

data

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Thursday 12th September, 2024 10:40

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

what are data?

- u/a: unit of analysis: what do you study?
- u/a=# of obs=# of rows=sample size
- dataset has variables, which are the *attributes* of u/as
- say students: age; counties: water area
- cols=vars, rows=obs; vars characteristics of obs
- if several layers: may have several u/as
- eg counties: #18; hospitals:#700; ex of attr?
- dataset is a matrix/spreadsheet/2D object

storage type: num v str

- strings are safer; eg string “0821” made into a number results in “821”, which is a mistake !
 - but then often need to make str into num to do the math/map
- be careful about it, triple check, there are often problems and it's non-intuitive

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

files

- .shp (along with bunch of others)
- .kml
- and other
- we'll cover them on “as is” basis
- if you bump into something weird, email listserv

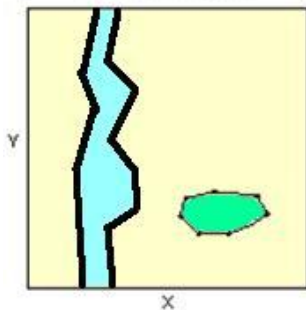
raster (picture) v vector (point, line, or polygon)

- raster (has resolution)
 - area covered by cells/pixels
 - each cell/pixel have values/colors
- vector (no resolution): all real world features:
 - points (dots/nodes): airports, cities, trees
 - lines (arcs): rivers, roads
 - polygons (areas): counties, cities

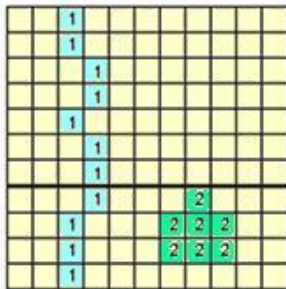
raster and vector



VECTORIAL



RASTER



gis data as layers of shapes with regular data

- data are organized by *layers*, eg roads, admin boundaries, etc; show example/draw a picture
- each layer: loc info (shapes)+often some regular data
 - ie data table with loc info (shapes) must underlie a map
 - (the data table often has some regular data, too)
 - thematic maps use different symbols/colors (themes) to show variation in regular data

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

some real skills

- this is where the real value come from:
 - to bring different vars together to produce new insight
- if you just map vars from same or similar data:
 - it has probably already been done!
 - just goog: “what you study, map” and see images
- but combining creatively variety of vars:
 - there is no such map in the world!
- eg https://scholarship.libraries.rutgers.edu/view/delivery/01RUT_INST/12643382240004646/13643522850004646

howto map regular (eg xls) data?

- it would likely have geo id:
 - hospital name/code, county name/id, etc
 - codes/ids are great: unique! (as opposed to names)
 - then google a shapefile that you can join with your data
- google “geo in you data, shapefile”
 - eg “NJ counties, shapefile”
- and then join the two to produce a map

“the join problems”: examples

- “Camden county” \neq “Camden”
- “Congo” \neq “Congo, Republic of”
- “Great Britain” \neq “United Kingdom”
- “Camden” \neq “CAMDEN”
- “Camden ” \neq “Camden” (space is a character !)
- “08012” \neq “8012”
- be very careful; check the tables to see if it merged right
- does it make sense? eg Camden richer than Cherry Hill?

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

don't trust anybody! neither yourself

- remember, always be critical
- triangulate your results: compare with other source
 - <https://researchmethod.net/triangulation/>
 - <https://conjointly.com/kb/measurement-error/>
 - just goog picture, eg 'nj counties property values map'
 - looks about right
 - (other definition of the prices, but correlation is important)
- show to others, ask for comments

triple check

- merging (joining) data is tedious and tricky
- be careful, double, triple check
- easy to make mistake

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

join/merge

Notebook: join/merge

DATA SOURCES

ex from past

data management takes time! value your time!

- producing maps fast; data management 50-95% of time
 - figuring out, understanding, cleaning, documenting, combining, etc
- so we start with data management like join/merge
- spend it on data you care about and will use in your career!
- think hard about data you'll use in your career
- otherwise you'll waste 100+ hours !!!

data ideas

- <https://www.dvrpc.org/data/>
- camden county <https://camdencountynj-ccdpc.opendata.arcgis.com/search?collection=Dataset> eg camden zoning :)
- NJ <https://gisdata-njdep.opendata.arcgis.com>
- Philly <https://www.opendataphilly.org>
- a lot!: <http://geocommons.com/search.html>
- just search for what you are interested in, say 'road'
- <https://www.policymap.com/maps>
- \$ to downld data, but click 'Source' and download by hand
- open gov, especially city data, just few examples
- <https://data.cityofchicago.org/> , <http://opencityapps.org/> ,
<http://www.opendataphilly.org/>

