EuroVis Workshop on Visual Analytics (2014)
M. Pohl and J. Roberts (Editors)

What's In a Name? Data Linkage, Demography and Visual Analytics

Feng Wang[1], Jose Ibarra[1], Muhammad Adnan[2], Paul Longley[2] and Ross Maciejewski[1]

[1]Arizona State University
[2]University College London

Abstract

This work explores the development of a visual analytics tool for geodemographic exploration in an online environment. We mine 78 million records from the United States public telephone directories, link the location data to demographic data (specifically income) from the United States Census Bureau, and allow users to interactively compare distributions of names with regards to spatial location similarity and income. In order to enable interactive similarity exploration, we explore methods of pre-processing the data as well as on-the-fly lookups. As data becomes larger and more complex, the development of appropriate data storage and analytics solutions has become even more critical when enabling online visualization. We discuss problems faced in implementation, design decisions and directions for future work.

1. Introduction

Family names (surnames) are a widely recorded marker for spatially-referenced population datasets. A surname can provide relevance to historical geography, genealogy and even population genetics. For example, work from Mateos et al. [MLO11] created global naming networks by generating linked forename-surname pairs revealing cultural naming practices for new and existing communities. Recent work from Cheshire and Longley [CL12] explored methodologies for identifying spatial concentrations of surnames. Their initial work focused on the development of an automated methodology for classifying the spatial distributions in surnames focusing on Great Britain [CLS10, LCM11]. Cheshire and Longley's work was later extended to 25 other countries (e.g., [CLYN13]), and an international surname mapping site (worldnames.publicprofiler.org) was created. This previous work in exploring demographics through names has primarily focused on classification methods and used visualization only as a means of displaying final results.

In this work, we extend the functionality of the worldnames profiler to explore not only the spatial distribution of names, but also linked demographic data. Our work focuses specifically on the United States, mining over 78 million records from the 2008 United States public telephone directories. Addresses are geocoded and then automatically linked to demographic data (specifically income distributions) from the United States Census bureau [U.S13]. Similar to the worldnames profiler, our tool (Figure 1) allows users to query surnames and see a density estimate distribution of the surname. Extensions include:

1. The ability to visualize and explore spatially similar names through a linked wordle of surnames where the size and color relates the spatial similarity of a surname;
2. The ability to visualize the estimated income distribution for a name based on census data, and;
3. The ability to explore the similarity between surnames based on income distributions through a linked wordle of surnames where the size and color relates the income distribution similarity of a surname.

While the visualizations provided are well known, the data linkage and integration of interactive analytic methods for comparing similarity is novel. Such a tool can provide unique insights into genealogy, demographics and social mobility. Furthermore, the challenge of distributing an online visual analytics tool for moderately large data provides an opportunity to explore the use of various data storage structures and distributed computing to enable interactive queries and visualization.

2. Names Profiler System

As georeferenced data has become increasingly available, more and more geographic visualization tools

