Geospatial analysis in historical research

Please join the Research IT Reading Group for a presentation by Daniel Viragh (Postdoctoral Fellow at the Magnes Collection for Jewish Art and Life) on the use of geospatial analysis in historical research. Daniel will be presenting The UC Berkeley Historical GIS Project, an ongoing work with a team of undergraduate research apprentices to build a historical geo-database of Budapest in 1896. Working from the bottom up, the team digitized data from an 1896 map of the city and a book-length listing of the city's commercial, industrial, and government resources. Daniel will provide a brief overview of the multi-stage process of cleaning and mapping the data. He will also discuss the research questions he addresses by working with geospatial analysis.

When: Thursday, April 23 from noon - 1pm
Where: 200C Warren Hall, 2195 Hearst St (see building access instructions on parent page).
Event format: The reading group is a brown bag lunch (bring your own) with a short ~20 min talk followed by ~40 min group discussion.

Please review the following in advance of the 4/23 meeting:

There will be no assigned reading, but please peruse Stanford University's Spatial History Project for some examples of finished geospatial analysis projects.

Presenting:

Daniel Viragh, Berkeley Historical GIS Project
historicalgis.berkeley.edu

Attending:

Alex Winton, Project Management
Aron Roberts, Research IT
Camille Villa, Research IT
Chris Hoffman, Research IT
Larry Conrad, CIO
Patrick Schmitz, Research IT
Perry Willet, CDL
Quinn Dombrowski, Research IT
Rick Jaffe, Research IT
Ron Sprouse, Linguistics
Scott Peterson, Library
Steve Masover, Research IT

Notes:

