Analysis of Immigration Via US Baby Names

Julia Argy, Raynor Kuang, Ezinne Nwankwo, Judson Woods

CS109a: Data Science

Fall 2016

Analyzing baby names can offer a quantitative perspective to the otherwise difficult problem of mapping immigration regions to states. This project seeks to identify names that are correlated with specific immigrant regions: Europe, Africa, Asia, Oceania and the Americas. From there, we are able to see where these names show up over time on a state-by-state level. While this analysis can track names that originate from these regions, it does not necessarily do so. Instead, it allows for examination of "immigrant" names by region. We use principal component analysis to reduce the dimensions of the data for each region mentioned above and then use (insert unsupervised learning method that we use here) to get groups of states that we then visualize. This analysis can then be used to understand the mechanisms of cultural assimilation, and how certain names are associated with different ethnic groups. It offers insight into important trends in immigration on a state-by-state level that may be useful for future local and federal policy decisions.

Introduction

Names can tell us a lot about society changing over time. On the most basic level, they have the ability to shape how a child is seen. However, they are not only important on an individual level, but can offer insight to larger societal trends. From names, we can derive a the type of people that are living in a society. This can lead to conclusions about social trends and public opinion being made. Previous work on this topic include exploring trends in American parenting choices based on baby name choice. The study seeked to analyse whether there had been a decline or increase in individualism in America by looking at the frequency of a child receiving what was considered a common name. In another paper by Barucca et al, the authors are studying the cross-correlation of girl names by each state. They want to examine the correlations between the trends of baby names in states over time.

For our project, we are looking specifically at immigration trends within the US from five geographic regions: Europe, Asia, Africa, Oceania, and Americas. We look to see where these names by region are most popular (which we determine by prevalence of each name) in each of the 50 states. From there, we are able to extrapolate where immigrants from each region move to on a state by state level.

Approach

1. First, we determined which baby names correlate to which of the 5 regions of immigration.
- The top 100 names were identified for each region.
- Top names determined by treating names as predictors for immigration from a region and computing R^2 scores for each name for each region.

2. For each region, we constructed matrices in which rows represent the US states (including D.C.) and columns represent the 100 names for that region.
- The value of each cell in the matrix is the overall count of that name in that state.

3. To reduce the dimensions of these matrices, we used principal components analysis (PCA).
- The two PCA components were then used to visualize the similarity of names between states over time on a per decade basis.
- Identified the most representative immigrant names for each state

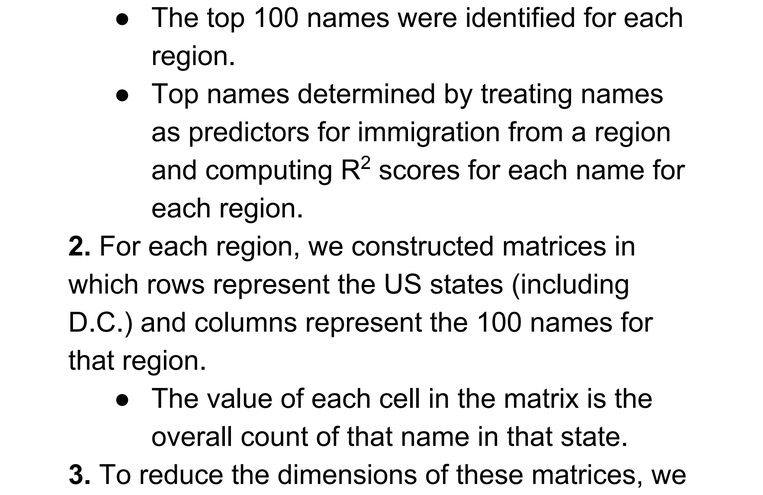


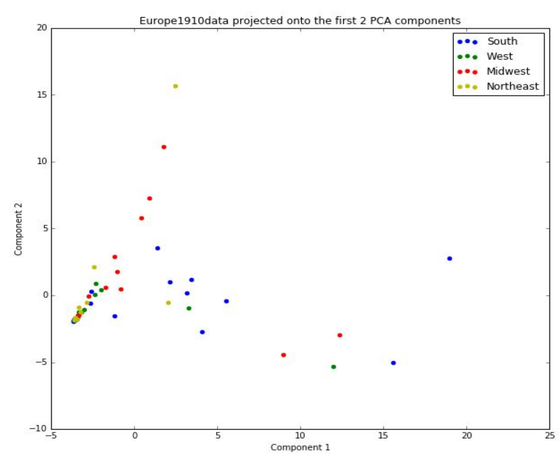
Figure 1. 1910 Uniquely European Names

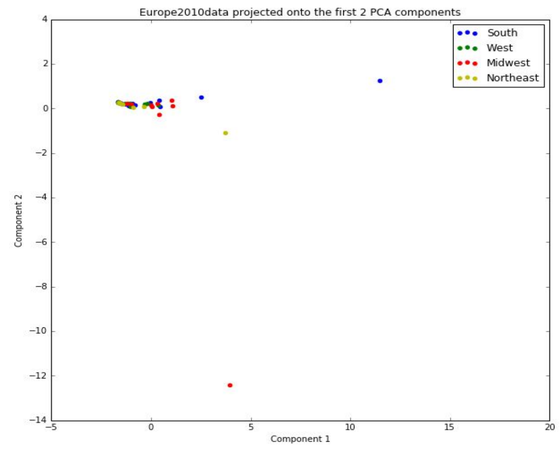
Data

Data Sources:
- *National Level*: name, year of birth, sex and number, are from a 100 percent sample of Social Security card applications after 1879
- *State Level*: From Social Security cards from 1910 onwards, 100% sample
- *Immigration Data*: From Census, happens every 10 years. Foreign born individuals include:
  - Naturalized U.S. citizens
  - Lawful permanent residents (immigrants)
  - Temporary and humanitarian migrants (such as foreign students and refugees and asylees)
  - Unauthorized migrants

