data

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Saturday 6th January, 2018 17:24

outline

old ps1 general comments

data in general

gis data specifically

the merge (or 'join' as qgis calls it)

Example: New Jersey Home Values

outline

old ps1 general comments

data in general

gis data specifically

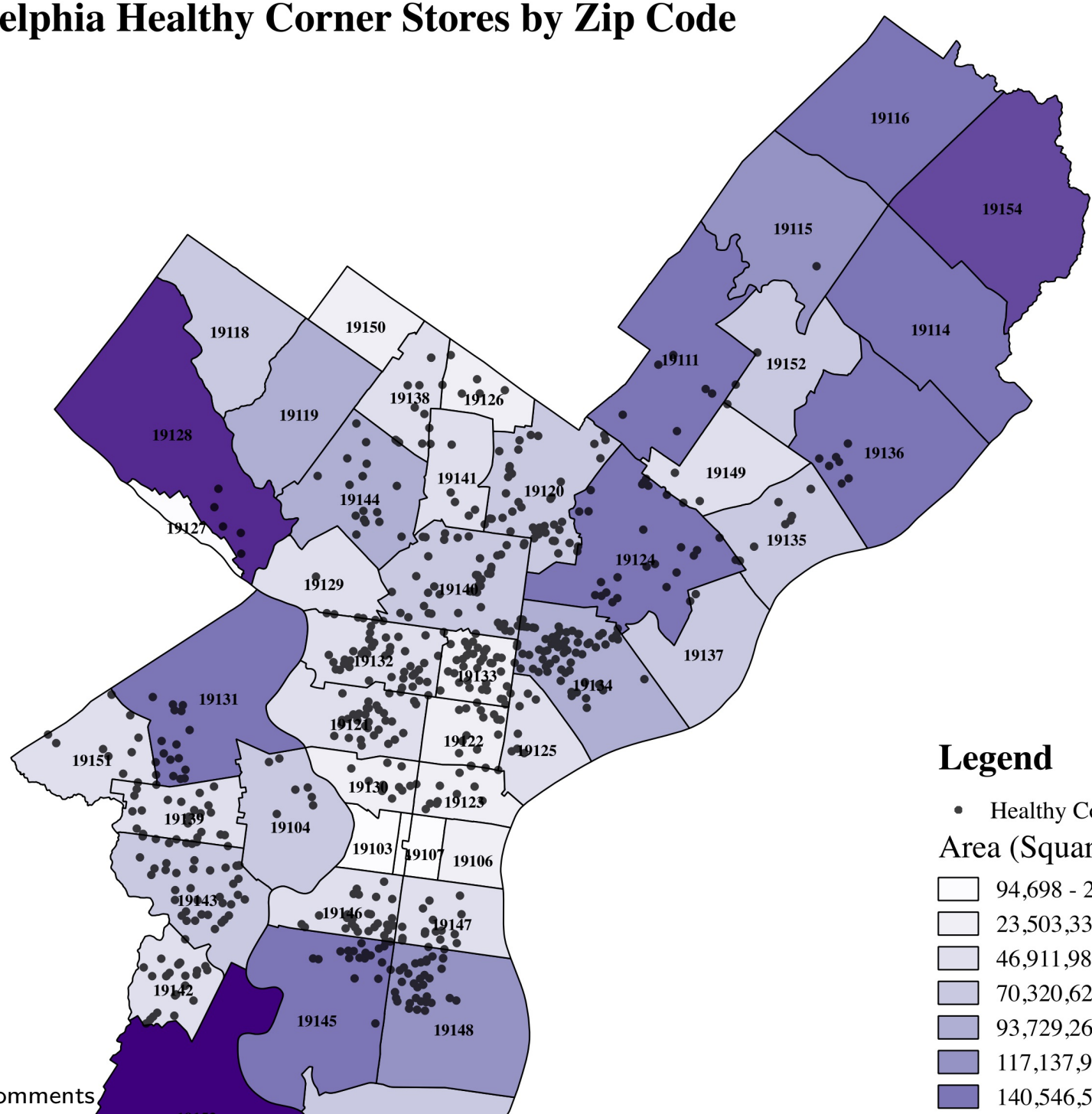
the merge (or 'join' as qgis calls it)

Example: New Jersey Home Values

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))
  - e.g. do not assume i have NJ counties
- send me the whole thing! you can just zip the whole project folder
  - if you just send me one .shp file, it won't run! (need .dbf .prj, etc)
- again, in journal you can ask me questions!

