misc: rules, tips, tricks, ethics

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Thursday 3rd November, 2016 14:36

outline

data

tips and tricks

some rules

ethics

an example of my research

research design again

outline

another look at data sources

- https://sites.google.com/site/adamokuliczkozaryn/gis_int/data_sources.csv
- essp nj.gov
- so many of them!
- http://www.nj.gov/dep/gis/listall.html

outline

difficulties

- print composer
- best way to find data (in this order): 1. google, 2. email me (nobody did!), specific websites (e.g. UN, FBI, etc), data_sources.csv
- display on pdf highlighted features
- selected features won't be colorized in print composer
  - open attr table-toggle editing-new col-tag features with say '1'
  - and then map this variable...
- or save selection as new shapefile, load back and color

workflow

- save the whole project (with many layers) and next time just open it
- can move layer up/down in layer window
- can have many layers with say different symbology of the same shapefile
- example–let's load nj_counties and produce several different symbologies and save whole project...and open it

misbehaving software

- most of the software sometimes misbehaves...
  - it crashes; refuses to do something, etc
- troubleshooting:
  - email me
  - do what you are doing in a different way-e.g try different dataset; different var; different approach etc (usually can do same thing in many ways)
  - shut it down and fire it up again
  - reinstall (last resort)
  - run it off apps.rutgers.edu

google it

- depressing, but whatever you are mapping, someone has already done it
- accept it, and make use of it!
- google and see images, say: 'nj counties contamination sites' `https://www.google.com/search?q=nj+counties+contamination+sites&tbm=isch`
- or "Philadelphia healthy stores map" (sometimes need word 'map' otherwise get pics of healthy food)
  - `https://www.google.com/search?q=philadelphia+healthy+stores+map&tbm=isch`
- get ideas, inspiration from these googled maps
- try to make your map better than the competition
- still, usually the key to be innovative is to join data!

google it

- and the related advantage of looking at maps that others made
  is that it serves as kind of literature review
  - rember from other classes: always do literature review first
  - here too, and look first at what others did
- there is never much glory from reinventing the wheel
  - it may be fun, and learning experience
  - but otherwhise useless

google it

- can't overestimate the usefulness of goog for finding data
- eg "what you are looking for, shapefile"
- eg "new jersey public schools, shapefile"
- tips:
  - may need to look for a higher level
  - eg NJ schools instead of Depford Twshp schools
- if you cannot find it, contact govt
  - eg city of Camden, state of NJ, etc
  - they'll be happy (not always!) you use their data
- again, may find only traditional data and need to merge with gis data

google it

- likewise, if you want to map it, and not sure how
- or perhaps just want to visualize it, google it
  - say 'census regions or divisions'
  - instead of reading about what this could be
  - a map is worth 1,000 words!
  - `https://www.google.com/search?q=us+census+divisions&client=firefox-a&hs=VPH&rls=org.mozilla:en-US:official&source=lnms&tbm=isch&sa=X&ei=sgUzVLSeOoeoyQTVh4LIBQ&ved=0CAgQ_AUoAQ&biw=1147&bih=1261`

join data

- the real value comes from joining data!
- again, a map about any single var was already made
  - but any 2 vars in a map or in 2 maps are rare
- there are so many data and variables out there
- use your creativity and imagination
- and you'll easily come up with something that no one did
- then join the data and make a map
- eg `http://people.hmdc.harvard.edu/~akozaryn/myweb/rel_inn.pdf`
  - see 2 maps at the end
- sure, gis is mostly a technical skill
  - but there's some art here too!

outline

quality

- GIGO: Garbage In, Garbage Out
- 'Cos it's in the computer, don't mean it's right

unknowns by Rumsfeld (be humble in your findings)

- There are known unknowns.
- That is to say there are things that we now know we don't know.
  - (these are benign, but be explicit about them)
- But there are also unknown unknowns. There are things we do not know we don't know.
  - (these are tricky: you can't do anything about unknown unknowns other that acknowledge that they exist; and never say you "proved" something)
  - (your statements are valid until disproved: all Swans are white, only until you see one day a black Swan)

what does it mean for you

- double, triple check
- ask yourself if it makes sense...
  - (Camden richer than Cherry Hill?)
- triangulate: use several datasources and/or several vars to measure the same thing
- are you getting similar results? why not?

