Aggregating Community Maps

Erin Chambers
erin.chambers@slu.edu
Saint Louis University
St. Louis, MO, USA

Moon Duchin
moon.duchin@tufts.edu
Tufts University
Medford/Somerville, MA, USA

Ranthony A.C. Edmonds
edmonds.110@osu.edu
The Ohio State University
Columbus, OH, USA

Parker Edwards
parker.edwards@nd.edu
University of Notre Dame
Notre Dame, IN, USA

JN Matthews
jnmatthews@uchicago.edu
University of Chicago
Chicago, IL, USA

Anthony E. Pizzimenti
apizzime@gmu.edu
George Mason University
Fairfax, VA, USA

Chanel Richardson
chanel@mggg.org
Tufts University
Medford/Somerville, MA, USA

Parker Rule
parker.rule@tufts.edu
Tufts University
Medford/Somerville, MA, USA

Ari Stern
stern@wustl.edu
Washington University in St. Louis
St. Louis, MO, USA

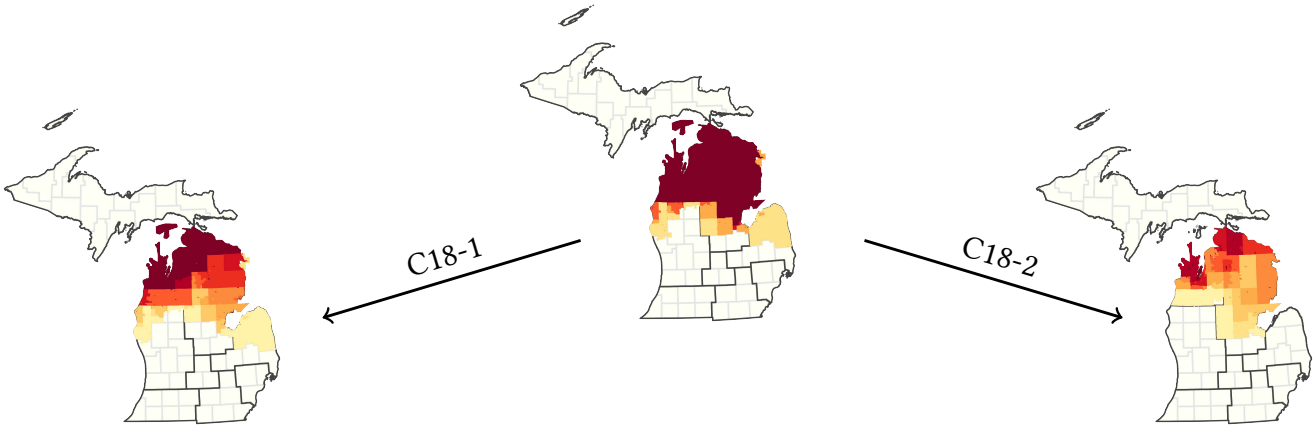


Figure 1: The "Upper Mitten" cluster of 71 community maps in Michigan, and its semantic splitting into subclusters C18-1 (recreation, environment, agriculture) and C18-2 (values, identity, religion).

ABSTRACT

This paper is motivated by a practical problem: many U.S. states have public hearings on "communities of interest" as part of their redistricting process, but no state has as yet adopted a concrete method of spatializing and aggregating community maps in order to take them into account in the drawing of new boundaries for electoral districts. Below, we describe a year-long project that collected and synthesized thousands of community maps through partnerships with grassroots organizations and/or government offices. The submissions were then aggregated by geographical clustering with a modified Hausdorff distance; then, the text from the narrative submissions was classified with semantic labels so that short runs of a Markov chain could be used to form semantic sub-clusters. The resulting dataset is publicly available, including the raw data of submitted community maps as well as post-processed community clusters and a scoring system for measuring how well districting plans respect the clusters. We provide a discussion of the strengths and weaknesses of this methodology and conclude with proposed directions for future work.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Information systems** → **Clustering**; **Geographic information systems**; • **Mathematics of computing** → **Graph algorithms**.

KEYWORDS

Geospatial data, clustering, semantic classification, regionalization, redistricting.

SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9529-8/22/11.
https://doi.org/10.1145/3557915.3560961

ACM Reference Format:
Erin Chambers, Moon Duchin, Ranthony A.C. Edmonds, Parker Edwards, JN Matthews, Anthony E. Pizzimenti, Chanel Richardson, Parker Rule, and Ari Stern. 2022. Aggregating Community Maps. In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3557915.3560961

1 INTRODUCTION

There are many settings in which a researcher might want to collect a large number of crowdsourced maps and combine them into a summarized spatial data product. One important application is in *redistricting*, or the drawing of boundary lines for electoral districts. In redistricting, *communities of interest* (or "COIs") is a term of art that refers to neighborhoods or regions that should be given weight in drawing those lines. A community of interest is a group of individuals with both shared residential terrain and some shared attributes that connect to policy interests or representational goals. In the context of redistricting, whether a COI is kept together in a single district—or how it is split up across multiple districts—is likely to affect whether residents obtain effective representation. State constitutions or legislative provisions refer to consideration of communities of interest in at least two dozen states [6, 12], and across the board the COI concept is considered as one of the traditional redistricting criteria (alongside federal requirements of population balance and the Voting Rights Act and conventional considerations like compactness, contiguity, and the preservation of political subdivisions like counties and cities). See [27] for a short COI explainer, including a review of selected COI case law from redistricting.

Communities may be defined by commonality across a number of dimensions, including economics or social class, geography, culture, media markets, and use of (or need for) public services. COIs commonly feature shared ties to an industry or employment sector that drives policy interests, such as farming, tourism, or resource extraction. They may be grounded in a neighborhood with infrastructural or service needs, or in a region where pollution or natural disasters animate environmental concerns. (Of course, this list is far from exhaustive.) The challenge is that the concept of communities of interest, because of its emphasis on community, does not lend itself well to routine or off-the-shelf models or formulas applied to standard data sources, such as the data products provided by the Census Bureau.[1]

A testimony-based approach—relying on members of the public to define their boundaries, articulate their shared interests, and advocate for consideration in the redistricting process—has been the main mechanism for ensuring that such voices get their due consideration. In past redistricting cycles, there has been a major gap between testimony and line-drawing: to our knowledge, there were very few examples from the 2010 Census cycle or earlier where testimony was converted to concrete maps or geographical polygons that distill public input in a manner that is clearly actionable for a commission or court.[2] That is, incorporating communities of interest into redistricting plans has been impeded by the absence of agreement about how to identify the relevant communities, either conceptually or procedurally, and how to measure compliance with a rule that counsels respect for COIs.

Evaluating proposed communities of interest for inclusion into a redistricting plan is ultimately an exercise of judgment by mapmakers and line-drawers that deserves transparency and clarity. Importantly, that does not and must not mean that the process should be fully quantitative and automated. But for the current redistricting cycle following the 2020 Decennial Census data release, the ability to take spatialized public testimony—a description of community needs in a format that is accompanied by a digitized map—was present in multiple states to a greater degree than before, partly due to a collection effort that will be described here. Importantly, this effort produced thousands of map submissions in each of the four states discussed in this article; the sheer volume of maps created a need to treat them in a clustered and summarized fashion rather than individually. The current article is dedicated to describing one proposed process and pipeline for producing an aggregated data product, and for measuring the alignment of a districting plan with those community clusters.

1.1 Related Work

The work here fits into the computing literature in two main ways. The first is the development of computational techniques, especially clustering techniques, for *regionalization*, or the identification of meaningful areas within a larger terrain—a long-standing topic in geography and planning.[3] The second is in leveraging algorithmic methods to bring mathematics and statistics to bear on redistricting, including in the evaluation of fairness for political redistricting plans.

The use of computer science and mathematical techniques to aid in fair redistricting is well established. For instance, [14] is a much-cited ACM paper from the 1970s, and [31] traces earlier (imagined and actual) computational approaches to redistricting back to the 1960s. The recent survey [2] gives a comparison of the fast-moving state of the art in algorithmic methods to build and compare districting plans. A selection of the ACM papers from the last five years that discuss these topics includes [4, 7–9, 18, 28, 36].

Our work uses the modified Hausdorff distance [11] as a way to compare geographic regions, which to our knowledge has not been previously used in redistricting or regionalization specifically. Hausdorff-style metrics are well studied for comparison and clustering of other types of GIS data, such as object location tags [22] and trajectories [3].

