Tracking Enrollment Demographics in Somerville Schools: Two Decades of Change

Sylvia Li

Introduction

As an education policy student, I am often curious about how school choice policies play out in real communities over time. While much of the academic literature focuses on theoretical frameworks or short-term studies, I wanted to understand the long-term demographic patterns that emerge when families are given choice within a public school system. Somerville, Massachusetts, presented an ideal case study because it operates an open enrollment system where families can apply to any school regardless of neighborhood, and it includes the fascinating case of the Healey School ("Enrollment Data (2005-24) - Arthur D Healey (02740075)"), which merged its "choice" and "neighborhood" programs around 2010 in pursuit of greater equity.

This project emerged from both personal curiosity and professional interest in understanding whether school choice policies inadvertently contribute to demographic sorting, even in progressive communities with good intentions. Rather than relying on district-provided summaries or annual reports, I wanted to systematically collect two decades of enrollment data to identify patterns that might not be visible in shorter-term analyses. The technical challenge of web scraping provided an additional learning opportunity to develop data collection skills that are increasingly valuable in education research, where comprehensive longitudinal datasets are often unavailable or expensive to access.

From a policy perspective, this analysis addresses critical questions about the effectiveness of choice-based reforms. Understanding how demographic patterns evolve over time can inform discussions about transportation policy, information dissemination, program placement, and resource allocation. While the technical methodology could easily be adapted to study other districts, the substantive findings contribute to broader debates about achieving equity and integration in public education systems that embrace parental choice.

Data Collection and Analysis Process

The data scrapping process is divided into 5 steps from extracting a single page to visualizing all the data across years.

Step 1: Single Page Data Extraction In the first step, I developed the foundational scraping technique by targeting a single school profile page from the Massachusetts Department of Education website. I extracted two critical tables: the race/ethnicity enrollment table and the gender enrollment table. The race/ethnicity table provided percentage breakdowns by demographic group, while the gender table contained the total enrollment figure needed for calculating absolute student numbers. I then formatted this data by using the first row as column headers, converted percentage strings to numeric values, and calculated the actual number of students in each demographic category by multiplying percentages by total enrollment.

Step 2: Function development and testing I transformed my single page scraping code into a reusable function called `get_demog_table()` that could process any MA DOE school profile page. This function included robust error handling to manage pages with missing data or formatting inconsistencies. I standardized the output format to include demographic group, category, school percentage, district percentage, state percentage, total enrollment, and abbreviated category labels. The function was tested on multiple schools to ensure reliability across different data formats and edge cases, establishing the foundation for large-scale data collection.

Step 3: Scrapping all the pages I identified all seven Somerville public schools and created a systematic scraping plan covering 20 years of data (2005-2024). This involved generating 140 unique URLs (7 schools × 20 years) and implementing respectful web scraping practices with random delays between requests to avoid overwhelming the server. I created a data structure to track each URL, school identifier, school name, and year, then systematically scraped all pages using the tested function. Progress tracking was implemented to monitor the success rate and identify any problematic pages during the collection process.

Step 4: Data integration and standardization In this crucial step, I processed all 140 scraped pages to extract demographic data using the standardized function. I combined individual school-

year datasets into a comprehensive dataset, ensuring consistent column naming and data types across all observations. The integration process included adding metadata (year, school ID, school name) to each observation and filtering out any failed extractions. I then saved the final dataset in both CSV and RDS formats, creating a clean, analysis-ready dataset with 686 observations covering demographic trends across all Somerville schools over two decades.

Step 5: Data Visualization and Trend Analysis I created a visualization showing demographic trends across all seven Somerville schools from 2005-2024. The analysis focused on the four largest demographic groups (White, Hispanic or Latino, Black or African American, and Asian) to ensure clear, readable charts. The visualization employed professional styling with consistent color coding and clear labeling for showing clear demographic patterns over time.

Data Overview

The final dataset contains 966 observations spanning 20 years (2005-2024) across seven Somerville public schools, representing one of the most comprehensive longitudinal views of demographic change in a single district available for analysis. Each observation captures the percentage and absolute numbers of students by racial/ethnic category for each school-year combination, along with comparative district and state percentages for context.

The data reveals substantial variation in both school size and demographic composition across the district. Total school enrollments range from approximately 200 students at some elementary schools to over 1,500 at Somerville High School, with most elementary schools serving 300-500 students. Demographically, the district has experienced significant change over the study period, with Hispanic or Latino students growing from roughly 30% to over 40% of district enrollment, while white student representation has declined from about 45% to 35%. Asian student enrollment has remained relatively stable at 5-8%, while Black or African American enrollment has fluctuated between 8-12% across different schools and years.

