UC Santa Barbara
GIScience 2021 Short Paper Proceedings

Title
An Individual-Centered Approach for Geodemographic Classification

Permalink
https://escholarship.org/uc/item/4xj1008p

Author
Tuccillo, Joseph

Publication Date
2021-09-01

DOI
10.25436/E2H59M

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

An Individual-Centered Approach for Geodemographic Classification

Joseph V. Tuccillo ORCID
Oak Ridge National Laboratory, United States

Abstract

Geodemographic classifications are an important tool to support public-service decision making. While people are the focal point of geodemographics, classifications are often built on variables that describe populations rather than individuals. Synthetic populations, model-based approximations of the individual makeup of small census areas, remain largely unused for geodemographic classification, yet they can provide a more direct and holistic understanding of localized resource needs than existing approaches. This paper develops a new method for performing individual-centered geodemographic classifications using synthetic populations. The building blocks of this approach are abstractions of the synthetic population attributed to each small census area via affinity matrices computed from similarities in both the size and attributes among individuals. Using a rank-1 spectral decomposition of an area's affinity matrix enables rapid computation of a dissimilarity metric which is compatible with cluster analysis techniques used in traditional geodemographic classifications. Using data from the American Community Survey (ACS), an example classification is developed for the Knoxville, TN, USA Public-Use Microdata Area (PUMA) to illustrate how distinctions can be drawn among small census areas in terms of specific types of representative individuals, providing a more tailored view of the groups that serve to benefit from spatial policy interventions. Beyond improving traditional public-domain geodemographic classifications, this approach provides a novel open-source alternative to commercial neighborhood segmentation products with added flexibility for custom research applications.

1 Introduction

Geodemographics is the study of spatial heterogeneity in demographics across social areas (i.e., neighborhoods, communities) comprising an urban, regional, or national system. Understanding who people are in the context of where they live is essential to support public service allocation in areas including health, education, and public safety (8; 6). To manage the complex task of measuring social composition, a practice known as *geodemographic classification* is used to group social areas based on their emergent properties. Each geodemographic class features a profile of population characteristics that distinguish it from others, providing tailored information about the groups expected to benefit from spatial policy interventions.

While geodemographics fundamentally involves the attributes of people, public-use geodemographic classifications seldom directly assess the central problem of "who people are". Instead, they rely on aggregate population statistics (i.e., median age, percent in poverty) to explain differences among social areas (16; 15; 6). Aggregating individual attributes makes it impossible to directly characterize the different types of people comprising in an area. Substantial information loss can result from aggregation, leading to a distorted representation of population characteristics (2; 13). This *cross-level* or *ecological* inference problem (1; 5) affects the soundness of decision support that a geodemographic classification can provide planners and administrators.

The cross-level inference problem in geodemographic assessments can be overcome with *synthetic populations*, realistic recreations of the makeup of small census areas consisting of geolocated individuals from public-use census microdata samples. Given that individual-level and aggregate data are simply different ways of measuring the same population, data fusion techniques like iterative proportional fitting and combinatorial optimization are used to bridge these two scales (7). The result extends a wide swath of individual attributes related to demographics, socioeconomic status, housing, and health to high spatial resolutions, providing a complete representation of individual attributes within an area unattainable via observational analysis, while also maintaining the privacy of census survey respondents (10; 9).

© Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).; licensed under Creative Commons License CC-BY 4.0

