MixMap: A user-driven approach to place-based semantic similarity

Grant McKenzie[a,b] and Sarah Battersby[b] and Vidya Setlur[c]

[a]Platial Analysis Lab, McGill University, Canada
[b]Tableau Research, Salesforce, Inc. Seattle, WA, USA
[c]Tableau Research, Salesforce, Inc. Palo Alto, CA, USA

ARTICLE HISTORY
Compiled March 28, 2025

ABSTRACT
What other locations are like my neighborhood? How? Why? The heart of many spatial analyses is in finding similarities or dissimilarities between locations. Discovering patterns and interpreting similarity is a complicated process that is based on both the spatial characteristics and the semantics or meaning that we assign to place. Human conceptualization of similarity in locations is multi-faceted and cannot be captured with a simple assessment of single numeric attributes like population density or median income; however, these quantifiable attributes are the basis for an initial pass of sense-making. *MixMap* facilitates the incorporation of similarity measures and spatial analytics to provide an information reduction (or semantic generalization) that brings the user closer to actionable insights. Through a preliminary evaluation of *MixMap*, we found that the tool supports the geospatial inquiry of determining similarity between regions, where participants can manipulate individual weights of the various attributes describing these locations. Based on feedback and observations from the study, we discuss potential implications and considerations for exploring the role of context and additional place-specific parameters for computing similarity, as well as understanding the nuances of semantics for place similarity in geospatial analysis tools.

KEYWORDS
Place, similarity, semantics, geovisualization, interactivity, data parameters, tool.

Corresponding author email: grant.mckenzie@mcgill.ca

1. Introduction

It has been estimated that 80% or greater of business datasets contain a spatial component (e.g., street address, latitude/longitude, state, country) (Garson, Biggs, & Biggs, 1992). The strong relationship of business data and location results in users frequently framing their questions and explorations around the spatial patterns in data. While many of these user queries and interactions tie to absolute locations, e.g., "how many customers are in California?", a wide range of important questions and avenues for exploration would benefit from additional flexibility in systems that are more in tune with semantics of place, e.g., "where should I expand my business given the success of our Sacramento store?" This is not just true for business-related questions; many decision-making opportunities involve the evaluation of relationships between locations to provide context. The relationship between locations may be the answer in itself - "what places are *like* this one," or the relationship may be a preliminary step in a larger analytic process - "what places are like this one so that I can use these locations in evaluating school district bussing policies." The key to this type of question is similarity. Quantifying similarity, however, is challenging. Locations are more than simply a count of attributes, and the ways in which people understand relationships between locations is strongly tied to the *character* or *semantic meaning* that we attach to the locations.

Sense-making about the world is often contextual; the relative importance of a location is based largely on how that location *compares* to other locations. The contextual evaluation is based on what is "similar" or "different" as well as a metric for *how* similar or different. These metrics may be based on ordinal interpretation of visual patterns, such as what regions are lighter or darker on a map, or quantitative metrics of indexed values representing multidimensional similarity scores.

Even with a well-designed map to visualize patterns in attributes, assessing the similarity between locations can be difficult. With a *single attribute* (e.g., percent of the population that is Black or African American, in Figure 1a), a reader can look for similarity between shades on the map as the indicator of similarity - i.e., all of the Census tracts shown in dark green could be considered similar. However, our understanding of regions and their relationships often depends on numerous attributes, and it is challenging to accurately identify the similarity between locations when the reader must visually interpret patterns and then mentally aggregate them to assess similarity. While there are methods for visualizing small numbers of variables on single maps (e.g., bivariate or trivariate choropleth maps), the complexity of larger numbers of variables, such as combining the layers in Figures 1a)-d), requires a different approach. Figure 1e)) shows the results of an approach that combines these different layers in a unique way. The methodology behind this approach is the focus of this paper and will be discussed in the sections that follow.

– Figure 1 near here –

There are numerous intertwined challenges in helping people identify and easily explore the similarity across multiple variables of interest. Outside of the general challenges of collecting appropriate data and calculating similarity, there is a broader issue of modeling similarity in a way that makes sense and allows people to tailor the calculation based on their intent. Spatial similarity as a concept is highly personal and influenced by what we can concretely measure, what we perceive about locations, and how we rate the importance of the individual elements used to assess similarity. Even for calculations using the same general inputs (e.g., the attributes exposed in *MixMap*), individuals may weight some of them to be more or less important than others when thinking about similarity.

Developing models for spatial and attribute similarity allows for improved recommendations – for possible amendments to a query (e.g., the region Los Angeles vs. the exact city boundary of Los Angeles in the query noted above) and also for expanding how we can guide users to related data. This process may be through recommendations of similar datasets or for analytical questions driven by the need to match characteristics of interest. For instance, if a company selects a region with a subset of its top donors to assess the characteristics and wants to find similar locations to target with their outreach or advertising activities. A place-similarity matrix can be used to recommend other regions or potential candidates with similar place characteristics (e.g., similar socioeconomic demographics, interests, etc.); these regions do not necessarily have to be near the original query location.

Contributions

This paper introduces *MixMap*, a tool that supports a user-driven approach for determining the similarity of geographic regions. Following a preliminary interview with a civic engagement and community organizing liaison and researcher, we identified a set of key design requirements for a place-based similarity tool. Given these requirements, we designed *MixMap* to meet the needs of a range of stakeholders, including community groups, data scientists, and urban planners. Specifically, our contributions are as follows:

- We developed an algorithm that computes a semantic similarity matrix for a selected geographic unit (e.g., block, tract, neighborhood) and a given set of attributes. For example: Specifying a region within Los Angeles would also recommend other regions in close proximity, but likely would also highlight regions such as Santa Cruz in Northern California, which are semantically similar but not within close proximity to one another.
- *MixMap* enables users to select an arbitrary location of interest and identify similar regions (in terms of characteristics of the location as well as the data points of interest within the newly created region). *MixMap* compares these user-defined search polygons to the underlying administrative geographic data (e.g., block, tract) to identify regions of interest.
- *MixMap* allows users to tune the similarity model by adding, removing, or re-weighting inputs to the model. These tuned parameters can be saved as a preset file for future analysis or sharing specific similarity calculations.
- An evaluation of the system provides useful insights for the development of semantic similarity interfaces for supporting geospatial inquiry involving place similarity.

2. Related Work

2.1. Geographic information retrieval

Place-based similarity analysis is useful in a range of applied domains. The task of searching for places similar to a selected location is one area that clearly benefits from the results of such analysis. Search related to geographic concepts, often referred to as Geographic Information Retrieval (GIR), has been a topic of research for decades, combining methods in data and information retrieval with the unique qualities of geospatial information. As much of the content with which we interact happens at, about, or in relation to geographic locations, having access to spatially-aware search is increasingly important (Purves, Clough, Jones, Hall, & Murdock, 2018). The difficulty often lies in the representation of geographic information. While data are often structured through digital gazetteers linking place names to coordinates, references to geographic content are increasingly identified in unstructured content such as images and text. A plethora of approaches has been developed to extract geographic content from unstructured text with the goal of identifying vague cognitive regions from geosocial content (S. Gao et al., 2017a), toponym disambiguation from housing advertisements (Hu, Mao, & McKenzie, 2019a), and even the development of a geospatial thematic search engine (Adams, McKenzie, & Gahegan, 2015a).

One of the emerging challenges in GIR is moving beyond a simple representation of location as geographic coordinates or a polygonal boundary to the multi-dimensional concept of *place* (Purves, Winter, & Kuhn, 2019). When describing a location on the Earth, one rarely references the geographic coordinates of a place, choosing to instead focus on the people, activities, and affordances (Jordan, Raubal, Gartrell, & Egenhofer, 1998; McKenzie & Adams, 2017). Researchers in this domain are taking a broader approach to understanding a location, stepping beyond the use of explicitly spatial information. For instance, McKenzie, Janowicz, Gao, Yang, and Hu (2015) leveraged the time of day that inhabitants of a city visit places of interest in order to better model the pulse of a city. Similar work by Silva, De Melo, Almeida, and Loureiro (2014) explored social behavior and participatory sensing of individuals to better understand city dynamics. Recent work has demonstrated that the mobility patterns of individuals between places are important to defining regions at multiple scales (Alessandretti, Aslak, & Lehmann, 2020).

The difficulty in quantifying places is often the complexity and nuance associated with its contributing dimensions. For example, place can be defined by its demographics, amenities, land classification, or any number of other features. This process makes places difficult to define in an absolute sense, with many choosing to instead define places in relation to other places. Similarity, therefore, becomes a key measure on which places are defined.

