MixMap: Exploring User-Driven Semantic Similarity of Places

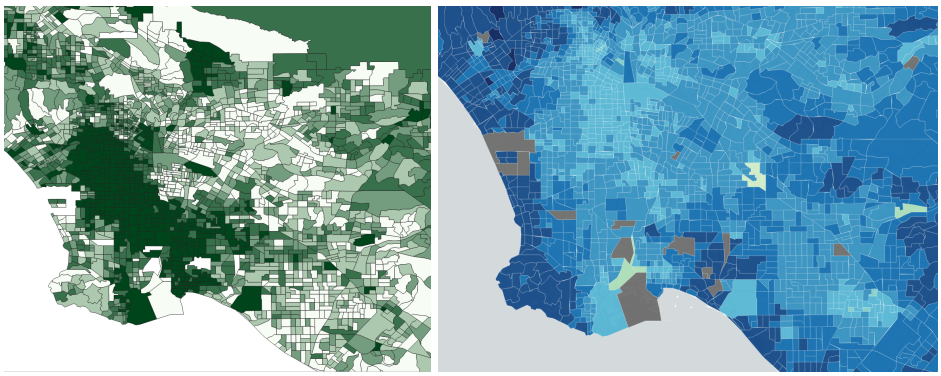
Grant McKenzie[ab*] **Sarah Battersby**[a] **Vidya Selter**[a]

[a] Tableau Research, USA
[b] McGill University, Canada
* grant.mckenzie@mcgill.ca

Keywords: place, similarity, semantics, geovisualization, interactivity, data parameters, tool

Introduction

Describing a place to someone can be a challenging task. The process is often subjective, relying on shared experiences and rarely involves description in an absolute sense. Most often, we describe a place in relation to another place, employing the concept of *similarity* (Rosch, 1978), relying on characteristics we know, or infer, about locations (e.g., Pismo Beach, CA is a small beach town similar to Carpinteria, CA). Identifying similarities between places is a useful task in a variety of domains such as retailers aiming to establish a new store location and community organizers wanting to better understand the impact of legislation on their communities through looking at similarly impacted communities. In fact, assessing similarity is a key component of geographic information retrieval (Adams and Raubal, 2014). The difficulty is in *quantifying* similarity between places. To address this challenge, we developed *MixMap*, a tool that supports a user-driven approach for determining the similarity of geographic regions.



(a) A single Census race attribute: Black or African American

(b) A single similarity measure comparing all Census race attributes

Figure 1: (a) Identifying similarity between regions based on a single Census race attribute (b) An approach that calculates similarity based on a comparison between all Census race attributes

In traditional demographic analyses, one might look at one attribute value at a time by comparing all possible geographic locations in terms of whether the value is higher or lower between

those locations (e.g., Figure 1a). The problem is that similarity between regions based on a small subset of variables, while possible to represent cartographically, is difficult, if not impossible, for a human to *mentally* consolidate into a single, coherent index reflecting similarity across all of the dimensions of interest (Slocum et al., 2009; Janowicz et al., 2011). However, multivariate analysis can be done computationally to provide streamlined, easy to interpret visualizations to facilitate interpretation (Figure 1b).

In developing *MixMap* we consulted with with civic engagement and community researchers to ground the work in practical problems regularly faced in their analytic work. Our approach makes the following research contributions (*RC*) to the cartographic and GIScience communities and aims to meet the needs of a range of user groups including, community organizers, academic researchers, and other data analysts.

RC1 An algorithm that computes a semantic similarity matrix for a selected geographic unit (e.g., block, tract, neighborhood) and a given set of attributes.

RC2 A tool called *MixMap* that enables users to select an arbitrary region of interest and identify similar regions based both on the characteristics of the region as well as the data points of interest within the newly created region.

RC3 A set of user affordances in *MixMap* wherein users can tune the similarity model, adjust the weights of each data dimensions, and add or remove regions.

Methodology

First, we provide a brief overview of our method for measuring similarity between regions (*RC1*). Next, we identify the design guidelines for a user-facing similarity tool (*RC2*). Finally, we briefly introduce the *MixMap* interface, our realization of the design goals (*RC3*).