There are numerous papers developing techniques that are applicable to geography aggregation, including for example [33], which formulates what they call the *cluster ensemble problem*, where they take a group of many different partitions of objects and attempt to build a new set of clusters via various "combiners," also called

[1] Law scholar Nicholas Stephanopoulos has sketched a kind of automated community detection based on demographic and economic attributes in the American Community Survey. As Mac Donald–Cain have argued, this would produce a very thin notion of community, unable to capture most of the kinds of shared interest commonly associated with COIs. [19, 32] Indeed Stephanopoulos also clearly endorses the value and interest of self-identified communities.

[2] There are a few notable instances of precise COI maps proposed by grassroots groups, including for example a set of "Asian American Neighborhood Boundaries" produced for New York redistricting by the non-profit AALDEF, available at `aaldef.org/uploads/pdf/intervenor-lee-attachment a.pdf`.

[3] For an enlightening discussion of regionalization and community in the context of redistricting, see Garret Dash Nelson, *The Elusive Geography of Communities* [24].

"consensus functions." Another is [17], clustering polygons via k-means on both geographic closeness and similarity in other quantitative aspects. That paper also employs Hausdorff distance, and defines a "boundary-adjusted Hausdorff distance" that rewards pairs of shapes for having shared boundary. One of the most relevant might be [1], which gives an algorithm to partition spatial data in a consensus-based approach by building a similarity graph, where spatial objects are vertices and edges are weighted by how often objects are put in the same region. They test on a synthetic dataset as well as a ecological marine units dataset.

There is a wealth of text-and-geography projects in the CS/GIS literature, including [29] (identifying language with geography in public datasets such as Twitter, Flickr, and OpenStreetMap); [23] (ML approach to flagging which features are frequently tagged in which locations); [25] (clustering by geography and group text data, then using the text to find and correct errors in the corpus); and [35] (deep learning on human mobility datasets to assess areas as well or poorly planned).

We know of no previous work that we can directly compare to the methods introduced here, which center on synthesizing a dataset consisting of thousands of personal maps drawn in a self-directed fashion. Several authors have outlined ideas for automated community detection through free-standing geospatial data sources (such as Stephanopoulos and Spielman–Singleton, drawing from Census Bureau data products [30, 32], or Makse, drawing from the results of statewide initiative votes [21]). Phillips–Montello [26] describe a hybrid method for combining census outputs with 107 regions drawn freehand in interviews with residents of Santa Barbara, California. Their method separates responses out by their current city council district and seeks to identify "cores" for re-drawing those districts, and does not readily generalize to a large volume of publicly submitted maps. Finally, the recent paper of Chen et al. [6] proposes a scoring system for measuring the alignment of a plan with a set of spatial polygons, but it does not contend with large numbers of overlapping maps drawn on many different scales.[4]

2 COI COLLECTION

The authors of this paper are the key contributors to the technical side of the collection and aggregation process described here. The effort produced many thousands of user-submitted maps, and we processed them into data products described in the following section, suitable for use with a score that is also described below. However, we emphasize at the outset that the intent of the current paper is to present the methodology in a scientific venue, rather than to offer the curated data product itself as the primary object of interest. On the contrary, the lessons learned from implementation suggest a suite of robustness checks, tweaks, and outright improvements, some ideas for which will be described below. The time window before the 2030 U.S. Census presents quite a long period for refinement and debate.

In the 2020-21 project described here, we partnered with both academic and non-academic collaborators in order to carry out community mapping in a way that responded to the needs of different organizations and was sensitive to local knowledge of geography and sociology. Our collaboration was initiated through an interdisciplinary project team we dubbed OPEN Maps (Ohio Public Engagement in Neighborhood Maps). After launching the effort in Ohio, we extended the collection project to Michigan, Wisconsin, Missouri, and to smaller-scale efforts in New Mexico, Texas, Indiana, Pennsylvania, Florida and numerous counties and cities.

2.1 The Collection Process

2.1.1 Mapping app. Our collection process leveraged a mapping application launched in 2018 by the MGGG Redistricting Lab, a data and democracy research group based at the Tisch College of Civic Life of Tufts University. **Districtr** (`districtr.org`) is an open-source webapp built on the Mapbox platform that functions as a purpose-engineered GIS. Users encounter familiar-looking paint tools and a mapping interface that closely resembles smartphone apps. Districtr has two modes: districting and community mapping. In districting mode, the dashboard and functions are designed to create painted regions that have balanced population, that combine for full coverage, and that are disjoint from one another. Community mapping mode provides users with a more detailed base map with neighborhood names, highways, and buildings visible. In this mode, the user can paint community areas of any size, which are allowed to overlap. They can name and describe these communities in text fields.[5] Importantly, both modes use geographic tiles as the base units, where depending on location the choices may range from fine units (census blocks) to coarse units (counties or community areas). That is, instead of drawing free-form on a map, users are selecting and assigning geographic units to their maps.

2.1.2 Partners. We worked with grassroots organizations to customize our collection effort in each state. In Ohio, redistricting is currently carried out by a politician commission which is in no way independent of the legislature. Outside community groups conducted public map collection in order to support what some would call a "shadow commission," intended to model best practices. In Michigan, by contrast, Districtr was contracted by the Michigan Department of State to support the Michigan Independent Citizens Redistricting Commission, which is fully empowered by law to draw the district lines. In Wisconsin, we contracted with the state Department of Administration to support the People's Maps Commission, a shadow commission created by Governor Tony Evers. And in Missouri, our partner was a non-governmental organization called the Fair Maps Missouri Coalition, which sponsored a mapping contest open to the public.[6]

One key benefit of collaborating with diverse partners in each state was to incorporate local knowledge into the participatory mapping design. For example, our Ohio team solicited input from geographers and organizers to suggest a zoning of the state into six regional areas that broadly made sense from the point of view of commuting and cultural exchange. They also advised us of which sizes of geographical unit to use for each level of mapping; that

[4] In particular, Phillips–Montello instruct respondents to draw their communities at or below the scale of city council districts. Chen et al. actually discard maps drawn larger than a district, leaving 16 non-overlapping COI maps from their collected dataset and therefore having no need for aggregation.

[5] Users can also mark, name, and describe landmarks or *points of interest*, but the process described here does not make use of point data.

[6] For the other state-level portals, the partners were the New Mexico Citizen Redistricting Committee (supported by the Thornburg Foundation), the office of Pennsylvania Governor Tom Wolf, Florida Rising, the Texas Civil Rights Project, and Common Cause Indiana. See `mggg.org/cois` for more information.

municipal boundaries would be an important and informative layer for users; that a Cleveland module should really include the key suburb of Euclid because important neighborhoods were likely to cross that municipal boundary; and so on. This tailoring increased the usability of the mapping tool.

2.1.3 Training and instructions to participants. In contrast to other COI collection efforts, we did not present participants with a script or formal checklist to walk through the mapping process.[7] Instead, we used a *train-the-trainers* model to teach community organizers how to collect maps, and how to train others to collect maps, and how to train others to train others, and so on. This was accomplished through dozens of videoconference meetings conducted in English, with a handful in Spanish, throughout the Spring and Summer of 2021. The user-friendly app and the remote trainings were crucial elements for operating in an extraordinary historical moment when COVID-19 contributed to a late Census and a compressed timeline for redistricting, in addition to creating unprecedented challenges in organizing in-person meetings.

The bottom-up, less-directed model means that participants had great latitude to interpret the instructions, and we received a wide range of disparate types of maps. Some were neighborhood-level; some were regional; some were pairs of counties that the author felt should or should not be kept together for political representation. Some respondents interpreted their task as one of displaying and narrating a named neighborhood, like the traditionally middle-class Black neighborhood of West Dayton. Others made much more personal maps, marking locations in their daily and weekly orbit from their church to their workplace to their favorite fast food restaurants and laundromats.

2.1.4 Submission portals. A final element in the collection process was to design an interface for submission of completed maps.[8] These were simple *public portal* websites that included background information on redistricting and the role of community maps through short YouTube videos and links to the relevant mapping modules in Districtr. Upon completion of a map in Districtr, a "Save" or "Share" button would send the user back to the portal with a submission form partly pre-filled. The user would then be prompted to add basic information like their name, to add their email address for verification purposes, to title their submission and comment on its relevance to the process. The portals also enabled submitters to tag their maps with keywords. We worked with partners in each state on translations where appropriate, developing content in Spanish, Haitian Creole, and Navajo.

Each portal featured a real-time gallery of public submissions, searchable by date, keyword, tag, or submission type. Users were also able to submit comments on other content, and each partner had their own approach to content moderation.

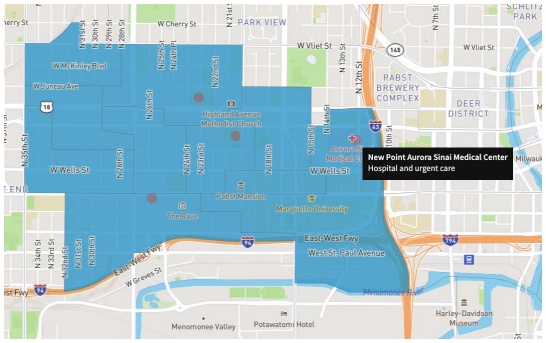
[7]For instance, a COI collection project based at Princeton (`representable.org`) used a highly guided intake form.

[8]In fact, most of the portals allowed for users to submit any of several forms of feedback: districting plans, community maps, and written testimony about any aspect of the process. In most states, written testimony remained the most popular mode of public submission by far.

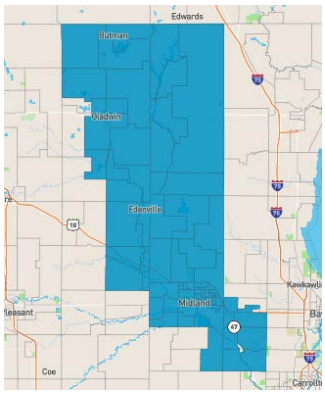
2.2 Sample COI Submissions

Here are two verbatim examples to start to give a flavor of the submissions we received through our portals.

- **Avant's map 7/21/21**: "Very racially diverse, police officers, bus drivers, teachers, blue collar workers, entrepreneurs live in the area. Some important places are the Milwaukee High School of the Arts, Milwaukee Academy of Science, Sinai Samaritan Hospital, u.s. Bank and New Life Center drug dealing. The neighborhood is great just a few things that needs to be addressed like housing upgrades, potholes in the streets, sex working, Develope a system to Form a block watch." (`portal.wisconsin-mapping.org/submission/c1896`)



- **Tittabawassee River Community**: "This community is closely linked by a shared waterway, the Tittabawassee River. The area has also experienced extreme flooding due to dam failures last year. The area is still working together to recover." (`michigan-mapping.org/submission/c373`)



Much more information on the COI collection from the four primary states in this project are linked from a post at `mggg.org/cois`.

2.3 Concerns about Data Quality and Gaming

Quality of COI data is inextricably tied to the outreach effort that supports data collection and depends heavily on any curation that takes place. (There is a significant geography literature on participatory mapping: see, for instance, [5] and its references.) Each line-drawing body must also decide on a level of moderation for the content, ensuring that COI testimony complies with state rules. For

instance, some states bar the use of communities of interest based on partisan identity, but members of the public may nonetheless submit maps with highly partisan descriptions.

Any data collection effort that relies on voluntary submissions or testimony comes with data quality challenges in the post-processing stage, and in particular carries the potential for bad actors to attempt to game the system. Relatedly, an important question for the aggregation protocol is whether importance is keyed to volume of submissions. For example, in our dataset, well over a hundred submissions painted an area that is nearly coextensive with an existing Congressional district in southern Michigan, using extremely similar language. It is highly possible that this was a coordinated 'astroturfing' effort on behalf of the sitting Congressman.[9] Coordination alone does not make the submissions invalid, of course, but we regard it to be a positive feature of our methodology that such a case does not receive 150× as much weight as more spontaneous testimony in other parts of the state.

There are other reasons to be wary of emphasizing COI submissions that support status quo districts. For example, sociologist Robert Vargas of the University of Chicago has argued that some Chicago neighborhoods have enjoyed *political stability* while others have been consistently *fragmented*—not only split, but split differently each time new lines are drawn. Vargas finds that "approximately 15% of city blocks have remained in the same political districts since the 19th century, and that these spaces are spatially clustered in neighborhoods home to local political, economic, and administrative elites." [34] This reminds us that mechanisms that favor the preservation of historical districts, though they may in some cases promote community interests, can easily reproduce inequality of representation.

Finally, public input processes that are aspirationally open to all will necessarily draw from a self-selected and partial tranche of the public, skewed toward the advantaged, well-resourced, and well-organized. Prioritizing the interests of the most vocal groups in the redistricting process may come at the expense of less well-connected groups. Therefore, we emphasize that *COI aggregation is only as good as the outreach effort that supports COI collection*: in this case, pointing members of the public to the hearings or collection portal and providing the resources and training to use them effectively. Effective outreach must proactively engage a wide range of stakeholders, so that the resulting maps do not over-weight the interests of the most vocal or connected groups.

3 THE AGGREGATION PIPELINE

3.1 Goals of Aggregation and Scoring

The main aim of synthesizing individual maps is to make the data surveyable, informative, and actionable for line-drawers. An important secondary aim is to give the districts themselves a multi-layered community character that has the potential to alert candidates (and the representatives ultimately elected) to salient identities and needs in their constituencies.[10]

The key to usability is therefore to reduce the number of objects and to provide the synthesized objects with simple descriptions. For instance, the Michigan dataset included 1225 individual polygons, far too many for the Citizens Redistricting Commission to weigh individually as they drew the lines for 13 Congressional districts. Since the submissions varied widely in their physical scope as well as their narrative content and valence, informal aggregation was extremely challenging. We sought to create a number of clusters in the dozens rather than the hundreds. We chose to allow their sizes to vary, but to design scores of COI preservation that handle large clusters differently from small clusters, and that are equally sensible for large districts and small districts.

Aggregating textual content poses its own difficulties. Some individual submissions had extremely sparse text while others ran to thousands of words; style ranged from matter-of-fact to extremely abstract and diffuse. In addition, proximal geography is no guarantee of harmonizing narratives or preferences. Conflicts between similarly situated communities—or between community preservation and other redistricting criteria—are surely unavoidable.

For an example of the tension between expressed desires and other good governance principles, some scholars have argued that respecting the boundaries of large and politically homogeneous communities of interest may come at the expense of drawing more competitive districts [13]. More generally, preferences themselves are often inconsistent within individuals as well as among collectives: many members of the public simultaneously hold the abstract desire for competitive districts and the particular desire to live in a district made up mainly of others who share their own views and outlook.

To illustrate conflicting content, consider an exurban area where one major share of the submissions cites reasons to be districted together with the nearby city, where residents go for employment and services. At the same time, another large portion counsels the reverse: that values and resource needs should keep the area together with rural counties and distinct from the urban core. It is impossible to follow both sets of preferences.

We have designed a two-tier aggregation process—first geographic, then semantic—in an effort to make trends visible and legible without submerging thematic conflicts. Clustering decisions were based on both supervised and unsupervised computational methods. In this section we outline the methods.

3.2 Geoclusters

Each mapped area can be regarded as a collection of smaller geographic units. In the discussion to follow, we will generally denote individual geographic units by lowercase letters $a, b, \ldots$; mapped areas (which are sets of geographic units) by uppercase letters $A, B, \ldots$; and clusters by script letters $\mathcal{A}, \mathcal{B}, \ldots$ (which, in a slight abuse of notation, are either sets of mapped areas or unions of mapped areas). The main units for geoclustering are *block groups*, which are defined and published by the U.S. Census Bureau every ten years. The block groups partition each state and are specified in shapefiles (as polygons, each with thousands of vertices). Clusters

[9] Per Wikipedia: "Astroturfing is the practice of masking the sponsors of a message or organization (e.g., political, advertising, religious or public relations) to make it appear as though it originates from and is supported by grassroots participants." `en.wikipedia.org/wiki/Astroturfing`. Retrieved on August 15, 2021.