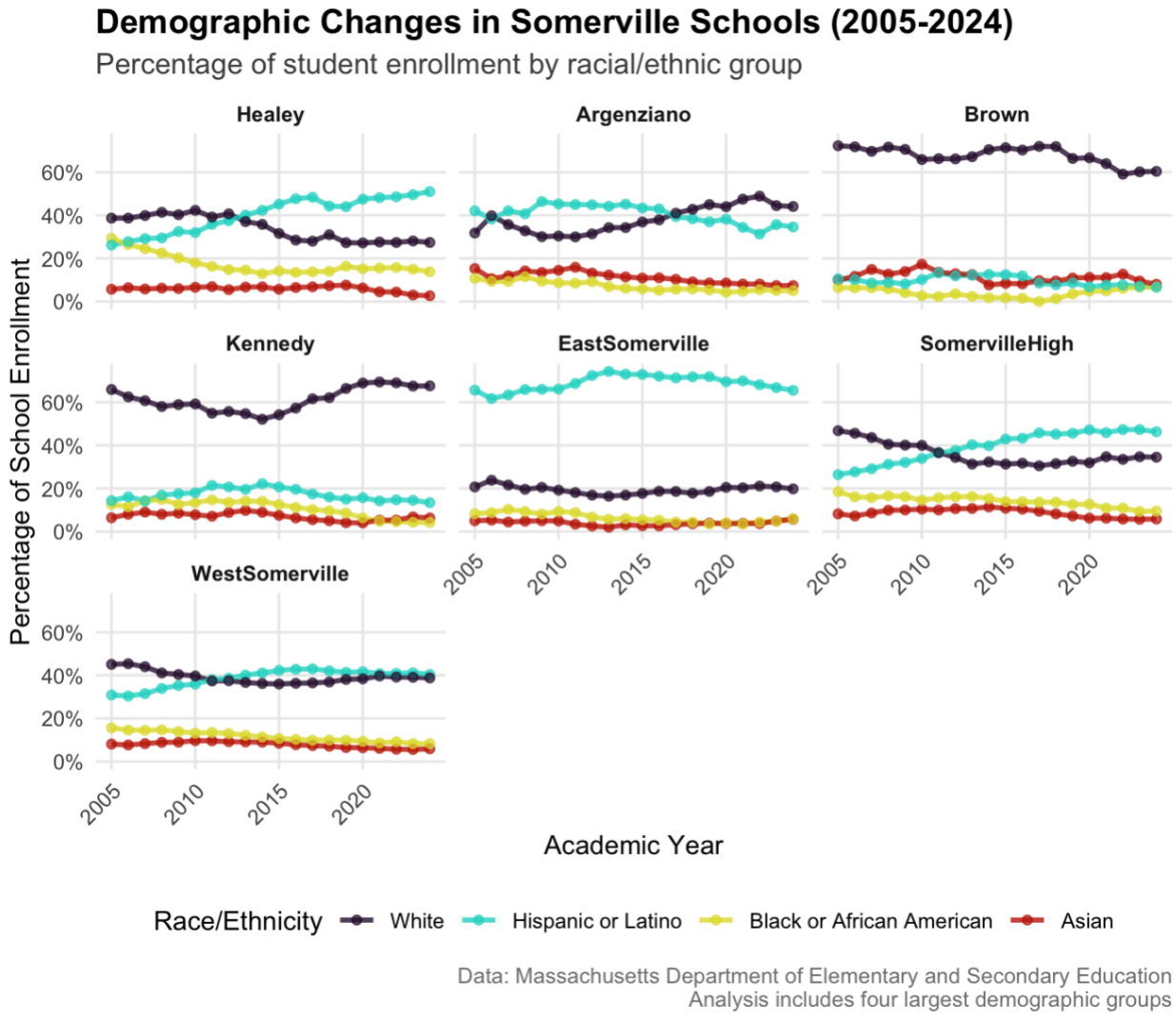
The longitudinal structure allows for analysis of both gradual demographic shifts and more dramatic changes following policy interventions. Missing data is minimal, occurring primarily in early years for some schools and likely reflecting either temporary closures, grade reconfigurations, or data reporting changes. The consistency of data collection methods over

time provides confidence in the ability to identify meaningful trends rather than artifacts of changing reporting practices.

Key Findings and Insights

Figure 1

Demographics Trends Over Time



Persistent Demographic Sorting Across Schools The data reveals significant and persistent demographic differences between Somerville schools, suggesting that despite the district's open enrollment policy, families' school choices continue to result in demographic clustering. Some

schools, like Brown and Kennedy, maintain predominantly white student populations (60-70%) throughout the study period, while others, like EastSomerville, serve predominantly Hispanic or Latino students (60-70%). This pattern indicates that structural factors beyond formal enrollment policies, such as transportation, information access, or neighborhood dynamics, continuing to influence school demographics.

The Healey School Transformation Success Story The Healey School demonstrates a remarkable demographic evolution that directly reflects its programmatic changes. Following the merger of its choice and neighborhood programs around 2010, the school achieved and maintained genuine diversity, with relatively balanced representation across major demographic groups. The Hispanic or Latino population increased from about 30% to nearly 50%, while the white population decreased from about 40% to 30%, creating one of the most demographically balanced schools in the district. This suggests that intentional integration efforts, combined with attractive programming, can successfully create diverse school communities.