DOI: 10.2312/eurova.20141143

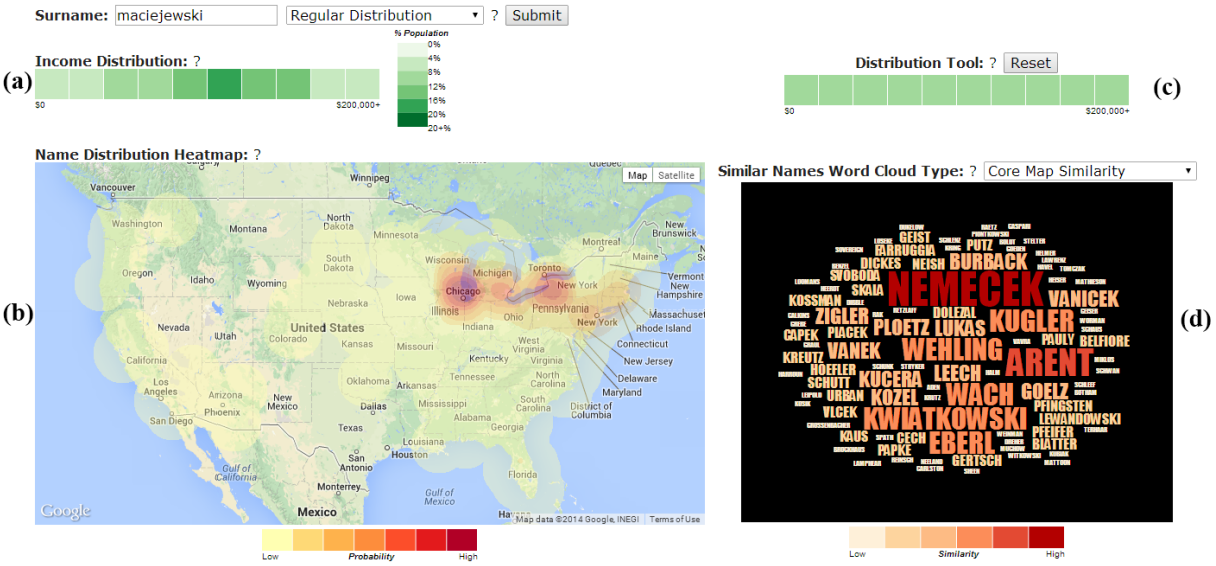

Figure 1: *The visual analytics interface to the United States name profiler. (a) A histogram encoded by color denoting the percentage of a given surname that is likely to map to an income range. (b) The spatial distribution of a surname. Users may look at a magnitude or probability distribution. (c) An income similarity toolbar. Users may search for names that are similar to a user defined income distribution. (d) The similarity wordle. The user may explore other surnames that have a similar spatial distribution or income distribution. Users can select a different similarity metric by changing the selected item in the dropdown.*

have been developed across a variety of domains (e.g., maritime analysis [MMME11, WvdWvW09], crime [MMCE10], healthcare [MBHP98, MHR*11], twitter analysis [MJR*11], movements [AAH*11] and various others [Wea09, GCML06, vLBA*12]). This work takes cues from Wood et al. [WDSC07] in developing a mashup for exploring surname distributions. We utilize publicly accessible telephone data that includes the geographic location of about 78 million people in the United States and link this data to the United States Census data. The goal of this work is to enable both novices and experts to explore name distributions and spatial relationships. We focus on three issues: aggregation, similarity and speed.

2.1. Density Estimation and Aggregation

This system estimates the probability density function of surnames to produce heatmap visualizations (Figure 1 (b)). We employ a fixed bandwidth kernel density estimation [Sil86] similar to other recent work [MRH*10, SWvdW*11]. Equation 1 defines the multivariate kernel density estimation.

$$\hat{f}_{\mathbf{h}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j=1}^{d}\frac{1}{\mathbf{h}_j}K\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{\mathbf{h}_j}\right)\right). \quad (1)$$

Here, $\mathbf{h}$ represents the multi-dimensional smoothing parameter, N is the total number of samples, d is the data dimensionality, and K is a kernel function. In our system, we used the Epanechnikov kernel:

$$K(u) = \frac{2}{\pi}(1-u^2)1_{\{u\leq 1\}}, \quad (2)$$

where $1_{\{u\leq 1\}}$ evaluates to 1 if the inequality is true and 0 for all other cases.

We provide views for visualizing both the magnitude (count of a surname in a given region) and probability distribution of the data (count of a surname in a given region divided by the population estimate of that region). For names with less than 100 records in the database, no aggregation was made to ensure data privacy.

2.2. Linking With Secondary Data Sources

In order to link surnames to income, we utilize the household income in the 2008-2012 American Community Survey 5-Year Estimates [U.S13]. Each surname's address can be mapped to a given census tract. We then solve a system of linear equations to estimate the probability distribution associated with a given surname. For surnames with over 1000 records, we use three matrices to represent the distribution of name records and income histograms. In matrix D, D_{ij} is the number of surname records for the ith census tract and the jth surname. B contains the income histograms of the census tracts. Specifically, each census tract reports the per-

Figure 2: *Heatmap comparisons for surname Alvarado. Subfigure A represents the L^2-norm comparison and Subfigure B represents the core comparison. The left most images are heatmaps of the population distribution of Alvarado. The wordle displays the most spatially similar names to Alvarado with the larger and darker names being the most similar. The right most images show heatmaps of the similar names to Alvarado based on the comparison type.*

centage of the population that falls within one of ten given income ranges. B_{ik} is the percentage of the population within a given income range in the kth income bin in the ith census tract. The linear system is then defined as:

$$DX = B \tag{3}$$

Since D is not a full rank matrix, we used a non-negative least square solver [LH95] to obtain a solution. For surnames with less than 1000 records, we take a weighted average of the income distributions of all the census tracts a given surname falls within. Finally, the income distribution of a surname is mapped as a 1D histogram, where color represents the % of the surname that is likely to fall within that income range (Figure 1 (a)).

2.3. Similarity Exploration

The third component of our system consists of a wordle that is encoded to show similarity between names with respect to either spatial distribution or income (Figure 1 (d)). For the spatial similarity [Coe07, AFC10], we explored two distance metrics: the L^2-norm (Euclidean distance) and the core distance. In order to allow for interactive rates of similarity matching, we first precomputed the density estimates at a fixed zoom level and resolution (170×90). The distance between two names is then calculated as the L^2-norm between the 2D density estimate array.

