International Cartographic Association
Association Cartographique Internationale

Putting mapper on a map: cartographic visualizations of topological data analysis

Jim Thatcher [a,*], David Retchless [b], Courtney Thatcher [c], Kristine Jones [d]

[a] *University of Washington Tacoma, jethatch@uw.edu*
[b] *University of Texas Galveston, retchled@tamug.edu*
[c] *University of Puget Sound, cthatcher@pugetsound.edu*
[d] *University of Washington, kristine.elizabeth.jones@gmail.com*

* Corresponding author

Keywords: Topological Data Analysis, Geovisualization, Cartographic Thought

Abstract:

Topological data analysis (TDA) combines approaches from mathematical topology with computational methods and has emerged as a promising area of research within mathematics and data science (Carlsson 2009). TDA methods have, in general, been shown to be particularly useful in data exploration through clustering and feature extraction (Chazal and Michel 2017); the mapper algorithm has been used in medical studies to identify individuals with certain types of breast cancer (Niolau et al. 2011) as well as subgroups of type 2 diabetics (Li et al. 2015). However, due to the nature of mapper's clustering and graph output, visualizing mapper data in a geographic context has been limited up to this point (although, see Palma 2020). Using U.S. census data, this abstract presents a novel approach for the cartographic visualization and analysis of clusters found when using mapper on spatially-bound datasets. First, we define mapper and explain the datasets chosen; second, we present the resulting 'communities' that emerged through TDA analysis; finally, we discuss how this approach can be integrated within and tested against existing methods of spatial analysis and data visualization. The mapper algorithm offers ongoing potential in the analysis of spatially-bound data, particularly in its ability to suggest non-deterministic clusters within data sets - something of particular interest to cartographers exploring more relational forms of spatial visualization (Bergmann and Lally 2021).

The mapper algorithm (Singh et al. 2007) takes a data set and returns a 1-dimensional representation of it in the form of a graph. Nodes in the graph are collections of "similar" data points, and edges exist between nodes that share common data points. Specifically, mapper takes a point cloud X with a metric (a way to measure the distance between points), and maps it to a lower dimension via a filter function, $f : X \rightarrow R^n$, for some small integer n (often 1 or 2). Common filter functions include projection, eccentricity, and truncated singular value decomposition (SVD). After applying the filter function, a cover of the image in R^n is chosen (a cover is a finite collection of open sets for which the collection contains every point in the image), and then mapper clusters each collection of points that are mapped to a single set in the cover using a standard clustering technique. The resulting mapper graph is created by placing a node for each cluster and placing an edge between clusters that share members.

Within the United States, electoral districting is recognized as a highly contentious political issue (Eagles, Katz, and Mark 2000; Makse 2014). While precise standards vary from state to state and even district to district, the National Conference of State Legislators lists twenty-seven states as requiring the preservation of "communities of interest" (COI). While exact definitions of COI are often left for legal interpretation, broad categories tend to include demographic, economic, and geographic factors that are assumed to produce populations with similar electoral and legislative interests (Grofman 1985; Morrill 1987). This abstract presents the results of visualizing mapper algorithm outputs analyzing American Community Survey (ACS) data for Mecklenburg County, North Carolina. This project was conducted as part of the Spatial Models and Electoral Districting National Science Foundation Research Experience for Undergraduates site (SMED REU). Mecklenburg County was chosen for several reasons, including that it contains Charlotte, the state's most populous city, has a relatively diverse population, and contains the 9th and 12 North Carolina Congressional Districts, which have featured prominently in a series of United States Supreme Court decisions on districting, such as Easley v. Cromartie (2001), Cooper v. Harris (2017), and Rucho v. Common Cause (2019).

Mapper was used to analyze and extract clusters from eleven American Community Survey data fields at the block group level (the smallest areal unit for which the Census Bureau releases detailed demographic information). Analysis and implementation was carried out using the giotto-tda python library (Tauzin 2021). The variables all relate to issues of race, education, and economic class that have historically been at the center of legislative districting challenges; each was taken from ACS 2018 5-year estimates and turned into a proportion of households within the areal unit. The project relied on American Community Survey data from the Census bureau because it provides more up-to-date and detailed information than the Decennial Census and given the known problems with under-representation of specific populations within Census data (Hogan et al. 2013), its use offers a robust test for mapper as a means of detecting similarities within necessarily incomplete data sets.

Abstracts of the International Cartographic Association, 3, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy.
https://doi.org/10.5194/ica-abs-3-288-2021 | © Author(s) 2021. CC BY 4.0 License.

