Geospatial Clustering for Balanced and Proximal Schools

Subhodip Biswas, Fanglan Chen, Andreea Sistrunk, Sathappan Muthiah, Zhiqian Chen, Nathan Self, Chang-Tien Lu, Naren Ramakrishnan
Discovery Analytics Center
Department of Computer Science, Virginia Tech
{subhodip, fanglanc, sistrunk, sathap1, czq, nwself, ctlu, naren}@vt.edu

Abstract

Public school boundaries are redrawn from time to time to ensure effective functioning of school systems. This process, also called *school redistricting*, is non-trivial due to **(1)** the presence of multiple design criteria such as capacity utilization, proximity and travel time which are hard for planners to consider simultaneously, **(2)** the fixed locations of schools with widely differing capacities that need to be balanced, **(3)** the spatial nature of the data and the need to preserve contiguity in school zones, and **(4)** the difficulty in quantifying local factors that may arise. Motivated by these challenges and the intricacy of the process, we propose a geospatial clustering algorithm called Geo$K$means for assisting planners in designing school boundaries such that students are assigned to proximal schools while ensuring effective utilization of school capacities. The algorithm operates on polygonal geometries and connects them into geographically contiguous school boundaries while balancing problem-specific constraints. We evaluate our approach on real-world data of two rapidly growing school districts in the US. Results indicate the efficacy of our approach in designing boundaries. Additionally, a case study is included to demonstrate the potential of Geo$K$means to assist planners in drawing boundaries.

Introduction

The last decade has seen rapid advancement in geographical information systems (GIS) frameworks for dealing with geospatial data. They are integral to the operation of modern economies and play a key role in urban planning, transportation, logistics, distribution services, policy making and policy implementation (Burrough et al. 2015). Complementary GIS data can provide a rich source of information for researchers, planners and policymakers to study, analyze and make decisions upon. Despite its ubiquity and utility, GIS has seen slow adoption among school planners. Recent works have highlighted the promise of using geospatial context for studying school district boundaries (Yoon and Lubienski 2018). In fact, the operation of a school district generates a vast amount of geospatial data that can be analyzed to answer important questions regarding the long-term planning and design of school boundaries (Kelly 2019).

In the US, public schools operate through school districts, geographical units for the local administration of schools, which are usually demarcated by the boundaries of a city or a county. Within a district, school boundaries are formed around each school property by aggregating smaller areas, called student planning areas (SPAs), into larger regions, called school attendance zones (SAZs), such that the areas inside a region are geographically contiguous. To delineate the SAZs, school planners consider factors such as population balance, compactness, proximity, stability, spatial contiguity and demographics, among others. These criteria are often conflicting in nature so finding the right balance between them becomes crucial to the planning process. An objective treatment of the school redistricting problem by a data-driven model can assist school planners by providing them with automated plans, from which planners can adapt their own plans in the way they see fit for equitable distribution of educational resources.

Motivated by this, we view the process of school redistricting through the lens of spatially-constrained clustering (Miller and Han 2009), also called regionalization (Duque, Ramos, and Suriñach 2007), and devise a geospatial clustering algorithm called Geo$K$means for solving it. The proposed method is based on the K-means approach (MacQueen 1967) but differs in the following aspects: (1) *Geospatial data*: Conventional point-based modeling does not apply to school boundaries since the data consists of polygonal (areal) spatial objects which have richer representations. Our method leverages the structural and topological information contained within these geospatial features. (2) *Constrained assignment*. Unlike K-means, which assigns data points to the nearest cluster, our method performs a series of checks to ensure that problem-specific constraints, such as capacity balance, compactness and spatial contiguity, are satisfied during assignment. (3) *Restrictive search and weighing*. The need to ensure the presence of one school per cluster induces restriction on feasible clustering configurations. Additionally, since the boundary of the cluster will form the SAZ of the school it contains, its centroid cannot deviate too far from the school's location. Otherwise students may have to travel unacceptably far to reach their assigned schools. To ensure this, we incorporate an adaptive weighing mechanism.

Background

School Redistricting as a Regionalization Problem

Regionalization, or spatially-constrained clustering, aims to aggregate a set of N areal units into K spatially contiguous regions while optimizing on a predefined criteria (Duque, Ramos, and Suriñach 2007). As an $\mathcal{NP}$-hard problem, the number of potential solutions can be enormous and it is difficult to find the global optimum (Megiddo and Supowit 1984). As such, different heuristic approaches have been proposed for solving regionalization problems, including classical clustering (Openshaw 1995), hierarchical clustering (Guo 2008), optimization on regional attributes (Bacao, Lobo, and Painho 2005), and graph-based techniques (Assunção, Neves, and others 2006).

In designing regionalization algorithms, spatial contiguity and shape receive primary consideration unlike many conventional redistricting approaches that fail to leverage the entire complexity of the structural and topological information contained within the geospatial features. Redistricting methods make simplified assumptions during modeling without considering the geography inherent in the GIS data and focus on achieving population and demographic balance even at the cost of contiguity (Belford and Ratliff 1972; Franklin and Koenigsberg 1973; Holloway, Wehrung, and others 1975; Lemberg and Church 2000). Hence, we treat school redistricting as a regionalization/ spatially-constrained clustering problem such that the distance of students to schools is minimized while ensuring effective capacity utilization and spatial contiguity. Next, we proceed to a discussion of how the problem-specific constraints can be incorporated into a clustering framework.

