REGAL: A Regionalization framework for school boundaries

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

ABSTRACT

Due to constant shifts in population and changing demographics, school boundary processes take place to make adjustments to school attendance zones. This spatial problem has multiple criteria like locations of schools, their capacity utilization, proximity, presence of geographical/ man-made barriers, etc. In this paper, we formulate the problem of designing school boundaries as a spatially-constrained clustering/ regionalization problem and propose an automated approach called REGAL for solving it. REGAL is two-stage framework that starts by creating a candidate solution with regard to domain constraints such as school locations and spatial contiguity. Then a local search method improves the quality of the candidate solution by optimizing population balance and compactness of school zones while satisfying problem constraints. Experimentally, we demonstrate the efficacy of the REGAL framework on actual datasets from two school districts in the US.

CCS CONCEPTS

• **Information systems** → *Clustering*; • **Theory of computation** → *Optimization with randomized search heuristics.*

KEYWORDS

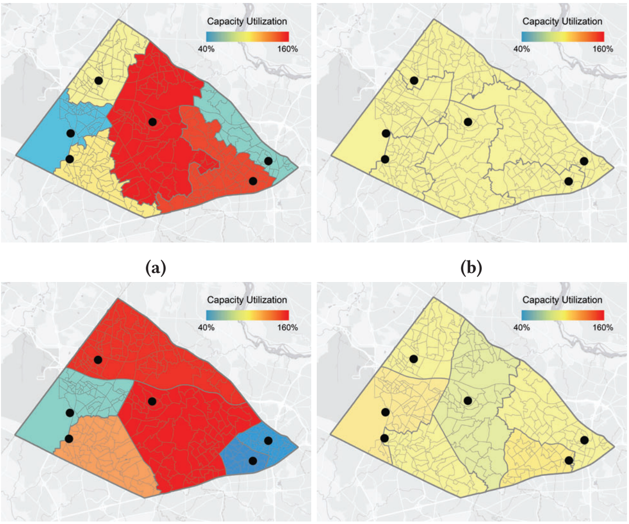
constrained clustering, local search, optimization, spatial clustering

ACM Reference Format:
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 *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359377

1 INTRODUCTION

In the US, public school systems function through school districts, geographical areas where schools share the same administrative structure [6]. Usually, the boundaries of a county or a city determine the jurisdictional area of a school district, within which each school has a designated geographical area known as *school attendance zone* (SAZ). A SAZ outlines where students in a particular neighborhood will attend public elementary, middle, and high school. Each SAZ spans across smaller-sized geographical areas called *planning units* or *student planning areas* (SPAs) [7]. The School Board may modify a SAZ to maintain or improve operational efficiency and/or to maximize instructional effectiveness. In general, adjustments may relieve facility crowding, ensure better utilization of existing space, better allocate program resources, reduce operating costs, and/or avoid underutilizing school facilities. This problem is an application of *districting* [4], where majority of existing works have focused on political districting [9], sales territory design [10], etc. Scant attention has been paid to the geography of schools [1].

Motivated by this, we revisit the school boundary formation problem by making use of the geospatial data of two school districts in the US. Starting from school locations, boundaries are formed by aggregating the smaller areas (SPAs) into larger regions (SAZs) such that the areas inside a region are geographically contiguous. While demarcating the SAZs, a school planner aims to balance factors such as school capacity, compactness, proximity, stability, spatial contiguity, demographics, etc. In Figure 1, we show some possible scenarios that can arise while designing SAZs. This is a *spatially-constrained clustering* problem, also called *regionalization* [2].



(a) (b)

(c) (d)

Figure 1: The boundaries of schools (black dots) are formed by grouping the smaller SPAs into larger, geographically contiguous SAZs. Considering factors like compactness and population balance, SAZs can fall into the following scenarios: (a) unbalanced schools with non-compact SAZs (b) balanced schools with non-compact SAZs (c) unbalanced schools with compact SAZs, and (d) balanced schools with compact SAZs (desirable).

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11…$15.00
https://doi.org/10.1145/3347146.3359377

Considering these factors, we formulate school boundary formation as a non-linear discrete ($\mathcal{NP}$-hard) optimization problem and propose a two-stage REGionalization Algorithm via Local search, called REGAL, for solving it. The first stage initializes a candidate solution (vector) that satisfies problem-specific constraints: SPA assignment must be mutually exclusive and into spatially contiguous SAZs; and, there must be one school of each level per SAZ. Then in the second stage, the compactness and capacity utilization of the candidate solution is improved by applying a local search method.

