Patrick M Kelley

Course: HIST3907B

Professor: Shawn Graham

Student ID: 100239483

US Elections: Urban vs Rural Voting Patterns | PARADATA DOCUMENT

Paradata document

US Elections: Urban vs. Rural Voting Patterns

PARADATA Document

Table of Contents

[Introduction 2](#_Toc416595566)

[Finding Data 2](#_Toc416595567)

[Wrangling Data 4](#_Toc416595568)

[Holes 6](#_Toc416595569)

[Analyzing Data 7](#_Toc416595570)

[Visualizing Data 7](#_Toc416595571)

[The Final Project 8](#_Toc416595572)

[The Future 8](#_Toc416595573)

[Conclusion 8](#_Toc416595574)

# Introduction

Information has physical place in the world. Layers.

Reason for choosing paper.

# Finding Data

My initial hunt for data to prove my case began with regular google searches. I worked in twenty to thirty minute blocks every few days for the month of January. I had racked up about four or five hours’ worth of searches by the end of the month.

Eventually, I was able to find county level data, but it was housed by individual state websites. This would have meant manual downloads from fifty different websites, each with their own idiosyncrasies and structure. It would be a massively manual process as the different locations of the data did not lend itself to an automatic solution such as an API. I thought that I might have to confine my project to a case study of one or two states only. Here, the availability of the data was constraining the scope of my research, as is so often the case with historical inquiry.

I was also able to find election results in archived newspaper webpages, often from the next morning issues after the elections. Often data taken from newspapers or news websites was incomplete or unreliable. For example, whole counties might have only “3 of 217 polls reporting”. The news outlets had gotten as much information as they needed to report an election win or loss, and then stopped gathering data. No final, comprehensive results could be found from such sources.

I simply could not believe that the data I was looking for did not exist. In February I put in two days of dedicated searching totalling perhaps eight or ten hours between them. Truth be told, I did find the data that I would ultimately use for my project during this search, but I did not realize what it was that I had found at the time. The format was not something that I had seen before or knew how to use. It was so arcane (to me) that I didn’t realize that I was holding the treasure that I sought before me.

Frustrated, I turned to the staff at Carleton University library MADGIC desk. A one hour session with the desk staff turned up no results. They were having the same problem that I had, and were similarly frustrated. What we were looking for ought to be readily available in the 21st century! They told me to leave it with them.

Three days later they got back to me with the link for that data that I would ultimately use. I had seen the link before and then realized my earlier oversight. The desk staff later confided in me that they had spent more than four hours trying to find it themselves, sometimes with two people working on it.

The links for the data are listed here:

<http://catalog.data.gov/dataset/2004-presidential-general-election-county-results-direct-download>

<http://catalog.data.gov/dataset/2008-presidential-general-election-county-results-direct-download>

<http://catalog.data.gov/dataset/presidential-general-election-results-2012-direct-download>

The data was zipped in a format that is native to LINUX systems. I had to download and use 7zip software to be able to open and to read them.

<http://www.7-zip.org/>

Once unzipped, I had shape files and the appropriate data for all of the election results. The shape files contained polygons for each county, and could be imported directly into mapping software.

As my intention was to correlate election results to population size or density, I needed to acquire population data. It was a short search and the data was found at a government website again. It was buried in an obscure pat of the website, but I knew what I was looking for and found it in about thirty minutes.

<http://www.census.gov/popest/data/intercensal/county/county2010.html>

<http://www.census.gov/popest/data/historical/2010s/vintage_2013/datasets.html>

I now had election results and population data from reliable resources as government websites. Now I had to begin putting the data files into a useful format.

The data was the “Dream Case” for data, with everything present. Or so I thought.

Original datasets can be found on my GitHub repo. A list of some of the relevant files and folders is as follows:

* The GitHub repo.
  + <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project>
* 2004 Election Data
  + <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project/2004Base>
  + The files named elpo04p020.\* are the original election datasets.
* 2004 and 2008 Population Data
  + <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project/2000sPopulation>
  + The file named CO-EST00INT-TOT.csv is the original population dataset. This original file was used for both 2004 and 2008.
* 2008 Election Data
  + <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project/2008ElectionUnzip/elpo08p020_nt00335.tar/elpo08p020>
  + The files named elpo08p020.\* are the original election datasets.
* 2012 Election Data
  + <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project/2012ElectionUnzip/elpo12p010g.shp_nt00887.tar>
  + The files named elpo12p010g.\* are the original election datasets.
* 2012 Population Data
  + <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project/2012Population>
  + The file named PEP\_2012\_PEPANNRES.zip is the original population dataset.

# Wrangling Data

The library staff were again very accommodating, recommending ArcGIS for the project. ArcGIS is a *very* expensive and very powerful software suite for mapping. Carleton University offers one free temporary 365 day licence to undergraduates with tuition. The library staff must be notified, who then contact the software provider to set up the license. The software provider then creates a key that can be used to download and install the software. Only one license will be provided to one machine per student, so I had to ensure that it was installed on a computer that was capable of running it. The process of acquiring the license took about two days, and the installation took about 2 hours. A detail of the software agreement and installation process can be found here:

<http://www.library.carleton.ca/services/arcgis-student-edition>

Once I had ArcGIS up and running, I had to get my data into the software so that I could manipulate it. There were many obstacles to making that happen. As with so many of the modules during the semester, I had no sense of what the proper workflow ought to be to get from raw data to a finished presentation. I had no idea what the correct path was, much less what the obstacles even *could* be until I began. Recognizing my limitations, I again approached the MADGIC staff, who helped me with a quick overview of the software. They showed me how to import the electoral shape files, and, at least in principle, how to join the data in the electoral maps to the population data.