2.2. Similarity

Assessing similarity and categorizing like entities are fundamental ways in which people organize information (Rosch, 1978). While the process is often done unconsciously (Lakoff, 2008), people regularly and explicitly *seek* similarity between entities. This seeking process happens for everyday activities like finding a gas station in an unknown location and targeting the search near locations similar to where they are found in a known place (e.g., near the highway), as well as for more specialized searches like finding school districts with similar populations to compare educational policies.

Similarity is a key component of geospatial information retrieval (Adams & Martin, 2014) and plays an increasingly important role as the field of place-based data analytics continues to gain momentum. Similarity assessment is important, whether one considers the simple visual assessment of patterns (e.g., interpreting a map) (Slocum, MacMaster, Kessler, & Howard, 2009), or more complex mental or statistical processes where multiple, possibly fuzzy attributes are used to identify locations that are *similar*. As earlier work has noted, approximation and explanation of similarity values is an area in need of research (Janowicz, Raubal, & Kuhn, 2011). Numerous cognitive approaches to assessing similarity have emerged, with many recognizing that as similarity assessments are a key component in reasoning and induction, the user has a governing role as they select criteria by which to assess similarity (Holt, 1999). More recent research has explored the use of context-dependent, user-defined weights on natural language-based place similarity (Adams & Raubal, 2014). Janowicz, Adams, and Raubal (2010) also emphasize the value of human-weighted / adjusted components used in similarity calculations, highlighting the need for further research in this area. Our work addresses a portion of this need through the development of a tool that allows users to control the weights of the socioeconomic and demographic dimensions on which similarity is assessed. While the data on which *MixMap* determines similarity is predefined, the amount by which each of the different dimensions contributes to an overall similarity model can be adjusted by the user. This approach of user-defined weights has been employed successfully for other similarity-based visualization tasks (McKenzie, Janowicz, & Adams, 2014a; McKenzie & Romm, 2021).

2.3. Heuristics for similarity

While much of our visual interpreting spatial patterns is in identifying similarities and differences between the visual encoding of locations on a map, the broader concept of 'similarity' is more challenging to quantify than simply finding like-shades for locations on a choropleth map. With a single attribute of interest, we can use simple heuristics to evaluate, such as whether a location of interest has a higher or lower value than our target and what the numeric difference is between the locations. However, evaluating similarity is more complex than looking at simple, single attributes; the interplay between numerous characteristics of interests is often where people find the greatest value in assessing similarity between locations.

To make the process tenable, people use various heuristics and personal weightings of input characteristics to assess similarity. Murphy and Medin (1985) note that "the relative weighting of a feature (and the relative importance of common and distinctive features) varies with the stimulus context and task." (p. 296). While they further emphasize that there is not a single correct answer for how similar one location is to any other, it is safe to assume that the task of determining similarity *requires* some objective simplification in order to use it as an input in their decision-making. For the judgment of similarity, people simplify to rely on a subset of salient properties to reduce the complexity of the process (Adams & Martin, 2014; Tversky, 1977), for instance, focusing on a simplified, binary high- and low-value visual interpretation of a series of maps for identifying patterns across multiple attributes as in Figure 1.

As noted earlier, the mental combination of multiple attributes can be difficult, and the result is likely to be skewed based on human visual limitations if the inputs are in map format (e.g., Figure 1) or by conscious or subconscious biases that we have in our perceptions of place. If there is an expectation that a similarity measure is defensible or reproducible, we need objective methods for computing similarity measures.

2.4. Geographic Proximity

While the similarity between locations can be a stand-alone measure based solely on attributes of place, the interplay of distance in the calculation is also quite interesting. As Tobler (1970, p. 236) has noted, "Everything is related to everything else, but near things are more related than distant things." While geographic research (and our everyday lives) have shown repeated demonstrations that nearby locations tend to be more similar, this is not always a factor that is key to decision-making. For instance, proximity might be weighted higher when one is interested in finding locations with specific, similar socioeconomic characteristics to target a local store advertising campaign but de-emphasized when looking for the most similar school districts with respect to student and family characteristics to compare specific policies in place across the districts. To address this in *MixMap* we opted to incorporate a *proximity* element so that users could selectively weight the importance based on their intent. In Figure 2, we see the impact of geospatial proximity on the similarity of all census tracts in a region to a single selected census tract (shown in yellow). The color density of each census tract in Figure 2a)) is determined based purely on proximity to the selected census tract. In Figure 2c)), the color density of census tracts is based solely on the similarities of educational distributions to the selected census tract. In Figure 2b)), these two attributes are combined. Each contributes 50% to the final similarity value. The methodology will be discussed in Section 4.3.

– Figure 2 near here –

3. Formative Study and Design Guidelines

To better understand which capabilities would be useful for our *MixMap* tool, we conducted a formative interview with a community organizing liaison and researcher who works with various teams on topics related to civic engagement. In this formative interview, we discussed and elicited examples of intent and how they might be supported by *MixMap* via applied examples and modes of interaction. The interview allowed us to design the system with a set of users in mind, including those in community organizations, policy experts, marketing professionals, and advocacy groups.

This initial interview was semi-structured, involving a set of open-ended questions that took place over the span of an hour. All authors were present for the interview alternating between asking questions and taking notes. The interview took place via a video-conferencing application and was recorded and transcribed for later analysis. The interview began with the authors demonstrating a simple mapping interface that showed differences and similarities between regions using a variety of socio-economic and demographic data and was conducted using *Tableau Desktop* with data from the U.S. Census Bureau (2019). The interviewee was asked a series of questions about the usefulness of such an application to their domain, what they liked and disliked about the approach and recommendations for an analytical tool that aimed to identify similarities between regions across a range of variables.

3.1. Design Goals

Out of this formative interview and through referencing the existing literature on place-based similarity (see Section 2), we compiled the following set of four design goals *(DG)* for *MixMap*.

DG1 **Configuring Similarity Characteristics.** Given the broad range of users and use cases for identifying similarities between regions, the ability to adjust the weighting of different socio-economic and demographic variables was mentioned several times as a feature of key design importance. 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.

DG2 **Filter and Focus Geographies.** Such a tool should offer a user the ability to filter the geographies on which the analyses are conducted. This feature could either be a manual process of selecting the geographies of interest or filtering based on some social or census attribute (e.g., population density) or physiographic property (e.g., regions on the coast). A system should also simplify the process of comparing and focusing on specific regions by enabling the ability to jump between regions of interest.

DG3 **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 turn details on or off, view data either in tabular format or cartographically, and offer manual selection as an option to view more details rather than have all the information presented by default in one view.

DG4 **Share and Collaborate.** The process of identifying similar and dissimilar regions through the adjustment of socio-economic and demographic weights is inherently a collaborative process. Such a process may involve many members of a community advocacy group, city planners, or racial equity researchers. The ability to share a configuration of weights as presets is essential to the usability of such a tool. Similarly, users should have the option to export data for offline analysis and print maps for inclusion in policy documents and reports.

These design goals are not exhaustive and represent input from one practicing researcher and through referencing existing literature on this topic. The goals served to provide initial scaffolding on which the *MixMap* tool was designed and developed.

4. System

In the sections below, we provide an overview of the *MixMap* tool, present the datasets, our approach for determining similarity, and an overview of each of the elements of the user interface (Figure 3).

– Figure 3 near here –

4.1. Overview

The *MixMap* tool consists of two components: a front-end interactive web platform and a back-end data store. A set of PHP web handlers pass data between the two components based on requests from the web client (for instance, when a user selects a Census tract with their mouse). The front end is an interactive web map built using the Leaflet framework[1] and a series of DOM controls built using D3 and JQuery frameworks. The data are stored in a spatially-enabled (PostGIS) PostgreSQL database and are linked to Census tract geometries using a unique Census geographic identifier. The geographic boundaries for the Census tracts are stored as GeoJSON and layered onto the Leaflet base map on page load.

4.2. Data

Socio-economic and demographic data were accessed from the 2019 American Community Survey (ACS) 5-year estimates at the Census tract level for the State of California, USA. Based on our formative interview, the granularity of the Census tract was deemed a logical resolution for analysis. The scaffolding of the tool is geography agnostic, allowing for these regions to be swapped out in place of higher (e.g., Census Block Groups) or lower (e.g., County) resolution geographies, if needed.

