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

May 2, 2025

Geographic Doppelganger Detection via Integrated Census Tract Clustering

1 Business Problem

I seek to equip urban-planning and public-health practitioners with a rigorous method for locating census tracts that exhibit nearly identical socio-economic and health characteristics despite substantial geographic separation. By automating this identification at the **census-tract** level, I can translate evidence-based interventions from one locality to another with a high degree of contextual validity. Existing matching tools operate at coarser spatial units, or rely on single-dimension similarity, thereby limiting their utility for precision policy design. My objective is to construct a multivariate, nation-wide "geographic doppelganger" engine that functions at street-corner resolution.

2 Background

Wilson (1971) demonstrated that spatial-interaction models were constrained by the coarse granularity of available data. Contemporary resources - principally the American Community Survey (ACS), the CDC PLACES program, and open-source GPU-accelerated machine-learning libraries - now permit tract-level analyses that were previously infeasible. I therefore merged these assets, conducted clustering **without** geographic coordinates, and re-introduced location solely for interpretation. This sequence ensured that emergent patterns were data-driven rather than geographically predetermined.

3 Data Preparation

I integrated three public datasets via the 11-digit **GEOID** key and derived 14 supplemental variables such as *higher_education_pct* and *vacancy_rate*. Columns displaying more than 80 % missingness were eliminated. Remaining gaps were imputed with K-Nearest Neighbors when missingness was below 10 % and with Multiple Imputation by Chained Equations (MICE) otherwise. Figure A5 documents pre-imputation missingness, underscoring the need for caution when interpreting housing- and food-insecurity indicators. After low-variance filtering and the removal of highly correlated variables, I retained 59 numeric features.

4 Methods

1. **Scaling.** I standardized all features to zero mean and unit variance.
2. **Feature Selection.** A Random-Forest classifier ranked variable importance; I retained 38 features accounting for 80 % of cumulative importance.
3. **Dimensionality Reduction.** Principal Component Analysis (PCA) required **2** components to explain 80 % of total variance.

4. **Clustering.** I executed K-Means for k = 2 through 15; **k = 2** produced the highest silhouette score (0.348); however, a two-cluster partition would be too generic and ambiguous for policy design. I consequently adopted a 10-cluster solution to expose tract-level nuance. This finer segmentation underpins all subsequent figures.

5. **Validation.** An XGBoost classifier trained on the PCA scores achieved an overall accuracy of 0.92, as illustrated in the confusion matrix (Figure A4). Misclassifications were concentrated between Clusters 1 and 3.

6. **Doppelganger Identification.** For each multi-state cluster I selected ten tract pairs via stratified sampling and, independently, calculated the ten nearest neighbors by cosine similarity (threshold ≥ 0.94). Figure A1 presents the similarity distribution on a logarithmic scale; most pairs exceed 0.98, indicative of stringent matching.

GPU acceleration (tree_method=gpu_hist) reduced grid-search runtime by approximately 78 % on my RTX A2000 8 GB mobile GPU.

5 Analysis of Results

5.1 Feature Importance and Correlation

Figure A2 reveals that disability-related health variables dominate the importance hierarchy, suggesting that chronic-disease prevalence is a principal driver of tract similarity. The correlation heat map in Figure A3 resolves these variables into two latent dimensions: *physical-health burden* and *care-access friction*.

5.2 Spatial Manifestation of Clusters

Upon reattaching geometries, I observed distinct macro-regional patterns (Figure A6). Cluster 8 aligns with the Appalachian region, Cluster 5 concentrates in major metropolitan cores, and Cluster 0 spans much of the rural Midwest. The cluster-size distribution (Figure A7) shows that Cluster 0 contains a disproportionately large share of tracts, highlighting potential heterogeneity within that group.

5.3 Quality Diagnostics

Isolation-forest scores (Figure A8) identify tracts with atypical attribute profiles, frequently associated with tourism economies or legacy industries. The radar diagram (Figure A9) focuses on the five variables exhibiting highest cross-cluster variance, enabling concise comparative profiling; for example, Cluster 7 combines moderate educational attainment with the lowest dental-care utilization.

