No Adults Allowed! Unsupervised Learning Applied to Gerrymandered School Districts

Divya Siddarth, Amber Thomas

1. INTRODUCTION

With more than 80% of public school students attending the school assigned to them by district[1], it is clear that school district boundaries play a critical role in determining the educational opportunities and resources provided to students. Unfortunately, as school district lines are constantly redrawn, historical and anecdotal evidence suggests that these boundaries are often artificially manipulated, or gerrymandered, into irregular shapes. Often, these manipulations directly result in deliberately exclusionary zoning processes that create artificial economic disparity or racial segregation between school districts. This leads to adjacent school districts with dramatically different resources and funding for educational programs, and thus to inequality among the education that public school students receive. Further, while congressional gerrymandering (in which county boundaries are drawn to benefit a particular political party over another) has been a much studied phenomenon, school district gerrymandering is far less well-documented. However, categorizing and documenting these districts is an important first step towards equalizing funding and other resources across districts instead of having pockets of underserved students. This project will add to the existing body of knowledge on this topic, and hopefully serve as an introductory point to further studies.

2. RELATED WORK

There has not been much significant computational work done in the field of school district gerrymandering. Much of the evidence cited for gerrymandering in school districts has come from a few qualitative studies of particular districts that look at demographic data of an area as compared to boundary lines and historical context. One study in particular focused on racial breakdowns and economic status measured by income in Richmond, Virginia[2], and the detailed work done in this study helped us narrow down the possible demographic variables used in the present study. Another study attempted to quantify gerrymandering of school attendance zones with shape data, but focused on drawing correlations between particular districts, rather than the country as a whole[3]. While congressional district gerrymandering has been studied in more detail, a review of the literature suggests that artificial boundaries here are also studied almost exclusively qualitatively, focusing on race as it pertains to party affiliation rather than as a separate entity[4,5]. There has been some sparse computational work on the subject, particularly with regards to the fairly recent idea that computer generated congressional district boundaries may serve as a way to eliminate gerrymandering[6]. Here, GIS is used to create completely regular polygon boundaries based on geographic compactness indices, similar to the polygon irregularity indices used in the present study. A q-state Pott's model, which heuristically attempts to yield districts that are contiguous, compact, and of equal population, has also been attempted, but this runs into difficulties with the Voter Right's Act[7]. Overall, there has been a lack of relevant computational work in this field and related fields.

3. Data And Featurization

3.1. Dataset

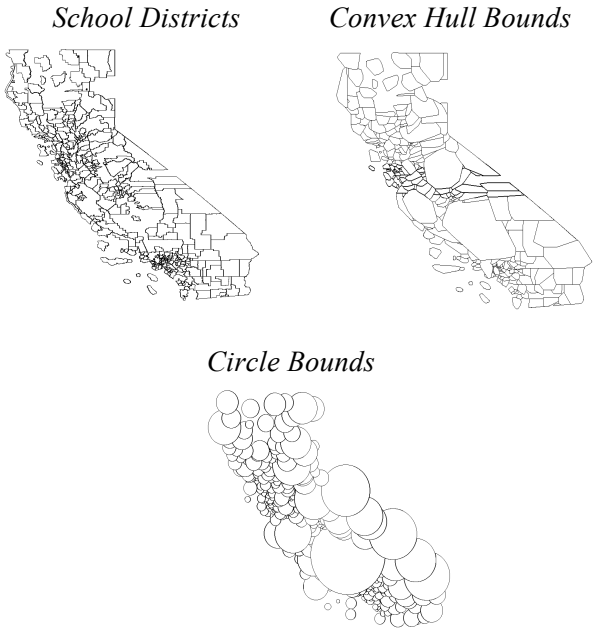
The dataset used to capture school district boundaries was provided by the Department of Education's TIGER/Line database, and was formatted as shapefiles, a common industry standard for representing spatial data in points, lines and polygons. Unified and elementary district information for 13,506 school districts was given, including districts in the continental US, Alaska, Hawaii, and US territories. Relevant raw features included in the shape file consisted of geographic id, state id, and name, all other features were calculated or extracted from inherent shapefile properties using ArcGIS software.

Census datasets were used to capture demographic data, and these data were divided along census tract lines, with data provided for each tract. Specific raw demographic features that were scraped from the census datasets consisted of the percentage of the population in a particular tract below poverty, and percentage racial breakdowns and mean household income for the given tract. 90% confidence intervals were given for each demographic estimate, which were used to calculate standard deviation.