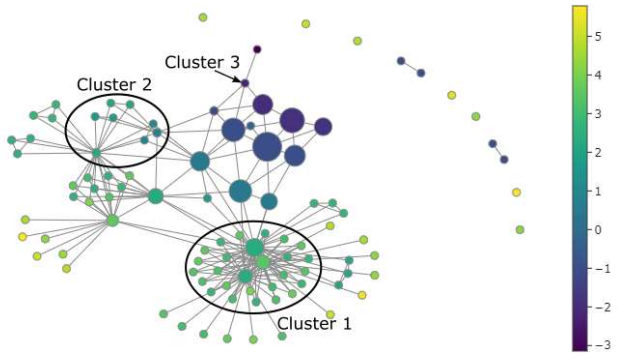

Figure 1. The graph output of the mapper algorithm results of block groups in Mecklenburg County, NC, USA

The graph output from the mapper algorithm (**Figure 1**) applied to the eleven ACS data fields. Each node is a collection of "similar" block groups. Edges exist between nodes that share block groups. The size of nodes indicates the size of the collection - larger nodes contain greater numbers of block groups, and smaller surrounding nodes are often subsets of these larger ones. The coloring of the nodes is determined by their position along the first truncated SVD component. Three collections of nodes of interest have been circled, though in this abstract we focus on the largest one (Cluster 1).

Of particular interest are two elements of mapper's outputs. First, its clusters are non-deterministic and non-exclusive, meaning that a single block group may be present in multiple clusters. Second, we considered the demographic data irrespective of geographic contiguity and only analyzed proximity after the initial clustering. Here, due to space, we present one of the six emergent clusters mapper extracted. **Figure 2** visualizes an emergent cluster against the three most recent districting plans for North Carolina; moving the graph output from **Figure 1** into a cartographic representation. This cluster is characterized by lower levels of wealth, education, and homeownership when compared to Mecklenburg County as a whole. In addition, this cluster is majority non-white. As can be seen from the visualizations, this cluster is mostly contiguous with a few outlier block groups; suggesting a geographic cluster as well as one based on demographic measures of similarity. The cluster is split relatively evenly between Districts 37 (~20% of the population), 38 (~35%), and 40 (~40%) in the 2011 plan. A similar split occurs in the 2019 plan, with District 40 at ~31%, 38 at ~39%, and 10% in each of the other three districts.

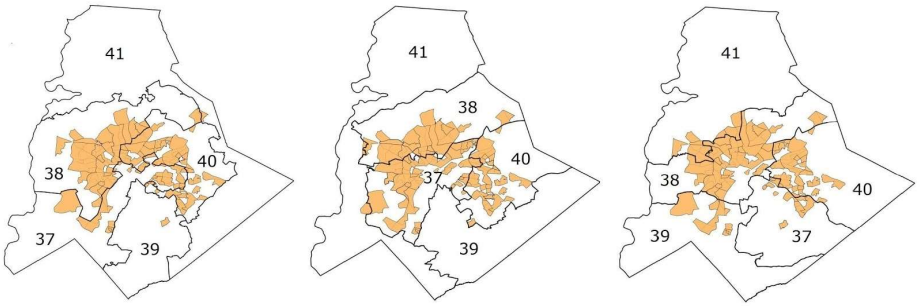

Figure 2. Cluster 1 visualized against the 2010 (left), 2018 (middle), and 2019 (right) districting plans.

Given the geographic and demographic contiguity, such consistent splitting of the cluster that avoids creating a majority-minority district suggests potentially racially-motivated gerrymandering. At the same time, with a population of ~226 thousand, the cluster is too large to be placed within a single district and doing so would likely constitute 'packing,' a form of gerrymandering. What these conflicting results suggest is twofold: first, mapper's outputs can detect potential communities of interest that may be hidden by other approaches to districting; second, while mapper outputs may be suggestive, further inquiry is required to understand the actual, lived contours of the region.

This result suggests the limits of knowledge posed by machine learning approaches and the need to augment such approaches with traditional methods of quantitative and qualitative inquiry. Continuing research in this area has begun to compare the clusters mapper detected with other ways of mapping communities, including: 1) clustering of landscape patches into regions with similar demographics and "neighborhood character" based on features observable via satellite imagery (Casey et al. 2017; Wurm and Taubenböck 2018); and 2) developing 'ground truthed' maps of communities' extents through qualitative interviews, participatory GIS, and resident-led neighborhood walkthroughs (Aronson et al. 2007; Dunn 2007; Rambo 2020). Through comparisons with these other techniques, we hope to further refine our understanding of how mapper's outputs may most productively inform electoral districting, public health, and other domains where management and policy depend on identifying communities of interest that are true to residents' lived experience.

Abstracts of the International Cartographic Association, 3, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy.
https://doi.org/10.5194/ica-abs-3-288-2021 | © Author(s) 2021. CC BY 4.0 License.