While straight-forward to implement, the single-core CPU implementation on a computer with a 3.4GHz Core i7-2600 needs 40 minutes to calculate the pairwise similarity for a single surname (there are 1.4M unique surnames in the dataset). While all similarities can be precomputed, our goal was to also explore other potential designs. Previous work by Cheshire and Longley [CL12] looked at what they called the core distance between density distributions. This distance was related to the distance between the centroids of regions between two distributions that cover approximately 55% of the data. We extract the five largest local maxima from each density estimate as our cores, and then compute the similarity as the smallest pairwise distance between the cores of each surname. In this manner, all core distances can be fetched and fit into local memory and pairwise correlations can be calculated. We need no more than 3.5ms to compute the distance of a pair of names. The time to compare one name with all the other names in the database is reduced from 40 minutes to 30-50 seconds. The top five maxima were chosen based on performance and

Figure 2 compares the results of using the L^2-norm and the core distance metric. For the surname Alvarado, Marquez is the most similar heatmap using the L^2-norm comparison and Herrera is the most similar heatmap using the core comparison. The wordle can also be mapped to income similarity which is calculated as the L^2-norm between all sets of surnames in the dataset. The smaller the L^2-norm the more similar the income distribution. The wordle in Figure 3 shows the most similar surnames to Wang with respect to the income distribution, where the largest and darkest colored

Figure 3: *Income comparison for surname Wang with the most similar surname, Loh, presented. The larger and darker colored names are most similar to Wang.*

Figure 4: *A user defined income distribution looking for names that are predominately wealthy. The larger and darker colored names are most similar to the defined income.*

names representing the most similar surnames. Users may also define an income distribution using the tool shown in Figure 1 (d). The wordle in Figure 4 shows the most similar surnames with respect to the user defined income distribtuion.

3. Experiments

Finally, our main research interest was in enabling interactive exploration of this modestly large dataset in a web environment where both data aggregation and similarity searches are a priority. Previous work on BigData infoVis has focused primarily on enabling data aggregation techniques as they form the basis for creating interactive maps, scatterplots and parallel coordinate plots. For example, Liu et al. [LJH13] addressed interactive scalability of big data systems through data reduction methods such as brushing and linking. Lins et al. presented Nanocubes [LKS13] as a method for efficient storage and querying of large datasets. However, the current nanocubes implementation supports only single spatial dimensions and some datasets use large amounts of memory. Both works primarily focused on the use of data cubes as a means of modeling and viewing data in multiple dimensions.

While data cubes have been shown to be extremely effective for enabling information visualization, it is important to note that the data in a data cube has already been processed and aggregated. Their primary functions lie in summarization of trends and operational reports. In our case where we want to enable similarity searches, and such calculations are not well supported within a data cube. For our current implementation, we primarily focused on preprocessing the data. Map aggregates were saved as images to reduce the data overhead, and pairwise similarity comparisons were generated and surnames were linked to their 100 topmost similar surnames. We use a single-core CPU implementation with a 3.4GHz Core i7-2600. Our program uses approximately 2GB of memory for the 73283 census data records and 78 million surnames in the database. The database takes about 14 GB of space in a MySQL database. The precalculated similarities can be returned within 30 ms and took 14 days to precalculate the similarities.

4. Conclusions

Surnames in our system tend to follow expected ethnic distributions, discounting names with a large populations, such as Smith. Figure 3 hints to potential ethnic patterns within surnames of similar origins. Wang is an Asian surname and the most similar name to Wang (Loh) is also of Asian origin. Similar patterns occur within the spatial distributions (Figure 2) and the income distribution tool (Figure 4).

While the visualizations presented in this work are standard, the implementation of a web-enabled system for large scale visual analytics is still challenging. Our design of precomputing similarities for a large number of categories is effective only under the case of static data. What this shows is the need for using high-performance computing as a method of quickly processing analytical queries. In this way we can move from putting the burden of finding similar data items on the user to placing this burden on the computational side. With regards to the name profiler system, anecdotal evidence suggests that the data matches users' mental models, and system users typically engage in exploration for 10 minutes or more. The current implementation can be tested at: http://goo.gl/gOGEVJ. A video demonstration can be viewed at: http://youtu.be/pANl4YJ1C5I.

5. Acknowledgments

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001 and by the Engineering and Physical Sciences Research Council UK EPSRC grant EP/J005266/1.

