data

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Saturday 7$^{\text{th}}$ September, 2019   00:05

outline

regular (not gis) data

gis data (has shapes, can make a map from it)

the 'join'

Example: New Jersey Home Values

census data [probably do one week later]

old ps comments

communication (forgot to cover last week)

◇ email is a preferred mode of communication; just email
   `gis_int@googlegroups.com`

· and everybody in the class

· including me and GA will get it

· messages will be marked with "[gis_int]" in the subject

◇ you can easily filter them to a specific folder, e.g. in gmail:

   `http://support.google.com/mail/bin/answer.py?hl=en&answer=6579`

ps0 comments

◇ i'll be just emailing comments to each of you individually

◇ how wass ps0? discuss?

◇ if you cannot find the right data, just email me

data management takes time! value your time!

◇ producing maps is fast

◇ most time (i'd say 50-95%) is data management:

· figuring out, cleaning, documenting, combining, etc

◇ so we start with data management

◇ but only about 20% of class is dat mgmt

· but it'll be about 80% of your time

◇ spend it on data you care about and will use in your career!

◇ note: join is difficult! start today/tomorrow on ps, ask Q!

data

◇ nj `http://www.nj.gov/dep/gis/listall.html`

◇ a lot of data here:

· `http://geocommons.com/search.html`

· just search for what you are interested in, say 'road'

· and see `https://www.policymap.com/maps`

· they make you pay to downlad data, but can see source and download by hand

open govt, especially city data

⋄just few examples

⋄trend is that more and more local, state, fed opens up

⋄http://phlapi.com/ , https://data.cityofchicago.org/ , http://opencityapps.org/ ,

http://www.opendataphilly.org/ , http://www.phila.gov/data/Pages/data.aspx

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

◇ if several layers: may have several u/as

◇ eg counties: #18; hospitals:#700; ex of attr?

◇ dataset is a matrix/spreadsheet/2D object

◇ cols are vars, rows are obs

◇ vars are characteristics of obs

◇ eg: edu, age, inc are vars

· and persons are obs–each row is a different person

storage type: numeric v string

◇ strings are safer; eg string "0821" made into a number results in "821", which is a mistake !

· that's why many software packages, incl qgis often store numbers as strings

· but then we often need to make them into numeric to do the math or mapping

◇ be careful about it, triple check, there are often problems and it's non-intuitive

outline

regular (not gis) data

gis data (has shapes, can make a map from it)

the 'join'

Example: New Jersey Home Values

census data [probably do one week later]

old ps comments

files

◇ .shp (along with buch of others)

◇ .kml

◇ and there's much more

◇ we'll cover them on "as is" basis

· if you bump into something else–let me know–we'll cover it

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

$\diamond$ raster (has resolution)

$\cdot$ area covered by cells/pixels

$\cdot$ each cell/pixel have values/colors

$\diamond$ vector (no resolution): all real world features:

$\cdot$ points (dots/nodes): airports, cities, trees

$\cdot$ lines (arcs): rivers, roads

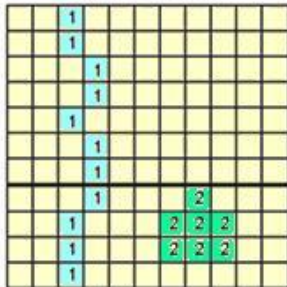$\cdot$ polygons (areas): counties, cities

raster and vector
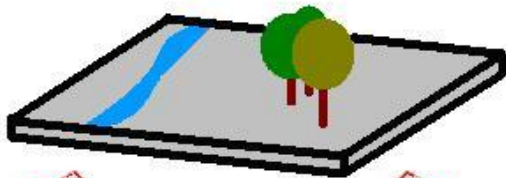
VECTORIAL
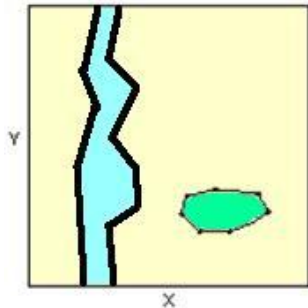
RASTER

◇

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: location info (shapes)+usually some regular data

· ie a data table with location info (shapes) must underlie a map

· (and the data table usually contains some regular data, too)

◇ often you want to produce thematic (choropleth) maps

· thematic maps use different symbols/colors to show variation in regular data

outline

regular (not gis) data

gis data (has shapes, can make a map from it)

the 'join'

Example: New Jersey Home Values

census data [probably do one week later]

old ps comments

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://sites.google.com/site/adamokuliczkozaryn/pubs/rel_inn.pdf

howto map regular (eg xls) data?