Philadelphia Healthy Corner Stores by Zip Code



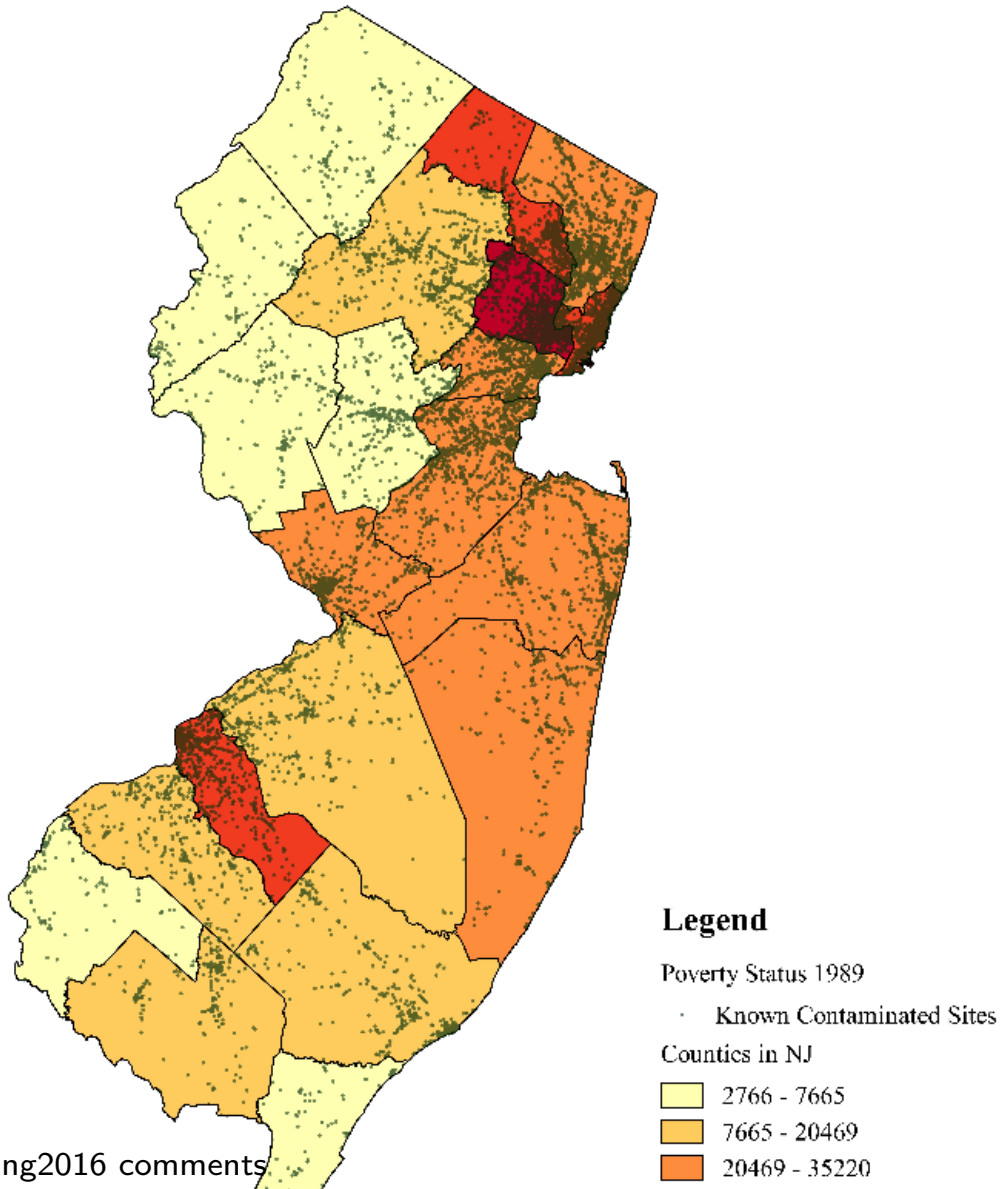
healthy corner stores

- very nice! pretty map; i like colors; neat legend!
- makes sense to label zipcodes; right proportions
- so map is perfect at this stage; but...for future:
  - colors denote polygon sizes–so same info twice
  - better could map population or even better yet:
  - e.g. educ, inc, median age, bmi, etc
  - dots could be little smaller or hollow so they overlap less
  - make another map with goog map and zoom in:
  - show more detail–then you can actually see
  - other businesses public transportation, schools, etc

healthy corner stores

- i do not think these are sq miles! sq ft or meters!
- there are also "Enhanced Healthy Corner Stores"
- could give them another symbol
- perfect description of what a healthy store is: say 5-50 sentences
- wonder about big healthy stores like wholefoods
- usually may want to put year on a map
  - (at very least in metadata/journal)