The ACS data used as the basis for the tool comprises five dimensions: *Age*, *Race*, *Income*, *Educational Attainment*, and *Mode of Commuting*. Each of these dimensions is a distribution across a set of individual socio-economic or demographic attributes. For instance, in the *Age* dimension, we have estimates for the number of people aged 0-10, 10-15, 15-25, etc., for each of the Census tracts in California. A full set of attributes associated with each dimension is shown in Appendix A. In order to compare values across regions, we normalized all attributes within a dimension. This was done by dividing each value in the attribute dimension by the

[1] https://leafletjs.com

sum of all values. The result is a numerical vector that sums to one for each dimension in each Census tract. All of our ACS data used as the basis of our similarity approach are exclusive and complementary, meaning that normalization is acceptable.

In addition to ACS data, Euclidean distance was calculated between all pairs of Census tract geometry centroids in our data. This process was to allow users to control the influence of proximity in identifying similar regions. The population density was calculated for each Census tract, as well as a Boolean value indicating whether or not a Census tract is within 20 miles of the coast. Both of these attributes are intended to serve as examples of how external data can be incorporated within the tool to allow more refined filtering (see Section 4.4.3 - Geographic Filters Widget).

4.3. Defining similarity

Given the set of ACS data split into five dimensions, each containing a normalized vector of binned socio-economic or demographic values, we calculated the pairwise similarity between all Census tracts for each dimension separately. To accomplish this, we calculated the Jensen-Shannon Distance (JSD). JSD is a method for measuring the dissimilarity between two probability distributions. The measure uses a relative entropy approach for two distributions, based on the Kullback-Leibler divergence (KLD) (Equation 2) but varies from KLD in that it is symmetric and the resulting measure is finite. JSD has been used successfully in assessing similarity for a wide range of applications, from predicting aesthetic rankings (Jin et al., 2018) to differentiating how health content is queried (De Choudhury, Morris, & White, 2014). In the geographic domain, JSD has been used for tasks such as differentiating places of interest (McKenzie, Janowicz, & Adams, 2014b) and assessing land use patterns (Nowosad & Stepinski, 2021). The JSD equation is shown in Equation 1 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 results of this analysis are a set of singular values that quantify the similarity between two Census tracts based on our five distributions of ACS data. This process is repeated for all pairs of Census tracts producing five similarity matrices, one for each of the ACS dimensions. Though JSD values are bounded between 0 (identical) and 1 (complete dissimilarity), the actual range of JSD values depends on the underlying input distributions. These ranges vary considerably between dimensions, with some reporting a maximum JSD of 0.5 while others report 0.9. Since the end goal is to determine a single, aggregate value on which to visually represent the similarity between regions, we need a way to combine the individual dimension JSD values to represent a single Census tract. Simply averaging the values is a potential approach but the difference in JSD ranges means that even an equally-weighted approach would weigh certain dimensions more than others. To mitigate this issue, we first normalize JSD values for each Census tract compared to all other geographies. This generates a range of 0 – 1 for all dimensions in all geographies. Finally, we convert the dissimilarity values to similarities by subtracting each JSD value from one. These five matrices of normalized JSD values form the foundation of *MixMap*.

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 in tabular form and is also translated to a color density for visualization. Figure 4 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 provide an opportunity for a user to determine the impact each of the five dimensions has on the overall similarity of the tracts (***DG1***). This process is realized through a series of user-defined weights, presented as sliders (Figure 3 Item B) 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 adjust the model to best meet their analytical requirements. Users 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 4 near here –

4.4. User Interface

Upon launching the application, a user is presented with a map showing Census tracts for the state of California in uniform gray. Map labels showing neighborhoods, towns, and cities (depending on zoom level) are overlaid on top of the Census tracts as a reference layer. A vertical panel on the left side of the screen invites a user to *Select a Census tract* by clicking the map. When a Census tract is selected, the identifier of the tract is sent via a web handler to the database to return a JSON response, the pre-calculated five dimension JSD similarity values. Each of the five JSD values for each Census tract is then multiplied by the normalized user-defined weights (evenly weighted on page load) and summed to produce the single similarity value for each Census tract.

4.4.1. Map Panel

These values are then translated to the map using an equal interval choropleth color scheme and applied to the Census tract layer on the map. Darker blue values indicate higher similarity. Equal interval classification best represents the similarity data being visualized since the calculation of similarity involve ratio (or percentage) values. Tooltip functionality (Figure 3 Item C) allows a user to hover their mouse over each Census tract on the map and receive information containing the Census tract identifier, county name, similarity rank, and percentage of similarity match to the selected Census tract. Within the Settings Menu, a user has the option to enable *Additional Tooltip Details*, which adds the similarity values for each of the individual dimensions to the tooltip (***DG3***).

The Map Panel is the heart of the *MixMap* tool as it is the method by which a user selects a Census tract of interest (Figure 3 Item A), starting the *MixMap* process of cartographically and tabularly presenting the similarity between the selected region and all other regions within the dataset. Users also have the standard ability to interact and explore the map through zooming and panning. As shown in Figure 5, the user can also select the *Draw Polygon* tool, which enables them to manually draw a region on the map in order to limit the similarity assessment to a specified subset of interest (***DG2***).

– Figure 5 near here –

From the Map Panel, users can also print the map, change the base map from standard map tiles to satellite imagery, and toggle map labels, county boundary layer, and the main Census tract layer.

4.4.2. Tabular Panel

As the map updates to depict regional similarity, another panel is presented below the map that provides descriptive content related to the similarity analysis (***DG3***). The descriptive text in this panel presents the number of highly similar Census tracts, the number of counties in which they are found, and the number of similar Census tracts in the same county as the selection (Figure 3 Item D). The text is embedded with hyperlinks allowing a user to zoom into the top counties or the county of the selected Census tract. In addition, a table listing the top five most similar Census tracts, their percentage of similarity match, county name, and distance and direction from the selected Census tracts are shown (Figure 3 Item E). Users are invited to click on a row in the table to highlight the tracts on the map or zoom to the selected tract by selecting the magnifying glass icon. Within this table, clicking the column header for *Similarity* toggles between descending and ascending orders, allowing a user to easily identify the top most, and least, similar tracts.

4.4.3. Side Panel

Once a Census tract is selected, a side panel also emerges offering a range of interactive tools to enable data exploration and analysis. The side panel consists of a series of widgets, including the *Mixer*, *Map Types*, *Presets*, *Location Bookmarks*, and *Geographic Filters*.

Mixer Widget

The *Mixer* (Figure 3 Item B) is the basis for the *MixMap* name and provides interactive functionality through which a user adjusts the importance (weight) of a socio-economic or demographic dimension in the overall contribution to the similarity value. These weights are represented by sliders, inviting a user to increase the weight by moving a slider to the right and decreasing the weight by moving the slider to the left. By default, the mixers are evenly weighted at an importance value of 50. As a user adjusts the mixer, the color density of the dimension label changes, the numerical representation of the weight changes (bounded between 0 and 100), and the tooltip associated with the mixer updates to inform the user the impact that their adjustment is having on the overall similarity model.

While the first five mixers in this widget are socioeconomic and demographic dimensions of the Census data, the last mixer is not. This *Proximity* mixer adjusts the weight of the Euclidean distance between two Census tracts in the mix. By increasing the weight of proximity in the Mix, those Census tracts physically closer to the selected Census tract will be deemed more similar than those further away. Adjusting the value of the proximity mixer to 0 removes the influence of geographic proximity altogether.

Once a user has identified a combination of weights that is useful for their analysis, they have the option to save the mix to a new preset. This option updates the *Preset* widget, generates a preset XML file for download and presents a unique URL to share with collaborators.

Map Type Widget

By default, *MixMap* presents Census tract similarity using a gradient-based choropleth map (Figure 6a))). This representation is useful in many circumstances, but may not be the most appropriate cartographic visualization in others. For this reason, we designed an alternative map type option, namely *Most / Least* (Figure 6b))). This option simplifies the map, presenting the similarity of each region as one of three options, highly similar (blue), highly dissimilar (red), or somewhere between (gray). Users can further refine what *highly* means by selecting the top similar/dissimilar 1000, 100, or 10 Census tracts to the selected tract. This cartographic approach is particularly useful for those users that prefer a Boolean-type (similar or dissimilar) visualization instead of a gradient.

– Figure 6 near here –

Preset Widget

When a user has created a mix by adjusting the sliders in the Mixer widget, they can choose to label and save the mix as a preset, and it appears as a button in this Widget (Figure 3 Item G). Multiple presets can be created to represent various scenarios and enable different types of analysis. These presets are also saved to a preset XML file and stored on the server with a unique ID that can also be appended to the *MixMap* URL in order to share presets with collaborators (***DG4***). Should the users have created or shared a preset through a preset XML file in a previous session, they can also upload this file through the settings menu, automatically adding buttons to the widget that adjust the mixers.