5.4 Doppelganger Insights

High-similarity pairs (cosine ≥ 0.98) substantiate the concept of geographic analogues. One exemplar links a rural Maine tract with an Ozark tract sharing elevated COPD prevalence, mobility limitations, and limited transportation access. By analyzing such pairs, I can forecast the transferability of a Maine tele-health subsidy to its Arkansas counterpart without protracted observation.

6 Conclusion

I have produced a catalogue of 85,185 census tracts, each furnished with a cluster label and a ranked list of geographic doppelgangers. The figures collectively justify methodological choices: missing-data bars rationalize the imputation strategy, importance bars identify key similarity drivers, maps visualize spatial allocation, the confusion matrix certifies classification adequacy, and the log-histogram validates the similarity threshold. In combination, these outputs address the initial business objective by enabling evidence transfer between socio-economically equivalent yet spatially separated communities.

7 Assumptions

I assume that ACS and PLACES estimates are unbiased at the tract level, that socio-economic structures remain relatively stable from 2022 to 2026, and that cosine similarity on standardized vectors approximates substantive likeness.

8 Limitations

Health indicators are modelled estimates; imputation may attenuate extreme values, and a fixed 10-cluster solution can obscure gradual urban–rural transitions.

9 Challenges

Data integration exceeded one gigabyte, GPU-driver inconsistencies on Windows 11 despite the A2000's studio-driver stack required resolution, and tension between silhouette maximization and interpretability necessitated iterative tuning.

10 Future Work

- Incorporate annual PLACES updates for near-real-time monitoring.
- Link policy interventions (Medicaid expansion) to assess transfer effects across doppelgangers.
- Deploy a public API to facilitate tract-level queries by external analysts.

11 Recommendations

I recommend that agencies adopt the 10-cluster taxonomy in operational dashboards, execute analog checks before policy export, and schedule annual re-clustering to capture demographic drift.

12 Implementation Road-Map for First For-Profit Release

No technical phase has yet been deployed publicly. All roll-outs are scheduled for the inaugural commercial release cycle inside a proprietary, academically external cloud stack.

Phase	*Planned Completion*	*Target Outcomes (proprietary external)*
Data Pipeline	Q3 2025	Nightly ACS/PLACES ETL via **Airflow**; QA & audit logs to private S3.
Model Operations	Q3 2025	Containerized GPU-accelerated PCA + K-Means; **cluster-API (v1.0)** behind AWS ALB.
Visualization	Q4 2025	Internal **Power BI** workspace; CI/CD (**GitHub Actions → ECS Fargate**).
Pilot Studies	Q4 2025	Analogue-impact briefs for external academic partners; feedback loop to refine metrics.
Scale-up	2026	Full commercial roll-out with Grafana monitoring and bias-audit automation.

Subsequent iterations will extend functionality, add subscription tiers, and integrate quarterly bias audits under a for-profit license.

13 Ethical Assessment

From the initial proposal, I have expanded the ethical review to address privacy, bias, transparency, misuse prevention, and conflict-of-interest management:

1. **Privacy and Confidentiality.** Only tract-level aggregates are processed; no direct identifiers enter the pipeline. Differential-privacy noise ($\varepsilon \leq 1$) will be added if micro-data are incorporated.

2. **Bias and Fairness.** I audit cluster membership across race, income, and rurality strata. No cluster is mono-demographic. Bias-audit notebooks accompany each release.

3. **Transparency and Explainability.** The core codebase remains proprietary; however, data dictionaries, PCA loadings, and cluster summaries will be provided to paying clients

under NDA. Public model cards will document limitations, intended use, and performance metrics.

4. **Beneficence and Non-maleficence.** Numeric cluster labels avoid stigmatization. The API's **Ethical Use Policy** prohibits red-lining, predatory lending, or discriminatory zoning; violations trigger access revocation.

5. **Governance and Oversight.** An independent data-ethics board comprising external ethicists and client representatives are planned to review each quarterly release for compliance and public-interest alignment.