[10] Some authors have gone so far as to argue that electoral districts should only unite similar communities, and that fusing unlike communities is as harmful as splitting coherent ones. Though there is some language in legal decisions to support this view, there is also a slippery slope from this homogeneity principle to polarization and "packing." We do not think that respect for communities in redistricting needs to, or should, be read this way.

are meant to identify submissions that are overlapping or geographically proximal.[11]

We primarily work combinatorially, building a dual graph to the selected geographic tiling (in this case, block groups) by defining a node for every tile and connecting two nodes with an edge if the tiles are rook-adjacent (i.e., if they have a shared boundary of positive length). The rest of the pipeline can now be described in terms of this graph, although the methods are mainly quite general and can be applied in the continuous as well as the discrete setting.

We can now define the distance $d(a, b)$ between geographic units to be the usual graph distance—the minimum length of an edge-path from a to b in the graph. Since block groups are constructed to have a generally similar population size, this notion of distance accounts for the effects of population density in a way that ordinary spatial distance (e.g., in miles) does not.[12]

We measured geographical dissimilarity between mapped areas using the *modified Hausdorff distance* introduced by Dubuisson and Jain [11], which is widely used to compare spatial regions (for instance, in medical imaging). Given two areas A and B, each of which is a set of nodes in the geography dual graph (i.e., a collection of block groups), the modified Hausdorff distance $MHD(A, B)$ is defined by

$$d(a, B) = \min_{b \in B} d(a, b), \qquad d(A, B) = \frac{1}{|A|} \sum_{a \in A} d(a, B),$$

$$MHD(A, B) = \max\big(d(A, B), d(B, A)\big).$$

That is, the modified Hausdorff distance between two areas is the average of the distances from each unit in one area to the closest unit in the other. If A and B overlap substantially or contain units that are mostly close together, then $MHD(A, B)$ will be small. In comparison with conventional Hausdorff distance, which uses maximum rather than average to define $d(A, B)$, modified Hausdorff distance is said to better capture the similarity of overlapping regions and to be more robust to outliers [11].[13]

Finally, the mapped areas were organized into geography-based clusters using complete-linkage agglomerative hierarchical clustering. The complete linkage defines modified Hausdorff distance between two clusters $\mathcal{A}$ and $\mathcal{B}$, regarded as sets of mapped areas, to be

$$MHD(\mathcal{A}, \mathcal{B}) = \max_{A \in \mathcal{A}, B \in \mathcal{B}} MHD(A, B).$$

The agglomerative clustering algorithm begins by placing each area in a cluster by itself, then repeatedly merges the two most geographically similar clusters, according to modified Hausdorff distance. The user decides where to stop the process (or, equivalently, where to cut the associated dendrogram). In our case, we aimed for a final set of 24-36 clusters, heuristically preferring to stop at a clustering in which clusters supported by very few submissions were rare. Figure 8 in Supplement C displays the dendrogram and cut height for the Michigan dataset. A visual inspection shows that our choice of 36 geoclusters for Michigan is robust in that it largely cuts long line segments.

It is natural to ask about the stability and robustness of the geoclusters obtained by this technique more generally, i.e.: Do the geoclusters change substantially when we add or remove a small number of submissions?[14] To test this, we ran the geoclustering procedure on 50 random subsamples of the Michigan dataset, each containing 90% of the submissions, and compared the resulting geoclusters to those obtained using all of the submissions.

For each subsample, let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ denote the restriction of the original clusters to the subsample and $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be the newly-computed clusters using only the subsampled data. For each $i = 1, \ldots, n$, the best match between $\mathcal{A}_i$ and one of the $\mathcal{B}_j$ was computed two ways:

(1) Following Hennig [15], compute the maximum Jaccard similarity $J_i = \max_{j=1}^{n} J(\mathcal{A}_i, \mathcal{B}_j)$, where

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}.$$

(2) Compute $\mathcal{MHD}_i = \min_{j=1}^{n} \mathcal{MHD}(\mathcal{A}_i, \mathcal{B}_j)$, where

$$\mathcal{D}(A, \mathcal{B}) = \min_{B \in \mathcal{B}} MHD(A, B),$$

$$\mathcal{D}(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} \mathcal{D}(A, \mathcal{B}),$$

$$\mathcal{MHD}(\mathcal{A}, \mathcal{B}) = \max\big(\mathcal{D}(\mathcal{A}, \mathcal{B}), \mathcal{D}(\mathcal{B}, \mathcal{A})\big).$$

This $\mathcal{MHD}$ is a second iteration of the modified Hausdorff distance construction, this time applied to the space of mapped areas equipped with MHD.

Figures 2-3 show box-and-whiskers plots of J_i and $\mathcal{MHD}_i$ for each of the $n = 36$ geoclusters in Michigan, taken over the 50 subsampling runs. We remark that Jaccard similarity may be somewhat pessimistic: $\mathcal{A}$ may differ from $\mathcal{B}$ only by areas that are nearby geographically—or even overlapping. In the figures we can observe that $\mathcal{MHD}_i$ is generally on the order of 1 block group. For comparison, the diameter of the Michigan 2010 block group dual graph is 86. We find this to be a highly satisfactory level of robustness.

After the initial geoclusters were formed, we conducted light manual processing to (a) remove a very small number of anomalous submissions, such as ones that paint the entire state or that are highly disconnected; (b) merge clusters that contained an especially small number of submissions into similar neighboring clusters; and (c) designate clusters that contained a large number of submissions as candidates for splitting into sub-clusters.

3.3 Semantic Subclusters

The next step in the methodology calls for using the textual content of the submissions to summarize each cluster and as a means to split the large clusters.

We first attempted to use off-the-shelf natural language processing tools (described in Supplement A), but found the results unsatisfying. We then designated a team to read many submissions

[11] There was a non-trivial technical hurdle to clear in converting all submissions to 2020 block groups, because the portals allowed for mapping in many different units, including blocks, precincts, and block groups from 2010. Where necessary we relied on the spatial transfer package called MAUP (`github.com/mggg/maup`).

[12] In terms of the scale of communities, a mile can be a long distance in a dense city but a short distance in a sparse rural area.

[13] Though this distance does reward overlaps, it can nonetheless produce the following effect: suppose area A is very large, containing small area B inside it, near its edge. Small area C is close to B but disjoint from A. Then B is regarded as closer to C despite being a proper subset of A. We deem this feature of MHD to be reasonable in this application.

[14] One may equally well ask whether the final product is very sensitive to the choice of the number of clusters, which would be an interesting question for future work.

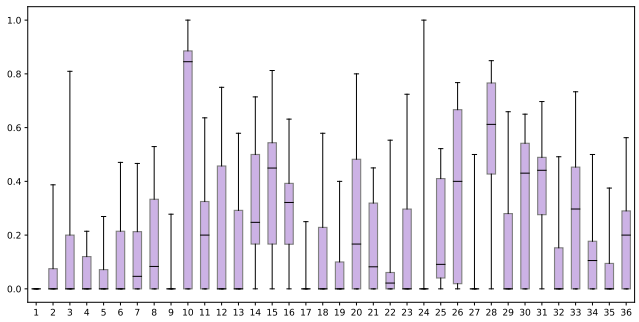


Figure 2: Maximizing J: How different are the Michigan clusterings built from a random 90% of the dataset, when matched to original clusters by Jaccard similarity, as in (1)? x axis is the cluster number from the original clustering; y axis shows $1 - J$, providing a measure of dissimilarity.

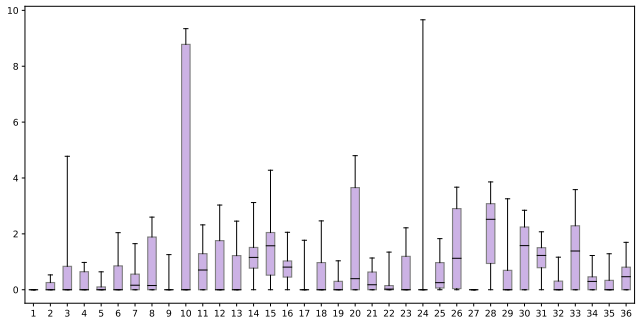


Figure 3: Minimizing MHD: Same as above, but for MHD matching, as in (2). y axis is now in units of graph distance.

