Building Alternative Indices of Socioeconomic Status for Population Modeling in Data-Sparse Contexts

Angela R. Cunningham
Oak Ridge National Laboratory, TN, USA

Joseph V. Tuccillo
Oak Ridge National Laboratory, TN, USA

Tyler J. Frazier
Oak Ridge National Laboratory, TN, USA

Abstract

Population modeling requires clear definitions of socioeconomic status (SES) to ensure overall estimate accuracy and locate potentially underserved subpopulations. This presents a challenge as SES can be measured in myriad ways and for divergent purposes, and the data required to calculate these metrics may be lacking, particularly in low and middle income countries (LMICs). To support more refined SES measurement, we explore improvements upon the Demographic and Health Survey's (DHS) Wealth Index (DHS-WI) using alternative characterizations of SES based on multiple correspondence analysis (MCA) and hierarchical clustering. We produce the MCA-derived metrics first on a full suite of household economic, demographic, and dwelling variables, then on a reduced set of occupant-only SES characteristics. We explore the utility of these metrics relative to DHS-WI based on their ability to 1) differentiate DHS household types and 2) identify mixtures of SES levels within DHS samples and mapped at high spatial resolution. We find that our full suite MCA yields more clearly defined SES segments and that our reduced MCA delineates occupant SES most clearly, suggesting potential pathways to improve upon the DHS-WI in future population modeling efforts for LMICs.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Demographic and Health Survey, multiple correspondence analysis, population modeling, socioeconomic status, spatial statistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.25

Category Short Paper

Funding This material is based upon the work supported by the U.S. Department of Energy under contract no. DE-AC05-00OR22725.

Acknowledgements Our thanks to Daniel Adams and Clinton Stipek. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy 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`).

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 25; pp. 25:1–25:7
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

To improve our ability to locate and address population-related challenges like climate change resiliency and disaster response, we need accurate and precise population estimates that attend to socioeconomic status (SES). SES – variously defined – is predictive of the density at which a population lives [15], as well as life expectancy, [8], lifetime mobility [2], and consumption patterns [9]. However, building appropriate SES definitions is challenging. Measures of SES variably incorporate income, accumulated wealth, education, occupation, cultural markers, demographic factors, resource/infrastructure access, national policies, or embeddedness within the international economy, and are shaped by data availability, and research and policy goals [2][9].

In this paper, we describe efforts to delineate and map SES groups from data recorded in Ghana's 2014 Demographic and Health Survey (DHS) [5]. Ghana's survey, like all DHS surveys, is a nationally representative sample collected by a local statistical service in concert with the United States Agency for International Development (USAID), largely for the purposes of monitoring child and maternal health in LMICs. The DHS provides its own Wealth Index (DHS-WI), a measure of household economic status derived from the household's assets, access to utilities, quality of water sources and toilet facilities, urban/rural status, and the materials with which the household's dwelling is constructed. DHS reduces these variables to a single metric through a principal components analysis (PCA), taking the first component of the PCA to score households, and dividing the population into quintiles [13]. While DHS-WI is widely used by development agencies and for validation purposes[3], reducing and flattening occupant and dwelling characteristics into a single metric via the first component PCA method complicates the investigation of subpopulation-place relationships whose understanding is central to our own work.

We explore alternative characterizations of SES using DHS data, first based on a full suite of household economic, built environment and demographic variables, and then on a reduced occupant SES only variable set (education and assets). We employ multiple correspondence analysis (MCA) and hierarchical clustering to generate new metrics to compare against DHS-WI, motivated by arguments that 1) PCA is inappropriate for non-continuous variables and 2) that too much useful information is discarded when relying on the first component alone [12][16]. We compare the MCA-derived metrics to the DHS-WI based on distinctiveness and diversity in class labels of respondent households, as well as based on spatial variation in class labels when mapped.

2 Methodology

We began by selecting and preprocessing our variables of interest from Ghana's 2014 DHS, which consists of 12,831 weighted observations drawn from 30 households randomly sampled from each of the country's 427 enumeration areas (referred to by DHS as "clusters")[6]. We extracted variables pertaining to the built environment (source of drinking water; condition of toilet facilities; provision of electricity; number of sleeping rooms; materials used to construct the floor, walls and roof), those that can be used as proxies of wealth (the presence of assets from cars and sewing machines to tables and telephones), the highest level of education in the household, and variables recording basic demographics (household size; age and sex of household head). Following [14], we aggregated building material variables into natural, rudimentary, and finished categories, water and toilet facilities into three quality grades, and binned quantitative variables. We implemented basic data cleaning tasks such as the imputation of missing data using methods specialized for MCA [7].