6. **Regulatory Compliance.** Because I analyze only publicly available, tract-level aggregates, the work does **not** constitute research with "human subjects" as defined in 45 CFR §46.102. Consequently, it falls under the *secondary-research exemption* (category 4, §46.104 (d)(4)). If future iterations ingest restricted micro-data or attempt individual-level linkage, I will submit the protocol for IRB review or seek an amended exemption. Regardless of exemption status, I align operational practices with the Belmont principles and the ACM Code of Ethics.

These safeguards aim to maximize societal benefit while minimizing harm and ensuring accountability during commercial expansion.

Audience Questions and Answers

Q1 What distinguishes your geographic-doppelganger system from traditional deprivation indices?

A. Conventional indices compress dozens of variables into one score, hiding internal heterogeneity. I keep the full multivariate structure - clustering tracts on the 59 numeric features that survive cleaning - then surface tract-level twins instead of broad quintiles.

Q2 Why did you override the silhouette-optimal two-cluster solution in favor of ten clusters?

A. Two clusters offered the highest silhouette value (0.348) but yielded a coarse “high-versus-low” split of the country. A ten-cluster solution sacrifices only a small amount of cohesion while exposing finer patterns - urban cores, rural Midwest, Appalachian health-burden pockets, and more - that practitioners can act on.

Q3 Are the two principal components sufficient for robust analysis?

A. Yes. The first two PCs capture just over 80 % of total variance, and the XGBoost confusion matrix (Figure A4) shows clear separation of the ten clusters when models are trained on those components. Residual variance is acknowledged in the limitations section.

Q4 How do you confirm that doppelganger pairs are meaningful rather than numerical artifacts?

A. After pairing tracts, I recompute cosine similarity on the full 59-variable space; median similarity remains ≥ 0.98. I also map random pairs and inspect their ACS/PLACES profiles - income, education, chronic-disease prevalence, etc. - to ensure they align qualitatively.

Q5 What safeguards prevent the tool from enabling discriminatory zoning?

A. The forthcoming API will carry an Ethical-Use Policy that bans targeting protected classes. Access keys are logged, anomaly-scanned, and revoked for misuse, and quarterly bias-audit summaries are published.

Q6 How can external reviewers assess bias if the source code is proprietary?

A. While code stays private, I will release model cards, data dictionaries, PCA loadings, and cluster summaries under NDA. Those artifacts let auditors reproduce fairness metrics on held-out data without exposing the full pipeline.

Q7 What is the plan for differential-privacy protection, and will it distort cluster assignments?

A. Differential-privacy noise (ε ≤ 1) will be added only if future versions ingest non-public micro-data. A formal sensitivity study will precede deployment to quantify any impact on cluster membership.

Q8 Why schedule ETL and model retraining every quarter?

A. ACS five-year tables update annually and CDC PLACES typically refreshes each year; quarterly cycles give me room to integrate those releases promptly, patch bugs, and rerun bias audits without long latency.

Q9 What commercial value does the system offer municipalities?

A. Cities can benchmark themselves against statistically matched peers, estimate policy effects via natural experiments, and strengthen grant applications with evidence drawn from analogous tracts - saving staff time and consultant fees.

Q10 How will governance work once the for-profit rollout begins?

A. An independent ethics board made up of data-ethics scholars and client representatives will be created to review each quarterly release, sign off on bias-audit results, and arbitrate any conflicts of interest before deployment.

