EDUCATION & WORK

Data and Methodology

Redrawing the Lines: How Purposeful School System Redistricting Can Increase Funding Fairness and Decrease Segregation

By: Jordan Abbott

Abstract

School district boundaries in the United States perpetuate educational inequity by linking local property wealth to school finance mechanisms. This creates systematic funding disparities that correlate with racial and economic segregation. This technical white paper presents a computational framework for optimizing district boundaries to simultaneously improve funding equity and reduce segregation.

Our approach employs a two-stage algorithmic pipeline centered on a critical innovation. First, we use spatial clustering according to a method known as SKATER (Spatial 'K'luster Analysis by Tree Edge Removal) to generate geographically coherent and maximally compact initial configurations that can be systematically varied across different numbers of districts. This initialization strategy dramatically reduces computational requirements, compared to traditional redistricting algorithms that may require hundreds of thousands of iterations. Our method achieves similar performance with far fewer iterations, approximately 2.5 times the number of census tracts per state. Following SKATER initialization, our second stage, Markov Chain Monte Carlo (MCMC) optimization, refines boundaries while enforcing hard constraints including contiguity, minimum enrollment thresholds, and infrastructure capacity limits. We systematically explore district configurations, which we allow to comprise a number of districts ranging from 25 to 175 percent of each state's current district count, and generate Pareto-optimal proposals on our optimization metrics.

Optimization targets three Theil indices, measuring disparities in property tax capacity (T-index for assessed property value per pupil), racial segregation (multigroup H-index), and economic segregation (binary H-index for poverty status), combined through weighted normalized aggregation. Default weights of 3:1:1 prioritize equality of property tax capacity per pupil and equally weight racial and economic segregation.

We implement three modeling scenarios: tract-level optimization, offering maximum potential optimization gains (the "blank slate model"); district-level consolidation, preserving existing administrative structures (the "merger model"); and universal county-level school district boundaries, which provide an interpretable baseline. The framework provides policymakers with quantified trade-offs between equity maximization and practical implementation constraints, establishing both theoretical frontiers and feasible pathways toward more equitable resource distribution.

1. Problem Formulation

The optimization of school district boundaries represents a high-dimensional combinatorial problem where geographic units must be assigned to districts while satisfying multiple objectives and constraints. Unlike legislative redistricting, which focuses primarily on population balance and compactness, school redistricting must simultaneously consider property tax capacity, demographic integration, infrastructure capacity, and political viability. This section formalizes the mathematical framework underlying our optimization approach.

1.1 Multi-Objective Optimization Framework

We formulate school district redistricting as a multi-objective optimization problem where census tracts serve as atomic geographic units that must be assigned to districts. Let $G = \{1, 2, \ldots, n\}$ represent the set of census tracts in a state, and let $D = \{1, 2, \ldots, k\}$ represent the set of districts, where k varies systematically. Each tract $i \in G$ must be assigned to exactly one district $d \in D$, creating a partition of the geographic space.

The optimization seeks to minimize an objective function combining three equity dimensions:

$$f(D) = w_1 \cdot T_{\text{val}}(D) + w_2 \cdot T_{\text{racial}}(D) + w_3 \cdot T_{\text{econ}}(D)$$

where T_{val} measures disparities in per-pupil property tax capacity, T_{racial} captures multigroup racial segregation, and T_{econ} quantifies economic segregation based on binary poverty status. Default weights are set to $w_1 = 3.0$, $w_2 = 1.0$, and $w_3 = 1.0$, reflecting a priority on addressing disparities in property tax capacity while maintaining focus on integration objectives. These weights can be adjusted to reflect different policy priorities or to explore trade-offs along the Pareto frontier.

1.2 Mathematical Formulation of Theil Indices

Theil's T for Disparities in Property Tax Capacity

The Theil T-index captures inequality in assessed property values per pupil across districts:

$$T_{\text{funding}} = \sum_i p_i \cdot \frac{x_i}{\mu} \cdot \log\left(\frac{x_i}{\mu}\right)$$

where:

$p_i = n_i/N$, the proportion of total students in district i

n_i = number of children in district i

N = total children in the state

x_i = assessed property value per pupil in district i

μ = state mean assessed value per pupil

This formula directly measures the capacity for local revenue generation, as property assessments form the tax base for school funding. The index equals zero when all districts have identical per-pupil property values, and increases with greater inequality.