2 PROBLEM STATEMENT

Areas. Let $\prod = \left\{\pi^{(1)}, \pi^{(2)}, \cdots, \pi^{(N)}\right\}$ represent the set of SPAs where $|\prod| = N$.

Attributes. Let each SPA be denoted by a tuple

$$\pi = (L, P, C),$$

where, $L = [(x_1, y_1), (x_2, y_2), \ldots, (x_t, y_t), (x_1, y_1)]$ is the set of geographic coordinates (latitude and longitude) defining the boundary polygon of a SPA, $P = \left(n_0,\, n_1,\, n_2,\, ...,\, ,\, n_{12}\right)$ is the vector of grade-wise (K-12) student population residing[1] in this SPA, and $C = \left(c_{\text{ES}}, c_{\text{MS}}, c_{\text{HS}}\right)$ is a vector containing the capacities of any elementary school (ES), middle school (MS) or high school (HS) present in this SPA. Most SPAs do not contain a school, and thus have $C = (0, 0, 0)$. Alternatively, we can aggregate grade-wise student counts into ES, MS and HS student populations such that $P = \left(n_{\text{ES}}, n_{\text{MS}}, n_{\text{HS}}\right)$, where

$$n_{\text{ES}} = \textstyle\sum_{g=0}^{5} n_g \quad n_{\text{MS}} = \sum_{g=6}^{8} n_g \quad n_{\text{HS}} = \sum_{g=9}^{12} n_g,$$

by assuming that every ES, MS and HS consists of grades K-5, 6-8 and 9-12, respectively. Mappings of grade levels to school levels are generally consistent across all the schools in a school district.

Spatial relationship. Let $\mathcal{G}(\prod) = (V, E)$ be the contiguity/ adjacency graph associated with SPAs $\prod$ such that each SPA $\pi^{(i)} \in \prod$ has a corresponding node $v^{(i)} \in V$, and there exists an edge $(v^{(i)}, v^{(j)}) \in E$ between two nodes if and only if their respective SPAs, $\pi^{(i)}$ and $\pi^{(j)}$, have an edge in common (rook contiguity). For most real world school districts, $\mathcal{G}(\prod)$ is a fully connected graph. This graph can be encoded as an adjacency list and is used while determining spatial contiguity of a region.

Valid partition. Let $\mho = (\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(K)})$ denote a partition of N areas into K regions[2] where

- $|\mathcal{S}^{(k)}| > 0 \quad \forall k \in \{1, 2, \ldots, K\}$,
- $\prod = \bigcup_{k=1}^{K} \mathcal{S}^{(k)} \quad \forall k \in \{1, 2, \ldots, K\}$,
- $\mathcal{S}^{(k)} \bigcap \mathcal{S}^{(k')} = \phi \quad \forall k, k' \in \{1, 2, \ldots, K\}$ and $k \neq k'$.

Any partition is considered *valid* if each of its constituent regions

- is fully connected, i.e. $\mathcal{G}(\mathcal{S}^{(k)})$ is connected $\forall k \in \{1, 2 \ldots, K\}$,
- contains one school inside it.

The set of all valid partitions together form the feasible search space Ω of the problem.

Criteria/Desirability. A valid partition $\mho \in \Omega$ is considered desirable if each of its constituent regions $\mathcal{S}$ have

- total residing student population roughly equal to the base capacity of the school inside it, and
- a geographically compact shape.

The desirability of a region $\mathcal{S}$ is dependent on the following:

(1) **Target balance**. In a region, the balance between the residing student population T and the base capacity c of the school they attend is calculated as

$$\mathcal{U}(\mathcal{S}) = \left|1 - \frac{T + \epsilon_1}{c + \epsilon_2}\right|, \tag{1}$$

where $|\cdot|$ indicates absolute value and ϵ_1, ϵ_2 are infinitesimally small constant such that $\epsilon_1/\epsilon_2 \gg 1$. This score normalizes across schools of widely varying capacity and identifies invalid regions (i.e., not containing a school) by assigning very high values.

(2) **Target compactness**. It is determined in a *non-linear* manner by comparing the area A of a region (shape) to the area of a circle with equal perimeter p as

$$\mathcal{V}(\mathcal{S}) = 1 - \frac{4\pi \mathrm{A}}{\mathrm{p}^2}. \tag{2}$$

The score ranges from 0 to asymptotically approaching 1, where 0 means perfectly compact (i.e., a circle). It is not possible for the score to reach 1 since any region must necessarily have non-zero area and perimeter.