Clustering with Supervision

Clustering is an unsupervised learning problem that is often posed as an optimization problem with an objective function. It can be solved via exact algorithms, approximation methods or heuristics. In some cases, *a priori* information, in the form of expert opinions, domain knowledge and geometric constraints, is incorporated in the clustering process to guide towards better partitioning of the data. This is called semi-supervised clustering (Chapelle, Schlkopf, and Zien 2010) or constrained clustering (Basu, Davidson, and others 2008).

Generally, there are three ways of incorporating domain knowledge in the clustering process: enforcing constraints (Wagstaff, Cardie, and others 2001), seeding (Basu, Banerjee, and Mooney 2002), and metric learning (Xing, Jordan, and others 2003; Basu, Bilenko, and Mooney 2004). For a regionalization problem like school redistricting, school locations can be regarded as seeds to the clustering process while the spatial contiguity of a region can be enforced by checking geographical adjacency between areas. Incorporating these domain-specific constraints into a distance-based technique like K-means, which tries to minimize a distance-based function, aligns with our goal of keeping students close to their assigned schools while ensuring balanced utilization of each school's capacity.

K-means and its Constrained Adaptations

Given a set of N data points $\mathcal{X} = \{x_1, \ldots, x_N\}, x_i \in \mathbb{R}^d$, the K-means algorithm creates a K-partitioning $\{\mathcal{X}_k\}_{k=1}^{K}$ of the dataset such that the mean squared distance between the datapoints and the K partition centers $\{\mu_1, \mu_2, \ldots, \mu_K\}$ is minimized, as captured by the following objective function:

$$\mathcal{F}_{K-means} = \sum_{k=1}^{K} \sum_{x_i \in \mathcal{X}_k} ||x_i - u_k||_2. \quad (1)$$

Each of the partition is a cluster and the membership of each data point/instance to a cluster is updated iteratively. The algorithm operates by executing the following steps until there is no change in the membership of the data points to clusters over successive iterations:

- Compute the distance of each data point $x_i, i = 1, \ldots, N$ to the center $\mu_k, k = 1, 2, \ldots, K$ of each partition/cluster using some distance function $d(.,.)$.
- Assign each data point x_i to the cluster $\mathcal{X}_k$ whose centroid μ_k is closest to it, i.e., $r = \underset{k}{\operatorname{argmin}}\ d(x_i, \mu_k)$.
- Recompute the centroid μ_r of every cluster $\mathcal{X}_r$ such that it is the mean of all the data points in the cluster.

The K-means algorithm has found widespread adoption due to its simplicity and ease of use. There are classical works that propose constrained adaptations of it (Bradley, Bennett, and Demiriz 2000; Wagstaff, Cardie, and others 2001; Basu, Banerjee, and Mooney 2002; Basu, Bilenko, and Mooney 2004). However, these classical variants are not suitable for the regionalization problem as they are designed to work with point-based data. Furthermore, they do not consider important problem-specific factors including population balance, contiguity and shape.

The Proposed Method: Geo$K$means

Geo$K$means deals with geospatial data in a constrained setting by adopting a hybrid approach. It starts with a subset of data points marked as seeds or initial clusters. Then, it repeats the following steps iteratively. Given a data point x, its nearest cluster $\mathcal{X}_r$ is selected based on geodesic distance and then a series of constraint checks are performed before assigning x to $\mathcal{X}_r$. If one of the checks fail, the next nearest cluster is selected and the constraints are likewise checked. This continues until a data point is assigned to a cluster. If no such cluster is found then x is marked as *unassigned* for the present iteration. This is the E-step. In the M-step, the cluster centroids are recomputed based on the constituent member instances. In the subsequent sections, we provide the preliminaries, quantify the constraints of the school districting problem and show how these constraints entail modifications in the basic K-means approach.

Preliminaries

Let $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ indicates a dataset where each data instance $x^{(i)}$ corresponds to a spatial (polygonal) unit (i.e. the SPA). Each SPA can be represented as

$$x^{(i)} = (\mathcal{A}, \mathcal{G}, \mathcal{C}),$$

where $\mathcal{A}$ is the set of coordinates (longitude and latitude) outlining the boundary of the SPA, $\mathcal{G} = (g_0, g_1, \ldots, g_{12})$ is the grade-wise student population residing within $\mathcal{A}$ and $\mathcal{C} = (c_{\text{ES}}, c_{\text{MS}}, c_{\text{HS}})$ is a tuple containing the capacities of the elementary, middle and high schools contained in the SPA. Usually, grade levels are distributed uniformly across each level of school (elementary, middle and high) throughout a district. Hence we can rewrite the student population counts as $\mathcal{G} = (g_{\text{ES}}, g_{\text{MS}}, g_{\text{HS}})$[1]. For brevity, we denote the school capacity and student count of SPA x as $c^x_{\mathcal{L}}$ and $g^x_{\mathcal{L}}$, respectively, where $\mathcal{L}$ corresponds to school level (ES for elementary school, MS for middle school and HS for high school). These N instances will be grouped together to form K partitions or clusters denoted by $\{\mathcal{X}_k\}_{k=1}^K$. Note that a cluster $\mathcal{X}_k$ correspond to the boundary of the k^{th} school, i.e., its SAZ.

Constraints

Given a dataset $\mathcal{X}$ of size N, a clustering (partitional) algorithm seeks to obtain a K-partitioning $\{\mathcal{X}_k\}_{k=1}^K$ such that the objective is minimized under the following conditions:

(C1) each instance is exclusively assigned to a cluster,

(C2) the number of clusters should be less than or equal to the number of areas, i.e., $K \leq N$, and

(C3) each cluster should be non-empty.