We conducted our MCAs on the full suite of demographic, occupant SES (assets/education), and aggregated dwelling variables and on a reduced set of only occupant SES variables using R's `FactoMineR` package [7]. We then segmented the individual (household) mappings from each MCA run using `FactoMineR`'s hierarchical clustering method. We set the number of components used by the clustering function to the number of dimensions required to capture over 75% of the variance, thus the full suite and reduced MCAs used seven and four dimensions, respectively. Instead of identifying the number of segments a priori, we allowed the function to select an appropriate number of segments based on minimizing within-cluster inertia for each partition, resulting in three household segments for each MCA run. We also derived quintiles from the first component of our MCAs for the sake of comparison with the DHS-WI.

We evaluated the MCA-derived labels against the DHS-WI based on 1) distinctness of the household types they encompass and 2) diversity in household types by DHS sampling cluster. To measure distinctness of households by segment, we calculated silhouette scores, averaged across all variables, built environment variables, occupant SES variables, and demographic variables. Very compact, well separated clusters would score 1; completely overlapped clusters would score -1.

To measure diversity of household types within DHS sampling clusters we calculated Shannon equitability scores, using R's `vegan` package [11] to calculate Shannon entropy and normalizing these figures by the natural log of the number of classes. A score of 0 indicates complete dominance of one label within a sampling cluster; 1, that all labels are present in equal measure. Taken together, these distinctness and diversity measures enable comparison between the DHS-WI and MCA-derived metrics based on both survey-specific and geographic properties of the DHS.

Finally, we compared spatial variation between the DHS-WI and the MCA-based metrics. For each metric, class prevalences (proportions) were described as planar point patterns (PPPs) modeled from 30 random displacements of the DHS sampling cluster centroids (GPS points) for Ghana (within 2km, and 5km-10km displacement zones for urban and rural areas, respectively) [1]. We then apply an edge corrected, absolute risk function to each PPP at 500m resolution to estimate gridded class prevalences that average as final estimates [4]. In future work, these estimates could be combined with gridded population totals to estimate counts of households by SES class; for now they simply describe the expected mixture of groups.

3 Results

Figure 1 visualizes the MCA-derived segments (full suite: panels A, B; reduced: panels C, D) (light gray hulls) relative to projected DHS-WI labels (red to blue: poorest to richest), based on active MCA variables (black text) and descriptive supplemental variables (gray text). For both MCA-derived metrics, dwelling and occupant variables expected to be associated with lower SES (natural or rudimentary building materials, poor quality water and toilet facilities; low education, lack of assets) fall to the left side of the factor maps (panels A and C), while higher household education and increased assets and dwelling quality variables fall to the right. As seen in the cluster maps (panels B and D), individuals assigned a DHS-WI from poorest to richest also track left to right. While the reduced MCA's three segments largely align along this left-right axis (64.5% of explained inertia), the full suite MCA's segments also split across the second dimension which appears to be influenced by demographics: household size (HS), having children (kids), head's marital status (h.married), and working age adult

