines, double digitizing of arcs and spikes
 - data in the wrong place; data at the wrong scale; distortion of spatial data related to distortion of the base image (from paper stress or angular rotation—remotely sensed imagery)

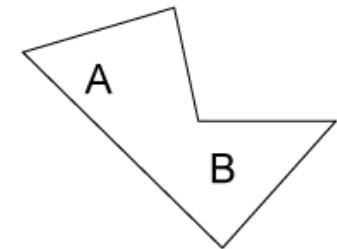
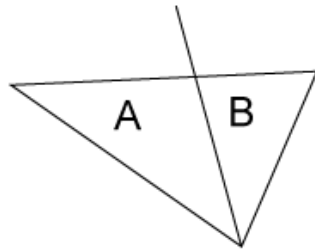
Digitizing errors (1) □



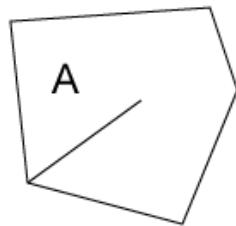
Sliver error, caused by leaving spaces between polygons.



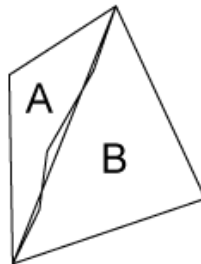
Line closing errors. Some systems avoid this by setting a tolerance within which a line is designated to close. But tolerances don't remedy all instances.



Missing segment

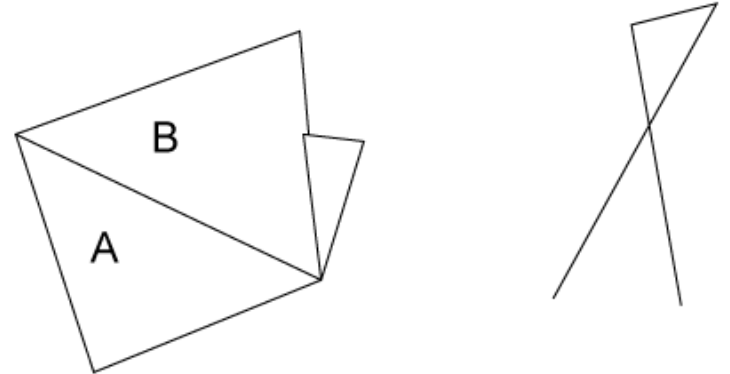


Poorly labelled segment

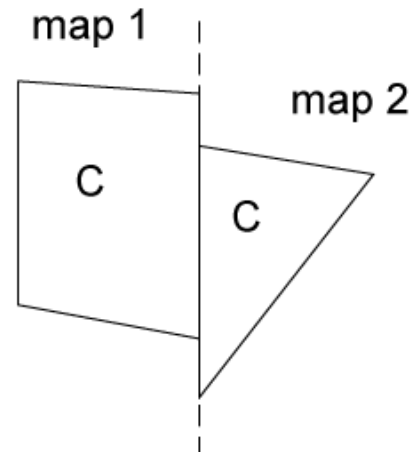


Twice-digitized line

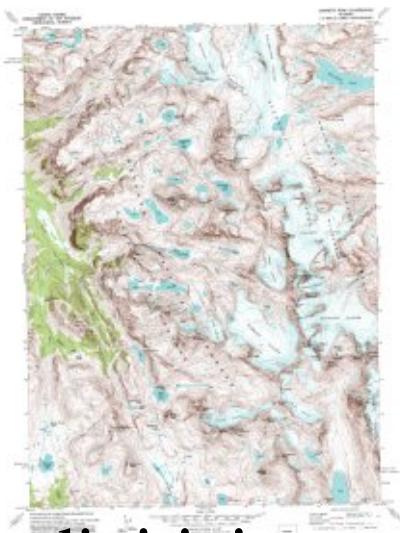
Digitizing errors (2) □



Errors caused by careless or poor digitizing.



Lack of alignment between polygons on adjacent map sheets.



Scanning

- Both digitizing and scanning are techniques that refer to an analogue original that invariably already contains errors.
- Scanning is, however, a means to vectorize raster data.

Downloading digital data from the net

- Main advantage is ease of acquisition.
- Problems include: resolving compatibility between multiple data sets; assuring data quality; relevance.

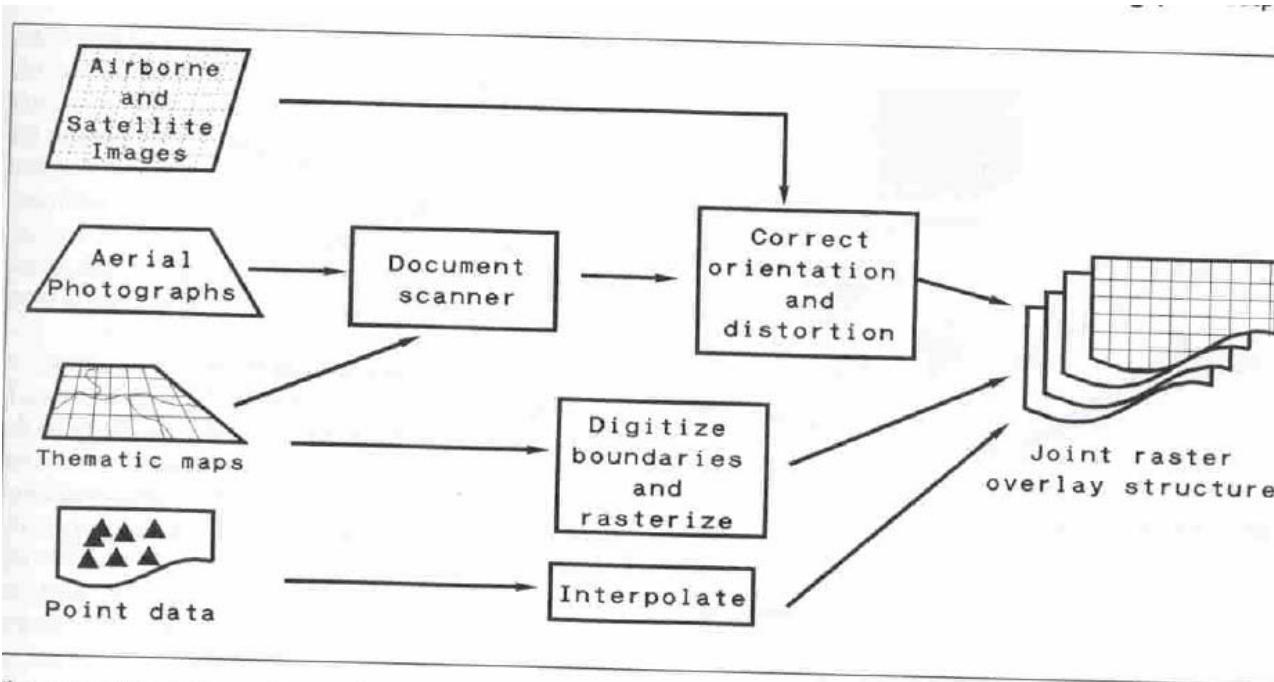


Figure 4.3. The capture and processing of spatial data to build a raster database

