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](#_Toc416628746)

[Finding Data 2](#_Toc416628747)

[Wrangling Data 5](#_Toc416628748)

[Holes 8](#_Toc416628749)

[Analyzing Data 10](#_Toc416628750)

[Visualizing Data 12](#_Toc416628751)

[The Final Project 12](#_Toc416628752)

[The Future 12](#_Toc416628753)

[Conclusion 12](#_Toc416628754)

# 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.

The shape files for 2004 and 2008 had shapes for bodies of water included in them, whereas the file for 2012 did not. I had to remove the bodies of water from them. I did this by eliminating shapes with no votes or population. I manually verified the shapes to be deleted before eliminating them.

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

The raison d'être for the project was to make a count argument against the prevailing notion in the public discourse that the main dividing line in US politics between the Republican and Democratic parties is the Southern states. Yet I have not established that such a discourse is prevalent. Given that the purpose of the project was to explore the value of data visualization as a method for exploring historical events, I feel that it is acceptable to simply accept this portion of the argument as an unverified assertion. There are many books and journal articles and news articles that I can point to, but I have not demonstrated my assertion in a systematic way.

In the project itself, I point to one example of a visualization method that could be used to make such a demonstration. Transcripts from various new sources could be compiled and analyzed using various tools of text analysis. Voyant Tools was the example that I provided in the project itself. Another could be topic modelling methodology. Given that one massive mapping data visualization project is quite enough, I will have to simply let the existence of this predilection in the discourse to remain as an assertion on my part. Instead, I have used a particular interview as a case study.

Real Time with Bill Maher from HBO held an interview with Mike Huckabee on March 27, 2015 while I was well in to the development of my project. It was so timely and on topic that I integrated it into my project. In it, Maher asserts that cities in the Southern states are different from the rural areas. Having a video like this to use was also a happy coincidence, as I was using video to make the presentation of my maps. I integrated it easily into the final presentation.

Previous Cartograms do demonstrate the predilection, and perhaps partially explain it. Looking at the exact same election results, but with only a state-level resolution, many people have previously produced cartograms. Resolved to the state-level, the maps do show a “Southern Block” of states. It is likely that one can fault the perception to people simply not looking at the data with a fine enough resolution!

Included in the elections results were votes for third party candidates, independents, and/or spoiled ballots. What exactly they were was not defined in the data, the votes were simply listed as “Other” votes. Overall they represented a very small percentage of votes. I chose to exclude these votes from the data when preparing the maps for display.

Firstly, as already stated, the number of votes was very small. In no circumstances could I find it having determined the result of an election.

Second, I did make some initial attempts to find a way in the spreadsheets to take the “Other” votes into consideration. I hoped to output a percentage indicator that would be used to display the range of results on the choropleth maps expressed as a percentage spread or as a margin of victory. However, the results would have had to be applied inside ArcGIS as a calculated field. While possible, it would have further increased the size of these files, and I was already having difficulty uploading them to GitHub. There were other challenges, but I will simply have to summarize them as: “It was technically challenging, and unnecessary given the scope and spirit of the project.”

Thirdly, I was told by MADGIC staff that the Cartogram software would have great difficulty in expressing anything that required multiple layers. I did find that this was not true in this case, but could not demonstrate it until after the entirety of the project was complete. Redoing the project nearly from scratch in order to include a variable with little impact I deemed to be surplus to requirements.

Finally, I will add that in retrospect it is tempting to create a few maps to express the independent vote as a choropleth map purely out of curiosity to see where the independent votes are most common. It would not be difficult to do, and would fulfill the spirit of the project of using data visualization as a methods of research to discover trends not previously known.

Creating my first cartogram was very difficult. The gap here was knowledge. It took multiple attempts and many hour for me to figure out how to use ScapeToad (<http://scapetoad.choros.ch/>) to create the cartograms. I was able to figure out after much trial and error that ScapeToad was outputting a simple shapefile that contained no colouration. Once I realized this, I was able to simply take ScapeToad’s output, import it back into ArcGIS, and then recreate the formatting there. Several hours of experimentation were required to finally make it work. But he output cartogram was well worthwhile.