Background in history: finished PhD in central European history in 2014
Lots of theoretical training, minimal data training
2 newspapers—criticize each other, but editorial offices w/in two blocks of each other
What kind of history can you reconstruct using spatial data in point form from newspapers and other sources
Assembled team of undergrad researchers, juniors and seniors in comp sci, economics, programming, Arc GIS
Using ArcGIS to recreate historical networks using data from an address book that lists economic, administrative and residential resources
1k pages of addresses in various categories
Budapest 1896: data is available, was massively transforming from agricultural base to heavy industry w/in 20 years
Expansion of road, rail networks
Had historical map from 1896, listing every block, street names and numbers, certain house imprints
Digitized address books from the city, with gaps, covers 30 years
For 2/3 of city, very detailed information; for non-industrialized parts, no information, no street names/numbers, just streams and planned streets
 Took city blocks, put in address ranges into ArcGIS
Address book: PDF file with info you have to decode
Every street with district, sub-district, from/to points
Can check data in map against address book
All in Hungarian, but once you explain words for “from” and “to” it’s okay to work with
Listing of every street, house number, house owner, and reference number
Requires a lot of time to make use of this data; also, not always consistent
Administrative offices / government offices — tells you where it is, when it’s open, who works there — organizational history of that part of Hungary, down to the lowest secretary
Could use this data to associate it with residential listings, show who lived and worked where
Commercial listings — functions like yellow pages
Residential listings, give you info about the person’s job/job title, in some other cases, says “barren” or “duke of something” or “house owner”
Interested in where people lived; students are doing density maps (where do coffee shop owners live vs. people who work there)
Women’s roles included widow, woman-teacher
2-3 comp sci students, 3-4 in geography-related fields, a few in environmental science
Students wrote Python code to make data intelligible; took OCR’d text from City Library of Budapest, wrote code to determine where commas/spaces occur, got 1/2 of 65k addresses this way
Comments aren’t always picked up, Python code has to guess between comma/period
Original data has districts as roman numerals in small font; came up with a dictionary to map roman numerals and OCR errors to districts
ARC likes the N American address order; code had to reorganize the address components
Fall 2014: create address geocoded, had to georectify historical map
Downloaded street network from Open Street Map, but working with someone else’s messy data was a problem
Had to clean street network, fix orientations
What happens at corners, or things with multiple addresses
A square in the middle of two streets crossing has a different name than anything else; have to tell ARC that’s not part of the same line
Did the work for the whole city in order to not cut out addresses available for geocoding
Spring 2015 goals: try to recreate production network for wood products in the city: identify producers and consumers that handle wood, use geolocator to identify producers and consumers in GIS
Glass, wood, metal — all available in address book
Wanted to identify local networks, see who buys/sells from whom
Tried to get lists of merchants in each of these categories, have students work on one set of merchants, see if thy live in certain parts of city
Text had to be cleaned for use; comp sci students reformatted data
Problems with OCR accuracy; OCR was spotty, were operating with 50-60% loss, not many points actually coded (10% from starting number)
Geocoded 34k people who were living in the city, identified their professions, marital status, nobility or not
Creating density maps; correlations between areas inhabited by rich / commercially involved folks vs. other areas for workers, lower-level people
Interested in most prominent industries in address book; pub workers vs pub owners, coffee shop owners vs workers; land owners living in central part of city, close to parliament and river; they own eastern Hungary (far away from where owners live)
House owners: own house and rent it out
Two ring roads; larger ring road is a boundary, outside is fields, people who commute to inside
Coffee shops line up on ring road, one of arteries
Pest developing much faster than Buda (relatively stagnant)
Very descriptive stage of project, a student asked “do we have a theorem to prove?”
Use of GIS for historical research is very new; trying to see how a city functions, how it comes into being as a mourn city
Interested in looking into using same approaches for other cities; similar patterns?
Street name changes?
Yes, had to create own geocoder; would be 50% success if we used today’s names
Ring road has 4 sections, renamed with every war and revolution
Street names have long evolution, from “arrow” to names of various ministers, Hitler, Stalin, and back to “arrow”
Was there a Hungarian language model for OCR?
Yes, but not very reliable. Diacritics were a problem; eliminated diacritics and ArcGIS was less prone to crash
Legal archives, or anything else to associate transactions / activity to people rather than just knowing occupations?
Many notaries who sign off on documents, docs are online in a shared database
If we had one particular person, could look him up and call up documents
Could run queries through this database for names that we identified; probably not within the next year, need to finish with address book first
Looks like you’re trying to extract meaning from existing data, without pre-determining historical questions. Is that new in historical research, frowned upon, etc?
Methodologically, was appealing because narrative coming out of this period was “Hungary, great empire, empire collapsed, since then it’s all gone awry”. Narrative gotten from other nationalities in Hungary was that Hungarians oppressed them, tried to develop economically, couldn’t, made own countries, got oppressed, now can finally develop since 90’s. Historical narrative is so skewed with nationalism that model that only stresses data points, or argues from “neutral” perspective, set of facts about where things were located, is useful
Unclear what story will come out, had to put narratives / methodological questions on hold; tech is so new that it’s worthwhile to explore it
Concerned about scope of facts being a source of bias?
Yes, worried that historians who deal with this material deal with primary/secondary sources of people in power at the time (ministers, bankers, industrialists) with a stake in status quo
Can’t see or don’t want to see global environment
Bias in this data set — people see address book as “the authorities” prying on lives
People don’t like the idea of giving an address or profession, sometimes they lie and sometimes they make it up
Some addresses are “New Pest next to the post office” — either it’s not industrialized and there’s no address, or guy doesn’t want to be found
Just threw out addresses you couldn’t code
Nobility likes to live in places like “English Queen’s Palace” — we don’t know where this is, but you can find it
How representative is the sample you actually are able to geocode?
Hard to say; probably a function of the OCR, lost that most of that data that way, that’s consistent throughout
7,000 jobs; top 50 jobs account for half the data
Used Google Refine to concatenate jobs that OCR split up, got better data for that
Distribution of widows?
Haven’t done it yet; female teachers live on the outskirts more
How do you make the call about how much/little data makes it legitimate?
If we got better Python code, could increase size of data set
What can we get with code now, see those results, try to improve it