Original Article

Sociological Methodology
2022, Vol. 52(1) 53–74

DOI: 10.1177/00811750211057572
http://sm.sagepub.com

Small-Area Analyses Using Public American Community Survey Data: A Tree-Based Spatial Microsimulation Technique

Nick Graetz[1,2], Kevin Ummel[2,3], and Daniel Aldana Cohen[4]

Abstract

Quantitative sociologists and social policymakers are increasingly interested in local context. Some city-specific studies have developed new primary data collection efforts to analyze inequality at the neighborhood level, but methods from spatial microsimulation have yet to be broadly used in sociology to take better advantage of existing public data sets. The American Community Survey (ACS) is the largest household survey in the United States and indispensable for detailed analysis of specific places and populations. The authors propose a technique, tree-based spatial microsimulation, to produce "small-area" (census-tract) estimates of any person- or household-level phenomenon that can be derived from ACS microdata variables. The approach is straightforward and computationally efficient, based only on publicly available data, and it provides more reliable estimates than do prevailing methods of microsimulation. The authors demonstrate the technique's capabilities by producing tract-level estimates, stratified by race/ethnicity, of (1) the proportion of people in the census-tract population who have children and work in an essential occupation and (2) the proportion of people in the census-tract population living below the federal poverty threshold and in a household that spends greater than 50 percent of monthly income on rent or owner costs. These examples are relevant to understanding the sociospatial inequalities dramatized by the coronavirus disease 2019 pandemic. The authors discuss potential extensions of the technique to derive small-area estimates of variables observed in surveys other than the ACS.

Keywords

spatial microsimulation, small area estimation, decision trees, American Community Survey

Context is central to sociological study. Across sociological methods, qualitative studies often focus on rich description and explanation within a specific place and time. Quantitative studies, on the other hand, try to extend empirical generalizability by using sample designs that are representative of broader geographic or temporal contexts. In the literature on social stratification and inequality, these quantitative studies

[1]Department of Sociology, Princeton University, Princeton, NJ, USA
[2]Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA
[3]Greenspace Analytics, Fort Collins, CO, USA
[4]Department of Sociology, University of California, Berkeley, Berkeley, CA, USA

Corresponding Author:
Nick Graetz, Princeton University, Department of Sociology, 106 Wallace Hall, Princeton, NJ 08544, USA.
Email: ngraetz@princeton.edu

are often based on large nationally representative cohorts (e.g., the Panel Study of Income Dynamics, the National Longitudinal Survey of Youth), in which sample sizes and limited geographic information often prohibit stratifying results by region, let alone by specific neighborhoods. Such work cannot easily engage the more spatially specific context explored by ethnographers. But quantitative sociologists now increasingly recognize the need to explore spatial context with far greater precision, as exemplified by the development of the "neighborhood effects" literature. From this perspective, place, broadly defined, organizes social life, labor relations, institutions, and more (Sampson 2008; Sharkey 2013; Sharkey and Faber 2014). This literature convincingly shows that quantitative sociology needs to develop its capacity for neighborhood-level research.

Doing so, however, will require adjusting quantitative sociology's conventional techniques in dialogue with traditions of geospatial analysis in economics and geography, as we will discuss. Because of the increasingly fractal nature of social and economic phenomena in the United States, researchers and policymakers are increasingly interested in local estimation. As Sharkey and Faber (2014) noted, if a "place effect" is identified, the primary questions become who is most affected and where they live. Where is the exposure most geographically concentrated in the United States, and what are its salient intersections with related social dimensions such as race and gender? Local estimates of social, economic, and demographic processes provide rich insight into the interaction of place and individual characteristics (Cagney et al. 2014; Sampson 2012; Sharkey 2013). One approach used by quantitative sociologists so far has been to use targeted, independent primary data collection projects in specific urban contexts, such as the Project on Human Development in Chicago Neighborhoods (Sampson 2012). However, this decentralized, unharmonized system of data collection and analysis cannot provide a comprehensive national picture of how place increasingly structures social, health, and environmental stratification in the United States. Using a handful of distinct local data sets forecloses systematic comparison across the country, and the prevalence of independent, localized data collection has led to the overrepresentation of certain urban populations, such as New York City and Chicago, in the place effects literature. The emerging alternative is administrative linkage projects based on restricted-use data sets such as tax returns. This work is important and compelling, but the research is time consuming, expensive, and exclusive, inhibiting its wider use (Chetty et al. 2016, 2018). In short, U.S. quantitative sociology's local analysis risks being stymied.