Multigroup Theil's H for Racial Segregation

For racial integration, we employ the multigroup entropy-based Theil H-index:

$$T_{\text{racial}} = \frac{E_{\text{state}} - E_{\text{weighted}}}{E_{\text{state}}}$$

where:

$E\ \_\ \text{state} = -\sum_r (\pi_r \cdot \log(\pi_r))$, the entropy of racial composition at state level

π_r = proportion of racial group r in total state enrollment

$E\ \_\ \text{weighted} = \sum_i (p_i \cdot E_i)$, the enrollment-weighted average of district entropies

$E_i = -\sum_r (\pi_{ir} \cdot \log(\pi_{ir}))$, the entropy within district i

π_{ir} = proportion of group r in district i

The index ranges from 0 (perfect integration, where every district mirrors state demographics) to 1 (complete segregation). This multigroup formulation avoids the limitations of binary indices and captures the full complexity of racial composition.

Binary Theil's H for Economic Segregation

Economic segregation uses a similar entropy-based approach with two groups:

$$T_{\text{economic}} = \frac{E_{\text{state}} - E_{\text{weighted}}}{E_{\text{state}}}$$

Applied to children in poverty versus not in poverty, this measure captures the concentration of economic disadvantage across districts. The binary formulation is appropriate, given the policy relevance of poverty thresholds for federal program eligibility.

1.3 Constraint Specifications

Hard Constraints

1. Geographic Contiguity: Each district d must form a connected component under rook adjacency (shared edges, not just vertices).

2. **Minimum Enrollment Threshold:** Each district must contain at least m students, where $m = 0.5 \times \text{min(current district enrollments in each state)}$

3. **Infrastructure Capacity:** For each district d: $\sum_d(\text{children}) \leq 1.25 \times (\text{capacity})$

2. Data Architecture

2.1 Geographic Foundation

Geographic Units

Census tracts serve as the fundamental geographic units for redistricting, providing a standardized nationwide framework with sufficient granularity to capture neighborhood-level variation. We use 2020 census tract boundaries from the Census Bureau, encompassing approximately 80,000 tracts. These polygons define the atomic units that cannot be subdivided during optimization.

Tract adjacency relationships are computed using rook contiguity, establishing edges only between tracts sharing boundaries rather than vertices. This adjacency matrix forms the foundation for contiguity constraint checking and move generation during optimization. Disconnected components (e.g., islands, water boundaries) are connected via minimum distance stitching between nearest tract centroids to ensure graph connectivity.

2.2 Property Valuation Data

Property valuation data is provided by the Center for Geospatial Solutions at the Lincoln Institute of Land Policy, pre-aggregated at our geographic units of analysis. This dataset provides total assessed property values from the most recent year available, which directly determine local education revenue capacity. Unlike market values or sale prices, assessed values reflect the actual tax base available to districts.

Despite broad coverage, approximately 8 percent of census tracts lack property assessment data. To maximize nationwide coverage, we implement a quantile regression approach using XGBoost trained on tracts with known values. The model leverages demographic, economic, and spatial features from the American Community Survey including median household income, educational attainment, housing characteristics, and spatial lag variables, capturing neighboring tract values. It is enriched by land use data provided by Landsat data from the United States Geological Survey. We train three separate models for the tenth, fiftieth, and ninetieth percentiles of the property value distribution, providing both imputed values and uncertainty bounds. The width of the prediction interval directly indicates imputation confidence. Narrow intervals suggest reliable predictions based on similar tracts, while wide intervals flag areas of uncertainty. The median (fiftieth percentile) serves as the imputed value, but we integrate only high-confidence observations. This approach ensures that imputations are restricted to cases where the model demonstrates high confidence, rejecting the quartile of most uncertain predictions. After imputation, states with less than 70 percent geographic or 75 percent population coverage were not included in our analysis.

2.3 Demographic Data

Demographic Composition

Demographic data come from two primary sources. Racial and ethnic composition for five categories (White non-Hispanic, Black, Native American, Asian, and Hispanic/Latino) are obtained from the American Community Survey 5-year estimates (2018–2022) for the population ages 5–17. Economic status, specifically child poverty rates, are derived from the Census Bureau's Small Area Income and Poverty Estimates program. This approach captures all school-age children regardless of current enrollment status.

Enrollment Capacity Estimation

School capacity constraints are derived from historical maximum enrollment data for each school from the past decade, which represent the practical capacity without new construction. Schools are geocoded to identify the census tract in which they are physically located, and capacity is calculated by aggregating to the tract level by summing all schools within each tract's boundaries and multiplying by 1.25. This sets an upper bound for the number of children that can be assigned to a simulated district, based on the physical capacity of its current school buildings. This geographic assignment ensures that redistricting respects the physical location of educational infrastructure and existing facility constraints.