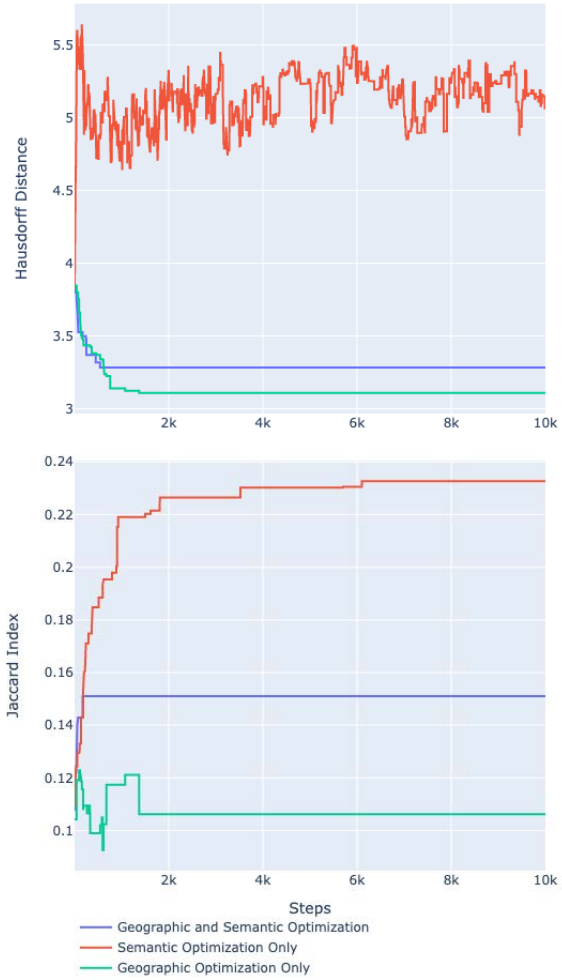


Figure 4: Trace plots of geographic and semantic similarity scores over time with $\beta = 10$ in the Grand Rapids geocluster in Michigan.

and develop a list of a few dozen thematic labels to use for manual classification.[15] Examples include urban, K-12, affordability, etc. (see Supplement B).

These labels then served as a foundation to build a metric for semantic similarity between submissions. We used the intersection-over-union score for the binary indicators, which we will again refer to as Jaccard similarity, to measure the semantic similarity across submissions. Namely, the score $J(A, B)$ of similarity for two submissions A, B is the number of labels possessed by both submissions divided by the number of labels possessed by either.

Within a larger geocluster, we looked for a splitting into two or three sub-clusters that would possess greater within-group similarity than between-group similarity. We used a standard style of local-search Markov chain to start with arbitrary groupings and improve them over a short run. The goal was to simultaneously drive down the intra-cluster geographic distances and drive up the intra-cluster semantic similarity. (See Figure 4.)

[15] Our team consisted of about a dozen human labelers working together with light training and some conversation. This team reviewed the submissions with the help of an annotation tool that would highlight keywords from a long list of about 100 and offer the user push-button access to the short list of labels, which could be selected in any combination.

Consider a fixed set $\mathcal{A}$ of submissions that define a particular geocluster. A candidate grouping can be denoted $(C_1, \ldots, C_k)$, where $C_1 \sqcup \cdots \sqcup C_k = \mathcal{A}$. For each cluster that was selected for subclustering, we fixed a choice of $k = 2$ or $k = 3$ subclusters (or bins) based on the volume and diversity of submissions. We defined geographic and semantic scores $g(C), \sigma(C)$ (to assess the degree to which entries in the same bin were similar) via

$$g(C_1, \ldots, C_k) = \sum_{i=1}^{k} \sum_{A,B \in C_i} MHD(A, B)$$

$$\text{and} \quad \sigma(C_1, \ldots, C_k) = \sum_{i=1}^{k} \sum_{A,B \in C_i} J(A, B),$$

where MHD is the modified Hausdorff distance described above and J is the Jaccard similarity score. Therefore, a good grouping has a low g and high σ.

At each step of the Markov chain, a new partition is proposed by flipping the sub-cluster assignment of a single submission at random. If the new proposed sub-clustering $(C'_1, \ldots, C'_k)$ is at least as good as the current state $(C_1, \ldots, C_k)$ in both scores, it is definitely accepted; otherwise it is accepted with probability

$$\exp\left(\frac{-\beta}{M}\right), \quad \text{for} \quad M = \min\left(\frac{g(C)}{g(C')}, \frac{\sigma(C')}{\sigma(C)}\right).$$

Here, β is a *temperature* parameter for the chain that can be tuned to improve performance.[16]

Figure 4 compares the behavior of this multi-objective chain to alternatives that optimize solely with respect to geographic or semantic similarity. The comparison is favorable for the method.

4 CLUSTERS AND SCORING

4.1 Basic Statistics on the Clusterings

One simple way to survey the outputs of the clustering and subclustering process is to examine the population, the land area, and the number of supporting submissions by cluster. This is shown for Michigan in Figures 5; corresponding figures for Ohio and Wisconsin can be found in Supplement C. The figures show that most clusters in Missouri contain 0-10% of the population of the state, while the range in Michigan for most clusters is 0-25%. The land area also varies widely.

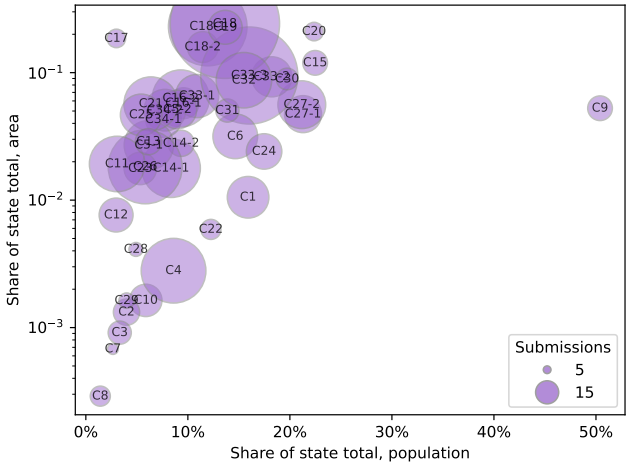


Figure 5: Population share of each cluster in Michigan (plotted on log scale) compared to its share of the statewide area. Circles are sized according to the number of total submissions in each cluster.

Certain features of the underlying collected data are visible. In Michigan, cluster C9 has half of the state's population, while being supported by under 20 submissions. This reflects the fact that several members of the public painted all of Detroit (and some surrounding terrain) as a single community of interest—possibly not in keeping with the more granular needs of the redistricting commission, but not in any way against the rules.[17] Likewise, C33 (Calhoun-Jackson) is extremely large in all dimensions; we chose to split it into three subclusters to create a more manageable final product. However, the unmistakable overlaps in the supporting areas made it hard for the geoclustering step to find a finer splitting. The text can be seen to have many overlaps as well, suggesting that this is likely the result of a coordinated campaign to influence the COI process of the commission.

The supplemental material contains a set of plots (Figure 11) showing an unexpected feature of the clusterings: there is a strikingly clear log-linear pattern in all four states when the clusters are reordered by population rank. In future work, it would be very interesting to explore whether this is more attributable to the method or to the input data.

4.2 Scoring Plans Against COI Clusters

Next we briefly discuss the creation of a score to measure the degree to which a districting plan (a partition of a state) might be said to "respect" a set of COI clusters.[18]

The most naive approach would be simply to count how many clusters are intact (that is, not split between districts; equivalently, having nodes which all have the same district assignment). However, this will clearly fail to be informative when clusters are larger in population than districts are mandated to be—as they often were in our datasets—which rules out the possibility of intactness.

Instead, we begin by flattening each cluster or sub-cluster to a polygon that is the union of the involved geo-units. Next, set a threshold value $0 < T < 1$ for intactness and say that one region A is *T-contained* in another region B if at least T share of the population of A is within B. We might consider a small cluster to be respected if enough of its population is assigned to a single district—that is, if it is T-contained in a single district. For a large cluster, on the other hand, respect for its meaningful community status might entail having a large number of districts be T-contained within the cluster. For example, consider a large-area cluster that runs along a lakefront border of the state and discusses interests in tourism and environmental issues. It would be desirable to have several state House districts be drawn totally within such a region.

For a given (plan,clustering) pair, we will award cluster C whichever score is higher: the binary indicator of whether the cluster is T-contained in a district, or the number of districts T-contained in the cluster (divided by the maximum number of districts that could theoretically fit in the cluster). This is then summed over the clusters to get an overall score for how well the plan respects the clustering.