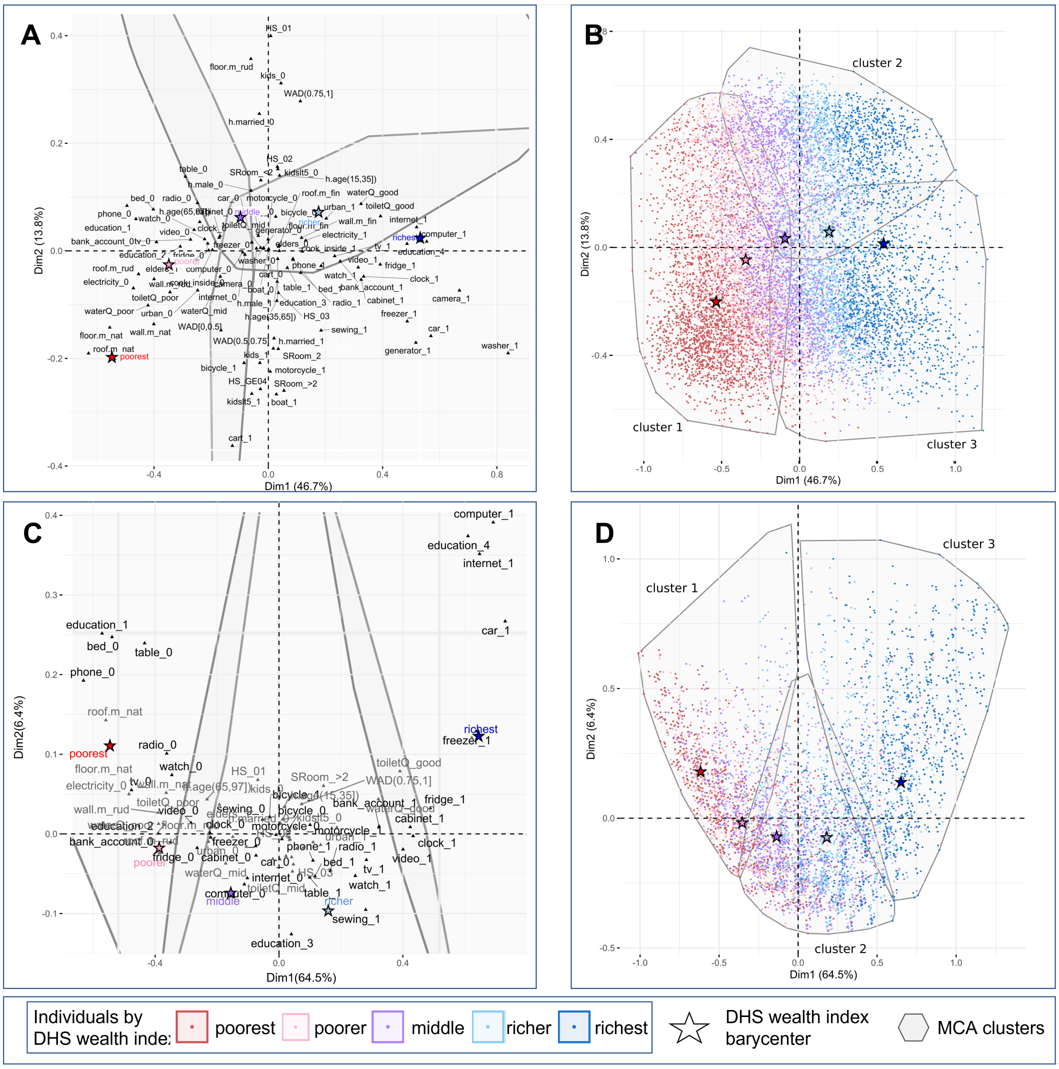

Figure 1 Visualizing segmentation results. A: factor map of full-suite MCA. B: cluster map of full suite MCA. C: factor map of occupant SES MCA. D: cluster map of occupant SES MCA.

proportion (WAD). Given that scholarly opinion is divided as to how such demographic factors should be used in the calculation of SES, our results invite further investigation of this dynamic [8][13].

Table 1 compares the DHS-WI and MCA-derived metrics (both quintiles and segments) based on diversity and distinctness criteria. In terms of **household type distinctness**, the MCA-derived metrics generally improve upon the DHS-WI: full suite MCA segments more cleanly distinguish household types than the DHS-WI when measured across all variables or the variable subsets, while the reduced MCA segments yield the highest silhouette score for the occupant SES variables (though lower ones for built environment variables). Comparing the metrics based on **diversity**, reduced occupant SES segments tend to be considerably more mixed within DHS sampling clusters than the DHS-WI or full suite MCA. Comparing the quintile and segment-based MCA metrics, we also note that allowing for the expression of additional MCA dimensions (via hierarchical clustering) generally improves distinctness/diversity over reliance purely on the first MCA component.

Table 1 Distinctiveness and diversity measures. Silhouettes measured from -1 (complete overlap) to 1 (complete separation), diversity from 0 (one class present) to 1 (all classes equally present).