data ideas



- NJ parcels

- `https://njgin.nj.gov/njgin/edata/parcels/#!/`

- `https://www.njmap2.com/parcels/parcels/`



- `https://www.arcgis.com/apps/webappviewer/index.html?id`

outline

regular (not gis) data: xls, csv, etc

gis data (has shapes, can make a map from it): shp, kml, etc

Notebook: dive into thematic/choloropleth maps

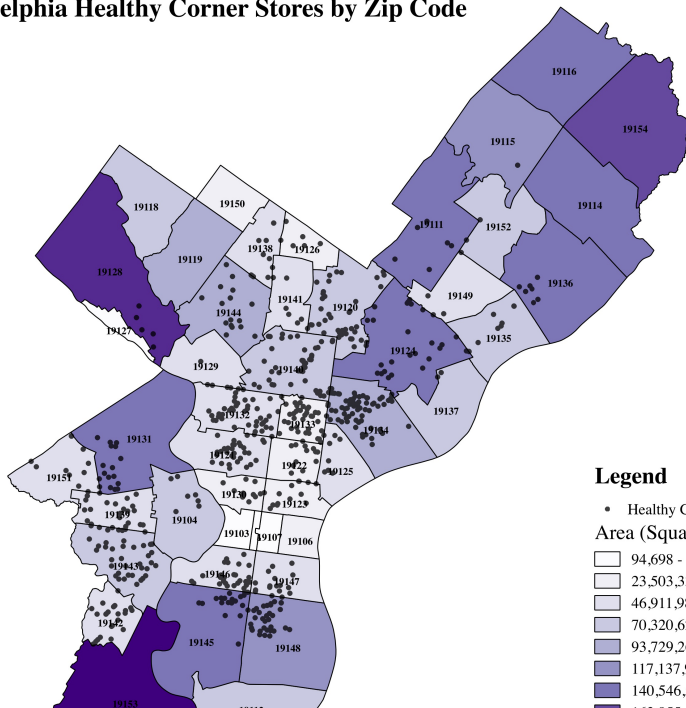
join/merge

Notebook: join/merge

DATA SOURCES

ex from past

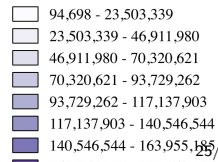
Philadelphia Healthy Corner Stores by Zip Code



Legend

- Healthy Corner Stores

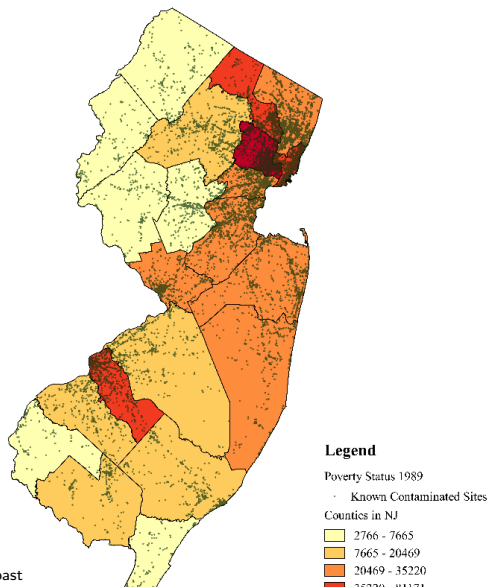
Area (Square Miles)



healthy corner stores

- makes sense to label zipcodes; right proportions
- these aren't sq miles! sq ft or meters!
- colors denote polygon sizes—so same info twice
- better could map educ, inc, age, bmi, etc
- dots could be little smaller or hollow so they overlap less
- make goog map and zoom in: show more detail
 - see environ: other businesses, pub transpo, sch, etc
- wonder about big healthy stores like wholefoods
- could denote big ones with big dots
- usually may want to put year on a map

Contaminations Sites in New Jersey 1992



contaminations

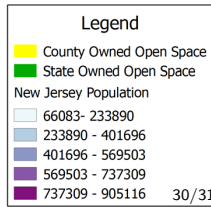
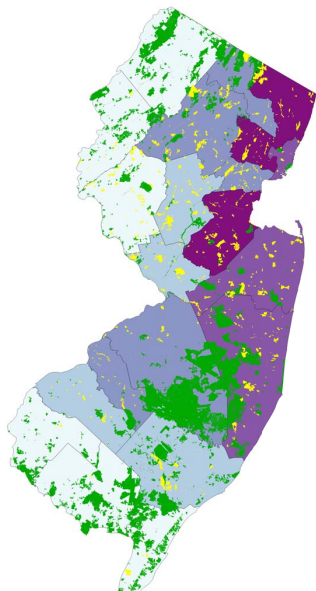
- perfect size and color for contaminated sites!
- doesn't overlap much but big enough to see
- and grayish good for contamination
- informative— NYC and Philly the worst
- excellent idea to relate poverty to contamination
- there is lit linking them! so nice test! [also can do race]
- could do poverty at municipal or census tract levels
- use space better! NJ should be bigger like Philly stores
- thousands must be set off by commas in legend
- very good to match contaminations and poverty by year!
- “poverty status”—guess counts; better %
- as in Philly map: zoom to Camden, have goog map in

contaminations

- http://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html?_r=0
- in couple classes we'll be making online maps like this
- but already now you can do sth similar
- see footnote: census and socialexplorer.com: download data
- map in qgis and bring in background from googmaps
- with openlayers plugin

open space

New Jersey Preserved Open Space



open space

- excellent idea for map—open space related to population
- great use of multiple layers
- great non-cluttered borders
- can use space better—portrait orientation, bigger NJ
- use commas for population
- say for which year it is
- pop den probably more meaningful
- on the other hand, we already see size from map
- and so we can sort out density