Measuring Similarity

To address *RC1*, our objective is to compute the similarity between regions across a wide range of attribute data. In this example, we use data at the Census tract level from the US Census' American Community Survey (ACS) (U.S. Census Bureau, 2019). Specifically, five data dimensions (i.e., `Age`, `Race`, `Income`, `Education`, `Commuting Behavior`) were used, each containing a normalized vector of binned socio-economic or demographic values (e.g., Age 5-10, 10-15, etc.). We determine the pairwise similarity between all Census tracts for each dimension separately. To accomplish this, we calculate the Jensen-Shannon Distance (JSD). JSD is a technique for measuring the dissimilarity between two probability distributions and uses a relative entropy approach for two distributions based on the Kullback-Leibler divergence (KLD) (Equation 2). JSD has been used in previous work for geographic tasks ranging from differentiating places of interest (McKenzie et al., 2014) to assessing land use patterns (Nowosad and Stepinski, 2021). Equation 1 shows the JSD equation, where CT_A and CT_B are normalized vectors of the same Census dimension (e.g., race distribution) for two different Census tracts, $M = \frac{1}{2}(CT_A + CT_B)$ and x is a single attribute value in the dimensional vector X.

$$JSD(CT_A \parallel CT_B) = \sqrt{\frac{D(CT_A \parallel M) + D(CT_B \parallel M)}{2}} \tag{1}$$

$$D(CT_A \parallel M) = \sum_{x \in \mathcal{X}} CT_A(x) \log \left(\frac{CT_A(x)}{M(x)} \right) \tag{2}$$

The result of this technique is a set of singular values that quantify the similarity between two Census tracts based on five dimensions of ACS data. This process is repeated for all pairs of Census tracts producing five similarity matrices, one for each of the ACS dimensions. The JSD values are bounded between 0 (identical) and 1 (complete dissimilarity).

The next step involves merging the JSD values for each of these independent ACS dimensions into a single similarity value for each pair of Census tracts. This single value is the basis on which similarity is assessed by the user both in tabular format and also translated to a color density for cartographic visualization. Figure 2 presents a graphical overview of the process from ACS distributions to a single similarity value using two sample Census tracts *A* and *B*. Rather than average the five dimension-specific JSD values, we instead allow a user to determine the relative importance of each dimension to the overall similarity of the tracts. The relative importance is represented as a series of user-defined weights shown as sliders in the *MixMap* interface. A weight is assigned to each of the dimensions, with all five weights summing to 1. The exposure of these weights invites a user to refine the model to best meet their analytical requirements. Users often have different preferences, objectives, and exploration goals, and the opportunity for an individual or group to govern the similarity assessment process empowers the user, enhancing the usability of the tool.

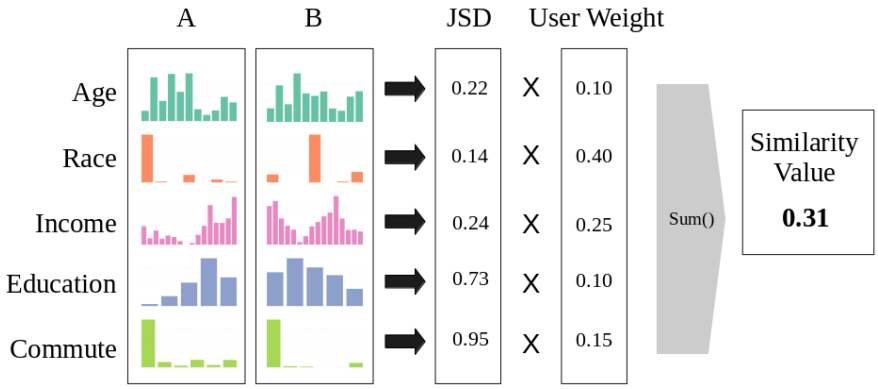


Figure 2: The process of computing similarity between two Census tracts *A* and *B*. Each Census tract consists of five dimensions of data, each a distribution of Census values (e.g., Ages 5-10, 11-15). JSD is calculated between each Census tract dimension individually, then multiplied by a user-defined weight, and finally summed to produce a single similarity value for the pair of Census tracts.