	DHS-WI	Full suite MCA quintile	Full suite MCA segments	Reduced MCA quintile	Reduced MCA segments
Household Distinctness (Mean Silhouette Width)					
All variables	0.0364	0.0635	0.1882	0.0639	0.1231
Built environment	-0.0163	-0.0075	0.0501	-0.0601	-0.0266
Occupant SES	0.0044	0.0567	0.0988	0.0953	0.2047
Demographic	-0.0912	-0.0430	0.1579	-0.0384	-0.0319
DHS Cluster Diversity (Mean Shannon Equitability)	0.5512	0.7037	0.4995	0.7895	0.7082

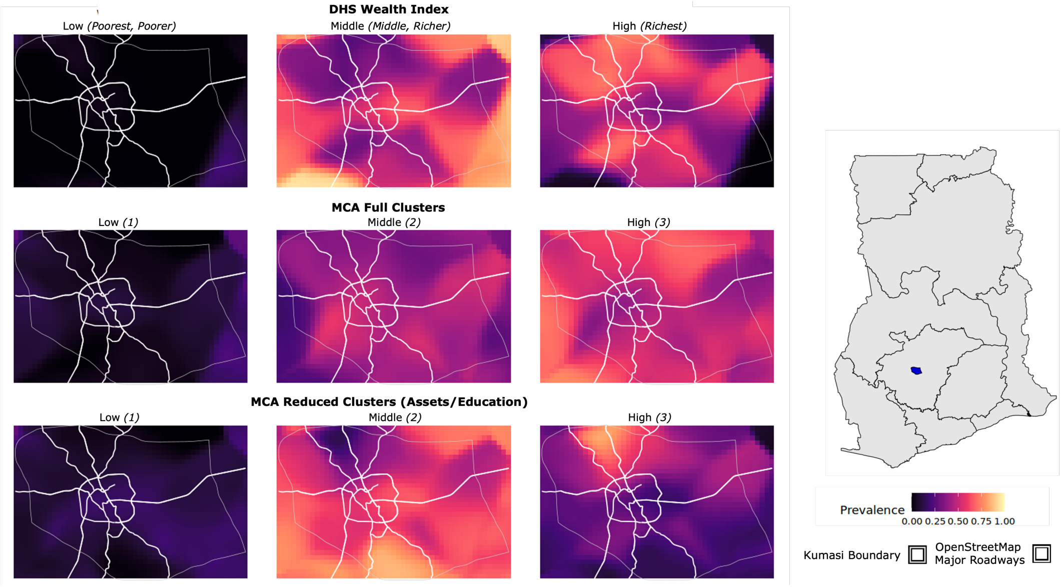

Figure 2 Prevalence estimates of SES classes for the DHS-WI, full suite MCA segments, and reduced occupant SES only MCA segments in Kumasi, Ghana.

Figure 2 displays spatial variation in the prevalence estimates of SES segments by metric for Kumasi, Ghana's second largest city. Consistent with the diversity measures in Table 1, the MCA-based metrics reveal increased mixing of SES classes. In particular, the reduced MCA segmentation shows an increased presence of households associated with low SES (segment 1) compared to the DHS-WI, which features virtually no households with "poorest" and "poorer" labels within Kumasi. Compared to DHS-WI and the full suite MCA, the mapped reduced MCA segments also show increased concentrations of middle-SES households (segment 2) in the south-central area, and high-SES households (segment 3) in the north-central area.

4 Concluding discussion

To explore improvements upon the DHS-WI, we developed several alternative SES classifications based on MCA. Comparing the DHS-WI to the MCA-derived metrics reveals that the former may be both 1) too broad in its scope (subjects/variables) and 2) too reductionist (one-dimensional) to clearly delineate household SES. Paring our metrics down from an all-inclusive MCA (full suite) to assets and education alone (reduced), we remained able to identify SES categories in alignment with an intuitive socioeconomic gradient. This suggests that in future efforts we can eschew mixing subsets of variables (i.e. built environment with occupant SES) whose relationships we would prefer to explicitly measure. Models leveraging such subpopulation-place relationships could potentially be applied in data sparse contexts to predict occupant SES where this information has not been observed but built environment characteristics have. Further, as demonstrated in Figure 2, these MCA-based metrics reveal more nuanced spatial patterns than the DHS-WI, including potential residential locations of urban poor populations, which in turn can lend detail to characterizations of the built environment extracted from remotely-sensed imagery [10]. Work remains in data engineering, testing modeling alternatives, and external validation, yet this work takes an important step in advancing our ability to map and model populations accurately and precisely.