References

[AAH*11] ANDRIENKO G. L., ANDRIENKO N. V., HURTER C., RINZIVILLO S., WROBEL S.: From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology* (2011). 2

[AFC10] ANSARI M. H., FILLMORE N., COEN M. H.: Incorporating spatial similarity into ensemble clustering. In *MultiClust KDD* (2010). 3

[CL12] CHESHIRE J. A., LONGLEY P. A.: Identifying spatial concentrations of surnames. *International Journal of Geographics Information Science 26* (2012), 309–325. 1, 3

[CLS10] CHESHIRE J. A., LONGLEY P. A., SINGLETON A. D.: The surname regions of great britain. *Journal of Maps 6*, 1 (2010), 401–409. 1

[CLYN13] CHESHIRE J. A., LONGLEY P. A., YANO K., NAKAYA T.: Japanese surname regions. *Papers in Regional Science 92* (2013), In Press. 1

[Coe07] COEN M. H.: *A Similarity Metric for Spatial Probability Distributions*. Tech. rep., CSAIL MIT, 2007. 3

[GCML06] GUO D., CHEN J., MACEACHREN A. M., LIAO K.: A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1461–1474. 2

[LCM11] LONGLEY P. A., CHESHIRE J. A., MATEOS P.: Creating a regional geography of britain through the spatial analysis of surnames. *Geoforum 42* (2011), 506–516. 1

[LH95] LAWSON C. L., HANSON R. J.: *Solving Least Squares Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1995. 3

[LJH13] LIU Z., JIANG B., HEER J.: imMens: Real-time visual querying of big data. *Comput. Graph. Forum 32*, 3 (2013), 421–430. 4

[LKS13] LINS L., KLOSOWSKI J., SCHEIDEGGER C.: Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2456–2465. 4

[MBHP98] MACEACHREN A. M., BOSCOE F. P., HAUG D., PICKLE L.: Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Proceedings of the IEEE Symposium on Information Visualization* (1998). 2

[MHR*11] MACIEJEWSKI R., HAFEN R., RUDOLPH S., LAREW S., MITCHELL M., CLEVELAND W., EBERT D.: Forecasting hotspots - a predictive analytics approach. *IEEE Transactions on Visualization and Computer Graphics 17*, 4 (2011), 440–453. 2

[MJR*11] MACEACHREN A. M., JAISWAL A., ROBINSON A. C., PEZANOWSKI S., SAVELYEV A., MITRA P., ZHANG X., BLANFORD J.: Senseplace2: Geotwitter analytics support for situational awareness. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology* (2011). 2

[MLO11] MATEOS P., LONGLEY P. A., O'SULLIVAN D.: Ethnicity and population structure in personal naming networks. *PLoS ONE (Public Library of Science) 6*, 9 (2011), 1–12. 1

[MMCE10] MALIK A., MACIEJEWSKI R., COLLINS T. F., EBERT D. S.: Visual analytics law enforcement toolkit. In *Proceedings of the IEEE Conference on Technologies for Homeland Security* (2010). 2

[MMME11] MALIK A., MACIEJEWSKI R., MAULE B., EBERT D. S.: A visual analytics process for maritime resource allocation and risk assessment. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology* (2011). 2

[MRH*10] MACIEJEWSKI R., RUDOLPH S., HAFEN R., ABUSALAH A. M., YAKOUT M., OUZZANI M., CLEVELAND W. S., GRANNIS S. J., EBERT D. S.: A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics 16*, 2 (2010), 205–220. 2

[Sil86] SILVERMAN B. W.: *Density Estimation for Statistica and Data Analysis*. Chapman & Hall/CRC, 1986. 2

[SWvdW*11] SCHEEPENS R., WILLEMS N., VAN DE WETERING H., ANDRIENKO G., ANDRIENKO N., VAN WIJK J. J.: Composite density maps for multivariate trajectories. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2518–2527. 2

[U.S13] U.S. CENSUS BUREAU: 2008-2012 American Community Survey 5-Year Estimates, 2013. 1, 2

[vLBA*12] VON LANDESBERGER T., BREMM S., ANDRIENKO N., ANDRIENKO G., TEKUSOVA M.: Visual analytics methods for categoric spatio-temporal data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct 2012), pp. 183–192. 2

[WDSC07] WOOD J., DYKES J., SLINGSBY A., CLARKE K.: Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1176–1183. 2

[Wea09] WEAVER C.: Cross-filtered views for multidimensional visual analysis. *IEEE Transactions on Visualization and Computer Graphics 16* (2009). 2

[WvdWvW09] WILLEMS N., VAN DE WETERING H., VAN WIJK J. J.: Visualization of vessel movements. *Computer Graphics Forum* (2009), 959–966. 2