Contaminations Sites in New Jersey 1992



contaminations

- nice map ! actually almost like a heatmap [thematic.pdf]
- perfect size and color for contaminated sites!
  - doesn't overlap much but big enough to see
  - and grayish is good for contamination
- informative–easy to see that it's bad close to NYC, Philly
- excellent idea to relate poverty to contamination
  - there is literature linking the two! so nice test!
- for future:
  - 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!

contaminations

- I do not understand "poverty status"–what does it mean?
- say a number is 3k v 8k–that many people in poverty?
- "persons for whom poverty status is determined, including both those above and below[???] the poverty level. This provides overall counts"
- so seems like counts of poor folks, fine, it is meaningful: counts of contaminations and poor folks
- but would be interesting also to see percent poor
- and definitely at lower level, at least municipality
- and as in Philly map:
  - zoom to Camden or Newark, have goog map in background
  - and explore further at micro area
  - perhaps show race or poverty by tract at lower level

contaminations

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

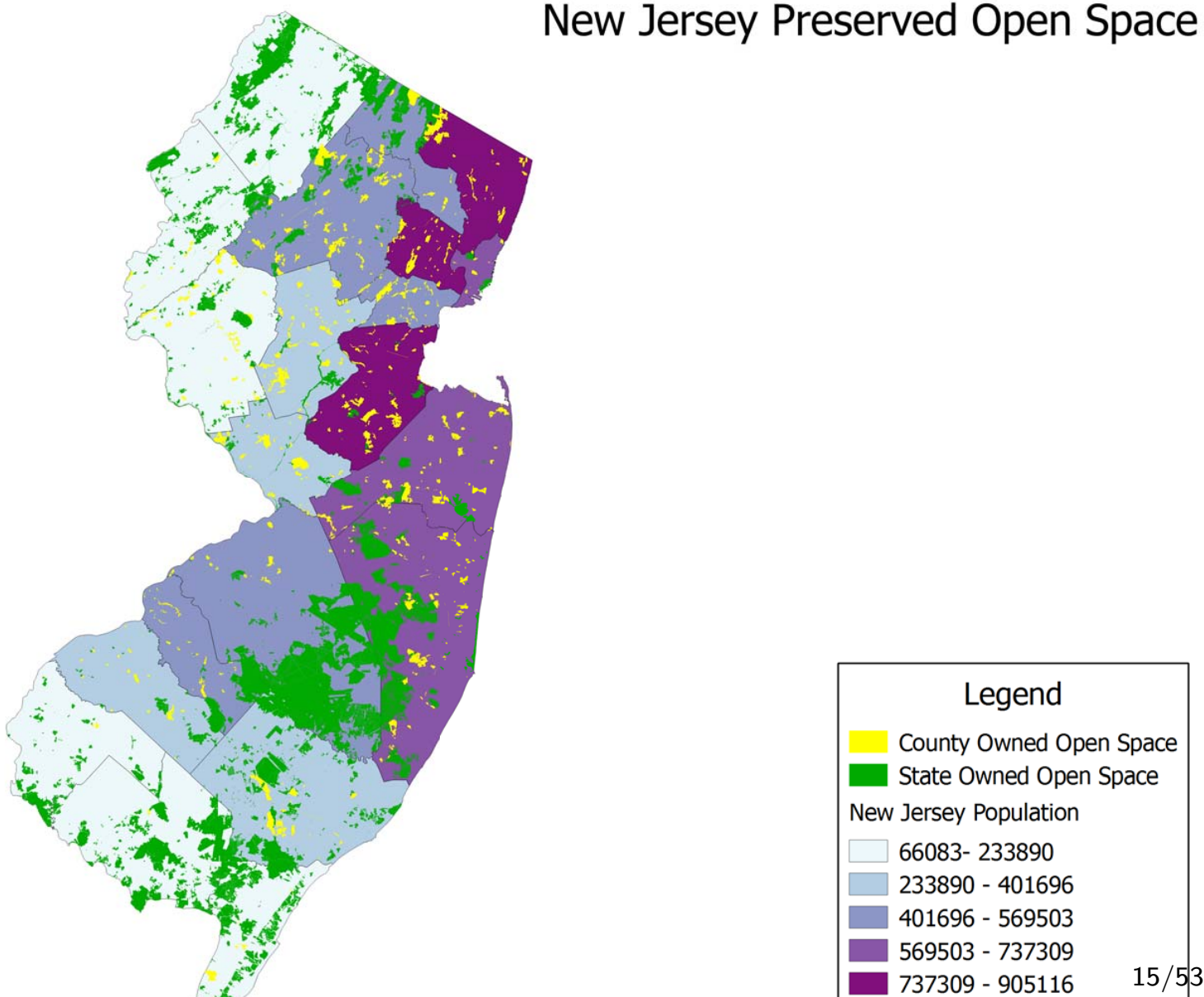
outline

ps2

- cannot ever have num of obs: 1 !!
- relabel in map composer layer to something meaningful
  - eg instead "NJ_legislative215": "NJ legislative districts"
- zoom in onto the map!! needs to be as big as possible!!
- whatever you have mapped, google it and see images
  - there will be maps by others that will inspire...
  - more on this in rulesTipsTricksEthics.pdf

ps2: open space

◇



ps2