Location Bookmarks Widget

The *Location Bookmark* widget stores locations of interest as buttons. Clicking a button zooms the map to a specified region (***DG2***). Three location bookmarks are added to the *MixMap* tool by default, but additional bookmarks can be added by uploading a preset XML file containing labels, geographic coordinates, and zoom levels.

Geographic Filters Widget

This widget allows users to filter the existing Census tracts through additional attributes. For instance, a user could decide that they are only interested in exploring Census tracts whose population density is less than 100 people per square mile (Figure 7), or only those Census tracts less than 20 kilometers from the coast (***DG2***). By default, the *MixMap* tool includes a sample set of additional variables such as those mentioned. Users can write their own SQL-type queries against these data, add them to the preset XML file, and upload them to the *MixMap* tool. When the file loads, only those Census tracts that meet the criteria specified in the query, will be displayed on the map. This is a powerful feature, allowing users to subset the data through external attribute filtering before digging into regional exploration and similarity analysis.

– Figure 7 near here –

5. Evaluation

We conducted an evaluation of *MixMap* with the following goals: (1) collect qualitative feedback on the usefulness of the tool for exploring similarity between geographic regions and (2) identify limitations and future opportunities. Because the main goal of our study was to gain qualitative insight into the algorithm's behavior, we encouraged participants to think aloud with the experimenter.

5.1. Method

Participants were assigned the *MixMap* interface with socio-economic and demographic data from the US Census 2019 American Community Survey (ACS) for Census tracts in California. Two staff members supported each session: one facilitator and one notetaker. Participants were first introduced to the study and asked about their backgrounds and role. They were then given instructions and spent most of the session interacting with the interface and observing the resulting visualizations and text responses. We then wrapped up the session during the last 5-10 minutes, getting their overall feedback about the prototype.

5.1.1. Participants

We recruited 18 volunteers (9 males, 9 females) from various roles across a data visualization company as well as research collaborators in related fields (e.g., spatial and/or socio-demographic data analytics). Participants were recruited on a first-come, first-serve basis. The participants had a variety of backgrounds: software engineers, technical writers, sales consultants, product managers, program managers, professors, and graduate students. Based on self-reporting, all were fluent in English and had basic experience with maps for navigation. 12 participants had experience with spatial analysis and visualization using tools such as *Tableau* (2021) and GIS software such as *Esri ArcGIS* (2021) and the open-source *QGIS* (2021). Four participants reported implementing map-related software engineering features on a development team. Two participants reported having limited experience with using maps and mainly used them for navigational applications and GPS tracking. We use the notation `P0X` to indicate participant IDs in the study results.

5.1.2. Procedure and Apparatus

All sessions were screen-recorded and audio-recorded. Field notes were expanded to a video log after the study through transcription of the videos. The video log (and raw video for reference) was then qualitatively coded to look for themes and trends. Due to COVID-19 social distancing protocols, all studies were completed virtually, and participants used their own computers. Participants used either the Chrome or Firefox browser to access *MixMap*. Each session took about 45 minutes and consisted of two parts. Participants were first shown a pre-recorded tutorial video explaining the *MixMap* interface and basic functionality. The study concluded with a short interview. The experimenter script, tutorial video, task descriptions, and dataset are included in the supplementary material.

Part 1: Closed-ended tasks

Closed-ended tasks were mainly intended to familiarize participants with the *MixMap* tool while also providing some consistent objectives for task comparison. For each task, participants were instructed to use *MixMap* to answer the questions but were not told how to do so. Participants were provided with a preset file that they loaded into *MixMap* for working through the tasks. Participants completed five closed-ended questions that included common visual analytic tasks, including:

Q1 *Similar and dissimilar tracts to coastal area tract*: To start, click on a Census tract on or near the coast. What are the top five most similar tracts? For the most similar, on what parameters are they most similar to? What are the five least similar tracts? For the least similar tract, on what parameters are they least similar?

Q2 *Geographic filter and location bookmarks*: Using the Geographic Filters widget, can you tell me what restrictions have been set on the geographies included in the map?

What Location Bookmarks are included? What do these bookmarks do? Why do you think some of the tracts are greyed out on the map?

Q3 *Similar tracts to Southern California*: With the preset XML file still in use, pick a location in Southern California. Where are other similar tracts located?

Q4 *Similar and dissimilar tracts to rural tract*: Now, let's reset the presets. Pick a rural (less populated) tract in Northern California. Where in California are the most similar tracts when compared to this selected tract? Based on what parameters are they most similar? On what parameters are they least similar? Where are the tracts that are most dissimilar?

Q5 *Manually restrict analysis area and save presets*: Manually restrict the analysis area to just those Census tracts in the San Francisco area. Which of these Census tracts in the selection are most and least similar to your selected Census tract? Adjust the sliders in the Mixer to emphasize categories of interest to you. Now save it as a new preset file.

Part 2: Open-ended exploration

Following the closed-ended tasks, participants completed an open-ended exploration task. This task enabled us to observe how participants would adjust weights, create and use preset files, and calculate similarity for focused views (i.e., geographic subsets) in a natural analysis flow. Instructions were: "Now that you have played with the prototype for a bit, please spend a few minutes exploring on your own. Remember to talk aloud as you interact with *MixMap*." The study concluded with an interview with the facilitator asking the participant the following questions:

Q6 Consider how the system responded to the interaction and chart construction. What was the most satisfying aspect of your experience? What would you suggest to improve?

Q7 Would you find *MixMap* helpful in your analytical workflow? How? Can you give me an example?

Q8 Do you have any other feedback on your experience? Do you have any questions for us?

5.1.3. Analysis Approach

We employed a mixed-methods approach involving qualitative and quantitative analysis but considered the quantitative analysis mainly as a complement to our qualitative findings. The primary focus of our work was a qualitative analysis of how the interface affordances in *MixMap* influenced people's analytical workflows. We conducted a thematic analysis through the open coding of session videos, focusing on the interaction behavior, strategies adopted, challenges using the tool, and insights gained by the participants. The quantitative analysis consisted of how often participants adjusted the sliders for the weights and whether they succeeded in completing the tasks or not. Given the remote nature of the study setup, we did not measure the time taken for task completion.

5.2. Results

We summarize people's reactions to the *MixMap* prototype and examine the impact of their behavior as participants interacted with the tool. Overall, participants were positive about their interaction with the tool and identified many benefits. All participants appreciated the responsiveness and interactivity of the interface as they experimented with the various weights, including viewing the updated details in the tooltips on the map - "This is more informative than a static demographic plot and I learned about the Census data more deeply. I can see being helpful for making policy decisions [*P*02]" and "It's nice to see options like age and race as they help me navigate the data with a small set of options as opposed to showing 100 different things in the data. I can get some preliminary conclusions very quickly [*P*10]." Participants were able to understand the provenance of the system behavior with the text explanations provided - "Really cool and the fact that it explained in natural language like this is what's going on, what you selected, these are the most dissimilar things. I thought that was really helpful. [*P*01]" and "Everything seems fairly easy to use and was self-explanatory [*P*17]". The choropleth map and the table provided an effective takeaway as to which places were most similar and dissimilar from each other. *P*07 commented, "It was really easy to get an immediate view of the most similar and likely similar areas, even when I was just exploring the data." The export option in the tool was also found to be useful to participants for saving their analyses for future exploration with *P*18 commenting, "I found it very satisfying to be able to make my own XML file. And so like basically have preset filters all saved in one place that I can use again."

Part 1: Closed-ended tasks

We describe how participants fared for each of the five tasks.

Q1 *Similar and dissimilar tracts to coastal area tract*: All participants were able to identify similar and dissimilar tracts compared to the coastal tract that they selected on the map. 15 of the 18 participants were able to find the parameters that the tracts were most similar or dissimilar to. Successful participants adopted one or more strategies such as using the tooltips to view information while hovering over tracts on the map (61.1% of the participants), referring to the table view (33.3%), examining the mixer setting (22.2%), and using the 'Most / Least' map feature (5.5%). Three participants struggled to find the parameters for the most similar but eventually used the table or tooltip to determine the parameters for the least similar tracts.

Q2 *Geographic Filter and Location Bookmarks*: All participants understood the functionality of these features and completed the task. The only exception was that P06 understood the general idea, but inverted the interpretation of the filter.

Q3 *Similar tracts to Southern California*: 17 of the 18 participants successfully completed this task, with a majority using the table (72%) to help answer the question. In addition, participants used the map (22.2%) and the 'Most / Least ' feature (11.1%).