This creates a score of concordance between a districting plan and a COI clustering that depends on the threshold parameter T, rising to a maximum of k (the number of clusters) as $T \to 0$ and dropping to a more modest maximum achievable value (which depends on the data) as $T \to 1$. In this way, a user does not need to

[16] Temperature is a parameter in a very well-studied family of optimization techniques known as *simulated annealing*. Broadly speaking, optimization algorithms can alternate between periods of running hot (unconstrained) and periods of cooling (tightening the constraints) to facilitate an explore-and-exploit regime. See [10] for an introduction in the context of redistricting.

[17] It is also notable that Michigan C9, the *Western Wayne County* cluster, contains several large-area maps devoted to describing the urban-rural differences to the West of Detroit. The situation is similar with Missouri C1 (*Greater St. Louis*) and C13 (*Greater Kansas City*).

[18] For a hands-on look at such a score, we have provided an interactive tool available at `desmos.com/calculator/gp5y3lflt2`.

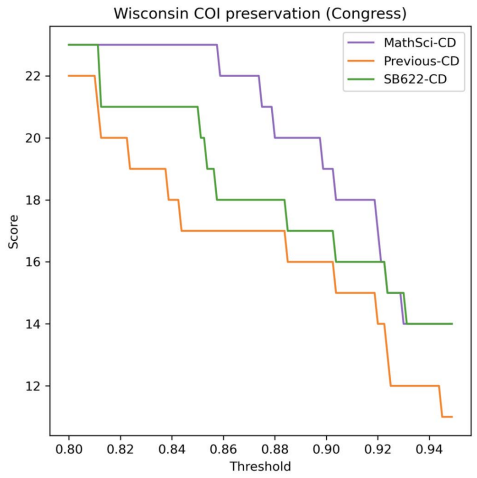


Figure 6: Trace plots for varying values of the threshold parameter T to score three competing Congressional plans in Wisconsin: the legislative proposal SB622, its predecessor from ten years prior, and a map proposed by an outside group.

hard-code an arbitrary threshold but can survey the comparison as a COI profile, with T varying. To illustrate this, Figure 6 shows the COI profiles for three different Congressional plans scored against the 36 clusters in Wisconsin.[19] We see that the "MathSci" congressional plan generally respects the COI clusters better than the benchmark plan "Previous" or the legislature's proposed replacement plan "SB622." This would be obscured if a fixed threshold like $T = .8$ or $T = .94$ had been chosen.

5 ACCESSIBILITY OF DATA

To ensure the accessibility of our processed dataset by varied stakeholders including mapmakers, researchers, and all members of the public, we have published data for Michigan, Missouri, Ohio, and Wisconsin in three primary formats. For each state in our dataset, the repos contain (1) a summary shapefile which flattens each geocluster into a polygon and (2) a shapefile containing individual polygons for each COI. We also have provided (3) complete databases of geographic and text submissions for each state.[20]

Those sources are suitable for further technical analysis via GIS and/or Python tools. We have also developed a visual interface within Districtr to make our analysis accessible to the general public and policymakers.[21] This Communities tool integrates COI exploration with redistricting, allowing users to simultaneously draw districts and visualize the geographic boundaries and supporting information for each COI (sub)cluster. An example in Michigan is shown in Figure 7.

[19]The figure is reproduced from the expert report of Moon Duchin in *Johnson v. Wis. Elections Comm'n*, No. 2021AP1450-OA, 2022 WL 621082 (Wis. Mar. 3, 2022).

[20]The Python pipeline used for our analysis, including raw data, can be found at `github.com/mggg/coi-states`. The processed dataset described here is available at `github.com/mggg/coi-products`.

[21]To access the Districtr visualizations, visit one of the state landing pages
- `districtr.org/michigan`
- `districtr.org/ohio`
- `districtr.org/missouri`
- `districtr.org/wisconsin`

and select a mapping module. Then navigate to the "Communities" tab.

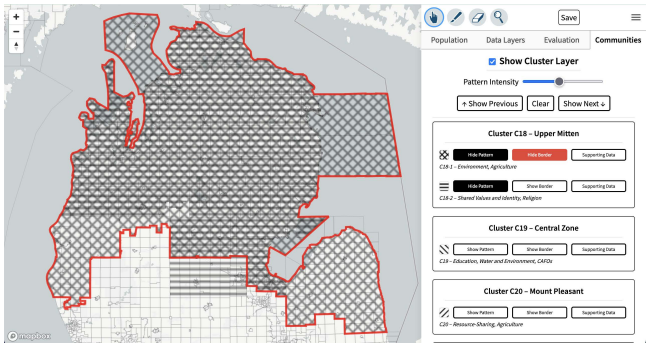


Figure 7: Subclusters ▨ C18-1 and ≣ C18-2 in Michigan, overlapping in the textured area ▩. Though the subclusters cover near-identical regions, they are semantically distinct.

6 CONCLUSIONS AND FUTURE DIRECTIONS

This paper represents one systematic effort at collection and aggregation, but there are many directions for future scientific exploration. Here we highlight ideas related to both the collection and aggregation process of community of interest data.

The protocol described here could clearly benefit from further stability testing (for instance, to see whether the semantic splits are robust) and from experimenting with variations in the post-processing steps.

Another promising direction of inquiry would be, rather than geographic clustering followed by semantic sub-clustering, to seek to handle spatial and text data simultaneously under the broad umbrella of *biclustering*. Here is one sketch of an approach. We can represent our data as a matrix, with one geographic unit per row and one topic tag per column, with a score representing how closely the dataset relates each geographic unit with a particular tag. Patterns in this matrix carry information about what interests matter where. Techniques for biclustering are well studied in the biological literature as a way to understand which groups of genes are correlated with biological conditions in the organism; see [20] for a survey of this topic. However, biclustering for geographic and textual data is a new application.

Finally, no amount of cleverness in clustering and processing can fully control for data quality shortfalls. One potentially valuable model would be for states to recruit and train a small interview staff to manage map intake over a key period of public input. Perhaps the strongest lesson learned from this experiment in synthesizing highly self-directed mapping is that a small imposition of structure on the front end would enormously reduce the challenges in meaningful aggregation on the back end.

ACKNOWLEDGMENTS

Thanks to the members and collaborators of the MGGG Redistricting Lab for their work on the COI data project, particularly Michael Altmann, Jamie Atlas, Luis Delgadillo, Nick Doiron, Jack Deschler, Max Fan, Maria Fields, Cyrus Kirby, Elizabeth Kopecky, Lucy Millman, Vievie Romanelli, Heather Rosenfeld, Robbie Veglahn, Valeria Velasquez, and Zach Wallace-Wright. Thanks to Vlad Kogan, Glennon Sweeney, Michael Outrich, Facundo Memoli, Matt

Kahle, and Duston Mixon from Ohio State University for many illuminating conversations in the building of the OPEN Maps project, and to Prentiss Haney and the Ohio Organizing Collaborative for the initial funding and concept development. Thanks to Voters Not Politicians and to Common Cause for deeply valuable advice and support on outreach and community engagement. We are grateful to Jerry Vattamala for providing pointers to the AALDEF COI maps from New York. Work in the four states principally discussed here was partly supported by contracts with the Ohio Organizing Collaborative, the Wisconsin Department of Administration, the Michigan Department of State, and Fair Maps Missouri. We thank the Jonathan M. Tisch College of Civic Life for its ongoing support.

REFERENCES

[1] Orhun Aydin, Mark V. Janikas, Renato Assunção, and Ting-Hwan Lee. 2018. SKATER-CON: Unsupervised Regionalization via Stochastic Tree Partitioning within a Consensus Framework Using Random Spanning Trees: Research Paper. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (Seattle, WA, USA) *(GeoAI'18)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3281548.3281554

[2] Amariah Becker and Justin Solomon. 2022. Redistricting Algorithms. In *Political Geometry*, Moon Duchin and Olivia Walch (Eds.). Birkhäuser Books, Chapter 16, 303–340. http://mggg.org/gerrybook

[3] Philippe C. Besse, Brendan Guillouet, Jean-Michel Loubes, and François Royer. 2016. Review and Perspective for Distance-Based Clustering of Vehicle Trajectories. *IEEE Transactions on Intelligent Transportation Systems* 17, 11 (2016), 3306–3317. https://doi.org/10.1109/TITS.2016.2547641

[4] Subhodip Biswas, Fanglan Chen, Zhiqian Chen, Andreea Sistrunk, Nathan Self, Chang-Tien Lu, and Naren Ramakrishnan. 2019. REGAL: A Regionalization Framework for School Boundaries. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Chicago, IL, USA) *(SIGSPATIAL '19)*. Association for Computing Machinery, New York, NY, USA, 544–547. https://doi.org/10.1145/3347146.3359377

[5] Robert Chambers. 2006. Participatory Mapping and Geographic Information Systems: Whose Map? Who is Empowered and Who Disempowered? Who Gains and Who Loses? *Electronic Journal on Information Systems in Developing Countries* 25, 2 (2006), 1–11.