3.2. Features

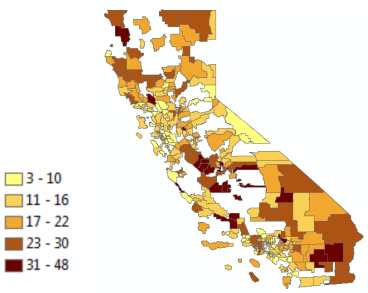
To get the data into a workable format, ArcGIS, a geospatial processing program, was used. First, feature vertices were generated from the boundary data of school districts by rasterizing the boundary shapes, and area and perimeter were calculated from these rasterized layers. To create a measure of irregularities in shape, minimum bounding geometries were then generated and layered on top of the original school districts. Both convex hulls bounds and circle bounds were generated for each district boundary (Figure 1). Area and perimeter of these bounds was calculated and paired with area and perimeter of the original districts.

FIGURE 1 – GENERATED MINIMUM BOUNDING GEOMETRY FOR CALIF. SCHOOL DISTRICTS



Demographic data was then combined with geographic data. First, it was necessary to mathematically project the geospatial coordinates of the data onto a flattened plane. Then each demographic dataset was overlaid separately in ArcMAP with the school district map boundaries. The GEO_id of the census tract was spatially correlated with the school district in which it fell, pairing the demographic information with the corresponding school district and creating a complete dataset of geographic and demographic data (Figure 2).

FIGURE 2 – GENERATED % BELOW POVERY MEASURES FOR CALIF. SCHOOL DISTRICTS

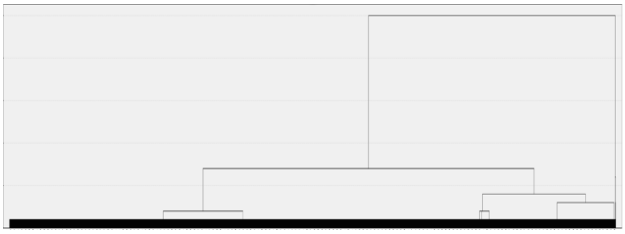


4. METHODS

4.1. Cluster Analysis

To determine the number of clusters to use in k-means clustering, we generated two dendrograms using using hierarchal clustering methods. In this process we generated a dissimilarity matrix that stored the dissimilarity between each school district pair. Then we began to build up the dendrogram by joining pairs at the lowest dissimilarity. Each time we joined, we would calculate the dissimilarity of the merged pair with the other districts and replace the old values in the matrix. To get the best sense of where to cut the tree, we ran the hierarchal clustering twice using two different formulas to calculate dissimilarity: complete linkage and Ward's method. Complete linkage sets the new dissimilarity value between the merge pair and each case to be the maximum dissimilarity within the pair for each case. We used complete linkage to visualize the maximum level of inter-district dissimilarity at each level of clustering.

FIGURE 3: COMPLETE LINKAGE DENDROGRAM



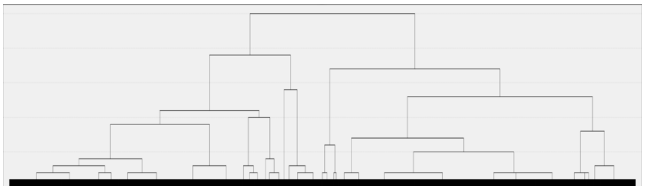
Visualization of the dissimilarity matrix. Recalculates the dissimilarity of each pair at each new level by taking the maximum dissimilarity of each pair in the join. Joins each pair of elements in the matrix at the maximum dissimilarity of each pair. This means that the dissimilarity between each pair within the cluster is less than or equal to the level of the join. The lower on the tree the values are joined, the more similar they are.

Ward's method estimates similarity by calculating the change in variance that would occur if a pair of clusters were merged. It joins the two clusters that minimizes the intra-cluster variance by comparing the instance of each variable to the grand mean for the variable
by calculating:

$$F = \frac{\sum_i \sum_j \sum_k |X_{ijk} - \bar{x}_k|^2 - \sum_i \sum_j \sum_k |X_{ijk} - \bar{x}_{ik}|^2}{\sum_i \sum_j \sum_k |X_{ijk} - \bar{x}_k|^2}$$

This is essentially the ratio of the variance between the sample means over the variance within the samples. By maximizing this value, we minimize the variance within clusters and maximize the significant distance between clusters.