outline

data

tips and tricks

some rules

ethics

an example of my research

research design again

integrity/honesty

- be explicit about problems in your data
  - eg non-merges, missing data, miscodings
- be explicit about problems in your models:
  - eg don't throw away variables from maps just because they contradict your story
  - discuss it: how, why; ask audience to comment/criticize
- instead of forcing data to tell your story, listen carefully; let data tell you her story !
- if you work for somebody: eg a bank or NGO: they will ask you to find something; use a disclaimer saying that

ethics

- everybody wants to sell something
- we academics or thinkers or students, too!
- we try to sell some idea or point of view
- rarely if ever anyone is 100% objective
- keep that in mind !
- and always present alternative/opposite points of view
- present the whole picture

ethics: bad examples

- cherry picking of vars or samples or timeframes, etc
- using only vars or operationalizations that fit your story
  - eg using year in which your find what you wanted to find
- classification bins: playing with bins to support your story
- in short: force yourself to be objective
  - because by default humans aren't

outline

happiness in Europe

- have a look at `https://sites.google.com/site/adamokuliczkozaryn/pubs/gesis3.pdf`
- first, on p.5 I show a histogram of happiness
  - (use Statist plugin or native histograms)
- then on p.6,7 two maps: quantiles, natural breaks/jenks
- note, that you can be creative, and calculate other interesting quantities sch as variation: p. 11

outline

important for paper

- think (and address them) when working on a paper

error of measurement

- keep in mind that measurement is always imprecise
  - and ask yourself how imprecise
  - see literature; eg happiness has been cross-validated:
  - PET scans, opinions of friends etc

error of measurement

- who produced data ?
  - eg Chinese data are less reliable than US data
  - natl govt reports lower pollution to look good intl
  - local govt reports higher pollution to show that it develops and produces a lot
  - [disclaimer: read it somewhere, may be inaccurate]
  - the point is to always think about quality of data!
  - and whether producer of data has motive to fake it
  - eg, in Chinese case triangulate with some intl data, say satellite images

...triangulate

- triangulation=use different measures for the same concept
- eg education:
  - years of schooling
  - highest degree obtained
  - avg SAT score
  - avg ranking of schools in the area
  - etc etc

think about incentives

- who is producing that data?
- again, you can measure a concept in many different ways
- organizations have an incentive to measure it in a way that benefits them

construct validity

- are you measuring what you say you are measuring?
- say you want measure ability, or IQ, but you only have data about education
- `http://www.socialresearchmethods.net/kb/constval.php`
- seven sins map

  `http://2.bp.blogspot.com/_R3SXJVojagU/SwLzZJL1E2I/AAAAAAAAAIE/7GbMzcZPDDk/s1600/sevendeadlysins.bmp`

external validity

- are your data representative ?
- how big is the sample ?
- eg I was geocoding WVS at province level only to find out it was unrepresentative

time matters IN MAPS, too

- we are exploring spatial variation
- but there is also time variation (MAP IT!)
- usually it is nice to show time changes in your maps
- eg can display a variable as a difference say
  - $POP10 - POP00$–which county gained most population
  - do it with nj_counties: calculator icon tool
  - and also do $\frac{POP10-POP00}{POP00}$
- other time issue is that things fluctuate over time
  - say due to business cycle
  - want a more reliable estimate? take an average
  - say avg. 5-yr unemployment rate

go places! fun and makes it more scientific!

- when you make maps and find things,
  - go and visit that place
  - eg I drove through MI from TX to NJ
- or map places where you grew up, live now, etc
  - always great to cmpare map to real world
  - AND this is a way to contribute!
  - you won't beat a guy with gis degree in general
  - but you'll beat him if you map sth you know about
  - (say crime if you are criminologist)
  - and/or if you map place you know about
  - and discuss your real-world experience