◇ it would likely have <u>geo id</u>:

· ISD 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

◇ beware of representativeness of your data for areas

· i spent months mapping provinces from WVS

· then emailed WVS and was told they're not representative

"the join problems": some 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

gis data (has shapes, can make a map from it)

the 'join'

Example: New Jersey Home Values

census data [probably do one week later]

old ps comments

figuring things out

◇ so say you've got housing prices for NJ counties

◇ then need to google matching gis data (shapefile)

· google: "NJ counties shapefile"

◇ both have county variable so you can join

◇ but both keys/ids need to be coded in exactly the same way

· characters and storage!

◇ and **you** need to figure this out

http://www.zillow.com/research/data
⬦ subset reposted on my website https://sites.google.com/site/
  adamokuliczkozaryn/gis_int/NJ-counties-Zillow-Home-Value-Index-TimeSeries.xls

⬦ adjust ID: make counties uppercase

· (or could drop 'County' from COUNTY LABEL variable)

· make col (var) names short: eg <5 alphanumeric chars

⬦ and clean up: dropped first row, excessive columns,$ (%,#,
  etc) and ","; cnty names upcase, saved as csv (first sheet)

⬦ https://sites.google.com/site/adamokuliczkozaryn/gis_int/all_homes.csv

· note missing val for Morris; think abt missing data!

⬦ nj counties data (same as alaways)

  https://docs.google.com/uc?id=1xJDhcRCkgv7k4tNCa72Oog5bohV6dTB2&export=download

excel note!!

◇ excel is clunky, and often adds special/weird characters!

◇ when save as csv, go to:

◇ tools-web options-encoding and select 'us ascii'

· other ideas: `https://www.webtoffee.com/`
  `how-to-save-csv-excel-file-as-utf-8-encoded`

install MMQGIS (just once) if not there already

◇ Plugins-Manage and Install Plugins:

· Search: MMQGIS

· and install

◇ now we can use MMQGIS to join and fix the data!

· [another way to do joins:

http://www.qgistutorials.com/en/docs/performing_table_joins.html]

MMQGIS: join; and text to float

◇ MMQGIS-Combine-Attributes Join From CSV File

◇ Input CSV: all_homes.csv

◇ CSV File Field: UPPER

◇ Join Layer: nj_counties

◇ Join Layer Attribute: COUNTY

◇ make sure notfound.csv is where you want it

◇ check notfound.csv: header and 'NEW JERSEY': makes
  sense!

· check the tables to see if it joined right; be very careful!

◇ MMQGIS-Modify-Text to Float (almost always need this!)

◇ highlight "Dec 2012" only (others are not clean:"$",",",")

missing value

$\diamond$ right click layer-Open Attribute Table

$\diamond$ note that now MORRIS has 0 for "Dec 2012"

$\diamond$ this is incorrect!

$\diamond$ hit pen icon at top left: "Toggle Editing Mode"

$\cdot$ and remove zero from that cell

$\diamond$ hit "Toggle Editing Mode" again and Save

and the thematic map

◇ nj_counties-Properties-Style and from drop-down: "Graduated"

◇ Column: "Dec 2012"

◇ Color ramp: i like Blues!

◇ many ways to classify [if time, discuss later]

◇ usually good: 'natural breaks/jenks' say 3-7

◇ and hit "Classify" button

◇ and hit "OK" to see the map–viola!

◇ zoom in as much as needed

printing to file: Project-New Print Layout

◇ left: blank icon "Add New Map" and draw a rectangle

◇ NJ is tall: on the right "Layout"and do "Resize layout"

◇ left: icon with arrows "Move Item Content" to adjust view

◇ right: "Item properties" change scale to adjust zoom

   and/or use mouse's wheel

◇ left: legend button "Add new legend" (legend needs fixing)

· right: **uncheck** auto-update and beautify it:

· drop items with minus sign; and edit by double clicking it

◇ top: on the left: Layout-Export as Image

· probably png is fine, just increase resolution to say 600dpi

· http://www.qgistutorials.com/en/docs/making_a_map.html and

· http://docs.qgis.org/2.0/en/docs/user_manual/print_composer/print_composer.html

don't trust anybody!