Both target scores are normalized to lie in the range [0, 1], and they reflect how far a region is from its target state of 0: the lower the score, the better its desirability.

Objective function. Given the target scores for every region, we define the objective of the problem as:

$$\mathcal{F}(\mho) = \sum_{\mathcal{S} \in \mho} w \times \mathcal{U}(\mathcal{S}) + (1 - w) \times \mathcal{V}(\mathcal{S}), \tag{3}$$

where $w \in [0, 1]$ is the weight parameter for balancing the above criteria. The value of w is empirically set based on design preference. Our goal is to obtain the partition $\mho^*$ that best minimizes the objective function $\mathcal{F}$. Hence the problem can be formulated as:

$$\mho^* = \underset{\mho \in \Omega}{\arg\min}\, \mathcal{F}(\mho). \tag{4}$$

3 THE PROPOSED FRAMEWORK: REGAL

REGAL solves the problem of school boundary formation in two stages. In Figure 2, we show the components of our framework. They are discussed in detail in the subsequent subsections.

3.1 Initialization

It starts by identifying SPAs that contain schools and marking them as *seed areas*. Each seed area is uniquely assigned to a new region (seeded region) in order to ensure only one school per region. This is the *seeding phase*. The seeding strategy is tailored to the application at hand and is guided by the constraints inherent to the problem.

The seeded regions are partial clusters that need to be grown using the adjacency relationship contained in $\mathcal{G}(\prod)$. To do so, we select a region $\mathcal{S}$ randomly at each step for growth. If $\mathcal{S}$ shares a

[1] The entire residing population is assumed to attend public schools.

[2] The value of K varies depending on whether we are considering ES, MS or HS boundary formation.

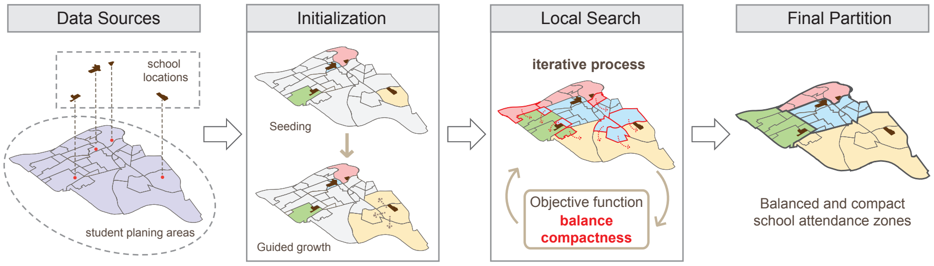


Figure 2: Outline of the REGAL framework.

boundary with unassigned areas, they are added to $\mathcal{S}$ in no specific order. The growth process is repeated until every SPA has been assigned to a cluster. It outputs a partition $\mho_0$ such that every region in it is spatially-contiguous. These areas are randomly assigned to clusters without consideration for the quality of the region. This results in some regions being far from their target state, having a distorted, non-compact shape, or having disproportionate student populations. Hence, a local search procedure is applied in the next stage to balance these clusters and improve solution quality.

3.2 Local search

It starts with the (feasible) solution returned by the previous stage and searches its immediate neighborhood for better solutions. Before searching, a neighborhood relation needs to be established to ensure that the neighboring solutions are feasible. The spatial contiguity constraint is imposed on neighboring solutions by using rook contiguity (neighbors sharing an edge).

Given a solution $\mho$, its neighboring solution set $\mathcal{N}^*(\mho)$ is constructed by altering the membership of areas located on the boundary of two regions, i.e., by moving an area from its present region (donor) to a neighboring region (recipient). Hence, every candidate solution has multiple neighboring solutions and each neighboring solution differs by a single assignment. How to move through neighboring candidate solutions is based on local improvement (at region-level) in the objective function. The algorithm keeps track of the best solution found in each iteration and terminates when there is no improvement in the functional value for a predefined number of steps. We integrate three well-known search methods within our framework as elucidated below.

- Stochastic Hill Climbing (SHC) [8]
- Simulated Annealing (SA) [5]
- Tabu Search (TS) [3]

In SHC, the search is somewhat akin to a steepest descent algorithm except that the order of picking up better or equally good solutions (uphill moves) from the neighboring solution set is random. Though SHC is a fast algorithm, it is prone to getting stuck in a local optimum due to its greedy nature. SA and TS avert this trapping by probabilistically allowing downhill moves[3].