FIGURE 4: WARD'S METHOD DENDROGRAM



Using these dendrograms and trial and error to estimate the ideal number of clusters, we chose to cluster our data into 9 groups. We used k-means clustering which initializes k cluster centroids randomly and then runs

{

For every i,

$$c^{(i)} := arg_j \min \left|\left|x^{(i)} - \mu_j\right|\right|^2$$

For each j,

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

}

until convergence. Each district is placed in the cluster that minimizes the squared Euclidean distance between our 16-feature vector and centroids.

4.1. Polygon Irregularity

We calculated polygon irregularity using 4 indices:

$$S = 1 - \frac{2\sqrt{\pi a}}{p}$$

The Schwartzberg index measures indentation by comparing the district perimeter to that of a circle with equal area. p is the perimeter of the district and a is the area of the district.

$$P = 1 - \frac{4\pi a}{p^2}$$

The Polsby-Popper measures indentation by comparing the area of the district to the area of a circle with the same perimeter. p is the perimeter of the district and a is the area of the district.

$$R = 1 - \frac{a_{district}}{a_{circle}}$$

The Reock index measures dispersion of the district by getting comparing the area of the district to its minimum bounding circle.

$$C = 1 - \frac{a_{district}}{a_{convex\ hull}}$$

The convex hull measurement measures dispersion by comparing the area of the district to the area of its convex hull, or minimum convex bounding geometry.

5. Results

We performed k-means clustering on our data, and the number of iterations totaled 143 to meet a convergence factor of 0.0001. We determined several interesting correlations arising from this clustering. Mean income is significantly negatively correlated to each of the polygon irregularity indices (p-values ranging from 0.0003 to 0.03): higher the income, the less irregular the polygons (Table 1):

TABLE 1– CORRELATION COEFFICIENTS FOR INCOME VS POLYGON IRREGULARITY INDICES

	Schwartz-berg	Polsby	Reock	Convex Hull
Correlation coefficient	-0.84	-0.91	-0.73	-0.93
p-value	0.004	0.0007	0.03	0.0003

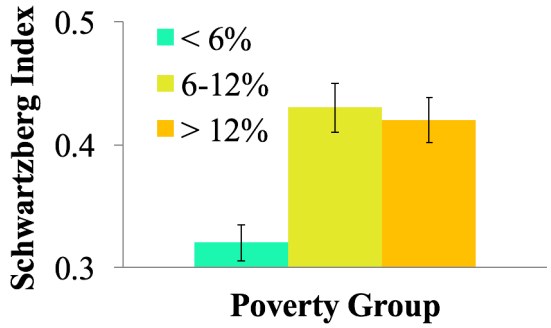
In order to arrive at a single measure of irregularity from the four different indices, we performed factor analysis to create a combined factor score. We found that the relationship between income and irregularity was preserved, with income still significantly correlated to the factor score (Table 2):

TABLE 2 – CORRELATION COEFFICIENT FOR INCOME VS COMBINED IRREGULARITY INDEX

	Factor Score
Correlation coefficient	-0.93
p-value	0.0003

We also found that the Schwartzberg index was associated with the below poverty index. We categorized percent below poverty into three groups – less than 6%, 6-12%, and above 12%. The Schwartzberg index for < 6% was significantly less than that for the 6-12%, with a p-value of 0.03. It was also significantly different from the index for the > 12%, with a p-value of 0.04 (Figure 5). However, the indices for 6-12% and >12% were not significantly different from each other. We did not find any significant associations between the racial breakdowns and the irregularity indices of the clusters.

FIGURE 5 – SCHWARTZBERG INDEX BY POVERTY GROUPS



6. Discussion

In this project, we look at the practice of school district gerrymandering, focusing on the use of geospatial and clustering techniques to examine boundary anomalies. In fact, the present algorithm evolved from several failed clustering attempts, the first of which involved generating feature vertices for each district boundary, calculating distance from each feature vertex to the centroid and thus capturing the shape, calculating angular distance to capture shape, and using this measure for clustering. We eventually used an updated algorithm that involved the minimum bounding geometry described. We found that many of our cluster centers

were quite close together, excluding cluster eight, which was more dissimilar to the other cluster centers.