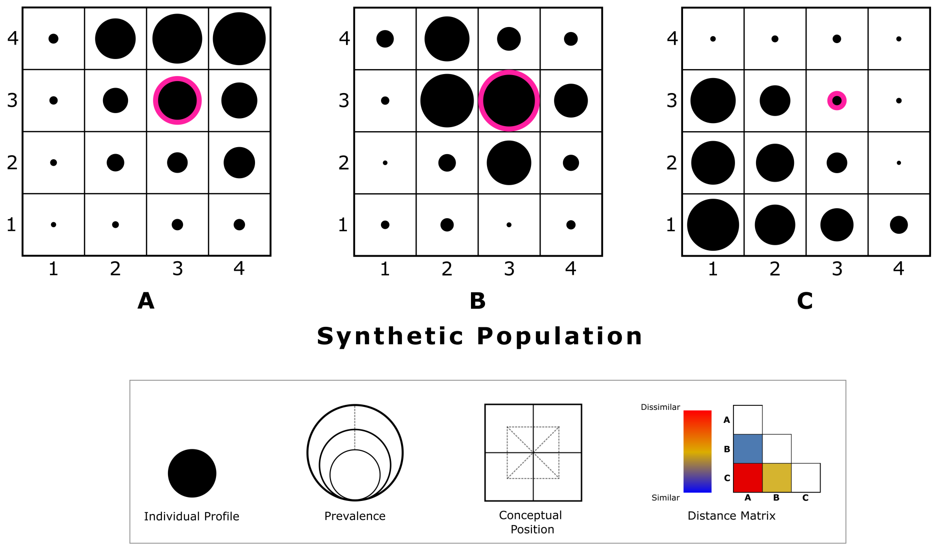

Figure 1 Conceptual illustration of representing and comparing synthetic populations by individual profiles.

While synthetic populations are often applied to study human activity through agent-based models (microsimulation) (4), their use toward characterizing the social fabric remains less explored. Synthetic populations are unwieldy, consisting of both individual and collective attributes, which poses a challenge for directly describing and comparing them. This paper develops an individual-centered approach for geodemographic classification, centered on a novel metric for efficiently comparing synthetic populations based on their latent properties. This metric can be used to compute dissimilarities and perform cluster analysis in a way that is compatible with traditional geodemographic techniques, enabling the creation of classifications tailored to a variety of planning needs.

2 A New Method for Comparing Synthetic Populations

Synthetic populations are more complicated to analyze than area-level population attributes because they are multidimensional *and* multiscalar. A synthetic population is simultaneously characterized by the attributes of people and the attributes of the collective. Counts of unique types of people belonging to an area's synthetic population (i.e., age over 60, in poverty, living alone; university student, employed in an unskilled job and living close to work) can be thought of as area-level attributes. Given a large number of study variables, thousands of unique types of people, or *individual profiles*, can characterize an area, leading to a highly fragmented view of its population. This in turn poses a challenge for geodemographic classification because measuring similarity and dissimilarity among social areas becomes less straightforward than traditional approaches.

The approach developed in this paper resolves this problem by abstracting the characteristics of synthetic populations in a way that facilitates more efficient comparison among them. This lower-dimensional representation is based not only on estimated sizes of individual profiles within a synthetic population, but also how alike they are.

2.1 Illustration

Figure 1 provides a simplified illustration of how *conceptual* (attribute) similarities and similarities in *prevalence* can be combined to characterize a community's synthetic population and compare it to others. The grid cells represent individual profiles organized by conceptual similarity in two hypothetical dimensions. (In practice, measuring conceptual similarity involves attribute matching across many dimensions. For example, individual profiles describing employed, highly-educated family-aged adults in married couple families who differ only in terms of commute length might be considered "conceptually similar" to one another yet "conceptually distinct" from seniors living alone and on a fixed income below the poverty line.) The symbol sizes represent estimates of the number of individual types in the synthetic population. Combining these factors results in a measure of "embeddedness" of a given individual type within the synthetic population. An individual profile that is both conceptually similar and similar in prevalence to a large number of other individual profiles belongs to a latent segment of the synthetic population with like characteristics.

Comparing the embeddedness of all individual profiles within a study area from its synthetic populations results in a dissimilarity metric useful for geodemographic classification. Figure 1 compares three hypothetical communities based on this approach. Individual profile 3-3 is highlighted as an example. In Synthetic Population A, individuals of type 3-3 are strongly embedded in the population. They and several their nearest neighbors in terms of conceptual similarity (3-4, 4-3, 4-4) among the most prevalent in the population. The converse exists for Synthetic Population C. Individual type 3-3 is not well embedded, being relatively small in size and conceptually distant from the most prevalent individual profiles in the community (1-1, 1-2). Along these lines, Synthetic Populations A and B are more similar to one another than they are to Synthetic Population C as individual profiles like 3-3 are highly embedded in each. As such, A and B would be more likely to be grouped together in a geodemographic typology, whereas C might be assigned a distinct class.