Design Guidelines

After identifying a technique for determining similarity between regions, we next focus on the design of a tool that implements this technique but also empowers a user with appropriate and relevant parameter tuning capabilities. Through our discussions with a civic engagement liaison, data visualization researchers, and referencing existing literature, we compiled the following set of four design guidelines (*RC2*) for the *MixMap* tool.

- *Configuring Similarity Characteristics.* The ability for a user to manually adjust the importance of each dimension individually places the user in control, allowing them to identify which aspects of the data matter most for their specific tasks.

- *Filter and Focus Geographies.* Such a tool should offer a user the ability to filter the geographies on which the analyses are conducted. This could either be a manual process of selecting the geographies of interest, or by filtering based on some social or Census attribute (e.g., population density) or physiographic property (e.g., regions on the coast).

- *Accessible Depth of Information.* The variety of use cases means that some users will be interested in a tabular representation and statistical details while others want to view a map and a bare minimum set of numbers. To accommodate a range of users, such a tool should offer users the ability to toggle the details provided to view data in either a tabular or map format.

- *Share and Collaborate.* The process of identifying similar and dissimilar regions through the adjustment of socio-demographic weights is inherently a collaborative process. The ability to share a configuration of weights as presets is essential to the usability of such a tool.

The MixMap Interface

With the previously mentioned design goals in mind, we developed the *MixMap* interface. A screen shot of the tool is shown in Figure 3 with alphabetical labels highlighting the various components of the tool (*RC3*). A live prototype of the tool is also available at *http://54.235.46.150/webapp/*.

The heart of the tool is the map interface (A), which cartographically depicts the similarity between a user-selected Census tract (orange) and all other Census tracts using an equal interval, density-based color palette. Hovering your mouse over a Census tract reports the similarity values in a *tool-tip* (B). The weights for each of the socio-demographic dimensions in the similarity model can be adjusted by a user through the sliders on the side panel (C). The cartographic styling of the map can be changed to a binary color palette showing the most and least similar census tracts (D). Previously saved data dimension weights (sliders) and location bookmarks can be loaded through the Presets widget (E). Finally, detailed information on the most and least similar census tracts is reported in natural language text and a sortable table in the bottom panel (F and G). A more detailed description of the tool is accessible via the *MixMap* tutorial video available on the prototype site.

Evaluation and Future Work

We conducted an evaluation of *MixMap* in order to collect qualitative feedback on the usefulness of such a tool and to identify limitations and future opportunities. A total of 18 participants were shown the interface containing ACS data at the Census tract level for the state of California. Participants completed a set of directed tasks as well as an open-ended exploration of the tool. Overall, participants were positive about their interactions with *MixMap*. Results suggest that participants found the parameter mixing to be intuitive in quickly exploring and understanding the effect of various parameters on the notion of place similarity. Participants were engaged in more sense-making behavior both during parameter tuning and when examining the system responses in the interface. Observations from the study helped to identify future directions of research and tool