Next, some domain-specific constraints are introduced for the school redistricting problem as stated below:

(D1) the data instances (SPAs) within a cluster should be spatially contiguous, i.e., geographically connected by some contiguity relation,

(**D2**) each cluster should contain one school,

(**D3**) the total student population in a cluster should be close to the capacity of the school it contains,

(**D4**) each cluster should be as compact as possible, and

(**D5**) the centroid of the cluster should not be too far from the location of the school it contains.

Constraints can be *hard* or *soft*. The output of the clustering algorithm should satisfy the hard constraints, (**D1**) and (**D2**), while soft constraints, (**D3**), (**D4**) and (**D5**), can be relaxed with an associated penalty cost. These soft constraints are used for adjusting cluster-level properties, like size or diameter, and can be violated within a pre-specified threshold. Next we show how these constraints are introduced as semi-supervision to the algorithmic framework.

Incorporating semi-supervision via constraints

Seeding: A *seed set* $\mathcal{S} \subseteq \mathcal{X}$ is the subset of data instances for which supervision is available as follows: for each seed instance $x \in \mathcal{S}$, the cluster $\mathcal{X}_k$, $k = 1, 2, \ldots, K$, to which it belongs is given. Usually for each cluster $\mathcal{X}_k$, there is typically at least one seed point $x \in \mathcal{S}$. In our case, the data instances/ SPAs[2] containing a school inside them form the seed set $\mathcal{S}$ and each of them is assigned to a unique cluster. The subroutine for seed set generation is shown below.

Algorithm 1: Seeding

```
Input  : Dataset $\mathcal{X}$, School level $\mathcal{L}$
Output : Seed set $\mathcal{S}$, Partition $\{\mathcal{X}_k\}_{k=1}^K$
Method:
$k \leftarrow 0, \mathcal{S} \leftarrow \phi$
for $i = 1, 2, \ldots |\mathcal{X}|$ do
    if $x^{(i)}$ contains school of level $\mathcal{L}$ then
        $k \leftarrow k + 1$
        $\mathcal{S} \leftarrow \mathcal{S} \bigcup \{x^{(i)}\}$
        $\mathcal{X}_k \leftarrow \{x^{(i)}\}$
return $\mathcal{S}$, $\{\mathcal{X}_k\}_{k=1}^K$
```

In Algorithm 1, the school level $\mathcal{L}$ can take values from ES, MS or HS, depending or whether it is elementary, middle or high school redistricting, respectively. The seed set $\mathcal{S}$ is used to initiate the K-partitioning of the data. To be considered a seed point, corresponding school-level capacity should be positive, i.e., $c_{\mathcal{L}} > 0$. In short, seeding ensures that constraints (**C3**) and (**D2**) are initially satisfied and helps to guide the subsequent clustering process.

Constrained assignment: In K-means method, each data point $x^{(i)}$, $i = 1, 2, \ldots, N$, is associated to a cluster based on proximity. This does not consider domain-specific constraints which are important in context of this problem. Hence the constrained assignment performs these checks in two groups as outlined next.

Neighborhood checks: These checks ensure that the spatial contiguity of each cluster is preserved based on the notion of neighborhood. We consider two data instances (polygons) to be adjacent if they share a common border of any length between them (i.e. rook's contiguity), as shown in Figure 1. We can construct a matrix $\mathbf{W}$ which encodes the adjacency relationship between every pair of instances, $x, y \in \mathcal{X}$ s.t. $x \neq y$, as

$$W_{x,y} = \begin{cases} 1 & \text{if } x \text{ and } y \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

Adjacency check: For an instance $x^{(i)} \in \mathcal{X}$ to be assigned to cluster $\mathcal{X}_k$ it should follow: $\exists y \in \mathcal{X}_k$ s.t. $W_{x,y} = 1$. The adjacency check ensures that spatial contiguity is maintained when an instance is added to a cluster.

It is equally important to verify if the contiguity is preserved when an instance changes cluster membership (from *donor* to *recipient*). Assume that cluster $\mathcal{X}_r$ in Figure 1 is composed of polygons $\{A, C, D\}$ while cluster $\mathcal{X}_d$ consists of polygons $\{B, F, E\}$. Suppose at some iteration t, F was chosen to be assigned to $\mathcal{X}_r$ based on proximity. If the move happens, the contiguity in $\mathcal{X}_d$ will be broken even though $\mathcal{X}_r$

[1] If the school district has grades $K(0) - 5$, $6 - 8$ and $9 - 12$ for elementary, middle and high school, respectively, then

$$g_{\text{ES}} = \sum_{c=0}^{5} g_c \quad g_{\text{MS}} = \sum_{c=6}^{8} g_c \quad g_{\text{HS}} = \sum_{c=9}^{12} g_c,$$

[2] We shall use these terms interchangeably henceforth.

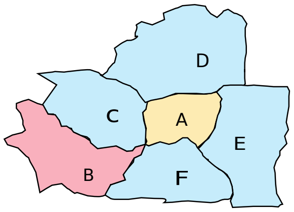


Figure 1: By rook's contiguity, polygon A is adjacent to all polygons except B as A and B intersect at a single point only.

will remain spatially contiguous. This is avoided by implementing the next check.