- 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 much more meaningful–i do not see why pop would be useful
  - on the other hand, we already see size from map
  - and so we can sort out density

outline

old ps1 general comments

data in general

gis data specifically

the merge (or 'join' as qgis calls it)

Example: New Jersey Home Values

map of a week

- actually a set of maps
- these are supposed to inspire you
- just see few, see all at home
- `http://twistedsifter.com/2013/08/maps-that-will-help-you-make-sense-of-the-world/`

tip of a week

- nice website with quick reference and howtos
- `http://www.qgistutorials.com`

listserv

- again very important, is everybody getting emails with [*gis_int*] in the subject line?
- if not, please send me email, and i will add you

looking ahead: paper

- today we'll talk about data and few datasources (more later)
- again, you will use your own data
- pick something that interests you...it'll be more interesting
- and work on it throughout the class
- use it for ps
- and finally for the paper
- as usual, if you are not sure what to do, email listserv

outline

old ps1 general comments

data in general

gis data specifically

the merge (or 'join' as qgis calls it)

Example: New Jersey Home Values

data management

- producing maps and spatial statistics is fast
- most time (I'd say 50-95%) is data management:
  - figuring out, cleaning, documenting, combining data, etc
- and we'll start with data management...
- say only 30% of class is data management
  - but it will be >75% of your time

layers

- data is organized by *layers* covering themes, e.g. roads, admin boundaries, etc etc
- show example/draw a picture