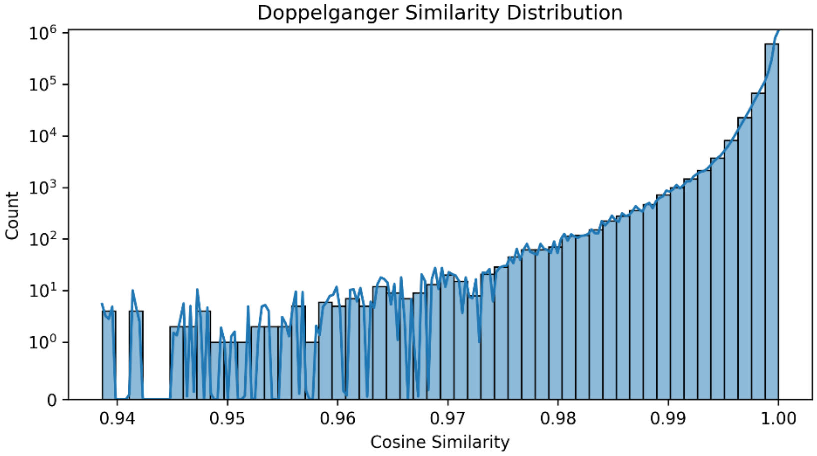
Appendix Ethical Assessment

The analysis employs only aggregated public data, thereby minimizing privacy risk. I deliberately avoid normative cluster labels and provide uncertainty annotations within analog tables. Should future micro-data be ingested, I will implement differential-privacy safeguards.