2.4 Data Integration

School District Mapping

Existing school district boundaries require careful processing due to the complex structure of American educational governance. We include only districts with full data coverage for more than 75 percent geographic coverage for their underlying census tracts. Where elementary and secondary districts overlap, we map elementary districts to their corresponding unified or secondary district to avoid double-counting. In cases where only elementary or secondary districts exist, we use those boundaries and their corresponding data directly. This process yields approximately 10,500 school districts with sufficient demographic, property assessment, and spatial data for analysis.

3. Algorithmic Framework

The optimization pipeline employs a three-stage approach: spatial clustering for initialization, simulated annealing for refinement, and systematic variation across district counts to explore the solution space. This section details the technical implementation of each stage and the mechanisms for constraint enforcement.

3.1 Stage 1: SKATER Initialization

The optimization begins with SKATER, which generates geographically coherent initial district configurations through constrained graph partitioning. This approach is critical to our computational efficiency. By starting from maximally compact configurations rather than random assignments, we reduce the required iterations from hundreds of thousands (typical in redistricting literature) to approximately 2.5 times the number of census tracts in each state. This reduction in computational expense allows us to complete the first school system redistricting analysis that is national in scope.

SKATER constructs a minimum spanning tree from the tract adjacency graph using edge weights based on scaled geographic coordinates. Each resulting partition forms a contiguous district by construction, eliminating the need for post-hoc contiguity repair that plagues random initialization approaches often used in MCMC optimization.

3.1.1 Capacity Repair Mechanism

When SKATER produces initial configurations violating capacity constraints, a repair mechanism attempts to restore feasibility before optimization. For each violating district, border tracts are evaluated for reassignment to neighboring districts with available capacity. The repair process attempts up to 50 chained explorations of 2,000 iterations each.

3.1.2 Status Quo Fallback Strategy

In cases where the SKATER initialization does not satisfy our constraints, and the chained repair strategy is unable to resolve the issue, we implement a fallback initialization approach. The system instead starts the MCMC optimization engine from the configuration of census tracts most similar to the status quo configuration of school districts. Because census tracts are not conterminous with existing school districts, we assign each tract to its geographic majority overlap district. The optimization process naturally varies the number of districts until it converges at three million iterations or reaches 25 percent of the status quo district count. This fallback mechanism ensures that the optimization can proceed, resolving the issue of invalid starting points.

3.2 Stage 2: MCMC Optimization

3.2.1 Algorithm Structure

Following SKATER initialization, Markov Chain Monte Carlo simulated annealing refines district boundaries to minimize our previously defined multi-objective function, improving our optimization objectives while maintaining all constraints.

3.2.2 Move Generation and Temperature Management

At each iteration, the algorithm selects a tract for potential reassignment. Border tracts are identified and preferentially selected, as they represent the only tracts that can change districts while maintaining contiguity. The selected tract is proposed for reassignment to a randomly chosen adjacent district.

The algorithm starts with a high temperature, its willingness to accept worse solutions, and it gradually becomes more selective over time, reducing this acceptance rate by 1 percent after each step, using the formula $T(t+1) = T(t) \times 0.99$. Beginning with full openness to any move $T_0 = 1.0$, the process continues until it has made 2.5 times as many successful changes as there are geographic units in the redistricting plan.

3.2.3 Constraint Validation

Each proposed move undergoes validation along three criteria before acceptance:

Population Threshold: The source district must maintain a minimum population, set at a given state's current lowest enrollment district.

Capacity Constraint: A district cannot exceed 125 percent of the sum of the maximum historical enrollment of schools geographically located within the new district.

Contiguity Preservation: Both source and target districts must remain contiguous after the reassignment, verified through connected component analysis.

3.3 Three Model Variants

The algorithmic framework is applied to two distinct redistricting models and a programmatic county-level merge, each offering different trade-offs between optimization flexibility and implementation feasibility.

Model 1: Blank-Slate Redistricting (Tract-Level Optimization)

This model uses census tracts as atomic units, providing maximum flexibility to create optimal boundaries. This model can completely reconfigure districts without regard to existing boundaries. It establishes the theoretical frontier for equity improvements.

Model 2: County-Based Redistricting

This model implements a programmatic, county-based consolidation, assigning all tracts within each county to a single district. No optimization is required, since it serves as a baseline showing what simple administrative consolidation achieves versus algorithmic optimization.

Both optimization models undergo systematic variation from 25 to 175 percent of current district counts. Each produces tract-to-district assignments with complete Theil index calculations, enabling direct comparison of equity impacts. The tract-level model demonstrates maximum theoretical improvements. The consolidation model balances feasibility with equity gains. The county benchmark validates the value of optimization over simple administrative boundaries.

Model 3: Redistricting by Merger (Optimized District Consolidation)

This model uses existing school districts as atomic units, preserving current administrative structures while allowing mergers. The same SKATER-optimization framework operates on a district adjacency graph rather than tract-level data. This provides more politically feasible solutions that maintain district identities while still pooling resources.