4 EXPERIMENTATION

In this section, we detail the experimental setup, including the dataset, model parameters, metrics, and discuss the results.

4.1 Dataset

For this study, we collaborated with two rapidly growing school districts, say *District A* and *District B*, located in the mid-Atlantic region of the US. District A was divided into 1315 SPAs and contained 188 schools– 138 ES, 26 MS and 24 HS. District B had 454 SPAs and 86 schools– 55 ES, 16 MS and 15 HS. The data consists of the following shapefiles:

- SPA: Geographical coordinates of the area and grade-wise count of student population
- School: Location coordinates, school type and capacity.

4.2 Parametric setup

We set the value of w in the objective function (Eq. 3) to 0.8 and 0.7 for *District A* and *District B* respectively. The local search procedures are run until there is no improving moves for 3 consecutive iterations and the best solution is returned. The parameters for search methods are set based on literature and are given below.

- SA: Using cooling rate $\alpha = 0.85$, we vary the temperature T from 0 to 10^{-9}.
- TS: Tabu list of length 80 was used.

4.3 Performance Metrics

Regional metrics. These metrics are used to assess the quality of an individual region $\mathcal{S}$ in the partition $\mho$.

- **Balance score (BS)**: If a region $\mathcal{S}$ contains a school with capacity c and has T students attending that school, then the balance score is computed as

$$\mathbb{B}(\mathcal{S}) = 100 \cdot \left(1 - \left|\frac{c-T}{c}\right|\right), \tag{5}$$

  If a school's attending population is equal to its capacity, it will have a perfect balance score of 100.
- **Compactness score (CS)**: If a region's shape has area A and perimeter p, it's compactness is calculated as

$$\mathbb{C}(\mathcal{S}) = 100 \cdot \left(\frac{4\pi \mathrm{A}}{\mathrm{p}^2}\right), \tag{6}$$

  Given this metric, a circle would be the most compact shape with a score of 100.
- **Quality score (QS)**: The quality of a region, say $\mathcal{S}$, determines how close $\mathcal{S}$ is to it's target state as

$$\mathbb{Q}(\mathcal{S}) = w \cdot \mathbb{B}(\mathcal{S}) + (1-w) \cdot \mathbb{C}(\mathcal{S}), \tag{7}$$

  where w is the weight parameter (see Equation 3). The higher the score, the better is the quality of the region.

[3] Accepting inferior solutions is a randomization move that helps to escape local optima and perform a more extensive search for the global optimal solution.

Table 1: Comparing the performance of local search techniques for designing the school attendance zones of both the districts.

District A									
Schools	Elementary School			Middle School			High School		
Models	ABS	ACS	AQS	ABS	ACS	AQS	ABS	ACS	AQS
Stochastic Hill Climbing	80.637 ± 0.309	35.159 ± 1.135	71.542 ± 0.360	88.892 ± 0.396	27.221 ± 2.466	76.558 ± 0.597	94.296 ± 1.040	23.320 ± 2.281	80.101 ± 0.933
Simulated Annealing	**80.845 ± 0.381**	36.295 ± 1.426	71.935 ± 0.453	88.822 ± 0.500	27.576 ± 2.267	76.573 ± 0.648	94.457 ± 1.136	22.329 ± 2.179	80.032 ± 1.026
Tabu search	80.819 ± 0.276	**36.637 ± 1.408**	**71.982 ± 0.307**	**88.994 ± 0.214**	**29.988 ± 2.559**	**77.193 ± 0.541**	**95.19 ± 0.542**	**25.447 ± 2.704**	**81.241 ± 0.681**
District B									
Schools	Elementary School			Middle School			High School		
Models	ABS	ACS	AQS	ABS	ACS	AQS	ABS	ACS	AQS
Stochastic Hill Climbing	89.298 ± 0.862	36.147 ± 1.298	73.352 ± 0.687	93.433 ± 1.062	32.370 ± 3.617	75.114 ± 1.284	95.843 ± 2.255	30.483 ± 2.992	76.235 ± 1.752
Simulated Annealing	**90.324 ± 0.883**	**36.976 ± 1.623**	**74.320 ± 0.777**	**93.835 ± 0.703**	33.398 ± 3.305	75.704 ± 1.123	**96.879 ± 1.905**	28.941 ± 2.827	76.497 ± 1.503
Tabu search	90.012 ± 0.877	36.303 ± 1.704	73.899 ± 0.791	93.777 ± 0.977	**33.844 ± 3.362**	**75.797 ± 1.173**	96.866 ± 2.100	**30.760 ± 3.176**	**77.034 ± 1.511**