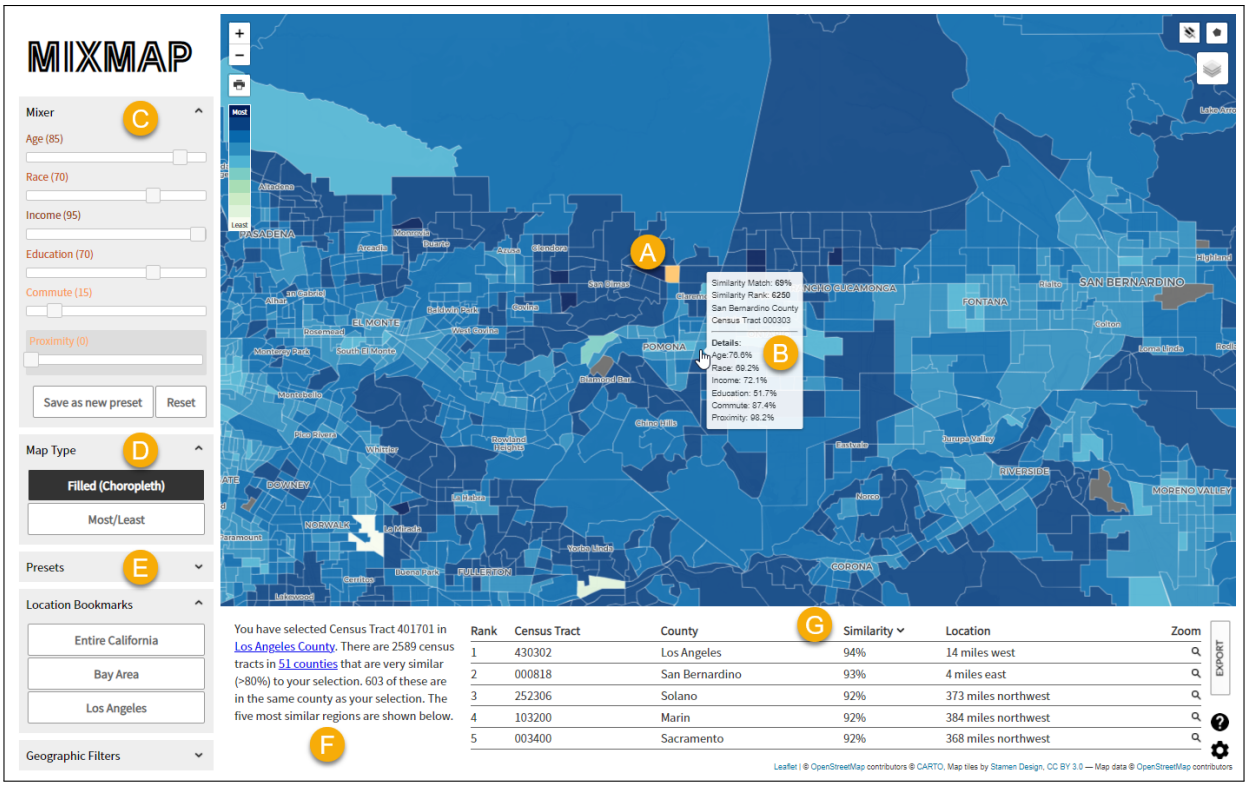


Figure 3: The *MixMap* interface showing similarity of US Census tracts in comparison to a selected location, highlighted in orange (A).

design, such as including additional datasets, support for more complex comparisons, and more detailed explanations on how similarity was determined.

Our next steps are to further investigate the recommendations of the evaluation participants and further engage with a variety of groups to determine the role that *MixMap* might play in their decision-making process. While a lot of interesting work remains to be done in the future, we believe that the insights learned from our work can identify unique opportunities for better understanding the nuances and semantics of comparing geospatial features for a variety of data-driven decisions. As quoted from Hofstadter (1979) - "to find similarities between situations despite differences which may separate them [and] to draw distinctions between situations despite similarities which may link them," may guide us towards more meaningful and intelligent analytical inquiry as we reason about places.

References

Adams, B. and Raubal, M. (2014). Identifying salient topics for personalized place similarity. *CEUR Workshop Proceedings*, 1142:1–12.

Hofstadter, D. R. (1979). *Godel, Escher, Bach: An Eternal Golden Braid*. Basic Books, Inc., USA.

Janowicz, K., Raubal, M., and Kuhn, W. (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, (2):29–57.

McKenzie, G., Janowicz, K., and Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, 41(2):125–137.

Nowosad, J. and Stepinski, T. F. (2021). Pattern-based identification and mapping of landscape types using multi-thematic data. *International Journal of Geographical Information Science*, 35(8):1634–1649.

Rosch, E. (1978). Principles of categorization. In Rosch, E., Lloyd, B., on Cognitive Research, S. S. R. C. U. C., Lloyd, B., and (U.S.), S. S. R. C., editors, *Cognition and Categorization*. John Wiley & Sons, Incorporated.

Slocum, T. A., MacMaster, R. B., Kessler, F. C., and Howard, H. H. (2009). *Thematic Cartography and Geovisualization, 3rd edition.* Pearson, Upper Saddle River, NJ, 3 edition.

U.S. Census Bureau (2019). 2019 American Community Survey 5-year estimates. http://data.census.gov.