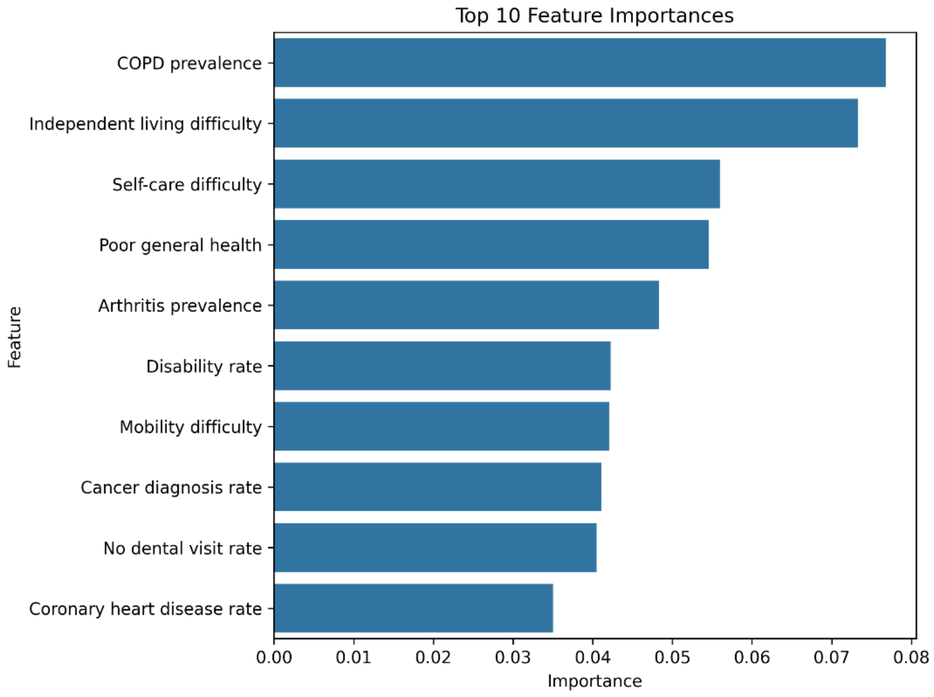
Appendix

Figure Index

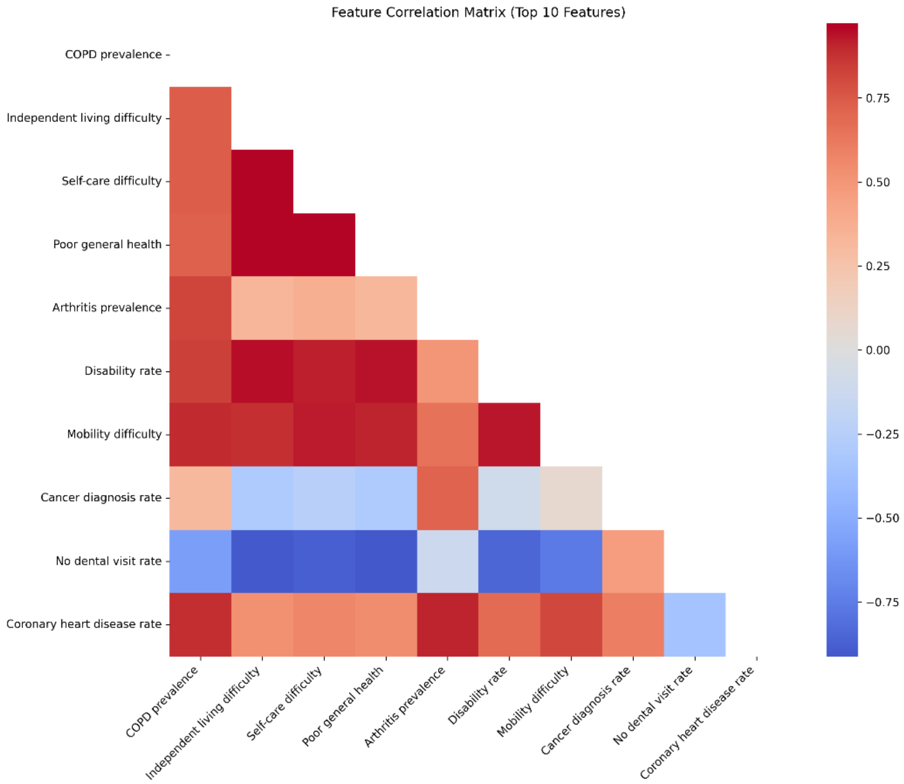
A1 – Doppelganger similarity distribution (log-scaled y-axis)