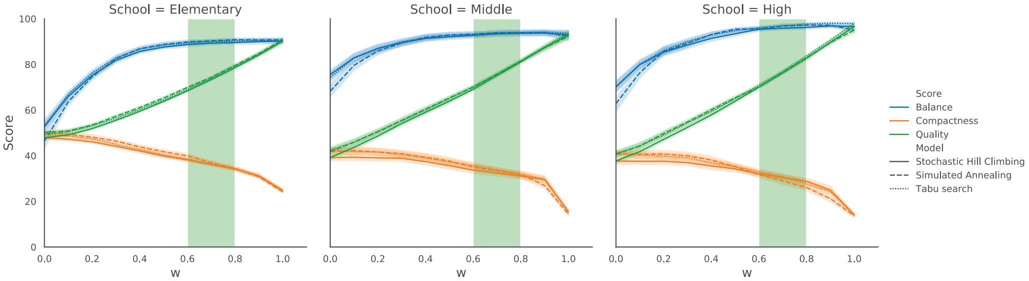


Figure 3: Error plots showing the variation of configuration metrics with parameter w for District B: ABS (balance) in green, ACS (compactness) in orange and AQS (quality) in blue. The desirable range for parameter w is marked by the green patch.

Configuration metrics. To evaluate the quality of an output configuration/ partition $\mho$ we use configuratioin metrics, which are mean of the earlier defined regional metrics. They are as follows:

- **Average balance score (ABS)**
- **Average compactness score (ACS)**
- **Average quality score (AQS)**

4.4 Results

In real-world, practical design considerations should guide the selection of parameters. The parameter w (see Equation 3) controls the relative importance of population balance and compactness objectives in the final partition which is expected to have good population balance (high ABS score) and compactness (high ACS score). To identify a suitable range for parameter w, every model was simulated 51 times by varying the value of w from 0 to 1 in steps of 0.1 and the error plots of the configuration metrics are obtained. For space limitation, we only show the plot for District B in Figure 3. We observe the ideal range for parameter w to lie in $[0.6, 0.8]$. For District A and B, we set w equal to 0.8 and 0.7, respectively. and report the mean and standard deviation of the configuration metrics in Table 1 for comparing the local search methods. We didn’t observe a significant difference in the performance of these techniques. The randomizing moves in TS and SA gives them edge over the greedy SHC. We also compared the plan generated by TS with the highest quality score with the existing ones in both school districts. Overall, we noticed the automated plans to improve the balance in schools without sacrificing the compactness of SAZs.

5 CONCLUSION

In this paper, we present school boundary formation as a discrete non-linear optimization problem and propose a regionalization framework called REGAL for solving it. Experimentally we demonstrated the efficacy of our approach on two real-life school datasets for designing school boundaries. The REGAL approach seems to be a useful tool for school planners to use during the school boundary process. One important implication of this work is that REGAL can be an effective tool connecting a mathematical model’s ability to handle complexity and human’s intuition and experience to solve a highly subjective spatial problem like school redistricting.

REFERENCES

[1] Felipe Caro, Takeshi Shirabe, Monique Guignard, and Andrés Weintraub. 2004. School redistricting: Embedding GIS tools with integer programming. *Journal of the Operational Research Society* 55, 8 (2004), 836–849.
[2] Juan C Duque, Raúl Ramos, and Jordi Suriñach. 2007. Supervised regionalization methods: A survey. *International Regional Science Review* 30, 3 (2007), 195–220.
[3] Fred Glover. 1990. Tabu search: A tutorial. *Interfaces* 20, 4 (1990), 74–94.
[4] Jörg Kalcsics. 2015. Districting problems. In *Location science*. Springer, 595–622.
[5] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220, 4598 (1983), 671–680. https://science.sciencemag.org/content/220/4598/671
[6] Horace Mann and William B Fowle. 1841. *Common School Journal*. Vol. 3. Marsh, Capen, Lyon, and Webb.
[7] Meredith P Richards. 2014. The gerrymandering of school attendance zones and the segregation of public schools: A geospatial analysis. *American Educational Research Journal* 51, 6 (2014), 1119–1157.
[8] Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.
[9] Justin C Williams Jr. 1995. Political redistricting: a review. *Papers in Regional Science* 74, 1 (1995), 13–40.
[10] Andris A Zoltners and Prabhakant Sinha. 1983. Sales territory alignment: A review and model. *Management Science* 29, 11 (1983), 1237–1256.