Q4 *Similar and dissimilar tracts to rural tract*: 13 of the 18 participants successfully completed this task and adopted one or more strategies to help answer the question: using the table (61.1%), map (11.1%), and the 'Most / Least' feature (16.7%).

Q5 *Manually restrict analysis area and save presets*: All participants were successful in manually restricting the Census tracts to the San Francisco area and saving a preset XML file. Three participants needed some guidance to find the tool for manual restriction, but once located, were able to complete the task.

Part 2: Open-ended exploration

The open-ended task demonstrated how *MixMap* was helpful for exploring the concept of similarity in a geospatial dataset. All participants used the mixers to adjust the various parameter weights for computing similarity. A majority of the participants used the 'Most / Least' feature (77.8%), and 44.4% used the polygon selection tool to select a specific geographical area for computing similarity. 44.4% of the participants exported their analyses and saved their parameter weights as a preset XML file. Some of the participants had specific analytical questions they wanted to explore. For example, *P*01 interacted with *MixMap* with socio-demographic questions around race, while others interacted with the mixers with an assorted set of questions in mind (*P*06, *P*11, *P*13, *P*14).

The study also revealed several shortcomings. While most participants were able to complete the tasks and found the mixings useful for exploring similarity and dissimilarity measures, others struggled to understand why certain tracts showed "no data" (e.g., airports or large parks) when trying to find similar and dissimilar tracts. Also, there were issues around feature discoverability, particularly with the 'Most / Least' feature and interacting with the summary table. Some participants did not understand how the mixer settings were used to compute the similarity scores for places and often relied on the visual feedback in the choropleth map.

6. Discussion and Future Work

An evaluation of *MixMap* confirmed our intuition that people found the tool useful for performing similarity analysis for geospatial data. 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 help identify the following opportunities for research and tools to help aid the understanding of the semantics of place similarity during exploratory data analysis:

Data enrichment

While participants appreciated scoping down the data attributes to a handful of parameters, they also expressed the need for adding additional data and parameters for computing place similarity - "I'd like to layer Census data with other types of data like a violent crime database and things like that, exploring additional factors for similarity [*P*01]" and "There are many more things in the Census like age, race, income, education, community proximity that I'd like to import and explore with [*P*02]." Importing predefined geospatial geometry could also be useful - "Having the ability to import catchments would be kind of cool and then being able to look at other store catchments. The catchments would be like little cookie cutters of pre-selected locations. [*P*14]"

UI enhancements

Participants provided feedback for adding additional enhancements to the *MixMap* interface. The tool expects users to directly click on a region of interest to compare it against other regions. However, when the geography is unfamiliar to the user, having some form of navigational help would be useful. *P*10 stated, "I think it would be nice if there was a search box where I could enter 'Santa Barbara'; that's where we went this summer. I don't think I can find it now on the map easily though." *P*01 suggested, "I'd like to be able to split the screen where I can have San Francisco on one side and see a different part of the map on the other." Several participants also requested enrichment in the tooltips to visualize distributions of attributes (e.g., age distributions) to support their understanding of the underlying factors in calculating similarity. Other UI enhancements include provisioning a back button for the user to find the previously selected location [*P*07] and being able to orient the user better based on their map selection, "when I select something that's really far away from where I'm zoomed in right now, having some sort of inset that helps orient me - this is where you're looking at on the map; this is where you selected, would be helpful. [*P*09]"

Provenance of similarity computation

Balancing the complexity of the underlying system with a simple and intuitive user experience can be challenging. Participants were unclear about how exactly the similarity measure was computed from the individual weights. Comments referencing the lack of system provenance for the similarity model included, "This similarity number, I'm not sure I understand that. A tooltip or an advanced option that shows the math would help [*P*07]." Future work should explore ways to intuitively communicate the underlying algorithm to help the user's understanding and mental model of how the system works.

Impact of parameter mixes on similarity maps

By default, the mixers are set to a flat mix, i.e., all five JSD dimensions are equally weighted. Adjusting the mix can have a significant impact on the resulting Census tract similarity rankings and map. For instance, if one selects the Census tract in which Stanford University is situated and adjusts the mix to heavily weight *income*, we see a substantially different similarity map than a mix that weights *education* quite heavily. While most socio-economic analysis finds that income and educational attainment are highly correlated (Charles & Hurst, 2003; Tinbergen, 1972), analysis through this platform demonstrates that there are regions of the United States where these two variables differ spatially. Future work should explore how this can be incorporated into human-in-the-loop interfaces to aid in understanding outliers or other elements of data that may skew model results.

Automate preset recommendations

While we focus on providing presets for general types of place similarity questions, extending these presets as recommendations to the user would be an interesting direction to pursue. Real-world location histories imply, to some extent, users' interests in places and bring us opportunities to understand the correlation between users and these places. Personalized parameter mixes can be recommended based on previous search history, properties of the underlying data, and common user preferences using techniques such as collaborative filtering.

Support for more complex comparisons

Place similarity can be nuanced and complex due to the semantic heterogeneity of place types for a variety of analytical tasks such as social sensing, urban planning, and other policy-making decisions (Janowicz, McKenzie, Hu, Zhu, & Gao, 2019). Extending *MixMap* to support additional perspectives of place similarity based on population density, temporal patterns, and local government regulations are interesting research directions to pursue. As *P*04 remarked, "I want to be able to study areas to compare how people get weather forecasts in different areas such as rural versus urban. This is particularly important in rural places. We've been talking a lot about migrant farm workers. They're not going to be captured by the Census data, but they are the most vulnerable population to climate change." Other ways of making comparisons more accessible would be through language - " I was talking to LA Sanitation, and they wanted to understand the prevalence of disease vectors in homeless encampments. So nearness and similar terms would be incredibly helpful in that analysis, so it's not the data scientists on their staff who are able to answer those questions [*P*05]." Finally, the ability to interpret arbitrary geographic definitions enables more expressive ways of specifying comparisons. *P*07 explained, “Can you find things that are like this (pointing to a user-defined selection on the map) because there are a lot of interesting, complex questions that come with the variability inside the selection.”

7. Conclusion

In this paper, we present a tool, *MixMap* that supports a user-driven approach for determining the similarity of geographic regions during an analytical workflow. Users are able to select an arbitrary location of interest. *MixMap* then compares the socio-economic and demographic characteristics of the region to these characteristics in all other geographic regions in a given administrative area, with the goal of identifying similar and dissimilar locations. *MixMap* allows users to tune the parameters of the similarity model that can be saved as a preset file for future analysis. A preliminary evaluation of the tool validated our premise that providing intuitive, configurable affordances for exploring place similarity can help users make relevance judgments on the geographic features that they are comparing against. 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, p. 26) - “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., & Martin, R. (2014, 01). Identifying salient topics for personalized place similarity. *CEUR Workshop Proceedings*, *1142*, 1-12.

Adams, B., & McKenzie, G. (2012, 08). Inferring thematic places from spatially referenced natural language descriptions. In (p. 201-221).

Adams, B., McKenzie, G., & Gahegan, M. (2015a). Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th international conference on world wide web* (pp. 12–22).

Adams, B., McKenzie, G., & Gahegan, M. (2015b). Frankenplace: Interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th international conference on world wide web* (p. 12–22). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. Retrieved from `https://doi.org/10.1145/2736277.2741137`

Adams, B., & Raubal, M. (2014). Identifying salient topics for personalized place similarity. *Research@ Locate*, *14*, 1–12.

Alessandretti, L., Aslak, U., & Lehmann, S. (2020). The scales of human mobility. *Nature*, *587*(7834), 402–407.

Charles, K. K., & Hurst, E. (2003). The correlation of wealth across generations. *Journal of political Economy*, *111*(6), 1155–1182.

Cocos, A., & Callison-Burch, C. (2017, April). The language of place: Semantic value from geospatial context. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 99–104). Valencia, Spain: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/E17-2016`

De Choudhury, M., Morris, M. R., & White, R. W. (2014, Apr). Seeking and sharing health information online: comparing search engines and social media. In *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1365–1376). New York, NY, USA: Association for Computing Machinery.

Esri ArcGIS. (2021). `https://www.esri.com/en-us/arcgis/about-arcgis/overview`.

Eynard, D., Inversini, A., & Gentile, L. (2012). Finding similar destinations with flickr geotags. In *Proceedings of the 27th annual acm symposium on applied computing* (p. 733–736). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2245276.2245415`

Fu, C., & Weibel, R. (2020). Towards measuring place function similarity at fine spatial granularity with trajectory embedding. *arXiv: Artificial Intelligence*.