# Analyzing Data

The first time that I was able to output a national level cartogram, I actually raised my hands in victory and cheered. It had been many, many hours in the making and I was ecstatic. Also, it was immediately obvious how correct I had been in my hypothesis. As I stared at the map smaller details started to become clear, further confirming the urban versus rural partisan trend. Many of the smaller, confirming details can be seen in the final maps. Even in large rural swaths of the Midwest, nearly any county with a slightly higher population can be seen to have a lighter degree of Republican support. Democratic support in urban areas and Republican support in rural areas can be seen throughout the country.

The shape files for 2004 and 2008 are very different from the shape file for 2012. Again the difference can be clearly seen in the project. The treatment of coast areas, especially of swampy regions, led to very large distortions in the output. If I could re-do the project I would attached the electoral results form the 2004 and 2008 files to the shapes for the 2012 file. I would have to do some additional due diligence to confirm that there were no changes to the electoral boundaries between the years. I would do this by joining data between electoral years and then deleting the old shape fields. When the cartograms were produced they looked very different from 2012, but the same pattern could still be clearly seen. Given that they still confirmed the hypothesis, it was not necessary to go to the extra lengths for the purpose of this project.

Along the theme of using data visualization as a method of inquiry, and not to simply use it to display findings, I did look at some of the counties that voted Democrat in conspicuously rural areas. I looked at a number of counties in the Midwest and also in Northern Alaska, which the interactive nature of ArcGIS made extremely easy to do. I simply had to point at click at individual counties to extract relevant info. In all cases they were counties that were largely comprised of national parks, such as Yellowstone or the wildlife preserves of the North. “Base” datasets were kept so that I could easily output choropleth maps illustrating this observation. This interesting finding bears further investigation, though it is beyond the scope of this project.

# Visualizing Data

Given that I was trying to see if my hypothesis held true over time, mapping and GIS was at a distinct disadvantage. As methods, they do not convey the passage of time well. I therefore chose to display the maps that I was creating in a video format. Originally I imagined a kind of time-lapse, but that would have been more appropriate for a longer time frame than only three elections. Nevertheless, the choice of using video did a good job of allowing me to jump between time frames, display methods, and different levels of resolution (i.e. from federal to state and back again, or from choropleth to cartogram). The use of the table as displayed on the Wikipedia page in the project video was also an effective method of showing change over time.

Cartograms did a highly effective job of displaying the bias towards parties as a function of population density, as explained in the previous section. The eye is immediately drawn to urban centres, especially when immediately proceeded by a choropleth of the same area and time. The message was striking.

Georectification using Map Warper from Harvard.edu was also an effective way to make the same point while capitalizing on the video medium. Instead of a stagnant image, I could animate mouse movements to directly point to areas that I was speaking about. It was also a great way to bring to bear the classwork directly on the project. Ultimately, however, the georectification did a great job of showing the same relationship, but the cartograms were much better a illustrating the degree of the relationship.

I was somewhat dissatisfied with the georectification. In a few places the rectification didn’t quite align. I believe that the primary cause of this was the different map projections. If I had a better grasp of projection methods, I would be able to fix this easily in ArcGIS by selecting a different projection scheme.

Also, the choropleth maps imported into Map Warper for rectification were jpg files, and they ended up being visibly pixelated. If it is possible to use vector graphics instead, I would definitely have switched to vector graphics if given another chance. The display as it stands was not as clean as I would have liked it to have been.

The colour palette chosen for display was very traditional: red for Republican, blue for Democrat. The colour scheme is so common that to have chosen different colours would have interfered with the message and confused the viewer. For better or worse, I was stuck with it.

The legend used in the video was actually created in Excel. I did this to make sure that the gridlines and text all nicely aligned, and to create it quickly. While successful, some of the shades of blue and red may not be exact matches. I could have cleaned this up in Inkscape, for example by sampling colours from the map and colour filling the legend, but the sheer volume of maps that I had created meant that I wasn’t including the legend on every single slide anyway. Given the opportunity to produce it again (and infinite time) I would have placed the legend on all slides.

I used a temporary free trial license for Camtasia to create the screen capture videos of the georectification. <https://www.techsmith.com/camtasia.html> I created a proof of concept video first using the output from module four, exercise eight.

I used iMovie to create and edit the video for the final output film.

# The Final Project

Output as video.

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

Correlation and trend, not in and of itself causative. Also makes obvious that there are other factors in play. E.g. 2004 vs. 2008. Trend is still there, but degree of change goes through major change.

# 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