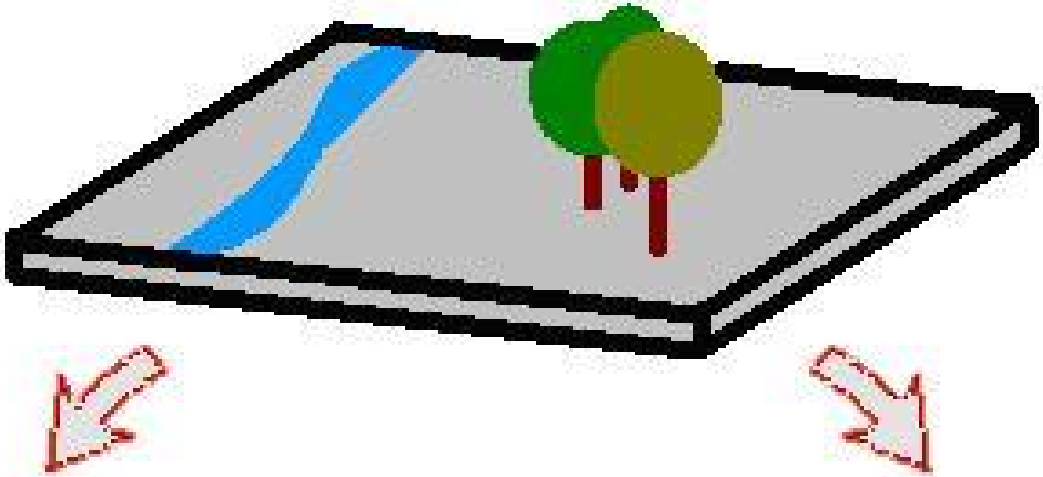
spatial and attribute data

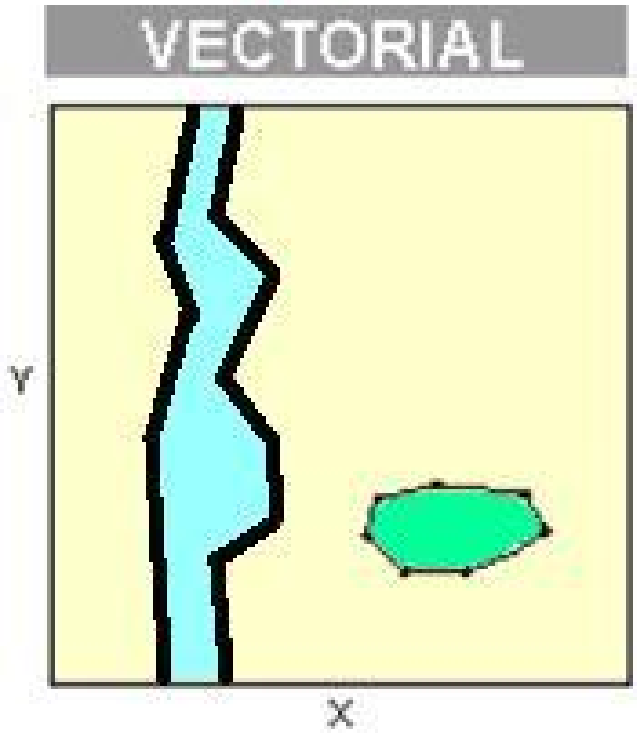
- spatial=location: where ?
  - coordinates, lat/lon
- attribute
  - what, how much, when
  - these are characteristics of a location
  - so the unit of analysis (U/A) is a location

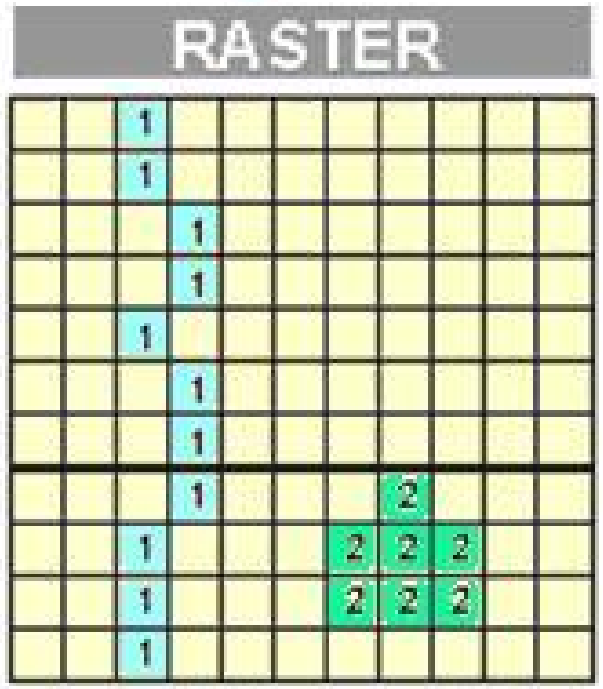
raster and vector

- 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







what is it?

- data=information
  - mostly numbers
  - raster data=pictures
- we'll just do numbers in this class
- dataset is a matrix
- columns are variables, rows are observations
- variables are characteristics or observations
- e.g., 'education', 'age', and 'income' are variables and persons are observations; each row is a separate person

u/a

- u/a: unit of analysis
- u/a=# of obs=# of rows=sample size
  - what do you study?
  - dataset has variables, which are the attributes of u/as
- say you study students or counties
  - then attributes could be age or water area
- if you have several layers, you may have several u/as
- e.g. counties: #18; hospitals:#700

numeric vs string

- strings format is characters: e.g. "Camden"
- numeric is a number, e.g. "22"
  - real (can have decimals), e.g. "22.01"
  - integer (no decimals), e.g. "22"
- cannot do any math with strings; e.g. no thematic map
- it is a storage format, not data recognition
  - storage type=how computer sees it, not you (human)
  - numbers can be stored as strings; strings cannot be stored as numbers (this is how computer sees it)

numeric vs string

- strings are safer; e.g. 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

metadata

- it's data about data
- i.e. documentation of data
- have it, use it
- e.g. codebook, variable definitions, source/url
- otherwise you'll get lost in the future
- ps will require you have "metadata" –see ps for details

outline

old ps1 general comments

data in general

gis data specifically

the merge (or 'join' as qgis calls it)