⋄ remember, always be critical

⋄ triangulate your results: compare with other source

· just goog picture, eg 'nj counties property values map'

⋄ looks about right (they have some other definition of the prices, but correlation is important)

⋄ show to others, ask for comments, present locally or at a conference

⋄ i mistakengly thought a lot of alcohol problems in Cape May

· but it is just tourists!

tip1

$\diamond$ merging (joining) data is tedious and tricky

$\diamond$ be careful, double, triple check

$\diamond$ easy to make mistake

tip2: missing vals

◇ tricky! pay extra attention to it!

◇ sometimes qgis makes ' ' to 0! esp MMQGIS: str to float

◇ sometimes qgis colors it yellow sometimes transparent:

· (i guess: ' '=transparent, 'NULL'=yellow)

◇ to make it stand out can change color ramp

· eg if NULL is white, make even number of classes say 2

· and say make color ramp GnRd

tip3: what if traditional data is in weird format

◇ same as with gis data

· if you see something else than .shp or .kml, email me!

· there are many data formats, and we cannot cover them all

· we'll do them if we bump into them–do let us know what you've found!

outline

regular (not gis) data

gis data (has shapes, can make a map from it)

the 'join'

Example: New Jersey Home Values

census data [probably do one week later]

old ps comments

census data: 5-yr ACS

◇ census is a good source of data, even at neighborhood level!

◇ for city/neighb lev probably want 5-yr ACS

◇ https://geomap.ffiec.gov/FFIECGeocMap/GeocodeMap1.aspx

◇ https://factfinder.census.gov/faces/nav/jsf/pages/
   searchresults.xhtml?refresh=t

◇ can search in top box but probably best select on the left
   from "Topics" eg: people-poverty-poverty

◇ then select "Geographies": eg census tracts (ie
   neighborhoods)

· go down to "All Census Tracts in Camden County" and hit
   "ADD TO YOUR SELECTIONS" and hit "CLOSE"

◇ and from "Show results from" pick "2015"

· click "S1701, POVERTY STATUS IN THE PAST 12

cont

◇ take note of margins of errors!!

· most precise is decennial census, but much fewer variables

◇ "Modify Table" and keep selected only the stuff you need

◇ ok, at top hit Download

· and check "Use" not "View"

· keep both checked: "Merge the annotations..." and
  "Include descriptive...", hit OK

· csv reposted `https://docs.google.com/uc?id=`
  `1MD-P2Iu0XWWkYAsIn0WCYfqZ15cJya8n&export=download`

again, always clean it up before getting into qgis

◇ open csv file, keep GEO ids (will use them for join)

· and just keep only needed vars and rename them:

· HC01_EST_VC01, Total; Estimate; Population for whom poverty status is determined: "tot"

· HC01_EST_VC53 Total; Estimate; ALL INDIVIDUALS WITH INCOME BELOW THE FOLLOWING POVERTY RATIOS - 125 percent of poverty level: "pov125"

◇ then calculate ratio of pov to tot: "prop"

◇ and drop row 2, the long name

· and save as csv

· clean csv reposted: https://docs.google.com/uc?id=
1Hw-3nugfIpSvvyai7Jy-lwA2IsRA0Pz0&export=download

get geo data

◇ census has geo data for any US geog!: `https: //www.census.gov/geo/maps-data/data/tiger-line.html`

◇ tracts: `https: //www.census.gov/geo/maps-data/data/cbf/cbf_tracts.html`

· doing 2015 because we have 2011-2015 data

◇ then note there are 2 similar IDs that would match census csv

· shp: `https://docs.google.com/uc?id=1KNe_ DSJQxiUiMVzKdVfHzYjUZSke2OnY&export=download`

join!

◇ load shp and then

◇ MMQGIS-Combine-Attributes join from CSV file

◇ MMQGIS: csv GEOid, shp: AFFGEOID

◇ and check notfound.csv—should be none

◇ MMQGIS: modify: text to float: tot pov125 prop

· (Ctrl and left click all three)

◇ right click layer-Properties-Style: "Graduated" map prop
with say Blues 5 jenks