Gao, S., Janowicz, K., Montello, D., Hu, Y., Yang, J.-A., McKenzie, G., ... Yan, B. (2017b, 01). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, *31*, 1-27.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., ... Yan, B. (2017a). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, *31*(6), 1245–1271.

Gao, T., Dontcheva, M., Adar, E., Liu, Z., & Karahalios, K. G. (2015). Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th annual acm symposium on user interface software &; technology* (p. 489–500). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2807442.2807478`

Garson, G. D., Biggs, R. S., & Biggs, R. S. (1992). *Analytic mapping and geographic databases* (No. 87). Sage.

Goodchild, M. F. (2011). Formalizing place in geographic information systems. In L. M. Burton, S. A. Matthews, M. Leung, S. P. Kemp, & D. T. Takeuchi (Eds.), *Communities, neighborhoods, and health: Expanding the boundaries of place* (pp. 21–33). New York, NY: Springer New York. Retrieved from `https://doi.org/10.1007/978-1-4419-7482-2`$_2$

Herskovits, A. (1997). Language, spatial cognition, and vision. In O. Stock (Ed.), *Spatial and temporal reasoning* (pp. 155–202). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-0-585-28322-7$_6$

Hockenberry, M., & Selker, T. (2006). A sense of spatial semantics. In *Chi '06 extended abstracts on human factors in computing systems* (p. 851–856). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1125451.1125618

Hofstadter, D. R. (1979). *Godel, escher, bach: An eternal golden braid.* USA: Basic Books, Inc.

Holt, A. (1999). Spatial similarity and gis: the grouping of spatial kinds. In *Eleventh annual colloquium of the spatial information research center (sirc05)* (pp. 241–250).

Hoque, E., Setlur, V., Tory, M. K., & Dykeman, I. (2018). Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *24*, 309-318.

Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., & Hitzler, P. (2013). A linked-data-driven and semantically-enabled journal portal for scientometrics. In H. Alani et al. (Eds.), *The semantic web – iswc 2013* (pp. 114–129). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hu, Y., Mao, H., & McKenzie, G. (2019a). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, *33*(4), 714–738.

Hu, Y., Mao, H., & McKenzie, G. (2019b). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, *33*, 714 - 738.

Janowicz, K., Adams, B., & Raubal, M. (2010). Semantic referencing–determining context weights for similarity measurement. In *International conference on geographic information science* (pp. 70–84).

Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., & Gao, S. (2019, 01). Using semantic signatures for social sensing in urban environments. In (p. 31-54).

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

Jin, X., Wu, L., Li, X., Chen, S., Peng, S., Chi, J., ... Zhao, G. (2018, Apr). Predicting Aesthetic Score Distribution Through Cumulative Jensen-Shannon Divergence. *AAAI*, *32*(1). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/11286

Jordan, T., Raubal, M., Gartrell, B., & Egenhofer, M. (1998). An affordance-based model of place in gis. In *8th int. symposium on spatial data handling, sdh* (Vol. 98, pp. 98–109).

Kanza, Y., Kravi, E., Safra, E., & Sagiv, Y. (2017, July). Location-based distance measures for geosocial similarity. *ACM Trans. Web*, *11*(3). Retrieved from https://doi.org/10.1145/3054951

Kim, J., Vasardani, M., & Winter, S. (2017, January). Similarity matching for integrating spatial information extracted from place descriptions. *Int. J. Geogr. Inf. Sci.*, *31*(1), 56–80.

Lakoff, G. (2008). *Women, fire, and dangerous things: What categories reveal about the mind.* University of Chicago press.

Leaflet JS. (2021). https://leafletjs.com/.

Mai, G., Janowicz, K., Hu, Y., & McKenzie, G. (2016). A linked data driven visual interface for the multi-perspective exploration of data across repositories. In *Voila@iswc.*

McKenzie, G., & Adams, B. (2017). Juxtaposing thematic regions derived from spatial and platial user-generated content. In *13th international conference on spatial information theory (cosit 2017).*

McKenzie, G., Battersby, S., & Selter, V. (2022). Mixmap: Exploring user-driven semantic similarity of places. In *Proceedings of the 24th international research symposium on cartography and giscience (autocarto 2022.*

McKenzie, G., & Hu, Y. (2017). The "nearby" exaggeration in real estate. In *Proceedings of the cognitive scales of spatial information workshop (cossi 2017), l'aquila, italy* (pp. 4–8).

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

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

McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A., & Hu, Y. (2015). Poi pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *50*(2), 71–85.

McKenzie, G., & Romm, D. (2021). Measuring urban regional similarity through mobility signatures. *Computers, Environment and Urban Systems*, *89*, 101684.

Montello, D. R., Friedman, A., & Phillips, D. W. (2014, September). Vague cognitive regions in geography and geographic information science. *Int. J. Geogr. Inf. Sci.*, *28*(9), 1802–1820. Retrieved from `https://doi.org/10.1080/13658816.2014.900178`

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

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

Ostermann, F., Huang, H., Andrienko, G., Andrienko, N., Capineri, C., Farkas, K., & Purves, R. (2015, 08). Extracting and comparing places using geo-social media. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3/W5*, 311-316.

Patroumpas, K., & Skoutas, D. (2020). Similarity search over enriched geospatial data. In *Proceedings of the sixth international acm sigmod workshop on managing and mining enriched geospatial data.* New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3403896.3403967`

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, *12*(2-3), 164–318.

Purves, R. S., Winter, S., & Kuhn, W. (2019). Places in information science. *Journal of the Association for Information Science and Technology*, *70*(11), 1173–1182.

QGIS. (2021). `https://www.qgis.org`.

Regalia, B., Janowicz, K., & McKenzie, G. (2019). Computing and querying strict, approximate, and metrically refined topological relations in linked geographic data. *Transactions in GIS*, *23*(3), 601–619.

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

Setlur, V., Battersby, S. E., Tory, M., Gossweiler, R., & Chang, A. X. (2016). Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th annual symposium on user interface software and technology* (p. 365–377). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2984511.2984588`

Silva, T. H., De Melo, P. O. V., Almeida, J. M., & Loureiro, A. A. (2014). Large-scale study of city dynamics and urban social behavior using participatory sensing. *IEEE Wireless Communications*, *21*(1), 42–51.

Slocum, T. A., MacMaster, R. B., Kessler, F. C., & Howard, H. H. (2009). *Thematic cartography and geovisualization, 3rd edition.* (3rd ed.). Upper Saddle River, NJ: Pearson.

Tableau Software. (2021). `https://tableau.com`.

Tinbergen, J. (1972). The impact of education on income distribution. *Review of Income and wealth*, *18*(3), 255–265.

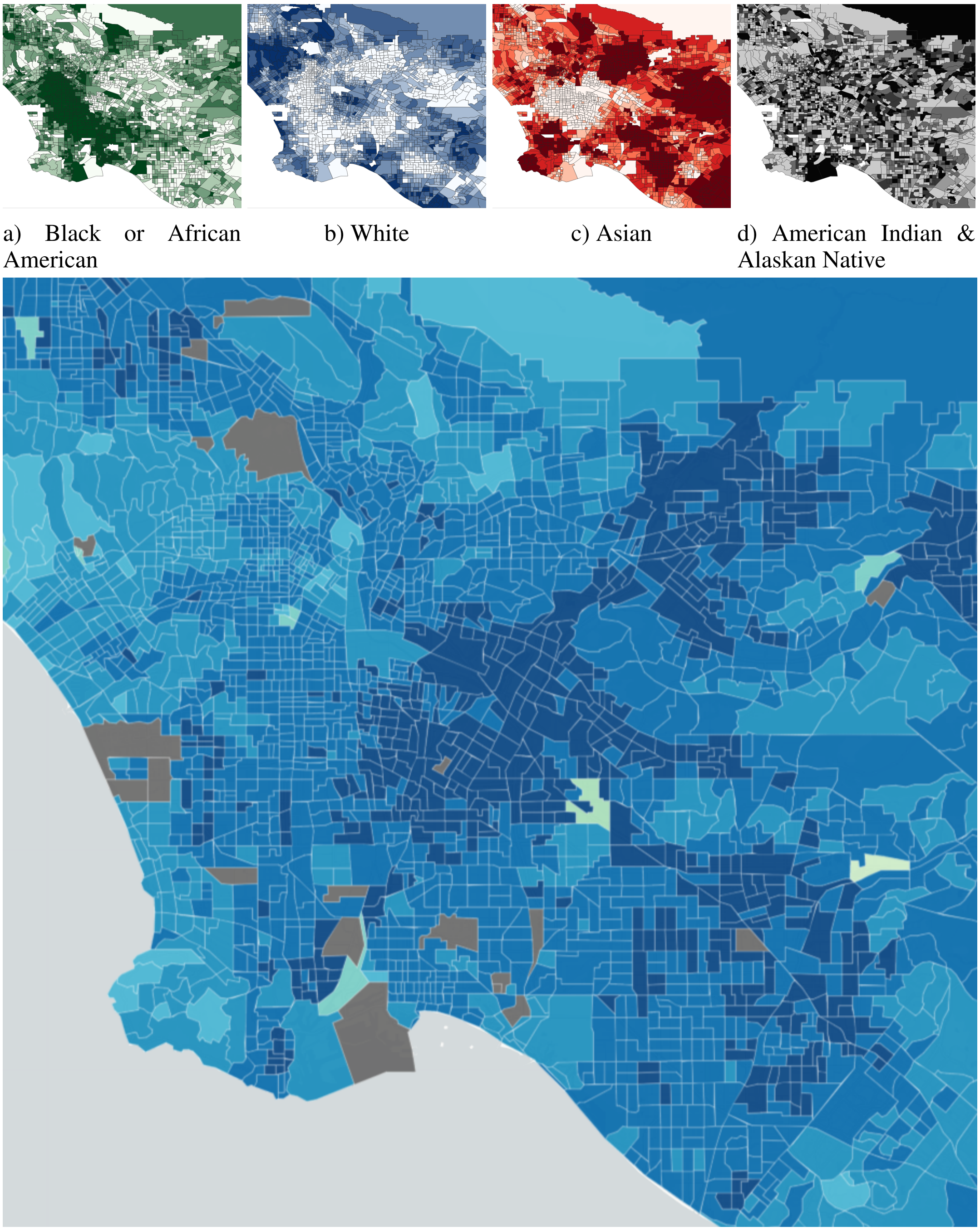
Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*(sup1), 234–240.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