Spatial contiguity check: This is performed if an already assigned instance changes cluster membership. Consider an instance x moving from (donor) cluster $\mathcal{X}_d$ to (recipient) cluster $\mathcal{X}_r$, thereby resulting in new clusters $\mathcal{X}_d^{'}$ and $\mathcal{X}_r^{'}$. If $\mathbf{W}(\mathcal{X}_d^{'})$ is the adjacency matrix corresponding to cluster $\mathcal{X}_d^{'}$, we can use depth first traversal on it to determine the number of connected components in $\mathcal{X}_d^{'}$. If spatial contiguity is to be preserved, it should be 1 in both $\mathcal{X}_d^{'}$ and $\mathcal{X}_r^{'}$. Since adjacency check has already been performed, we only check $\mathcal{X}_d^{'}$ for contiguity so as to satisfy the hard constraint (**D1**).

Feasibility check: During the cluster assignment process, we can infer the quality of the resultant cluster configuration by using the state function defined below:

$$\mathcal{F}\left(\mathcal{X}_k\right) = w \times \mathcal{F}_1\left(\mathcal{X}_k\right) + (1-w) \times \mathcal{F}_2\left(\mathcal{X}_k\right), \quad (3)$$

where $w \in [0,1]$, is a weight parameter, $\mathcal{X}_k$ is the cluster under consideration, and $\mathcal{F}_1$ and $\mathcal{F}_2$ are cluster balance and compactness functions, as elucidated below:

- The balance function, $\mathcal{F}_1$, measures how well a cluster $\mathcal{X}_k$ balances the residing student population with respect to the capacity of the school it contains. It is calculated as

$$\mathcal{F}_1\left(\mathcal{X}_k\right) = \left| 1 - \frac{\sum\limits_{x\in\mathcal{X}_k} g_{\mathcal{L}}^x + \varepsilon_1}{\sum\limits_{x\in\mathcal{X}_k} c_{\mathcal{L}}^x + \varepsilon_2} \right|, \quad (4)$$

  where $\sum_{x\in\mathcal{X}_k} g_{\mathcal{L}}^x$ is the student population of school level $\mathcal{L}$ residing in $\mathcal{X}_k$, $\sum_{x\in\mathcal{X}_k} c_{\mathcal{L}}^x$ is the capacity of the school contained in it, ε_1 and ε_2 are infinitesimally small constants such that $\varepsilon_1/\varepsilon_2 \gg 1$. For invalid cluster configurations (e.g. those with no student population or which do not contain a school), $\mathcal{F}_1$ will take values ≥ 1.

- The compactness function, $\mathcal{F}_2$, of a cluster $\mathcal{X}_k$ quantifies how tightly its area $A_{\mathcal{X}_k}$ is packed into its boundary of perimeter $P_{\mathcal{X}_k}$. It is calculated by comparing the area of the cluster to the area of a circle with equal perimeter as

$$\mathcal{F}_2\left(\mathcal{X}_k\right) = 1 - 4\pi \cdot \frac{A_{\mathcal{X}_k}}{P_{\mathcal{X}_k}^2}. \quad (5)$$

  For a perfectly compact shape like a circle, this value will be 0 and will approach 1 asymptotically as the shape becomes less compact.

The value of the state function indicates how close a cluster is to its ideal state: $\mathcal{F}\left(\mathcal{X}_k\right) = 0$. We expect the value of the state function to decrease as the cluster grows: it becomes more compact while achieving better balance. The weight parameter w determines the relative importance of balance and compactness in evaluating a cluster's state.

If an instance x moves from donor cluster $\mathcal{X}_d$ to recipient cluster $\mathcal{X}_r$, resulting in new clusters $\mathcal{X'}_d$ and $\mathcal{X'}_r$, the move is considered feasible, if it satisfies

$$\Delta\mathcal{F}_{r\leftarrow d} = \mathcal{F}\left(\mathcal{X'}_r\right) + \mathcal{F}\left(\mathcal{X'}_d\right) - \mathcal{F}\left(\mathcal{X}_r\right) - \mathcal{F}\left(\mathcal{X}_d\right) \leq 0 \quad (6)$$

This check is introduced for balancing the soft constraints (**D3**) and (**D4**).

There are situations where no instance can be assigned to a cluster by failing the feasibility check. This stalls convergence. In such a case, we relax the condition by associating each unassigned instance to the adjacent cluster that causes minimum constraint violation (Equation 6) so as to minimize the violation of these constraints.

Adaptive weighing: For ensuring each student is assigned to a nearby school, the constraint (**D5**) needs to be satisfied. To do so, we use a weighted center $\gamma_{\mathcal{X}_k}$ that lies between the centroid $\mu_{\mathcal{X}_k}$ of cluster $\mathcal{X}_k$ and the location $\sigma_{\mathcal{X}_k}$ of the school inside it. It is calculated as

$$\gamma_{\mathcal{X}_k} = \alpha \times \sigma_{\mathcal{X}_k} + (1-\alpha) \times \mu_{\mathcal{X}_k}, \quad (7)$$

where α is the ratio of the cluster's student population to the capacity of the school it contains. It is calculated as

$$\alpha = \frac{\sum\limits_{x\in\mathcal{X}_k} g_{\mathcal{L}}^x}{\sum\limits_{x\in\mathcal{X}_k} c_{\mathcal{L}}^x}.$$

Assignment to cluster is performed based on $\gamma_{\mathcal{X}_k}$. We assume that initially when α has value near 0, the weighted center is close to the centroid $\mu_{\mathcal{X}_k}$, allowing freedom for cluster growth. As the cluster grows, the value of α tends to 1, $\gamma_{\mathcal{X}_k}$ starts approaching the school's location $\sigma_{\mathcal{X}_k}$.