◇ move around and say zoom in on Camden

__outline__

regular (not gis) data

gis data (has shapes, can make a map from it)
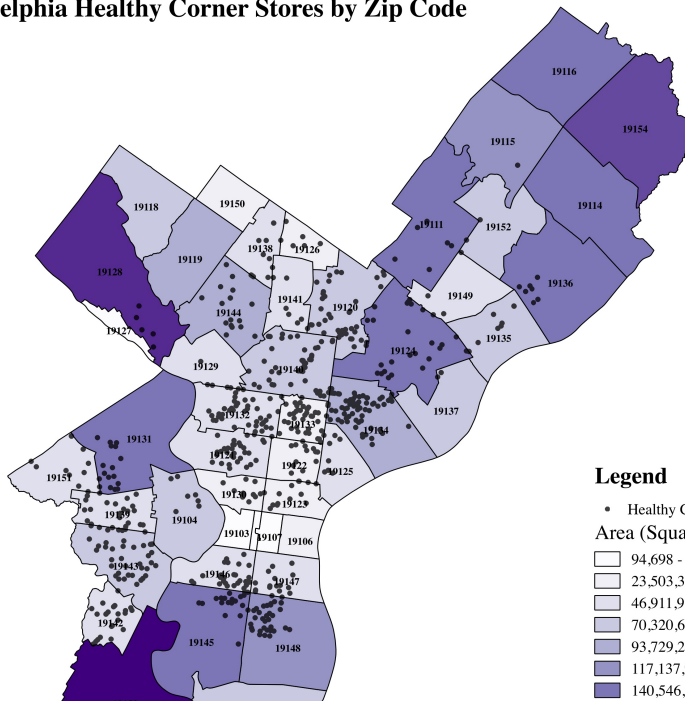
the 'join'

Example: New Jersey Home Values

census data [probably do one week later]

old ps comments

general comments

◇ please no ms word! txt or pdf

◇ remember to specify u/a and num of obs

◇ need to email me *all* data you've used

· (incl data you used for joining (toady's class))

· eg do not assume i have NJ counties

◇ send the whole thing! can just zip the whole project folder

· or share good drive, dropbox.com etc

· .shp file won't work! (need .dbf .prj, etc)

◇ again, in journal you can ask me questions!

Philadelphia Healthy Corner Stores by Zip Code

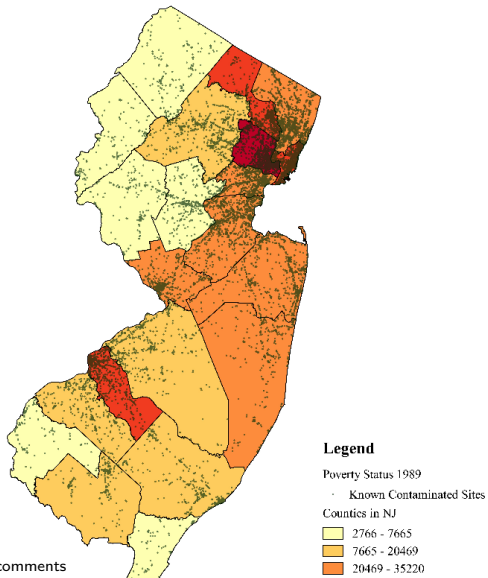
Legend

- Healthy Corner Stores

Area (Square Miles)

- 94,698 - 23,503,339
- 23,503,339 - 46,911,980
- 46,911,980 - 70,320,621
- 70,320,621 - 93,729,262
- 93,729,262 - 117,137,903
- 117,137,903 - 140,546,544
- 140,546,544 - 163,955,185

old ps comments

42 / 55

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 dentote big ones with big dots

◇ usually may want to put year on a map

· (at very least in metadata/journal)

Contaminations Sites in New Jersey 1992

Legend

Poverty Status 1989

· Known Contaminated Sites

Counties in NJ

	2766 - 7665
	7665 - 20469
	20469 - 35220

old ps comments

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 map

◇ 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 back

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
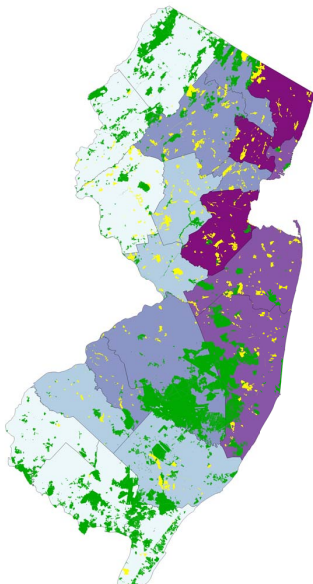
Legend
- County Owned Open Space
- State Owned Open Space

New Jersey Population
- 66083- 233890
- 233890 - 401696
- 401696 - 569503
- 569503 - 737309
- 737309 - 905116

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