District-Wide Hispanic Growth Reshaping the Landscape Across nearly all schools, there's a clear trend of increasing Hispanic or Latino enrollment, reflecting broader demographic changes in the Greater Boston area. This growth is most pronounced at EastSomerville and Healey, but is visible across the district. Meanwhile, white enrollment has generally declined at most schools, though some schools like Brown and Kennedy have maintained relatively stable white majorities. This demographic shift represents both an opportunity for increased diversity and a challenge for ensuring equitable access to all school programs.

Limited Evidence of "White Flight" but Clear Stratification Contrary to simple "white flight" narratives, the data shows a more complex pattern. While some schools experienced declining white enrollment, others maintained stable demographics, and the changes appear more related to broader demographic trends than to specific policy changes. However, the persistence of significant demographic differences between schools suggests that Somerville's school choice system may inadvertently maintain a form of stratification, where different schools serve markedly different populations despite being part of the same district.

High School Integration Challenges Somerville High School shows interesting demographic trends, with increasing diversity over time but still maintaining distinct patterns. The high school serves as a convergence point for students from all elementary schools, yet demographic differences persist even at this level. This suggests that the patterns established in elementary schools may have lasting effects on students' educational experiences and that district-wide integration efforts may need to address both elementary and secondary levels comprehensively.

Policy Recommendations

Proactive Integration Planning Somerville should develop explicit demographic goals for each school and implement proactive strategies to achieve balanced enrollment. The success of the Healey School's integration demonstrates that intentional efforts can create diverse school communities without sacrificing educational quality. District leaders should consider implementing controlled choice policies that take demographic factors into account during the enrollment process, ensuring that popular programs and schools serve diverse populations rather than inadvertently becoming segregated by race or class.

Transportation and Information Equity The persistent demographic differences between schools suggest that structural barriers may be limiting meaningful choice for some families. The district should conduct comprehensive analysis of transportation patterns, information dissemination methods, and application support services to ensure that all families have genuine access to all schools. This might include expanded bus routes, multilingual enrollment support, and targeted outreach to underrepresented communities at high-performing schools.

Program Distribution and Resource Allocation Rather than concentrating specialized programs at single schools, Somerville should consider distributing attractive programs across multiple schools to reduce incentives for demographic sorting. The district should also ensure that resource allocation formulas account for the additional supports needed at schools serving higher concentrations of students with language or economic needs, preventing a two-tiered system from developing within the choice framework.

Continuous Monitoring and Adjustment Implement annual demographic monitoring with clear protocols for intervention when schools become significantly unbalanced. The 20-year

view provided by this analysis should become standard practice, with regular policy adjustments based on emerging patterns. District leaders should also engage with families and community members about integration goals, building support for policies that promote diversity while respecting family preferences.

Challenges Encountered During Data Collection

There are several challenges I encountered during my web scraping process. First. The inconsistent web page formatting across years was one of the most significant technical challenges was handling the evolution of the Massachusetts Department of Education's website structure over the 20-year study period. Earlier years used different table formats, column names, and data organization schemes, requiring flexible parsing algorithms that could adapt to multiple layouts. Some years included additional demographic categories or changed category definitions (such as the introduction of "Multi-Race" classifications), necessitating careful mapping and standardization to ensure longitudinal consistency. This experience highlighted the importance of building robust data collection systems that can handle structural changes in source materials.

Moreover, managing respectful web-scraping at scale by collecting data from 140 individual web pages required balancing thoroughness with ethical scraping practices. Implementing appropriate delays between requests, handling server errors gracefully, and managing failed requests without losing progress proved more complex than initially anticipated. The Massachusetts DOE website occasionally returned incomplete pages or temporary server errors, requiring retry logic and careful validation of extracted data. This challenge emphasized the importance of building patience and error handling into automated data collection workflows, especially when working with government websites that may have varying performance characteristics.

Another big challenge was handling the data quality validation and missing values. Since we need to ensure data accuracy across such a large collection required developing systematic validation procedures that could identify problematic extractions without manual review of each page. Some school-year combinations returned unexpected values, missing tables, or formatting anomalies that required individual attention. Determining whether missing data represented actual missing information, temporary school closures, or scraping errors required cross-

referencing with external sources and developing decision rules for handling ambiguous cases. This experience reinforced the critical importance of building comprehensive data validation procedures into any large-scale collection effort.

Finally, creating a consistent dataset from 20 years of potentially varying source formats required careful attention to maintaining comparability over time. Changes in demographic category definitions, school names, grade configurations, and reporting standards all threatened the integrity of trend analysis. Developing mapping schemes that preserved historical accuracy while enabling meaningful comparison across years proved intellectually challenging and required deep engagement with the nuances of educational data reporting. This process highlighted how seemingly straightforward data collection projects can reveal complex methodological considerations about longitudinal research design.

References

"Enrollment Data (2021-22) - Arthur D Healey (02740075)." *Mass.edu*, 2021, profiles.doe.mass.edu/profiles/student.aspx?orgcode=02740075&orgtypecode=6&&fycode=2022.