Termination: To check for converge, we use a counter T, initially set at 0. T is incremented by 1 whenever the following occurs (in order): there is (a) no decrease in the number of unassigned data instances (given not all are assigned), (b) no decrease in the objective criteria, i.e., MSSC, or (c) no change in cluster membership. It is reset otherwise. We terminate the algorithm when T exceeds some threshold $T_{\max}$.

Putting all together: The complete pseudocode of our Geo$K$means is shown in Algorithm 2. The code is available at https://github.com/subhodipbiswas/GeoKMeans.

Experimentation

Data

We use GIS data from two US school districts located in the mid-Atlantic region. Both districts have recently seen rapid growth in population and have undergone several school boundary processes, thereby making them suitable choices. Summary statistics for the districts are provided in Table 1. For our study, we used the following geographic data:

- SPA: Geometric coordinates and grade-wise student count
- School: Location, capacity and level

Algorithm 2: Geo$K$means

```
Input  : Dataset $\mathcal{X}$, Adjacency matrix $\mathbf{W}$, School level $\mathcal{L}$
Output : Final Partition $\{\mathcal{X}_k\}_{k=1}^{K}$
Method:
$\mathcal{S}$, $\{\mathcal{X}_k\}_{k=1}^{K} \leftarrow$ seeding $(\mathcal{X})$
terminate $\leftarrow$ False, $t \leftarrow 0$
$\widetilde{\mathcal{X}} \leftarrow \mathcal{X} - \mathcal{S}$
while not terminate do
    // E-step (constrained assignment)
    $\mho \leftarrow []$
    for $x \in \widetilde{\mathcal{X}}$ do
        $flag \leftarrow False, s \leftarrow 0$
        while $s < |\mathcal{S}|$ do
            $r \leftarrow \arg\min_{l} ||x - \gamma_{\mathcal{X}_l}||^2$
            if $\exists y \in \mathcal{X}_r$ s.t. $W_{x,y} = 1$ then
                if $x$ is unassigned $||$
                $x : \mathcal{X}_r \leftarrow \mathcal{X}_d$ preserves contiguity then
                    Compute $\Delta\mathcal{F}_{r \leftarrow d}$
                    if $\Delta\mathcal{F}_{r \rightarrow d} \leq 0$ then
                        Assign/move $x$ to $\mathcal{X}_r$
                        $flag \leftarrow True$
                        break
            Remove $\mathcal{X}_r$ from $\mathcal{N}_x^*$
            $s \leftarrow s + 1$
        if not flag && $x$ is free then
            $\mho \leftarrow \mho \bigcup \{x\}$        // Unassigned instances
    if $|\mho| > 0$ then
        Repeat E-step $\forall x \in \mho$
    // M-step
    $t \leftarrow t + 1$
    Update the clusters $\{\mathcal{X}_k\}_{k=1}^{K}$
    terminate $\leftarrow$ check_termination()
if $|\mho| > 0$ then
    Repeat the main-loop $\forall x \in \mho$ without feasibility check
return $\{\mathcal{X}_k\}_{k=1}^{K}$
```

Table 1: School district data summary

District	# SPA	# Schools		
		ES	MS	HS
X	454	55	16	15
Y	1315	138	26	24

Metrics

Given the schools' boundaries in a partition, we adopt three performance metrics to evaluate their quality:

- *Balance:* This is a mean percentage score that tells us how well a cluster balances it student population as compared to the student capacity of the school it contains.

$$\frac{1}{K}\left(\sum_{k=1}^{K} 100 \times \left|1 - \left|1 - \sum_{x \in \mathcal{X}_k} g_{\mathcal{L}}^{x} \Big/ \sum_{x \in \mathcal{X}_k} c_{\mathcal{L}}^{x}\right|\right|\right). \quad (8)$$

  A score of 100 indicates full utilization of the capacity while scores below 100 indicate schools operating at over capacity and/or being under-utilized.

- *Compactness:* This computes on an average how tightly packed the perimeter $P_{\mathcal{X}_k}$ of a cluster is with respect to the circumference of a circle whose area is equal to the area $A_{\mathcal{X}_k}$ of the cluster.

$$\frac{1}{K}\sum_{k=1}^{K} 100 \times \left(P_{\mathcal{X}_k} \Big/ \left(2\pi\sqrt{A_{\mathcal{X}_k}/\pi}\right)\right)^{-1} \quad (9)$$

  This metric is a percentage score, with a circle (perfectly-compact) achieving the value of 100.

- *Proximity:* This is the geodesic distance (in miles) that students needs to travel on average to reach their assigned school. This weighted measure is calculated as

$$\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\sum_{x \in \mathcal{X}_k} g_{\mathcal{L}}^{x} \times d_{\mathrm{mi}}(\sigma_{\mathcal{X}_k}, \mu_x)}{\sum_{x \in \mathcal{X}_k} g_{\mathcal{L}}^{x}}\right). \quad (10)$$

  This score is a rough approximation of the travel-time of a student in the event of such data being unavailable.

Models

We compare the performance of the following models.

Regionalization methods: These are heuristics that start with an initial random spatially-contiguous partition of the dataset and improve them locally. We consider the average performance over 51 independent runs.