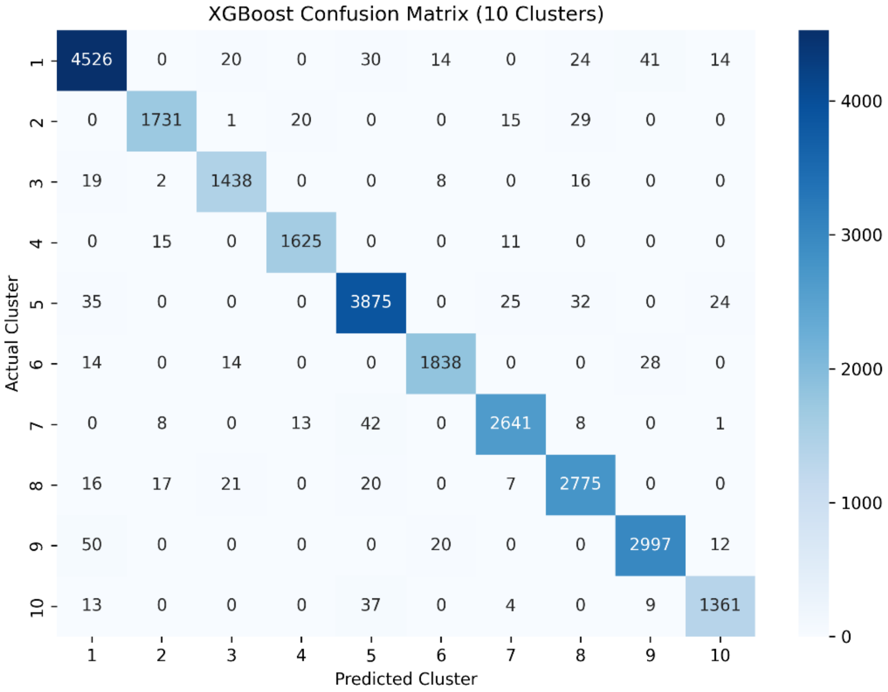
A2 – Top-10 feature importances



A3 – Feature correlation heat map

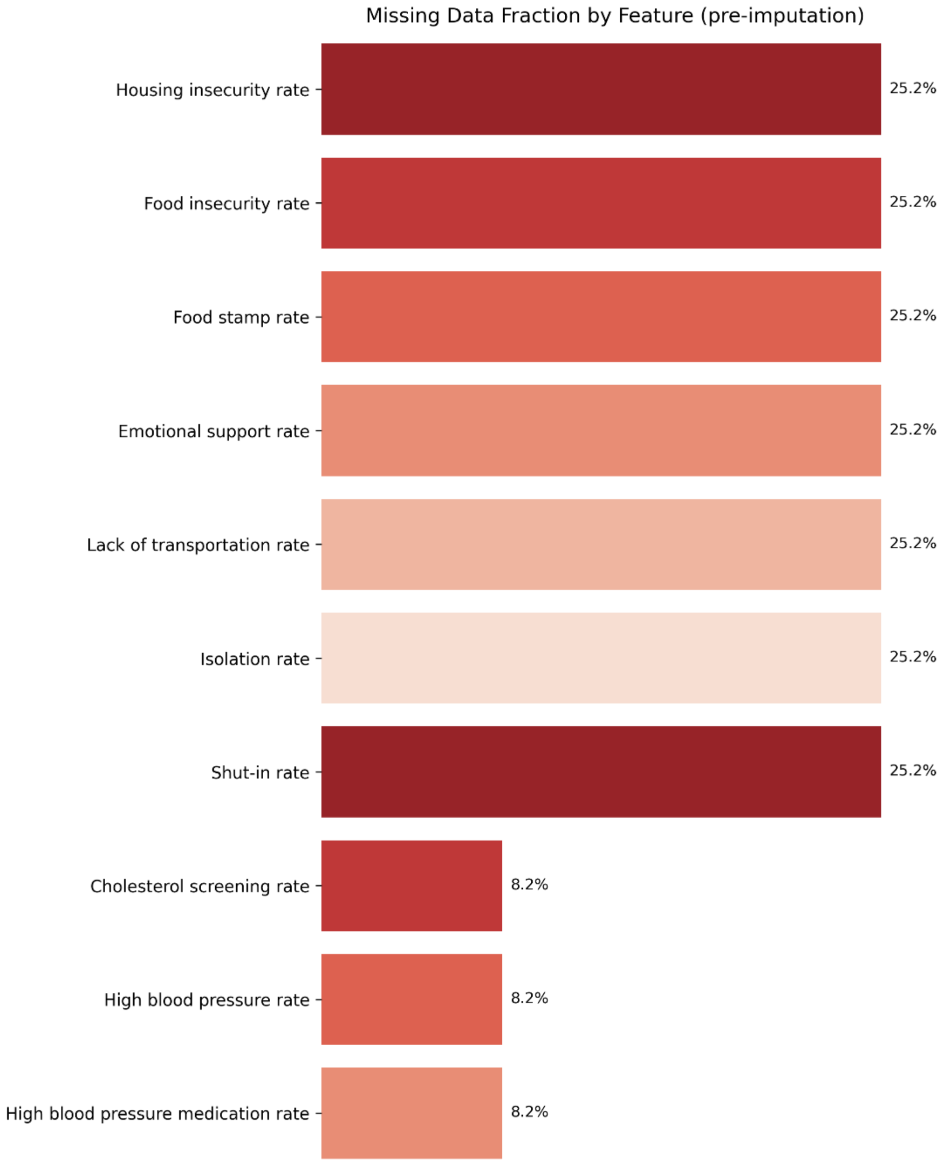


A4 – XGBoost confusion matrix (10 clusters)