4. Output Specification

4.1 Output Format

The optimization framework produces a comprehensive data structure capturing results across all dimensions of analysis. For each state, model type (“Blank Slate” tract-level optimization, county-based redistricting, or merger-based redistricting), and k value (number of districts), we store both the initial SKATER-generated assignment and the optimized configuration after simulated annealing.

Each configuration includes complete tract-to-district assignment vectors alongside a comprehensive metric suite: the three Theil indices (tax capacity, race, and poverty status), Polsby-Popper compactness scores, and additional metrics. This structure enables systematic comparison across varied numbers of districts, revealing that different district configurations can achieve more equitable resource distribution. The format remains consistent across all three models, which facilitates direct comparison of their relative performance.

4.2 Pareto Frontier Construction

Rather than selecting a single "optimal" solution, we identify the set of Pareto-efficient configurations that represent different trade-offs among competing objectives. A configuration enters the Pareto frontier only if no other configuration performs better on all objectives simultaneously. The Pareto selection evaluates configurations across four criteria: minimizing the three Theil indices and maximizing geographic compactness (Polsby-Popper).

The frontier reveals critical trade-offs that cannot be resolved through technical optimization alone. Some configurations achieve significant funding equity but maintain racial segregation, while others integrate diverse populations at the cost of funding disparities. These trade-offs require policy decisions about relative priorities rather than technical solutions.

4.3 Configuration Selection

While the Pareto frontier presents all efficient options, practical implementation requires selecting a single configuration. We implement a weighted scoring system that allows prioritization among objectives while maintaining technical rigor in plan selection.

4.3.1 Normalization

Within each state, metrics are scaled 0–1, using normalization to ensure comparability across metrics with different natural scales. For the Theil indices, lower values are better, so the minimum value receives a score of 0, where the associated plan demonstrates the most even distribution on its optimization metrics, and the maximum receives 1. For compactness, higher values are better, so the scale is inverted. Normalization allows meaningful aggregation of metrics with varying distributions.

4.3.2 Weighted Aggregation

Normalized scores are combined using policy-determined weights that reflect four relative priorities:

- Equality of per-pupil tax capacity (Theil T): weight = 3
- Racial integration (Theil H): weight = 1
- Economic integration (Theil H): weight = 1
- Compactness: weight = 0.5

These default weights prioritize tax base equity while maintaining focus on integration objectives and geographic coherence. The weighted score for each configuration equals the sum of each normalized metric multiplied by its corresponding weight. The configuration with the lowest weighted score is selected as the recommended plan for each state. This selection is performed only among Pareto-efficient configurations, ensuring the chosen plan is not dominated by any alternative.

4.4 Flexibility and Sensitivity

The framework's key strength is its flexibility to accommodate different policy priorities. Stakeholders can adjust weights to explore how different value systems affect optimal configurations. This approach transforms a complex multi-objective optimization problem into a structured decision process. Rather than claiming to identify a single "best" solution, we provide a menu of high-quality options and a transparent mechanism for selection based on explicit policy priorities. This preserves the role of democratic decision-making while ensuring that choices are informed by rigorous analysis of trade-offs.

5. Technical Limitations

5.1 Algorithmic Limitations

The local search nature of simulated annealing cannot guarantee global optimality. While SKATER provides high-quality initialization and our iteration count ensures thorough exploration, the final configuration may represent a local optimum. The fixed weight selection (3.0:1.0:1.0) represents one point on the Pareto frontier; other weight combinations would yield different optimal configurations.

5.2 Data and Modeling Assumptions

Our analysis uses the most recent available demographic data but acknowledges that populations shift over time. School-age populations can change through migration, aging, and birth rate variations, potentially altering optimal boundaries. Infrastructure capacity, based on historical maximums, may not reflect building conditions or modernization needs.

We also recognize that redistricting itself could induce demographic and economic changes. Property values might adjust in response to new district assignments, as school quality perceptions significantly influence real estate markets.

5.3 Implementation Realities Not Modeled

Political feasibility remains the largest unmeasured barrier to implementation. Communities have strong attachments to existing districts based on identity, tradition, and property values. Our model cannot capture the political capital required to implement boundary changes, especially those crossing municipal or racial lines. Though transportation costs are mitigated by the capacity constraint, they are not fully modeled. Proposed boundaries might require longer bus routes or violate natural boundaries or geographic features, including highways, mountains, or rivers.

Computational limitations force several simplifications. We use rook contiguity rather than actual road networks, potentially creating districts that are technically contiguous but practically disconnected. The census tract as an atomic unit, while computationally necessary, may split natural neighborhoods.