- AZP: Automatic Zoning Procedure is a classical regionalization approach initially designed for reorganizing census geographies in the UK (Openshaw and Rao 1995).
- SARA: Simulated Annealing Redistricting Algorithm operates by partitioning a set of populated zones into spatially-contiguous regions so as to minimize the population difference between the regions (Macmillan 2001).

To ensure these algorithms generate feasible solutions, i.e., one school per partition, we employed seeding technique.

Geo$K$means and its counterparts: These variations are instantiated by selectively activating the checks in the constrained assignment step.

- SKM: This is identical to a K-means with seeding enabled but constrained assignment deactivated.
- CKM: Adjacency check added to SKM.
- SCKM: Spatial-contiguity check added to CKM.
- *GeoKM* : Adding the feasibility check to SCKM results in our proposed algorithm.

Since each variant starts with identical seed sets, the difference in their performance is due to the constraint checks.

Parametric Setup

Default parameter settings are used for SARA and AZP. For Geo$K$means and its variants we set the threshold for stagnation at 5. The value for weight paramter w (Equation 3) was set to 0.9.

Table 2: Model performance for redistricting in both school districts.

District X

Models	Elementary School			Middle School			High School		
	Balance	Compactness	Proximity	Balance	Compactness	Proximity	Balance	Compactness	Proximity
AZP	84.99	54.32	1.20	89.14	47.27	2.44	86.54	45.38	2.98
SARA	70.06	49.97	1.07	86.06	41.12	2.14	78.74	36.47	1.70
SKM	79.58	48.32	**0.41**	78.79	44.54	**0.95**	73.01	40.49	1.01
CKM	79.46	52.59	0.55	85.93	45.45	1.04	86.69	45.66	**0.97**
SCKM	79.69	52.87	0.56	89.31	44.79	0.95	87.10	42.64	1.09
GeoKM	**90.26**	**59.02**	0.68	**93.94**	**58.10**	1.19	**94.85**	**51.79**	1.27

District Y

Models	Elementary School			Middle School			High School		
	Balance	Compactness	Proximity	Balance	Compactness	Proximity	Balance	Compactness	Proximity
AZP	88.28	49.03	0.77	88.98	36.77	1.80	88.88	38.13	1.84
SARA	69.53	48.88	0.67	83.12	36.65	1.58	78.74	36.47	1.70
SKM	67.44	45.42	**0.26**	80.54	35.94	**0.69**	87.79	36.50	**0.81**
CKM	75.59	49.31	0.48	87.35	39.41	1.08	89.08	41.32	1.76
SCKM	82.02	46.93	0.51	85.77	38.60	1.39	86.41	38.68	1.11
GeoKM	**91.23**	**56.78**	0.41	**93.08**	**41.13**	1.10	**94.64**	**48.35**	1.19

Results

All algorithms provide an approximate solution to the school redistricting problem as it is NP-hard in nature. For elementary schools, the problem is more challenging because of the low number of instances per cluster, i.e., N/K is 8.25 and 9.53 for Districts X and Y, respectively. Also, the arbitrary distribution of the elementary schools (assumed centers) contradicts basic clustering assumptions made by distance-based methods. The middle and high school cases present relatively easier problems as they are more uniform in their distribution. Next, we proceed to discuss the performance of the baseline algorithms for different cases, compare plans generated by our algorithm with existing plans and discuss the utility of automated plans to school planners.

How effective is Geo$K$means in making plans?

We independently ran simulations for every algorithm at all school levels (ES, MS and HS) and tabulated the evaluation metrics in Table 2. We observe that SKM outputs partitions with proximal schools but at the cost of balance, which rarely crosses 80 in District X. Except for District Y's high schools, which are distant and well-separated, the partitions generated by SKM are highly imbalanced and therefore would not be considered viable plan for adoption. This is expected since SKM tries to assign SPAs to schools only based on distance via the adaptive weighing technique. Other factors like adjacency, spatial contiguity and population balance are ignored. On accounting for these factors via constraint checks, both CKM and SCKM show an overall improvement in balance scores at the expense of schools being farther. This is akin to real-life boundary planning process, since adding in more considerations constrains the set of feasible plans.

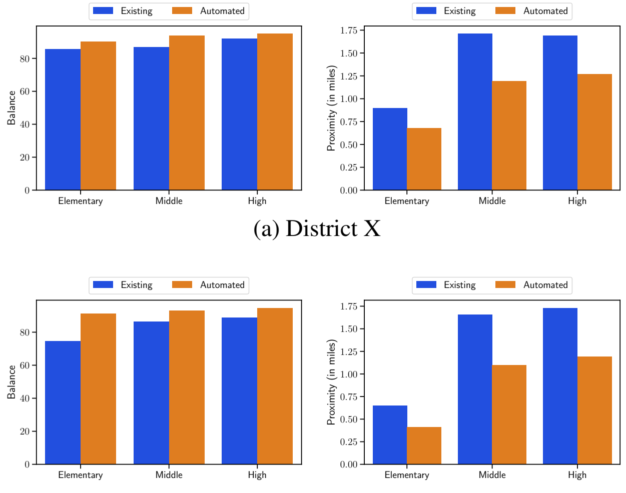
Though CKM and SCKM perform better than the unconstrained SKM, they are not noticeably better in balancing the elementary school population. The haphazard distribution of elementary schools coupled with high variance of student population in SPAs poses a challenging problem scenario for distance-based method like K-means that implicitly assumes uniform distribution of population across the school district. Given the land-use patterns of a county/school district, there are pockets of residential areas with high student population density. Balancing the students in such areas is difficult without explicitly accounting for them. The feasibility check is found to be useful in such scenarios. On activating this check, Geo$K$means improves by 12.6% (81.11 to 91.33 for District X) and 11.23% (82.02 to 91.23 for District Y) in balance score over its nearest-performing self-variant. Such improvements are also noticed for middle and high school cases, especially in District Y.