There was no primary key (a field containing a value unique to each record) between the two datasets on which join the electoral results to the population data. I had to construct one in the population data for each election year. I did so based on the FIPS field, which was present in the electoral results. The FIPS field in the electoral results had to be changed to a numerical field so that it could be joined. I had to reverse engineer a FIPS field for the population data sets by appending several different fields and inserting a variable number of characters into them, depending on the contents of the initial fields. For example, field A might contain ‘1’ or ‘48’, and field B might contain ‘4’ or ‘101’. Both fields had to be combined so that they were exactly five characters in length, resulting in values of ‘01004’ or ‘48101’ in our respective examples. This was not easy to do, and I solved the problem through sorts, concatenates, and altering data types in Excel. Ultimately I was able to construct a FIPS primary key in the population files. Once created, I was then able to successfully join the electoral results to the population tables inside of ArcGIS.

Although there were multiple iterations of the project, the folders with “\*Base\*” in their names contain the original shape files joined to the revised population data files. These shape files include Alaska and Hawaii. Considering that my intention was to create a Cartogram of the United States, I chose to exclude Hawaii and Alaska from the map. Including those two states would have had a serious impact on how the Cartogram was run. I tried it once later in the project just to see what the result looked like, and it produced a shape file that was nearly unintelligible.

In retrospect, I could have output Hawaii and Alaska as two separate shape files, performed the necessary colouration and cartogram deformations, and then used an application like Illustrator or Inkscape to create a larger graphic, to bring the two states back into the big picture, and to append the legend that I later developed. Ensuring that the hypothetical Alaska and Hawaii cartograms would have been deformed in a way consistent with the rest of the contiguous United States would have been difficult to ensure.

Ultimately the shape files in the “Base” folders do not appear in the finished project, but they do form the basis of all used images. The shape files for individual states, for example, were extracted directly from the “Base” shape files. They have been left with the project files on GitHub because of the significant role that they played in the ultimate form of the project.

Some of the notes online in GitHub provide examples of some of the wrangling that was necessary.

* <https://github.com/Prytanis/HIST3907B-ResearchNotes/tree/master/Project/Notes>
  + ProjectArcGIS-Notes-01.md provides an example of some of the steps that had to be executed for each iteration of map in ArcGIS.
  + ProjectArcGIS-Notes-02.md provides an example of some of the SQL that I had to develop in order to exclude or select individual states from the Base map. The SQL statement here excludes Alaska and Hawaii, allowing me to create a map of the contiguous United States.

Ultimately I would recompile the “Base” file several times, ultimately trying to strip out unnecessary information. The final version of the “Base” files are much, much smaller than they were on my first run-through. Originally each file contained ten full years worth of population estimate, for example. As I became more comfortable with SQL and Excel, I was able to eliminate extraneous years from each data file, keeping only the relevant fields. This made the files much smaller, and allowed me to upload the files to GitHub, which had originally rejected them as being too large.

# Holes

Raster vs. Vector files. If I could do it again, I’d try to use raster files. Jpgs pixelated in display on georectification portion. Not as clean as it could have been.

Previous cartograms vs. what I’ve put out.

Whole assertion that public discourse is wrong had to be addressed. Bill Maher video was a happy coincidence.

Elimination of independent and spoiled ballots (no way to distinguish between them anyway).

Started pretty solid attempt to find way to include it by calculating raw spreads. Between Dems and Republicans. See file.

Cartogram software / ease of display.

Didn’t significantly add to development of thesis, makes it very hard to display.

Would be interesting to display % of vote by county across the entire country as choropleth by itself.

# Analyzing Data

Means of doing research, not just of displaying findings.

Thesis confirmed.

Hated the shape files for 2004/2008. Would append info to shape file from 2012 if could for cleaner read. Don’t know if county boundaries changed. Difficulty / uncertainty of append. Implies unclear comparison.

Also included bodies of water, had to remove.

Found possible correlation with national parks. Difficulties in expressing in final project, as per other note. Yellowstone. Alaska. Made possible by interactive nature of ArcGIS. “Base” datasets were kept partially so that Alaska maps could be output later purely as choropleth maps to illustrate point.

Other note: Chose not to procede with national park thing because would need ways to highlight just the parks using these methods. Very time consuming.

# Visualizing Data

Eye tracking privileges top left corner.

Does not handle time well. Chose to use multiple iterations.

Implies use of table on Wikipedia in project.

Use of georectificaiton in project: alternate way of displaying same information as in cartogram. Georectificaiton only shows relationship, not degree.

Colour palatte: very traditional.

Jing / Camtasia for screen capture. iMovie for editing. Proof of concept with Module 4, exercise 8 screencap.

Legend created with Excel.

# The Final Project

Output as video.

Better shows maps in proper context. Zoom. Transparency. Comparison over time.

# The Future

Raises Questions:

WHY? Why are cities different? Hypothesis: hard to dislike neighbours. Better access to media / internet. Higher levels of education. Up to statisticians and sociologists.

National park thing…

Mathematical Function for predictive value.

# Conclusion