Example: New Jersey Home Values

gis or spatial data

- point: X,Y
- line: at least 2 X,Y
- polygon: at least 3 X,Y
- draw

some theory: data, layers

- gis data is (usually) regular data + (always) location info (lat/long)
- there is always a data table (usually regular data + location info) that underlies a map
- most of the time you want to superimpose different layers of gis data

  e.g. roads, cities, state boundaries, schools
- often you want to produce thematic (choropleth) maps

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

some theory: gis files

- gis data may be in many formats
- gis data have location info that allows mapping
- gis data can be points, lines, polygons
- usually, you want to overlay several layers...
- the most popular format is called “shapefile” .shp (comes with .dbf and others...)

shapefiles

- probably most popular
- it is actually 3 (or more) files:
  - .shp spatial data/coordinates (“main one” load this one)
  - .dbf attribute data
  - .shx other stuff
  - .prj projection
  - just manage it with gis soft, e.g. qgis

kml

- another popular format: google .kml (basically xml)
- this is Google Maps format
- it is a type of XML, a plain text/ASCII format
  - we'll cover it in onlineMappingFusionTables.pdf

other gis data

- 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

some gis data

- see data_sources.csv–i will be adding more there later
- you can open .csv with excel...

outline

references

- http://www.qgistutorials.com/en/docs/performing_table_joins.html
- http://maps.cga.harvard.edu/qgis/wkshop/join_csv.php
- WARNING!!! merging often doesn't work
- usually (as a rule) there are problems
- warning ! there will be lots of frustration–this is normal here

some real skills

- anybody can load a shapefile and make a map
- today's class gives you serious data management/gis skills
- dealing with real data, you' ll often have to do a join
- in fact, producing a thematic map is easy and fast
- on the other hand, you will usually spend majority of your time on data management–even say over 90% of the time
- this is where the real value come from: to bring many different datasets together to produce new insight

some real skills

- no matter what you're mapping
- likely such map already exists
- just google "what you study, map"
- and see images;
- but if you many variables and map it...
  - then there is no such map in the world!
- especially if you use innovative and unique vars
- eg `http://people.hmdc.harvard.edu/~akozaryn/myweb/rel_inn.pdf`
  - see 2 maps at the end

setup

- to produce maps/generate spatial statistics we need:
  - spatial data (gis data), i.e. mappable data: .shp, .kml, etc
  - attribute data (regular/traditional data)
- so far we had all that in spatial data file
- (we searched internet a lot to find such file)
- but most of the time you have some great data say in excel
- and you want to map it
- you need to merge it with gis data on common/unique/key/id variable
  - in this case (mapping) this variable is always location

howto map it

- ok you have some data, and it would very likely have some geo id:
  - ISD name/code, county name/id, etc
  - (codes/id's are great: unique! (as opposed to names))
  - then get a shapefile that you can merge with your data
- google "geo in you data, shapefile" e.g. "NJ cities, shapefile"
- and then merge the two to produce a map
- beware of representativeness of your data of geo...
- i spent months coding provinces from WVS; then emailed them and found out that they are not representative...

what is key/id var?

- U/As are in rows
- variables are in columns
- key/id variable: ssn, county code, zip code
- you can merge or join 2 datasets on the key variable
- draw a picture of merging

the "merging" 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 look right on the map? e.g. Camden richer than Cherry Hill?

outline

old ps1 general comments

data in general

gis data specifically

the merge (or 'join' as qgis calls it)

Example: New Jersey Home Values

figuring things out

- ok, so you've got gis data (shapefile) with NJ counties
- and you got housing prices for same counties
- both have the same (key/id) variable so you can merge
- but both keys need to be coded in exactly the same way (characters and storage)
- and you need to figure this out