One of our main findings was the clear correlation that we found between mean income and polygon irregularity; more specifically, the less irregular the districts, the higher the income of the district. This indicated to us the possibility that, as the income went up for a district, there was less need to create irregularities or gerrymander the school district boundaries, since there was likely no lack of resources to give to all schools. This is of particular concern when thinking about how low income areas are therefore far more likely to be subject to gerrymandering – those students that may need the equalizing factor of quality education the most are also those for whom resources are most likely to be taken away. This hypothesis is supported by the finding that, at the lowest percentages below poverty in a district, there is the least Schwartzberg irregularity. It was also interesting to note that we did not find relationships between the racial breakdowns of a district and the irregularity of the school district boundaries. Much of the traditional literature on gerrymandering, which is done qualitatively by examining particular districts, focuses on race and not income as a factor in redistricting. This project indicates that focusing more on income differences and poverty breakdown in trying to create equitable school districts may be more valuable as an approach.

7. Future

Much of the current project revolved around consolidating and compiling geographic and demographic data regarding school districts from several different sources and data types. Now that this rich dataset and our clustering analysis is available to us, our next step will be to create a model that can predict whether or not school districts, and more intriguingly, school attendance zones (smaller zones within districts), were subject to gerrymandering. Perhaps inclusion of features such as dependence on government assistance, or type of district (urban vs rural), would be included. For this to be possible, we would need to add to our current dataset with location tags for each school district, so our cluster analysis could more closely examine patterns regarding the location of districts within each cluster. We would love to create a predictive model that could be consulted when district lines are redrawn, to ensure that redistricting legislation is not due to gerrymandering. It would also be interesting to examine how the existence of gerrymandering in particular districts, predicted by our model, relates to those school district's budgets and earmarks from federal and state government. From that point, we could delve more deeply into examining resource allocation at the governmental level as it relates gerrymandering.

Appendix

APPENDIX 1.: CLUSTER CENTER INFORMATION FOR CLUSTERED DISTRICTS

	1	2	3	4	5	6	7	8	9
Schwartzberg	0.429	0.415	0.399	0.410	0.420	0.443	0.439	0.280	0.280
Polsby	0.656	0.640	0.622	0.635	0.646	0.670	0.666	0.475	0.609
Reock	0.549	0.553	0.550	0.549	0.545	0.543	0.540	0.531	0.541
Convex Hull	0.211	0.211	0.203	0.209	0.206	0.213	0.210	0.126	0.186
% white	88.8	91.1	73.3	84.5	86.3	90.0	90.4	80.4	82.9
% black	3.1	1.3	16.1	3.7	5.8	3.6	2.8	1.0	2.9
% asian	2.6	0.5	0.6	6.1	0.7	0.8	1.4	12.7	9.5
% other	5.5	7.1	10	5.7	7.2	5.6	5.4	5.9	4.7
Mean income	8.5e04	8.7e+04	4.4e+04	1.1e+05	5.3e+04	6.2e+04	7.2e+04	2.6e+05	1.5e+05
% below pov.	8	12	27	6	18	13	10	3	5

8. REFERENCES

1. Wright, J. Skelly. "Public School Desegregation: Legal Remedies for De Facto Segregation." New York University Law Review 40.2 (1965): 285-310.
2. Genevieve Siegel-Hawley. "Educational Gerrymandering? Race and Attendance Boundaries in a Demographically Changing Suburb." Harvard Educational Review: 83.4 (2013) 580-612.
3. Richards, Meredith and Kori Stroub. "An Accident of Geography? Assessing the Gerrymandering of School Attendance Zones." Teachers College Record 117.7 (2015) 1-32.
4. Lublin, David. "The Paradox of Representation: Racial Gerrymandering and Minority Interests in Congress." *Princeton University Press:* New Jersey, 1997.
5. Chen, Jowei and Jonathan Rodden. "Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures". Quarterly Journal of Political Science: 8 (2013) 239–269.
6. Altman, Micah, Micahel McDonald. "The Promise And Perils Of Computers In Redistricting." *Duke Journal of Constitutional Law and Public Policy:* 5.69 (2010), 70-111.
7. Chung-I Chou & S. P. Li. "Taming the Gerrymander—Statistical Physics Approach to Political Districting Problem." *Physica A: Stat. Mechanics & Its Applications* 799 (2006).
8. National Center for Education Statistics (2015). *Education Geographic and Demographic Information: School District Boundaries.*
9. United States Census Bureau (2015). *American Factfinder*.