Data Limitations:
- State and national data do not span the same years
- Census data exists by decade - reduces our sample size to 15 when looking at correlations
- Data by count, not percent - disproportionately favors later years where more children are being born

Results





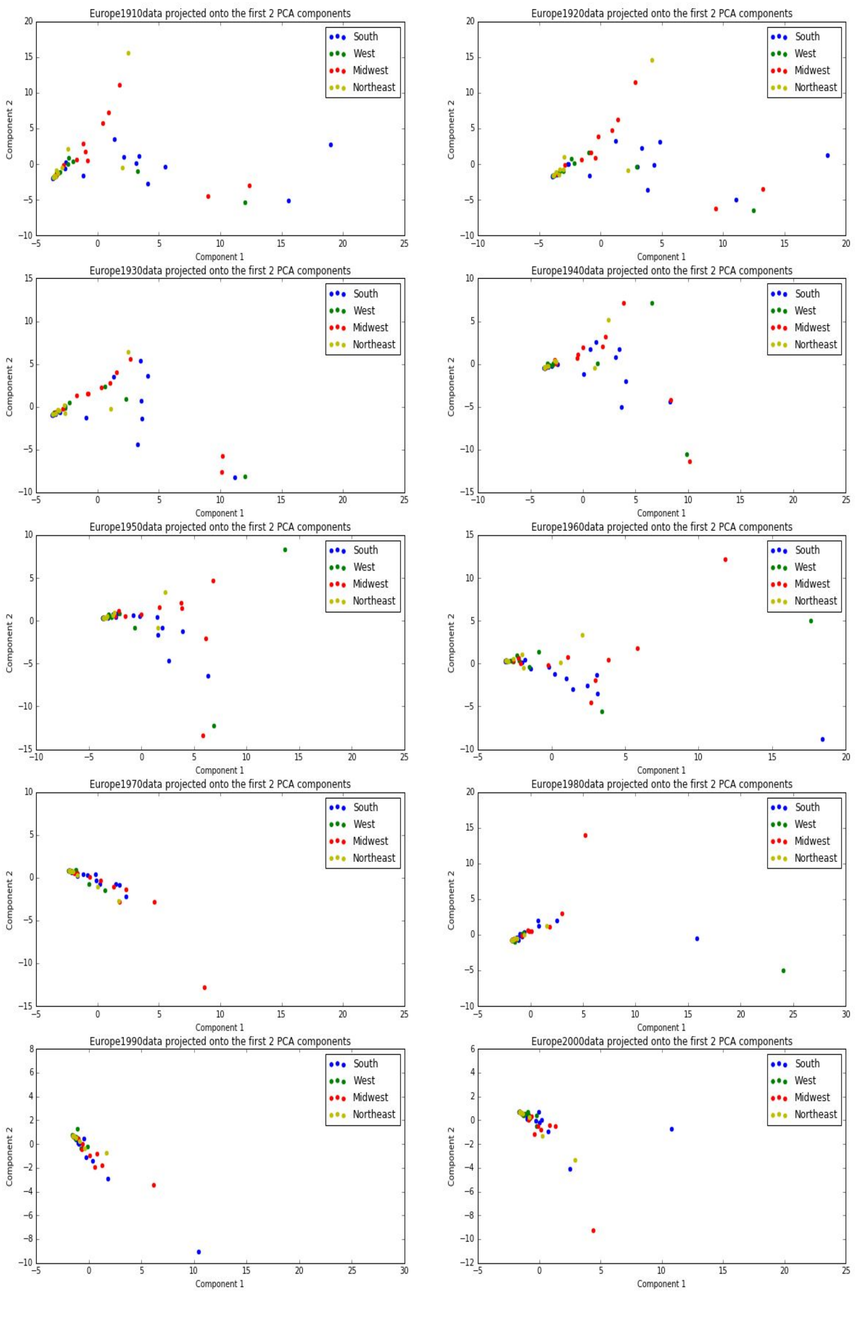


Figure 2. PCA Analysis for Europe by Decade

Conclusions

From the analysis performed, we discovered that generally the most representative names on a statewide basis are women's names. This may indicate that there is more variability in women's names than men's names, since the name with the maximum PCA components in each state tends to be a name more likely to be given to a woman. Also, the most representative names for each state are clustered somewhat by the regions of the US, a result that should be expected because immigrants coming from the same nation or adjacent nations are likely to settle in close proximity to each other and have similar naming patterns.

From the plots of the PCA components for the representative European immigrant naming data, we can see that the dispersion of the points on the plots decreases as time progresses. This could either indicate that the between-state variation of immigrant nationalities has declined over time or that the variation in immigrant names overall has decreased due to assimilation. We can also see from these plots that there is a higher degree of regional clustering over time for the Northeast and West regions of the US. This result may stem from the fact that the immigrants that settle in these regions tend to migrate from a select few nations rather than from multiple nations.

Citations and Links

Twenge, Jean M., Emodish M. Abebe, and W. Keith Campbell. "Fitting In or Standing Out: Trends in American Parents' Choices for Children's Names, 1880-2007." *Social Psychological and Personality Science* 1.1 (2010): 19-25.

Barucca, Paolo, et al. "Cross-correlations of American baby names." *Proceedings of the National Academy of Sciences* 112.26 (2015): 7943-7947.