2.2 Abstracting Synthetic Populations

An area's synthetic population is represented by an affinity matrix scored among all individual profiles in the study population that combines a matrix of pairwise conceptual similarities C with another consisting of prevalence similarities P as

$$A = C \times (1 + P)$$

When individual attributes have binary representation, the conceptual similarity matrix C consists of pairwise affinities (i.e., Hamming or Jaccard distances). For mixed type representation, Gower distance may be used.

The prevalence similarity matrix P is computed as

$$P = 1 - (D/max(D))$$

where D is a matrix of pairwise Manhattan distances among the count estimates of all individual profiles within study population.

The affinity matrix A is computed by upweighting the conceptual similarities by the local prevalence similarities among individual profiles. When A_{ij} is high, individual profiles i and j are characteristically similar to one another and exist in comparable measure within the population. Conversely, a low value of A_{ij} occurs when individuals are distinct from one another and mismatched in size.

To facilitate comparison among synthetic populations, a rank-1 approximation of the affinity matrix A is generated using spectral decomposition to the compute eigenvector centrality for the individual profiles. The results of this procedure are such that each individual profile is assigned an "embeddedness" score measuring the degree to which it represents the area's population. Higher values denote highly representative individual profiles, whereas lower ones indicate an those that are distinct from the area's population at large. Converting each synthetic population to a vector enables computation of area-level dissimilarities that can then be converted to a geodemographic classification using cluster analysis techniques.

3 Proof of Concept

A proof of concept for the individual-centered geodemographic approach introduced in Section 2 was performed on a sample dataset for Knoxville, Tennessee obtained from the American Community Survey's (ACS) Public-Use Microdata Sample (PUMS), containing the majority of the city's incorporated area (roughly 180,000 residents).

3.1 Data

Microdata and summary statistics for population synthesis were obtained from the ACS 2014 - 2019 5-year PUMS and Summary File across topics including basic demographics (age 60+, age under 18, marital status), socioeconomic status (race, employment, poverty, college education or higher, professional occupation), school enrollment (in school, K-12 student, post-secondary student), and worker mobility (living within 30 minutes of work). Synthetic populations were created at the block group level (census units of roughly 600 - 3000 people).

3.2 Methods

Population synthesis was performed using UrbanPop, an open-source spatial microsimulation framework developed by Oak Ridge National Laboratory (ORNL) (11; 3). UrbanPop relies on Penalized Maximum-Entropy Dasymetric Modeling (P-MEDM) an iterative proportional fitting (IPF) method specialized for uncertain census datasets like the ACS (12). UrbanPop generated 30 residential simulations from the P-MEDM occurrence probabilities, and synthetic populations based on unique individual profiles were computed from the median of the simulation estimates.

With the synthetic populations in hand, the 113 block group synthetic populations for Knoxville were then compared using the approach from Section 2. To handle the large number of unique individual profiles (n = 253), a fast spectral decomposition method provided by the Sparse Eigenvalue Computation Toolkit as a Redesigned ARPACK (Spectra) library was used (14). Dissimilarities were then organized into a dendrogram using the single-linkage (nearest neighbor) method. A suitable number of block group clusters was found by evaluating dendrogram cuts between k = 2 and k = 10 clusters based on a combination of internal consistency (percentage explained inertia) and distinctness (average silhouette width).