[6] Sandra J. Chen, Samuel S.-H. Wang, Bernard Grofman, Richard F. Ober, Kyle T. Barnes Jr., and Jonathan R. Cervas. 2022. Turning communities of interest into a rigorous standard for redistricting. *Stanford Journal of Civil Rights and Civil Liberties* 18, 1 (2022), 101–189.

[7] Aloni Cohen, Moon Duchin, JN Matthews, and Bhushan Suwal. 2021. Census TopDown: The Impacts of Differential Privacy on Redistricting. In *2nd Symposium on Foundations of Responsible Computing (FORC 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 192)*, Katrina Ligett and Swati Gupta (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 5:1–5:22. https://doi.org/10.4230/LIPIcs.FORC.2021.5

[8] Vincent Cohen-Addad, Philip N. Klein, Dániel Marx, Archer Wheeler, and Christopher Wolfram. 2021. On the Computational Tractability of a Geographic Clustering Problem Arising in Redistricting. In *2nd Symposium on Foundations of Responsible Computing (FORC 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 192)*, Katrina Ligett and Swati Gupta (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 3:1–3:18. https://doi.org/10.4230/LIPIcs.FORC.2021.3

[9] Vincent Cohen-Addad, Philip N. Klein, and Neal E. Young. 2018. Balanced Centroidal Power Diagrams for Redistricting. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Seattle, Washington) *(SIGSPATIAL '18)*. Association for Computing Machinery, New York, NY, USA, 389–396. https://doi.org/10.1145/3274895.3274979

[10] Daryl DeFord and Moon Duchin. 2022. Random walks and the universe of districting plans. In *Political Geometry*, Moon Duchin and Olivia Walch (Eds.). Birkhäuser Books, Chapter 17, 341–381. http://mggg.org/gerrybook

[11] Marie-Pierre Dubuisson and Anil K. Jain. 1994. A modified Hausdorff distance for object matching. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, Vol. 1. IEEE Computer Soc., Los Alamitos, CA, 566–568. https://doi.org/10.1109/ICPR.1994.576361

[12] Brennan Center for Justice. 2010. Communities of Interest. *Brennan Center* (2010). https://www.brennancenter.org/sites/default/files/analysis/6%20Communities%20of%20Interest.pdf

[13] James G. Gimpel and Laurel Harbridge-Yong. 2020. Conflicting Goals of Redistricting: Do Districts That Maximize Competition Reckon with Communities of Interest? *Election Law Journal* 19, 4 (2020), 451–471.

[14] Robert E. Helbig, Patrick K. Orr, and Robert R. Roediger. 1972. Political Redistricting by Computer. *Commun. ACM* 15, 8 (aug 1972), 735–741. https://doi.org/10.1145/361532.361543

[15] Christian Hennig. 2007. Cluster-wise assessment of cluster stability. *Comput. Statist. Data Anal.* 52, 1 (2007), 258–271. https://doi.org/10.1016/j.csda.2006.11.025

[16] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://doi.org/10.5281/zenodo.1212303

[17] Deepti Joshi, Ashok Samal, and Leen-Kiat Soh. 2009. A Dissimilarity Function for Clustering Geospatial Polygons. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Seattle, Washington) *(GIS '09)*. Association for Computing Machinery, New York, NY, USA, 384–387. https://doi.org/10.1145/1653771.1653825

[18] Harry A. Levin and Sorelle A. Friedler. 2019. Automated Congressional Redistricting. *ACM J. Exp. Algorithmics* 24, Article 1.10 (apr 2019), 24 pages. https://doi.org/10.1145/3316513

[19] Karin MacDonald and Bruce E. Cain. 2013. Community of Interest Methodology and Public Testimony. *U.C. Irvine Law Review* 3, 3 (2013), 609–636.

[20] S.C. Madeira and A.L. Oliveira. 2004. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1, 1 (2004), 24–45. https://doi.org/10.1109/TCBB.2004.2

[21] Todd Makse. 2012. Defining Communities of Interest in Redistricting Through Initiative Voting. *Election Law Journal* 11, 4 (2012), 503–517.

[22] D. Min, L. Zhilin, and C. Xiaoyong. 2007. Extended Hausdorff distance for spatial objects in GIS. *International Journal of Geographical Information Science* 21, 4 (2007), 459–475. https://doi.org/10.1080/13658810601073315 arXiv:https://doi.org/10.1080/13658810601073315

[23] Pradeep Mohan, Shashi Shekhar, James A. Shine, James P. Rogers, Zhe Jiang, and Nicole Wayant. 2011. A Neighborhood Graph Based Approach to Regional Co-Location Pattern Discovery: A Summary of Results. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Chicago, Illinois) *(GIS '11)*. Association for Computing Machinery, New York, NY, USA, 122–132. https://doi.org/10.1145/2093973.2093991

[24] Garrett Dash Nelson. 2022. The Elusive Geography of Communities. In *Political Geometry*, Moon Duchin and Olivia Walch (Eds.). Birkhäuser Books, Chapter 11, 221–234. http://mggg.org/gerrybook

[25] Maria Angela Pellegrino, Luca Postiglione, and Vittorio Scarano. 2021. Detecting Data Accuracy Issues in Textual Geographical Data by a Clustering-Based Approach. In *8th ACM IKDD CODS and 26th COMAD* (Bangalore, India) *(CODS COMAD 2021)*. Association for Computing Machinery, New York, NY, USA, 208–212. https://doi.org/10.1145/3430984.3431031

[26] Daniel W. Phillips and Daniel R. Montello. 2017. Defining the community of interest as thematic and cognitive regions. *Political Geography* 61 (2017), 31–45.

[27] Heather Rosenfeld and Moon Duchin. 2022. Explainer: Communities of Interest. In *Political Geometry*, Moon Duchin and Olivia Walch (Eds.). Birkhäuser Books, Chapter 16, 235–245. http://mggg.org/gerrybook

[28] Zachary Schutzman. 2020. *Trade-Offs in Fair Redistricting*. Association for Computing Machinery, New York, NY, USA, 159–165. https://doi.org/10.1145/3375627.3375802

[29] Christian Sengstock and Michael Gertz. 2011. Exploration and Comparison of Geographic Information Sources Using Distance Statistics. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Chicago, Illinois) *(GIS '11)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/2093973.2094017

[30] Seth E. Spielman and Alex Singleton. 2015. Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach. *Annals of the Association of American Geographers* 105, 5 (2015), 1003–1025.

[31] Alma Steingart. 2022. Law, computing and redistricting in the 1960s. In *Political Geometry*, Moon Duchin and Olivia Walch (Eds.). Birkhäuser Books, Chapter 8, 163–177. http://mggg.org/gerrybook

[32] Nicholas O. Stephanopoulos. 2012. Spatial Diversity. *Election Law Journal* 160, 5 (2012), 1379–1477.

[33] Alexander Strehl and Joydeep Ghosh. 2003. Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* 3, null (mar 2003), 583–617. https://doi.org/10.1162/153244303321897735

[34] Robert Vargas. 2016. *Wounded city: Violent turf wars in a Chicago barrio* (1st ed.). Oxford University Press, New York, NY.

[35] Dongjie Wang, Yanjie Fu, Kunpeng Liu, Fanglan Chen, Pengyang Wang, and Chang-Tien Lu. 2022. Automated Urban Planning for Reimagining City Configuration via Adversarial Learning: Quantification, Generation, and Evaluation. *ACM Trans. Spatial Algorithms Syst.* (mar 2022). https://doi.org/10.1145/3524302 Just Accepted.