References

1 Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: methodology and applications with R.* CRC press, 2015.

2 Abhijit V. Banerjee and Esther Duflo. What is middle class about the middle classes around the world? *The Journal of Economic Perspectives: A Journal of the American Economic Association*, 22(2):3–28, 2008. doi:10.1257/jep.22.2.3.

3 Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119, 2022. doi:10.1073/pnas.2113658119.

4 P.J. Diggle. *Statistical analysis of spatial point patterns.* Hodder Education, 2003.

5 Ghana Statistical Service, Ghana Health Service, and ICF International. Ghana 2014 Demographic and Health Survey [Dataset]. GHPR72FL.DTA, 2015. URL: https://dhsprogram.com/pubs/pdf/SR224/SR224.pdf.

6 Ghana Statistical Service GSS, Ghana Health Service GHS, and ICF International. Ghana demographic and health survey 2014. Technical report, Ghana Statistical Service – GSS, Rockville, Maryland, USA, 2015. URL: http://dhsprogram.com/pubs/pdf/FR307/FR307.pdf.

7 Francois Husson, Julie Josse, Sebastien Le, and Jeremy Mazet. FactoMineR: multivariate exploratory data analysis and data mining, 2022. URL: https://CRAN.R-project.org/package=FactoMineR.

8 Charles I. Jones and Peter J. Klenow. Beyond GDP? Welfare across countries and time. *American Economic Review*, 106(9):2426–2457, 2016. doi:10.1257/aer.20110236.

9 Homi Kharas. The unprecedented expansion of the global middle class: an update. Global Economy and Development Working Paper 100, Global Economy and Development at the Brookings Institution, 2017. URL: https://www.brookings.edu/research/the-unprecedented-expansion-of-the-global-middle-class-2/.

10 Dalton Lunga, Jacob Arndt, Jonathan Gerrand, and Robert Stewart. Resflow: A remote sensing imagery data-flow for improved model generalization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10468–10483, 2021.

11 Jari Oksanen, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Matt Barbour, Michael Bedward, Ben Bolker, Daniel Borcard, Gustavo Carvalho, Michael Chirico, Miquel De Caceres, Sebastien Durand, Heloisa Beatriz Antoniazi Evangelista, Rich FitzJohn, Michael Friendly, Brendan Furneaux, Geoffrey Hannigan, Mark O. Hill, Leo Lahti, Dan McGlinn, Marie-Helene Ouellette, Eduardo Ribeiro Cunha, Tyler Smith, Adrian Stier, Cajo J. F. Ter Braak, and James Weedon. vegan: Community Ecology Package, 2022. URL: https://cran.r-project.org/web/packages/vegan/index.html.

12 Mathieu J. P. Poirier, Karen A. Grépin, and Michel Grignon. Approaches and alternatives to the Wealth Index to measure socioeconomic status using survey data: a critical interpretive synthesis. *Social Indicators Research*, 148(1):1–46, 2020. doi:10.1007/s11205-019-02187-9.

13 Shea O. Rutstein and Kiersten Johnson. The DHS wealth index. DHS Comparative Report DHS Comparative Reports No. 6, ORC Macro, Calverton, Maryland, 2004. URL: https://www.dhsprogram.com/publications/publication-cr6-comparative-reports.cfm.

14 Jeroen Smits and Roel Steendijk. The International Wealth Index (IWI). *Social Indicators Research*, 122:65–85, 2015. doi:10.1007/s11205-014-0683-x.

15 Dana R. Thomson, Forrest R. Stevens, Robert Chen, Gregory Yetman, Alessandro Sorichetta, and Andrea E. Gaughan. Improving the accuracy of gridded population estimates in cities and slums to monitor SDG 11: Evidence from a simulation study in Namibia. *Land Use Policy*, 123, 2022. doi:10.1016/j.landusepol.2022.106392.

16 Pierre Traissac and Yves Martin-Prevel. Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: the case for multiple correspondence analysis. *International Journal of Epidemiology*, 41(4):1207–1208, 2012. doi:10.1093/ije/dys122.