SARA and AZP operate by swapping polygons located on the boundary of clusters to improve balance without accounting for proximity. In all the possible cases, we notice that AZP outperforms SARA in terms of balance scores but at the cost of proximity. Adopting such a final partition (plan) in real-life would burden the school district with increased transportation costs as most of the students' residences will lie outside walking distance. On the other hand, with a proximity-based assignment, Geo$K$means generates better-balanced partitions with compact boundaries and nearby schools in all possible cases. In comparison to AZP, Geo$K$means improves proximity in District X by 43.3% (1.20 mi to 0.68 mi) for elementary schools, 51.2% (2.44 mi to 1.19 mi) for middle schools, and 57.4% (2.98 mi to 1.27 mi) for high schools.

Automated plans vs existing plans

School districts undergo boundary change processes when the need arises to redraw school attendance zones in response to present needs and the predicted forecasts. Usually these existing plans are balanced and reflective of the present scenario. To test the utility of our approach in generating real-life plans, we plot the balance and proximity of automated plans generated by Geo$K$means with existing plans from both districts in Figure 2.

Compared to the existing plans, the automated plans im-



(a) District X

(b) District Y

Figure 2: Bar plots comparing the automated plans against the existing plans in terms of capacity balance on left (higher is better) and distance to schools on right (lower is better). Overall, the automated plans achieve better performance.

prove the population balance in schools while enhancing the proximity of the schools to students. Figure 2a shows that existing and automated plans for District X are fairly well balanced. The proximity to elementary, middle and high schools improves by 32.2% (0.90 mi to 0.68 mi), 31% (1.71 mi to 1.19 mi) and 24.8% (1.69 mi to 1.27 mi). In Figure 2b, we observe identical trends. There is a 19.7% increase in capacity balance (74.73 to 93.08) along with 36.9% reduction in distance (0.65 mi to 0.41 mi) for elementary schools in District Y. Middle and high schools see a 33.7% (1.66 mi to 1.1 mi) and 31.2% (1.73 mi to 1.19 mi) decrease in distance.

Geo$K$means in real-life planning: A case study

For the purpose of case study on District X, we highlight the differences between the existing plan and the automated plan for elementary and middle schools in Figure 3 via choropleth map and distribution plots of the metrics. We notice that the automated plans have comparatively higher balance scores (darker is better). The distribution plots also reflect improvements in balance and proximity values in the automated plans. Geo$K$means is particularly good at balancing the growing population in the southeastern part of District X. However, there are occasional lighter patches in the automated plan as well, particularly in the western part. On further analysis of the land-use patterns of District X, we notice that the western section of the district and parts of the central section are zoned for low density housing and agricultural uses. Planners and politicians alike are driven to preserve the district's rural nature in these areas. The demographic data reveals that both have seen a steady decline in student populations over the years and have underutilized schools. In the southeastern part, rapid new home sales in the area's residential developments have led to unprecedented overcrowding in nearby schools.

District X planners anticipate the opening of four new schools in the next five years. Until each school is built and opened, the existing plan remains obscured. As each school is built and opened and populations fluctuate, district planners must reassess the situation. Hence, an automated plan can serve as an alternative suggestion from which planners can borrow ideas during actual boundary processes. This is particularly helpful when planners have projected estimates of future student enrollment and wish to make long-term plans by simulating future scenarios.

Conclusion

In this article, we propose a geospatial clustering technique called Geo$K$means which integrates the proximity-based assignment of traditional clustering algorithms with a constrained assignment mechanism. Through extensive experimentation on two real-world school district datasets, we demonstrated the advantage of our approach for designing school boundaries. Our results show the improvements of automated plans over existing plans and how they can serve as a guideline for planners during boundary processes. As such Geo$K$means can be applied to a plethora of zone-design problems where proximity plays an important role.

We have also identified some challenges and propose future research directions. Firstly, our framework works with fixed geometries. Having the ability to further fragment a polygon may yield better results, especially in areas with high population density or unbalanced schools. Secondly, to make the algorithmic plan-making more akin to real-life process we can incorporate other local factors/constraints like past rezonings, geographic barriers, political boundaries, etc. Lastly, we would like to be able to develop an interactive framework that can incorporate the feedback from actual stakeholders into the clustering process.

Acknowledgments. This research is supported in part by National Science Foundation grants DGE-1545362 and IIS-1633363. We acknowledge Susan Hembach, Colin Flynn, Pranita Ranbhise, Paul Ngo and Jessica Gillis for providing us with the data and helpful insights regarding the school districting process. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any school board, NSF, or the U.S. Government.

