Geographic Doppelganger Detection via Integrated Census Tract Clustering

Jacob Pickrel-Smith

Department of Data Science, Bellevue University

DSC680-T301: Applied Data Science (2255-1)

Professor Amirfarrokh Iranitalab

April 12, 2025

Geographic Doppelganger Detection via Integrated Census Tract Clustering

Topic

This project proposes the development of a tract-level clustering and geographic doppelganger detection framework. The system integrates demographic, housing, socioeconomic, and health data at the census tract level across the United States. The goal is to identify census tracts in different geographic regions that are socioeconomically similar. These identified pairs or groups of tracts, termed "geographic doppelgangers," provide a foundation for comparative analysis in urban planning, health equity strategies, and transferable policy modeling.

Business Problem

Urban planners, public health analysts, and policymakers often need to identify regions that share socioeconomic similarities but are geographically distant. Traditional approaches tend to operate at higher levels of aggregation, such as counties, or ignore complex interactions among demographic variables. This project aims to overcome those limitations by building a tract-level clustering solution. The methodology clusters census tracts using detailed multivariate data while explicitly ignoring their physical location. Once clusters are established, geographic information is reintroduced to identify tracts that share the same characteristics but exist in different states or counties. This enables cross-regional comparisons, supports targeted interventions, and improves the generalizability of policy applications.

Datasets

Several datasets were combined to build a comprehensive and analytically rich dataset:

1. The **American Community Survey (ACS) 5-Year Estimates (2022)**, which provides tract-level data on population, income, educational attainment, housing characteristics, and commuting behavior (U.S. Census Bureau, 2022).
2. **TIGER/Line shapefiles**, accessed using the pygris Python library, provide geospatial boundaries for census tracts, allowing calculation of land and water area (U.S. Census Bureau, 2022).
3. The **CDC PLACES dataset** supplies local health estimates, including chronic disease prevalence, mental health indicators, and access-to-care metrics (Centers for Disease Control and Prevention, 2024).

These sources were integrated using a consistent GEOID identifier. The final combined dataset was saved as integrated_tract_data_2022.csv, along with multiple derived outputs including cluster assignments, PCA components, and profile summaries.

Methods

The integration of datasets was handled through a custom Python class, IntegratedTractData, which merged ACS, TIGER/Line, and CDC PLACES data into a unified GeoDataFrame. Derived metrics such as higher education percentage, poverty rate, vacancy rate, and tract area in square kilometers were computed to enhance the analytical value of the dataset.

In the preprocessing phase, columns with more than 80% missing values were dropped. Remaining missing values were imputed using a combination of K-Nearest Neighbors (KNN) and Multiple Imputation by Chained Equations (MICE) (Pedregosa et al., 2011). Outliers were handled via interquartile range (IQR) capping and Isolation Forest anomaly detection. Highly correlated and low-variance features were identified and removed to reduce redundancy and improve model performance.

Feature selection was performed using a Random Forest classifier trained on synthetic cluster labels generated from an initial KMeans clustering. Features were ranked by importance, and the subset that explained at least 80% of the cumulative importance was retained. Dimensionality reduction was then applied using Principal Component Analysis (PCA), retaining only those components that accounted for more than 80% of total variance.

Clustering was conducted using KMeans with cluster counts ranging from 2 to 7. Optimal cluster number was selected using the Silhouette score and Elbow method. The best-performing clustering was then applied to the full dataset, and results were saved in both tabular and geospatial formats for further use.

To identify geographic doppelgangers, cluster labels were overlaid with geographic boundaries. For each cluster containing tracts from multiple states, pairs of tracts were sampled across states to find analogs. These matched tracts serve as high-fidelity candidates for comparative policy design.

Model evaluation was conducted using XGBoost (Chen & Guestrin, 2016), trained to predict cluster labels from the selected features. Grid search was used to test different numbers of clusters, and model performance was assessed using accuracy, F1 score, classification reports, and confusion matrices.

Ethical Considerations

All analysis was conducted using publicly available, anonymized data at the census tract level. The use of aggregated health and demographic statistics ensures individual privacy is maintained. Care was taken not to attach normative or qualitative labels to clusters, preventing misinterpretation of socioeconomic patterns. The geographic doppelganger concept, while analytically useful, is presented with caution to avoid stigmatization or inappropriate policy transference.

Challenges/Issues

Several challenges were encountered during the project. The large volume of tract-level data required thoughtful batching strategies and robust logging mechanisms to track ETL and transformation pipelines. Aligning health and socioeconomic data from different sources and years introduced temporal consistency issues. Choosing an appropriate number of clusters also required balancing interpretability with statistical rigor. Finally, ensuring that geographic doppelgangers reflected meaningful similarities required careful validation.

References

U.S. Census Bureau. (2022). American Community Survey 5-Year Estimates. https://www.census.gov/programs-surveys/acs

U.S. Census Bureau. (2022). TIGER/Line Shapefiles. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

Centers for Disease Control and Prevention. (2024). PLACES: Local Data for Better Health. https://chronicdata.cdc.gov/PLACES

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD.