Working with and digitizing geospatial data

Radost Holler[a] Paul Schäfer[a]

[a] *Bonn Graduate School of Economics*

Hackathon, Open Source Economics

October 14, 2019

Outline

What is GIS?

What is GIS?

Geographic Information Systems (GIS) is a computer-based tool that analyzes, stores, manipulates and visualizes geographic information, usually in a map.

... i.e. if you use any computer program to work with spatial data, you use GIS.

What programs?

Full blown GIS software:

- ArcGIS (costly, but can be obtained freely as student of Uni Bonn)
- QGIS (open source)

Both programs have a python plugin, so they can, **in general**, be combined with python.

What programs?

Full blown GIS software:

- ArcGIS (costly, but can be obtained freely as student of Uni Bonn)
- QGIS (open source)

Both programs have a python plugin, so they can, **in general**, be combined with python.

Of course you can use python, R and Stata as well...

- Python: There is a “pandas” derivative for geo-data: **geopandas**: http://geopandas.org/
- R: the sp package enables loading spatial data

Data Structure

Data Structure

- vector data
  - geometry: points, polylines, polygons
  - attribute table
  - shapefile, spatial lite
- raster data
  - cells or pixels
- coordinate reference systems

Geometry

The geometry of vector data consists of sets of coordinate pairs (x, y)

The exact structure depends on the type of the geometry:

- Points: (x,y)-coordinates
- Polyines: ordered sets of coordinates (nodes)
- Polygon: a set of closed polylines – last coordinate pair coincides with the first pair

Coordinate Reference System

Coordinates are only meaningful given a coordinate reference system...

- Where on the world is the origin?
- What is the unit of measurement (latitude/longitude, meters, kms, feet)?

Angular Coordinates

- angles: latitude, longitude
- visualization: https://www.rspatial.org/raster/spatial/6-crs.html#angular-coordinates
- need to estimate the angles by using a model of the world, models are called **datums**
- most common datum: WGS84

Projections

- Another way of defining coordinates is by first projecting the earth on a two dimensional planar.
- This is always necessary to display the data.
- there multiple ways of projecting, which one to use depends on where on the world you are, and what you would like to highlight.

Projections

- Another way of defining coordinates is by first projecting the earth on a two dimensional planar.
- This is always necessary to display the data.
- there multiple ways of projecting, which one to use depends on where on the world you are, and what you would like to highlight.

i.e. information contained in geospatial data: coordinates and a reference datum and most of the times a projection

Using ArcGis’ python package

Using ArcGis' python package

- ArcGis contains a built in python package called "arcpy"
- this is not open source !!!
- and is only compatible with python 2.7
- the documentation is horrible, but it is sometimes better then the pointing and the clicking

Digitizing Maps

Process of Digitizing

The process of digitizing geospatial data, i.e. maps:

1. Scan your map (more difficult than you'd think)
2. Georeference your map
3. Vectorize the data from your map.

Georeferencing

The art of mapping pixel coordinates to geographic coordinates.

What do I need for georeferencing?

- a good non-skewed scan of the map
- a good guess of the projection of my scanned map
- a digital reference map

Digitizing Choropleth Maps

Overview

1. Preparations
2. Map Colors to Data Values (Machine Learning)
3. Aggregate to Polygon Level
4. Manual Post Processing

Software Choice

- ArcGis:
  - Proprietary
  - Black Box algorithms
  - Expensive
  - Many Bugs, Very Annoying, Terrible Interface
- However:
  - Gets sufficiently good results with limited technical knowledge
  - Problems are not generalisable enough to avoid doing manual steps
  - Best / Second Best available software to do necessary manual steps
  - Alternatives: QGIS, Define Training Sample in ArcGis and Estimate Model in Python, Google Earth Engine

Preparations: The Raw Data

- Raw data often comes as pieces of a scanned map →show
- We need to put the pieces together and link to geographic coordinates
- Geo-reference first then us the mosaic tool to join the pieces
- https://pro.arcgis.com/de/pro-app/tool-reference/data-management/mosaic.htm

Machine Learning: Training Data

1. Enable the spatial analyst extension and the image classification toolbar (Customize →Toolbars / Extensions).
2. Select the Draw Polygon tool from the image classification toolbar and draw a polygon around all areas which belong to one class
3. Open the training sample manager. Mark all of the rows you just created (not ones you created earlier) and click on the merge training sample button. Give the class an expressive name.
4. Repeat for all classes and specific non-data elements
5. Save the training data-set, by clicking on the save icon in the training data manager.
6. Create a .csv file with mappings from class names to integers

Machine Learning: Estimation

1. Click on the create signature file in the training sample manager and save the signature file under an expressive name.
2. Go to the classification toolbar and click on classification, maximum likelihood classification. Leave all options at the standard choice and enter the file names.
3. Run and Pray!

Aggregation

1. Use the set Null tool to set the classes which are not data to the nodata value.
2. Use the zonal statistics as table tool to calculate the mode (majority) of your classified values for each admin area.
3. Right click on the attribute table of your shape file, select join and join it to the newly created attribute table.
4. Export the new shapefile (http://desktop.arcgis.com/en/arcmap/10.3/manage-data/feature-classes/creating-feature-classes-creating-a-new-feature-cl.htm), be care full to select shapefile in the pull down menu of the save dialog.