3.3 Results

The geodemographic classification shown in Figure 2 reveals key differences in the individual profiles distinguishing each block group cluster. For example, Clusters 1, 2, and 6 each represent areas with increased prevalence of K-12 students. While Clusters 1 and 2 feature a common exemplar of white K-12 students in married couple families, Cluster 6 differs in that it features more minority K-12 students not in married-couple families and in poverty. Cluster 1 also tends to feature more employed people in professional occupations who are in married-couple families than Clusters 2 and 6. Clusters 3 and 5, meanwhile, describe the University of Tennessee campus and adjacent neighborhoods, with exemplars characterized by adult post-secondary education students living in poverty (employed and unemployed/full-time students). Cluster 4 differs most clearly from the others by aging populations, both married and unmarried.

4 Discussion and Conclusion

Using synthetic populations to represent the social structure of small census areas produces new geodemographic classifications that more directly capture differences among individual residents of those areas. Representing small areas based on centrality or "embeddedness" of individual profiles within each synthetic population enables the identification of cluster-specific exemplar segments that can help to tailor policy and public service provision within a wider administrative area (city,

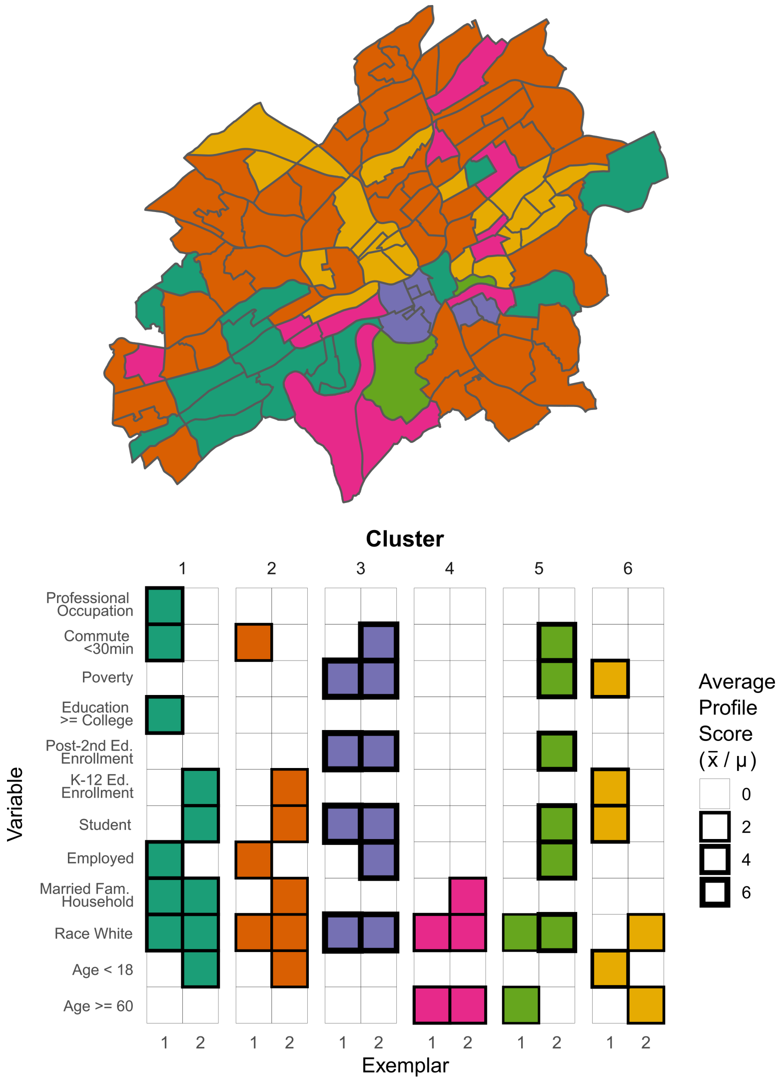

Figure 2 Individual-centered geodemographic classification for Knoxville, TN. Profiles consist of two exemplar segments distinguishing each cluster. The "average profile score" compares the mean proportion of the segment within the cluster ($\bar{x}$) to its mean proportion across all block groups in Knoxville (μ).

county, region). The proof of concept shown for Knoxville, TN (Section 3) reveals sections of the city with underserved K-12 students (Cluster 6), university undergraduates dependent upon outside employment for financial support (Clusters 3 and 5), and aging residents (Cluster 4), each of which corresponds to a distinct set of public service priorities.