U.S. Census Bureau. (2019). *2019 American Community Survey 5-year estimates.* Retrieved 2021-11-24, from `http://data.census.gov`

Wang, X., Zhao, Y.-L., Nie, L., Gao, Y., Nie, W., Zha, Z.-J., & Chua, T.-S. (2015). Semantic-based location recommendation with multimodal venue semantics. *IEEE Transactions on Multimedia*, *17*(3), 409-419.

Figure Captions



a) Black or African American

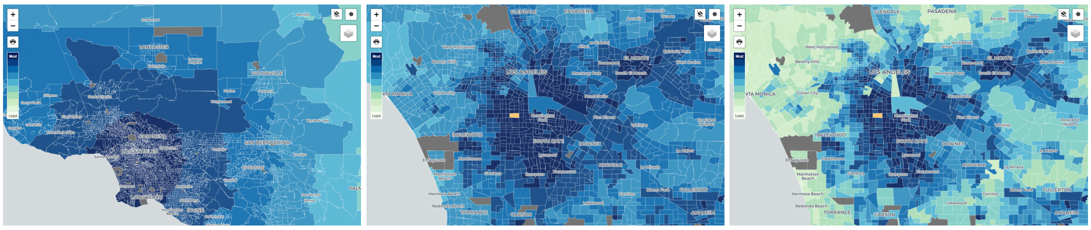
b) White

c) Asian

d) American Indian & Alaskan Native

e) A single similarity measure comparing race distributions.

Figure 1: Maps showing the relative percent of different races in a location of interest (top) would need to be mentally combined to determine overall similarity; however, combining these into a single similarity measure (bottom) makes the pattern easier to interpret.

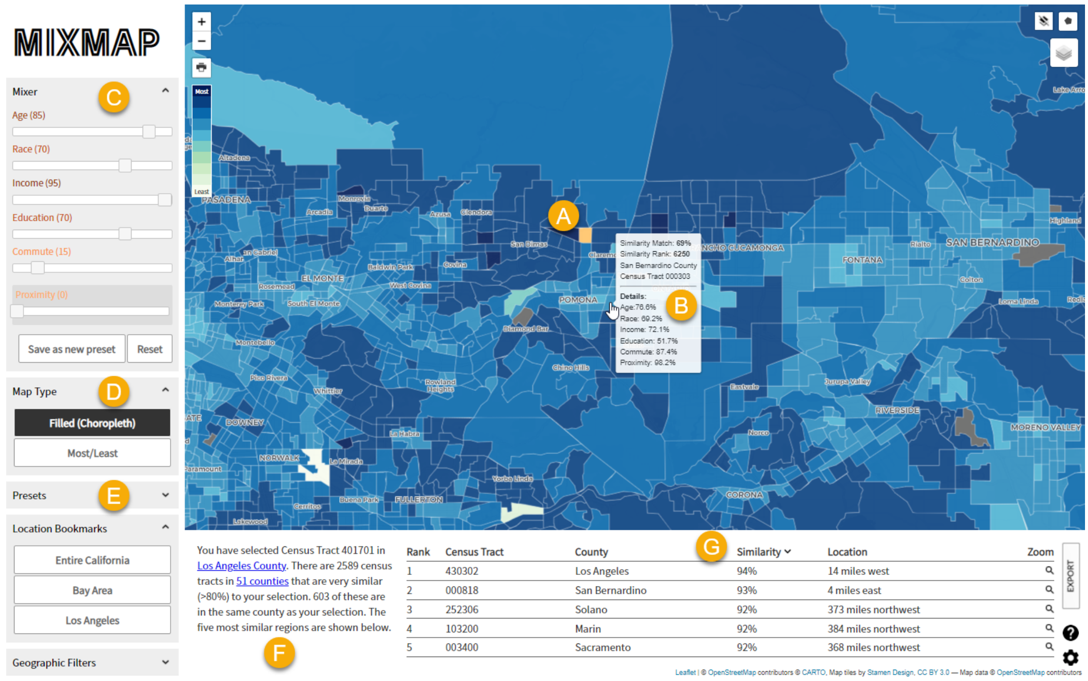


a) Proximity only

b) Proximity and Education

c) Education only

Figure 2: The impact of proximity on similarity. The maps show a) proximity as an attribute by itself, b) proximity and educational attainment equally weighted, and c) educational attainment by itself.



Mixer

Age (85)

Race (70)

Income (95)

Education (70)

Commute (15)

Proximity (0)

Save as new preset | Reset

Map Type

Filled (Choropleth)

Most/Least

Presets

Location Bookmarks

Entire California

Bay Area

Los Angeles

Geographic Filters

Similarity Match: 69%
Similarity Rank: 6250
San Bernardino County
Census Tract 000303

Details:
Age:76.6%
Race: 69.2%
Income: 72.1%
Education: 51.7%
Commute: 87.4%
Proximity: 98.2%

You have selected Census Tract 401701 in Los Angeles County. There are 2589 census tracts in 51 counties that are very similar (>80%) to your selection. 603 of these are in the same county as your selection. The five most similar regions are shown below.

Rank	Census Tract	County	Similarity	Location	Zoom
1	430302	Los Angeles	94%	14 miles west	🔍
2	000818	San Bernardino	93%	4 miles east	🔍
3	252306	Solano	92%	373 miles northwest	🔍
4	103200	Marin	92%	384 miles northwest	🔍
5	003400	Sacramento	92%	368 miles northwest	🔍

Leaflet | © OpenStreetMap contributors © CARTO, Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 3: The *MixMap* interface showing similarity of US Census tracts in comparison to a selected location, highlighted in orange (A). Details on similarity are included in a tooltip (B). *MixMap* allows users to weight specific characteristics of interest (C) to tailor the calculation of similarity according to the user's question and interest, provides multiple map types for visualization (D), as well as features to allow users to save and re-use preset files (E) with specific mixer settings and filters. The interface provides interactive text description (F) and a sortable summary table (G).