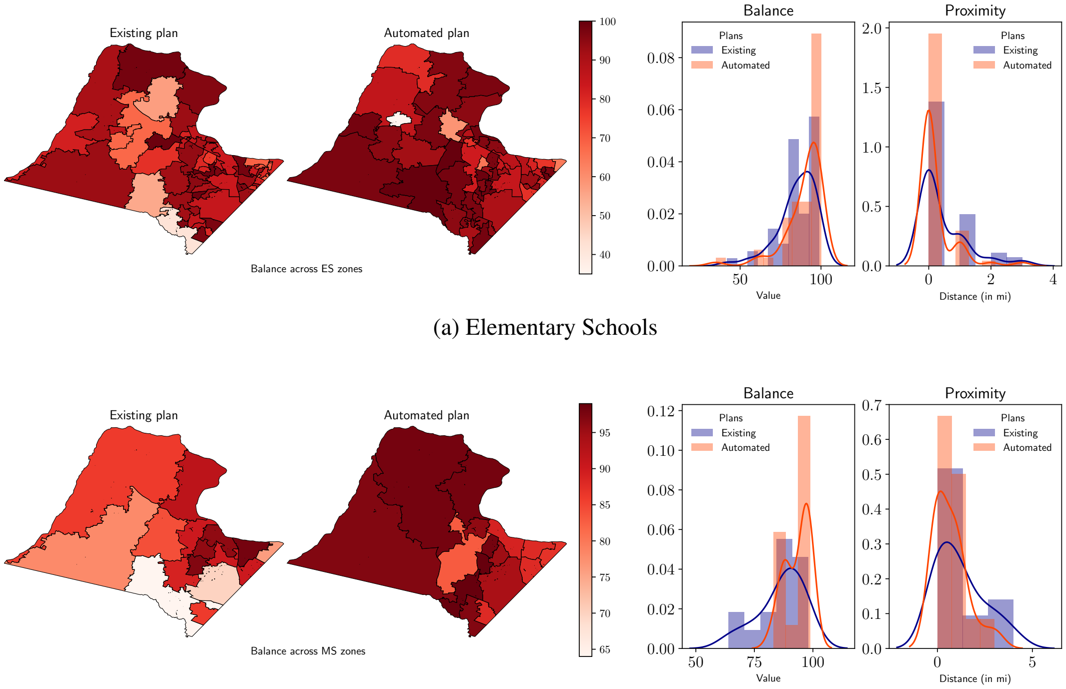
References

Assunção, R. M.; Neves, M. C.; et al. 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science* 20(7):797–811.

Bacao, F.; Lobo, V.; and Painho, M. 2005. Applying genetic algorithms to zone design. *Soft Computing* 9(5):341–348.

Basu, S.; Banerjee, A.; and Mooney, R. 2002. Semi-supervised clustering by seeding. In *Proceedings of 19th International Conference on Machine Learning*, ICML.

Basu, S.; Bilenko, M.; and Mooney, R. J. 2004. A probabilistic framework for semi-supervised clustering. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.



(a) Elementary Schools

(b) Middle Schools

Figure 3: Chloropeth maps on the left showcase the capacity balance (darker the better) for the automated plan and the existing plan. The automated plan showcases better balance in both elementary and middle school cases. The distributional plots on the right showcase the capacity balance and proximity scores. We see that the automated plan, shown in red, has better capacity balance while maintaining a low commute distance.

Basu, S.; Davidson, I.; et al. 2008. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press.

Belford, P. C., and Ratliff, H. D. 1972. A network-flow model for racially balancing schools. *Operations Research* 20(3):619–628.

Bradley, P.; Bennett, K.; and Demiriz, A. 2000. Constrained *k*-means clustering. Technical Report MSR-TR-2000-65, Microsoft Research, Redmond.

Burrough, P. A.; McDonnell, R.; McDonnell, R. A.; and Lloyd, C. D. 2015. *Principles of geographical information systems*. Oxford University Press.

Chapelle, O.; Schlkopf, B.; and Zien, A. 2010. *Semi-Supervised Learning*. The MIT Press, 1st edition.

Duque, J. C.; Ramos, R.; and Suriñach, J. 2007. Supervised regionalization methods: A survey. *International Regional Science Review* 30(3):195–220.

Franklin, A. D., and Koenigsberg, E. 1973. Computed school assignments in a large district. *Operations Research* 21(2):413–426.

Guo, D. 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science* 22(7):801–823.

Holloway, C. A.; Wehrung, D. A.; et al. 1975. An interactive procedure for the school boundary problem with declining enrollment. *Operations Research* 23(2):191–206.

Kelly, M. G. 2019. A map is more than just a graph: Geospatial educational research and the importance of historical context. *AERA Open* 5(1):1–14.

Lemberg, D. S., and Church, R. L. 2000. The school boundary stability problem over time. *Socio-Economic Planning Sciences* 34(3):159–176.

Macmillan, W. 2001. Redistricting in a GIS environment: An optimisation algorithm using switching-points. *Journal of Geographical Systems* 3(2):167–180.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1(14):281–297.

Megiddo, N., and Supowit, K. J. 1984. On the complexity of some common geometric location problems. *SIAM journal on computing* 13(1):182–196.

Miller, H. J., and Han, J. 2009. *Geographic data mining and knowledge discovery*. CRC press.

Openshaw, S., and Rao, L. 1995. Algorithms for reengineering 1991 census geography. *Environment and planning A* 27(3):425–446.

Openshaw, S. 1995. Classifying and regionalizing census data. *Census users' handbook* 239–270.

Wagstaff, K.; Cardie, C.; et al. 2001. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML.

Xing, E. P.; Jordan, M. I.; et al. 2003. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, NeurIPS.

Yoon, E.-S., and Lubienski, C. 2018. Thinking critically in space: toward a mixed-methods geospatial approach to education policy analysis. *Educational Researcher* 47(1):53–61.