In addition to overcoming the cross-level inference problem affecting open-source classifications built on aggregate data, this approach provides greater support for custom geographies/social variables than proprietary geodemographic products like ESRI Tapestry and Claritas PRIZM, which leverage individual data but often apply a "one size fits all" approach toward neighborhood targeting. This enables evaluation of the outcomes of spatial policy interventions at analytic scales and with features most appropriate toward specific planning applications (i.e., transportation, hazards, health).

Though for expository purposes the example in this paper was carried out for a single small study area (PUMA), this approach is also scalable to larger study extents. Future work will focus on developing regional and national-level classifications to understand spatial heterogeneity among large numbers small census areas. Scaling efforts will increase the computational and analytic intensity of this approach, particularly in terms of scoring similarities among larger volumes of individual profiles and characterizing the geodemographic classes. To address such challenges, these efforts will explore incorporating techniques including distributed processing, feature agglomeration (to handle increased numbers of individual profiles), and multilevel classification (to generate global/local geodemographic profiles).

References

1 Christopher H Achen and W Phillips Shively. *Cross-level inference.* University of Chicago Press, 1995.

2 Giuseppe Arbia. *Spatial data configuration in statistical analysis of regional economic and related problems*, volume 14. Springer Science & Business Media, 2012.

3 HM Abdul Aziz, Nicholas N Nagle, April M Morton, Michael R Hilliard, Devin A White, and Robert N Stewart. Exploring the impact of walk–bike infrastructure, safety perception, and built-environment on active transportation mode choice: a random parameter model using new york city commuter data. *Transportation*, 45(5):1207–1229, 2018.

4 Richard J Beckman, Keith A Baggerly, and Michael D McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, 1996.

5 Wendy K Tam Cho and Charles F Manski. Cross level/ecological inference. *Oxford handbook of political methodology*, pages 547–569, 2008.

6 Christopher G Gale, Alexander D Singleton, Andrew G Bates, and Paul A Longley. Creating the 2011 area classification for output areas (2011 oac). *Journal of Spatial Information Science*, 2016(12):1–27, 2016.

7 Kirk Harland, Alison Heppenstall, Dianna Smith, and Mark H Birkin. Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, 15(1), 2012.

8 Richard Harris, Peter Sleight, and Richard Webber. *Geodemographics, GIS and neighbourhood targeting*, volume 8. John Wiley & Sons, 2005.

9 Robin Lovelace, Morgane Dumont, Richard Ellison, and Maja Založnik. *Spatial microsimulation with R.* Chapman and Hall/CRC, 2017.

10 Douglas A Luke and Katherine A Stamatakis. Systems science methods in public health: dynamics, networks, and agents. *Annual review of public health*, 33:357–376, 2012.

11 April M Morton, Jesse O Piburn, Nicholas N Nagle, HM Aziz, Samantha E Duchscherer, and Robert N Stewart. A simulation approach for modeling high-resolution daytime commuter travel flows and distributions of worker subpopulations. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2017.

12 Nicholas N Nagle, Barbara P Buttenfield, Stefan Leyk, and Seth Spielman. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95, 2014.

13 Stan Openshaw. Ecological fallacies and the analysis of areal census data. *Environment and planning A*, 16(1):17–31, 1984.

14 Yixuan Qiu. Large-scale eigenvalue decomposition and svd with rspectra, 2019. URL: `https://cran.r-project.org/web/packages/RSpectra/vignettes/introduction.html`.

15 Seth E Spielman and Alex Singleton. Studying neighborhoods using uncertain data from the american community survey: a contextual approach. *Annals of the Association of American Geographers*, 105(5):1003–1025, 2015.

16 Dan Vickers and Phil Rees. Creating the uk national statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):379–403, 2007.