Zillow housing prices

- nj counties data (same as alaways) `http://people.hmdc.harvard.edu/~akozaryn/myweb/bounds_nj_shp.zip`
- then "traditional" (non-gis) data in excel from `http://www.zillow.com/research/data/`
- I reposted on my website `https://sites.google.com/site/adamokuliczkozaryn/gis_int/NJ-counties-Zillow-Home-Value-Index-TimeSeries.xls`
- and cleaned up: dropped first row, excessive columns,$ and ","; cnty names upcase, saved as csv (first sheet)
- `https://sites.google.com/site/adamokuliczkozaryn/gis_int/all_homes.csv`

an attribute/mapping variable

- you need to take care of the variable you'll map
- e.g. drop decimals, dollar signs
- change yes/no to 1/0, etc (though can map strings as categorized–see nj colleges data: type of institution)
- and values must be numeric, not strings!
- numbers can be stored as strings!
- and this is what typically happens when you join csv data
- so need to either tell it to load as string (.csvt file)
- or convert it to numeric with calculator: toreal()

create .csvt, load .csv

- typically, qgis reads csv numbers as strings!
- create .csvt (use word, but save as text! not .doc!)
  - or safer use text editor such as notepad!:
  - "String","String","Real","Real","Real","Real","Real"
  - need as many items as many columns in csv
  - one line, quotes necessary, case sensitive, no spaces!
  - it's very picky! again, best use text editor, *not* word
  - .csvt defines format (again, cannot map a string)
  - make sure you've saved .csvt and *not* .csvt.txt
- bring .csv (not .csvt) just like any vector data
  - .csvt and .csv must have same name and be in same dir
- all_homes.csv-Properties-Fields: Strings?

mapping csv/csvt data

- nj_counties-properties-joins-"+"
  - join layer: all_homes (csv)
  - join_field: UPPER (csv)
  - target field: COUNTY (shp) (always joint to geog data)
- and have a look: nj_counties-open attribute table
- and let's map Dec2012 prices, say 5 natural breaks
- missing Morris cnty: qgis2 leaves it transparent
  - (older qgis or perhaps if 'NULL' instead of '': yellow)
- if transparent then load nj_counties again and put underneath and pick some distinct color for it
- always must have clear color for NULL and always say it in legend!

merging without .csvt

- del join: nj_counties-properties-joins: "-"
- drop all_homes.csv: all_homes.csv-remove
- drop or rename csvt
- use excel to put '0' for 'Dec 2012' for 'MORRIS' in csv
  - and remember that the '0' is missing for mapping later!!

merging without .csvt

- bring csv into qgis again
- all_homes.csv-Properties-Fields; col 'Type name': 'String'
- nj_counties-properties-joins-"+"
  - join layer: all_homes (csv)
  - join_field: UPPER (csv)
  - target field: COUNTY (shp)
- and have a look: nj_counties-open attribute table
- try mapping Dec2012 prices: nj_counties-properties-Style
  - cannot select 'Dec2012 price'– it's a string!

toreal()

- under layers select nj_counties and click calculator icon
- 'Create a new field': 'Output field name': 'd12'
  - remember keep these names short! qgis likes short
- 'Output field type': 'Decimal number (real)'
- 'Output field width': 10; 'Precision': 5
- 'Functions':Conversions-to_real
  - Fields and Values: 'all_homes_Dec_2012'
  - and close ")"
- or you can just type this into 'Expression' box:
  - 'to_real("all_homes_Dec 2012")'
- and now you should be able to map this new var
- 'Add class' for 0-0 and make it distinct for missing val!

don't trust anybody!

- remember, always be critical
- triangulate your results: compare with other source
  - just goog picture, eg 'nj counties property values map'
  - `http://www.trulia.com/home_prices/New_Jersey/`
- 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 aclohol problems in cape may
  - but it is just tourists!

tip0

- merging (joining) data is very tedious and tricky
- be careful, double, triple check
- very easy to make mistake
- if stuck try a different method (eg toreal() instead of .csvt)

tip1: have short names without special chars

- somehow qgis doesn't like long var names; best few, say 2-5 chars
- if your names are lengthy and/or contain special chars such as "("
- there may be problems...
- for instance, you may not be able to calculate toreal()

tip2: missing vals

- tricky! pay extra attention to it!
- sometimes qgis makes it yellow, sometimes transparent...
  - (i guess: ''=transparent, 'NULL'=yellow)
- to make it stand out can change color ramp
- e.g. if NULL is white, make even number of classes on 2 color ramp (say BlueRed)

tip3: what if traditional data is in weird format

- same as with gis data...
  - if you see something else than .shp or .kml, email listserv
  - 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!