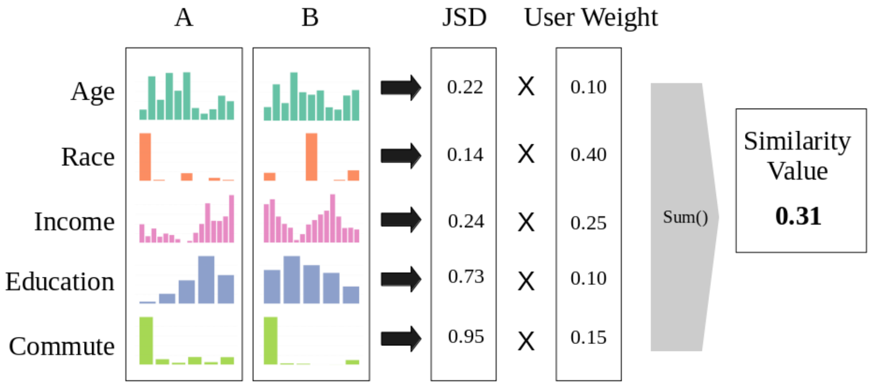
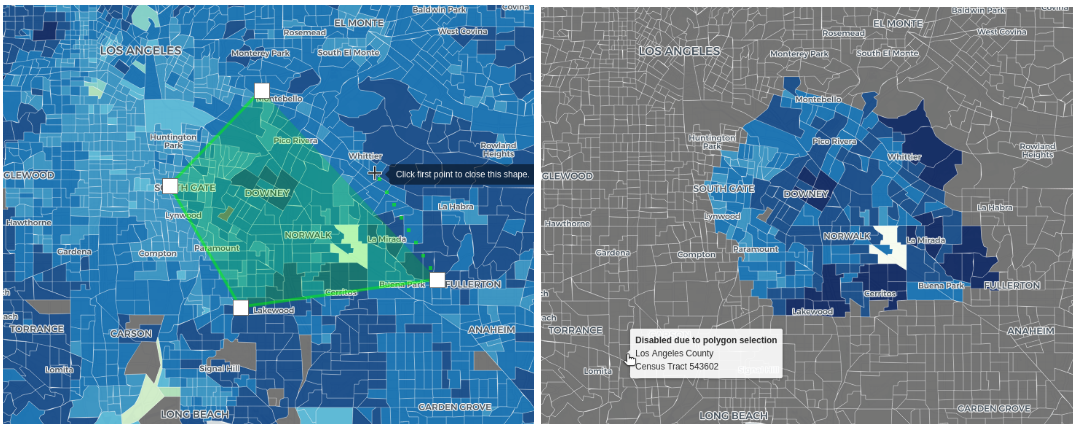


Figure 4: The process of computing similarity between two Census tracts *A* and *B*. Each Census tract consists of five dimensions of data, each a distribution of Census values. JSD is calculated between each Census tract dimension individually, then multiplied by a user-defined weight, and finally summed to produce a single similarity value for the pair of Census tracts.



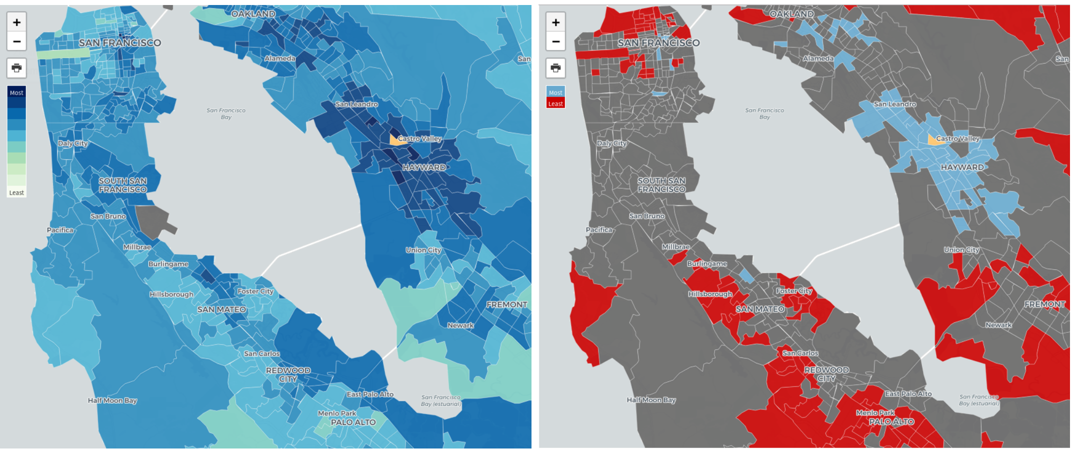
Click first point to close this shape.

Disabled due to polygon selection
Los Angeles County
Census Tract 543602

a) Draw polygon functionality

b) Subselection of Census tracts

Figure 5: The *Draw Polygon* tool allows a user to subselect a set of Census tracts for analysis. Those census tracts that intersect with the drawn polygon are selected. The tooltip shown in b) informs a user that a census tracts outside of the selected region is disabled.



a) Equal interval choropleth map

b) Most / Least color palette

Figure 6: Differing cartographic visualizations for the same underlying similarity data

Figure 7: The results of enabling a geographic filter restricting the analysis to only those Census tracts with a population density of less than 100 people per square mile

Appendix A. Census Variables

Table A1.: Variables from the US Census 2019 American Community Survey 5-year data

Variable Name	Variable Label	Census Concept
B06001_001E	Estimate!!Total:	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_002E	Estimate!!Total:!!Under 5 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_003E	Estimate!!Total:!!5 to 17 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_004E	Estimate!!Total:!!18 to 24 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_005E	Estimate!!Total:!!25 to 34 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_006E	Estimate!!Total:!!35 to 44 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_007E	Estimate!!Total:!!45 to 54 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_008E	Estimate!!Total:!!55 to 59 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_009E	Estimate!!Total:!!60 and 61 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_010E	Estimate!!Total:!!62 to 64 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_011E	Estimate!!Total:!!65 to 74 years	PLACE OF BIRTH BY AGE IN THE UNITED STATES
B06001_012E	Estimate!!Total:!!75 years and over	PLACE OF BIRTH BY AGE IN THE UNITED STATES
DP05_0037PE	Percent!!RACE!!Total population!!One race!!White	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
DP05_0038PE	Percent!!RACE!!Total population!!One race!!Black or African American	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
DP05_0039PE	Percent!!RACE!!Total population!!One race!!American Indian and Alaska Native	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
DP05_0044PE	Percent!!RACE!!Total population!!One race!!Asian	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
DP05_0052PE	Percent!!RACE!!Total population!!One race!!Native Hawaiian and Other Pacific Islander	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
DP05_0057PE	Percent!!RACE!!Total population!!One race!!Some other race	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
DP05_0058PE	Percent!!RACE!!Total population!!Two or more races	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
B19001_001E	Estimate!!Total:	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_002E	Estimate!!Total:!!Less than $10,000	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_003E	Estimate!!Total:!!$10,000 to $14,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_004E	Estimate!!Total:!!$15,000 to $19,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_005E	Estimate!!Total:!!$20,000 to $24,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_006E	Estimate!!Total:!!$25,000 to $29,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_007E	Estimate!!Total:!!$30,000 to $34,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_008E	Estimate!!Total:!!$35,000 to $39,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_009E	Estimate!!Total:!!$40,000 to $44,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_010E	Estimate!!Total:!!$45,000 to $49,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_011E	Estimate!!Total:!!$50,000 to $59,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_012E	Estimate!!Total:!!$60,000 to $74,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_013E	Estimate!!Total:!!$75,000 to $99,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_014E	Estimate!!Total:!!$100,000 to $124,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_015E	Estimate!!Total:!!$125,000 to $149,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_016E	Estimate!!Total:!!$150,000 to $199,999	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B19001_017E	Estimate!!Total:!!$200,000 or more	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)
B06009_001E	Estimate!!Total:	PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES
B06009_002E	Estimate!!Total:!!Less than high school graduate	PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES
B06009_003E	Estimate!!Total:!!High school graduate (includes equivalency)	PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES
B06009_004E	Estimate!!Total:!!Some college or associate's degree	PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES
B06009_005E	Estimate!!Total:!!Bachelor's degree	PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES
B06009_006E	Estimate!!Total:!!Graduate or professional degree	PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES
DP03_0019PE	Percent!!COMMUTING TO WORK!!Workers 16 years and over!!Car, truck, or van – drove alone	SELECTED ECONOMIC CHARACTERISTICS
DP03_0020PE	Percent!!COMMUTING TO WORK!!Workers 16 years and over!!Car, truck, or van – carpooled	SELECTED ECONOMIC CHARACTERISTICS
DP03_0021PE	Percent!!COMMUTING TO WORK!!Workers 16 years and over!!Public transportation (excluding taxicab)	SELECTED ECONOMIC CHARACTERISTICS
DP03_0022PE	Percent!!COMMUTING TO WORK!!Workers 16 years and over!!Walked	SELECTED ECONOMIC CHARACTERISTICS
DP03_0023PE	Percent!!COMMUTING TO WORK!!Workers 16 years and over!!Other means	SELECTED ECONOMIC CHARACTERISTICS
DP03_0024PE	Percent!!COMMUTING TO WORK!!Workers 16 years and over!!Worked from home	SELECTED ECONOMIC CHARACTERISTICS