[36] Archer Wheeler and Philip N. Klein. 2020. The Impact of Highly Compact Algorithmic Redistricting on the Rural-versus-Urban Balance. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* (Seattle, WA, USA) *(SIGSPATIAL '20)*. Association for Computing Machinery, New York, NY, USA, 397–400. https://doi.org/10.1145/3397536.3422249

A ATTEMPTED NATURAL LANGUAGE PROCESSING APPROACHES

We attempted to apply several popular methods for text classification and topic modeling—including, but not limited to, latent Dirichlet allocation and bespoke keyword-based analysis—to text accompanying COI submissions. However, we found that these methods did not produce particularly meaningful clusterings. We hypothesize that this is attributable to several properties of the COI datasets:

(1) **Size.** At the time of aggregation, our largest state-level dataset contained over 1200 submissions, many of which included minimal text; high-quality natural language processing (NLP) models are typically trained or tuned on datasets that are orders of magnitude larger.
(2) **Semantic blurriness.** Broadly speaking, state-of-the-art text classification and topic modeling methods perform best when applied to datasets that can be carved at obvious categorical joints. In contrast, all text in the COI datasets essentially pertains to redistricting. Picking out subtle categorical differences between COI submissions—which often span multiple redistricting subtopics and use idiosyncratic or ironic phrasing to complain about perceived political slights and advocate for regionally specific policies—is a subtle and ambiguous task for expert human labelers and therefore a near-impossible task for a simple keyword-based model. This holds even in a regime where topics are seeded with carefully selected keywords.
(3) **Named entities.** We hoped that applying basic NLP techniques to COI submissions would produce clusters with distinct themes corresponding to varying community concerns: for instance, one cluster might primarily contain submissions related to education, while another might contain submissions related to industry. However, we found that naive clustering algorithms tended to find commonalities along geographical lines, rendering the NLP-based clusters somewhat redundant with respect to our geographic clusters. This is likely because of the overpowering appearance of regionally specific named entities (for instance, city, town, and school names) in the COI submissions.

These problems are somewhat tractable. We found that existing named recognition algorithms in spaCy [16] are reasonably effective at identifying (and therefore filtering out) the names of municipalities. We anticipate that by the next decennial Census, improved COI collection processes will yield larger datasets, and off-the-shelf algorithms for topic modeling and text classification will be more accurate and easy to use for non-experts. Nonetheless, we caution against the naive application of NLP-based clustering and interpretation—setting aside data quality issues, we found that constructing lists of seed words for topics or otherwise guiding a topic modeling algorithm towards actionable and interpretable topics required normative judgments that were less direct, auditable, and transparent than those required for manual semantic clustering.

B SEMANTIC LABELS

After surveying a large sample of submissions, our labeling team came up with the following coarse content labels for what public mappers indicated as being important about their neighborhood or region. Each had several dozen related keywords that triggered the suggestion of that label in our annotation tool, which could then be confirmed by the team member reviewing the submission. These labels were the basis of semantic similarity scoring.

- Agriculture
- Cities
- Community engagement
- Cost of living – Services – Healthcare
- Culture
- Diversity
- Economy – Commerce – Industry
- Environment
- Ideology
- Infrastructure
- Elderly
- Environment
- Family – Children
- K12
- Named neighborhood
- NIMBY
- Policing
- Poverty
- Recreation – Tourism
- Religion
- Suburbs
- Technology
- University
- Violence
- Vulnerable populations

C FURTHER VISUALIZATIONS

Michigan and Missouri population-versus-area plots were presented above; the other two aggregation states, Ohio and Wisconsin, are shown here. In addition, we show that rank-ordered population statistics consistently exhibit a log-linear pattern.

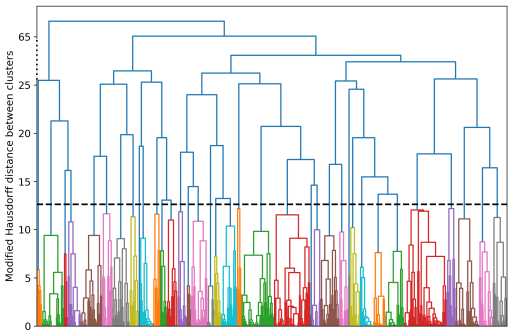


Figure 8: Dendrogram of geoclusters from Michigan submissions. Colors and dashed line show where it was cut. The y axis is compressed above 25.

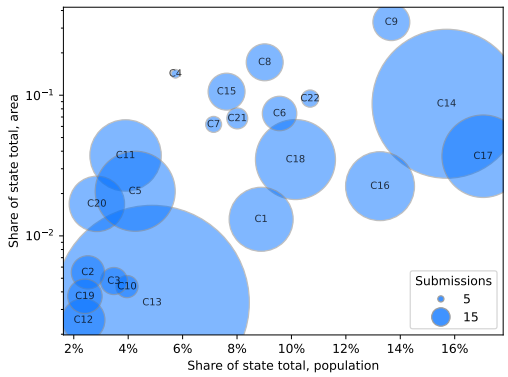


Figure 9: Population, area, and # of Ohio submissions.

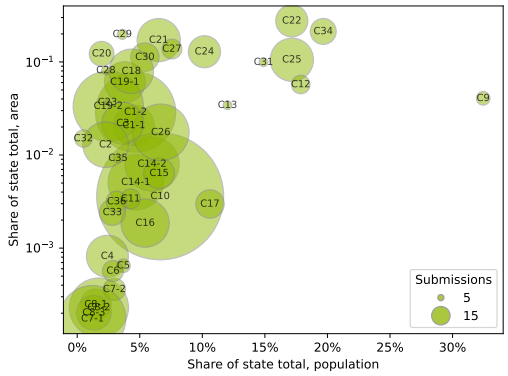


Figure 10: Population, area, and # of Wisconsin submissions.

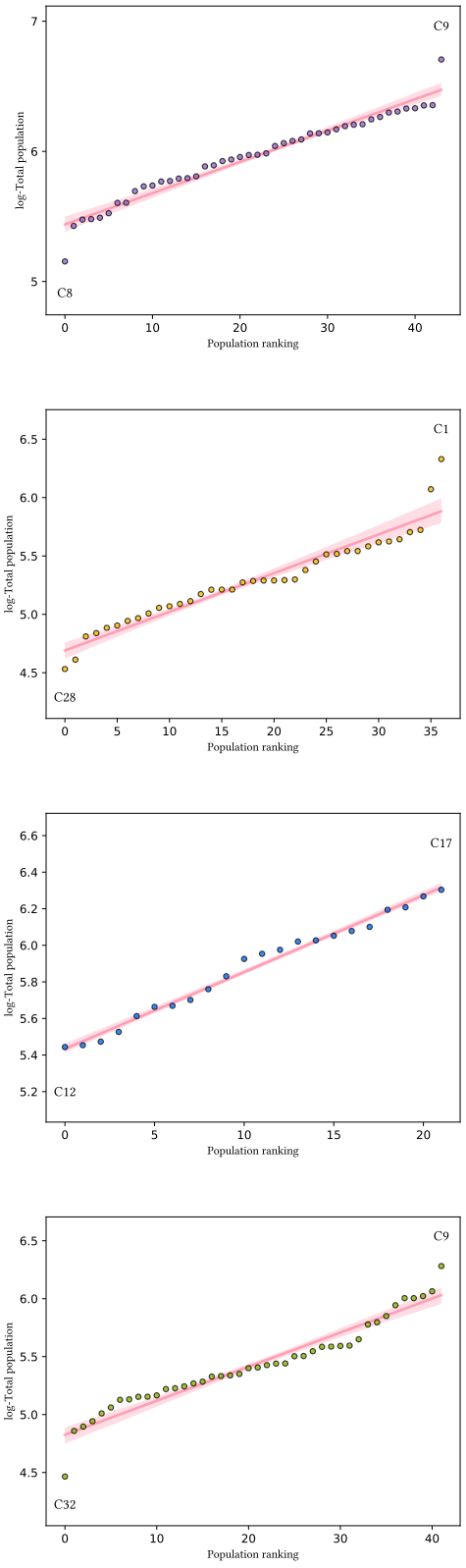


Figure 11: A surprising log-linearity trend in the total populations of the raw geoclusters (MI, MO, OH, WI, respectively).