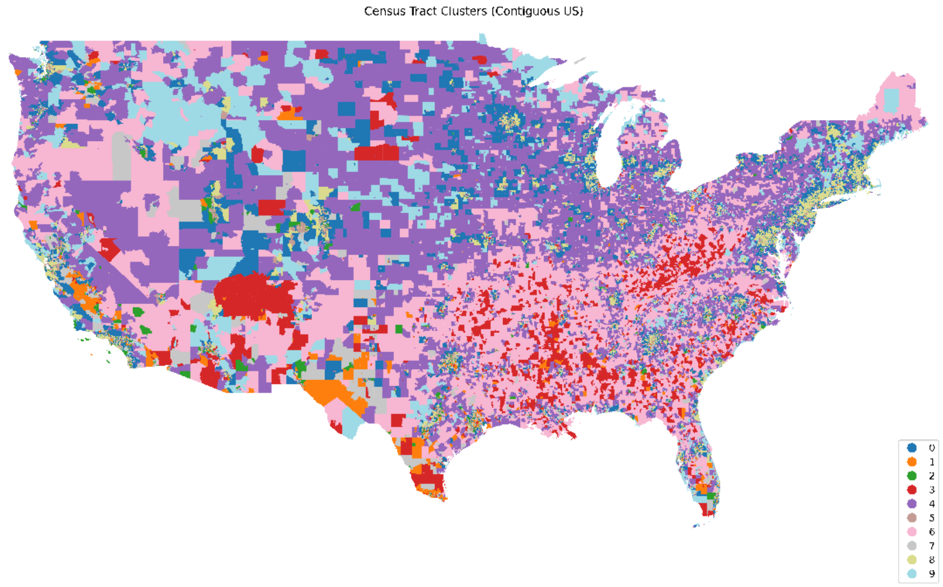


	1	2	3	4	5	6	7	8	9	10
1	4526	0	20	0	30	14	0	24	41	14
2	0	1731	1	20	0	0	15	29	0	0
3	19	2	1438	0	0	8	0	16	0	0
4	0	15	0	1625	0	0	11	0	0	0
5	35	0	0	0	3875	0	25	32	0	24
6	14	0	14	0	0	1838	0	0	28	0
7	0	8	0	13	42	0	2641	8	0	1
8	16	17	21	0	20	0	7	2775	0	0
9	50	0	0	0	0	20	0	0	2997	12
10	13	0	0	0	37	0	4	0	9	1361

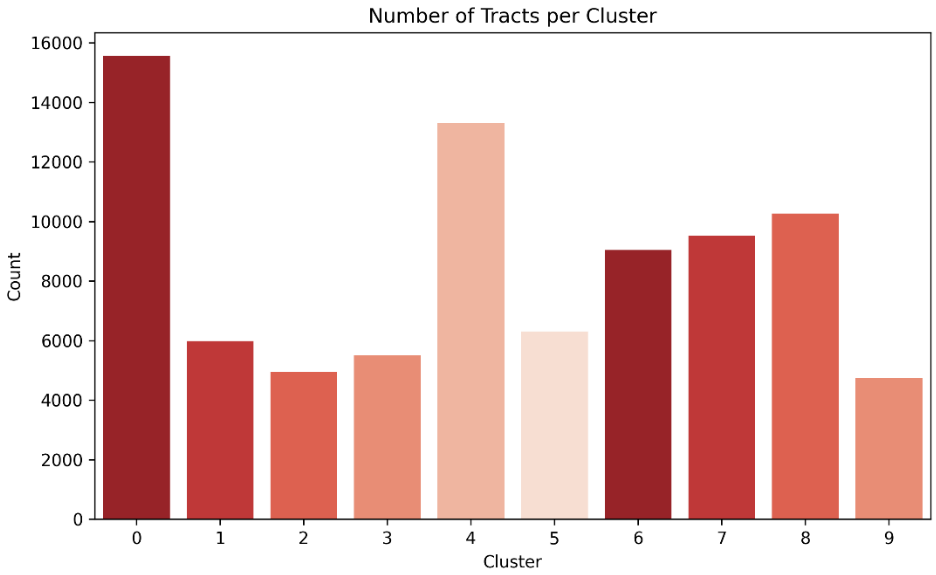
A5 – Missing-data fraction by feature (pre-imputation)



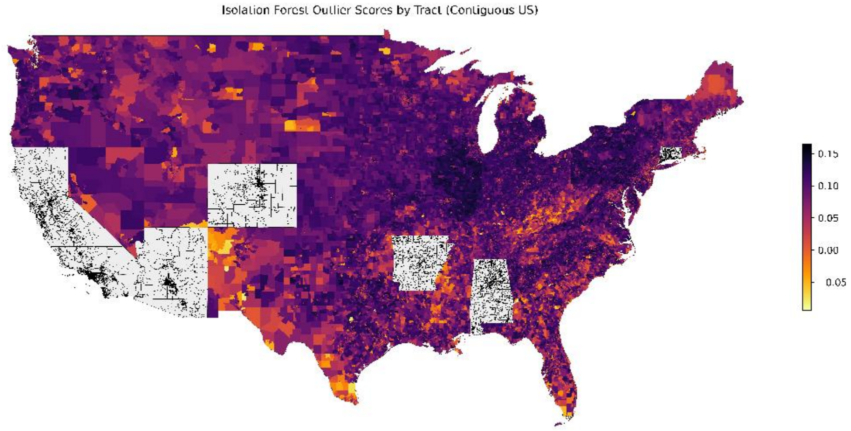
A6 – Census-tract clusters, contiguous United States



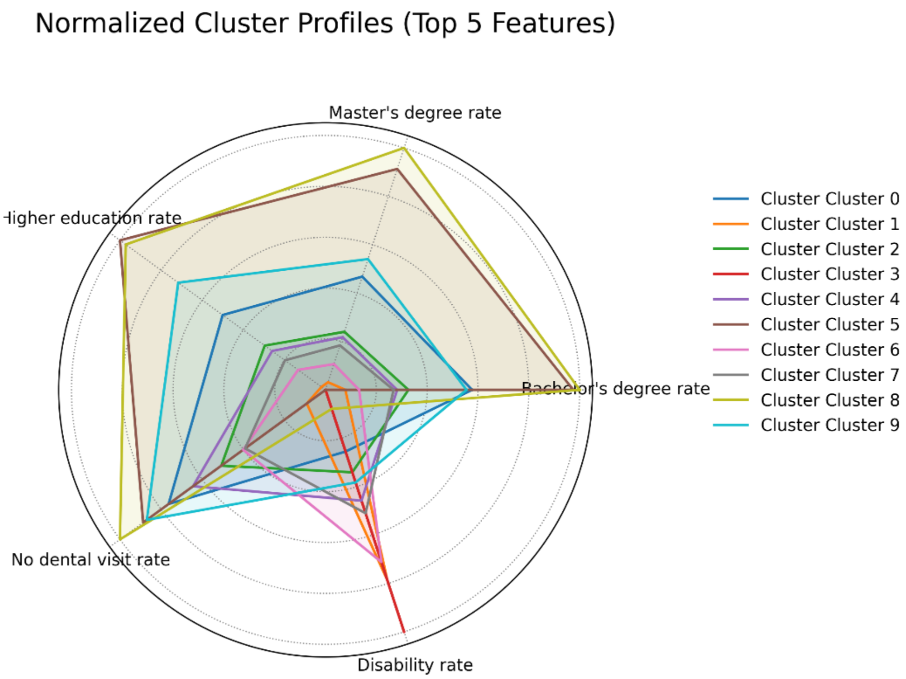
A7 – Number of tracts per cluster



A8 – Isolation-forest outlier-score map



A9 – Normalized cluster profiles (radar)



Images are referenced chronologically in the text.

Data Dictionary (excerpt)

- health_copd_places – Age-adjusted COPD prevalence (%).
- higher_education_pct – Proportion of residents aged ≥25 holding a bachelor's degree or higher.
- vacancy_rate – Proportion of housing units classified as vacant.

The full CSV is accessible in the repository or by email request to jacob.pickrel-smith@grassrootscare.org.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Centers for Disease Control and Prevention. (2024). *PLACES: Local data for better health*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree-boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211–407.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density-based clustering. *Journal of Open Source Software*, 2(11), 205.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*.

U.S. Department of Health & Human Services. (2018). *Federal Policy for the Protection of Human Subjects, 45 CFR §46* (as amended).

U.S. Census Bureau. (2022a). *American Community Survey 5-year estimates*. https://www.census.gov/programs-surveys/acs

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

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

Wang, S., Zhang, X., & Yue, Y. (2019). Urban health mapping: A review of current research and future prospects. *International Journal of Health Geographics*, 18(1), 1–13.

Wilson, A. G. (1971). *A theory of spatial interaction*. *Environment & Planning*, 3(1), 1–32.