An alternative to the strategies used so far is to make better use of all the publicly available information provided by the American Community Survey (ACS), the largest survey of U.S. households. The ACS provides extensive information on household demographics, finances, employment, health insurance, migration, ancestry, linguistics, housing conditions, and more. Given the ACS's uniquely large sample size and breadth of collected variables, it is indispensable for complex research questions. But to protect respondents' confidentiality, the ACS does not report all its data at the census-tract (or "neighborhood") level. As a result, research questions that pivot on neighborhood-level dynamics in the United States often go unanswered because the necessary information is not tabulated or made public. Sociologists using the ACS

often try to infer nuanced answers to complex research questions by mapping several crude tabulated indicators, or simply use data at a higher level (e.g., metropolitan-area microdata) to answer specific research questions, thus forgoing analysis of local context. There are various methods for conducting "small-area" analysis using data such as those provided by the ACS, but these have not been fully used in sociology.

We focus on the method of "spatial microsimulation" (SM), whereby individual- or household-level microdata originally sampled from a large geographic area are reweighted to produce a new sample that is thought to be representative of the population in a specific place ("small area"). The new, "local" sample is designed to be representative with respect to a set of "constraint variables" selected by the analyst. These variables provide a set of known, local population totals ("population constraints"), usually sourced from a census. The original microdata are reweighted ("calibrated") to match these totals in aggregate, for example, the number of women aged 25 to 35 years, the number of people with college degrees, or the number of people employed full-time. With this locally representative sample in hand, any number of place-specific research questions can be asked and answered. Researchers can layer on additional place-specific data sets, from surface temperatures that are rising because of climate breakdown to data on local police stops, to ask complex questions. Researchers can also use synthetic microdata to predict counterfactual scenarios or simulate policy experiments, and they can fuse synthetic microdata with other data sets to predict new variables at the individual- or small-area level (Zhang et al. 2014). SM is a powerful approach, and we believe we can improve upon it.

In this article, we focus on the foundation: demographic microsimulation with ACS data. We propose a general framework, tree-binned SM (TBSM), to provide estimates for any small area and for any variable derived from the ACS microdata. We extend a standard reweighting method for SM by proposing a novel decision tree framework for selecting optimal population constraints and automated "binning" of constraint attributes (e.g., income categories), removing the need for the researcher to arbitrarily select constraints. Our focus on an improved method for selecting population constraints is essential to mitigate multicollinearity and interaction effects among candidate variables, and the issue of constraint selection is perhaps the weakest link in extant microsimulation techniques. Importantly, our proposed methodology and associated ACS-tailored code base allow maximum flexibility, providing an automated process to derive small-area estimates across the United States, relying only on publicly available data. We validate the ability of our TBSM model to replicate census-tract estimates for various variables and locales, and we compare these results with those derived from a conventional SM model. We discuss how sociologists can use this method to answer timely, complex questions related to labor, housing, family, and social policy, at the spatial scale of the neighborhood, thus allowing us to explore fine-grained variation across local contexts. We use two examples and discuss their relevance to both existing sociological inquiry that uses only national data and practical implications for understanding inequitable consequences of the coronavirus disease 2019 (COVID-19) pandemic: (1) the proportion of people in a census-tract population who have children and work in essential occupations and (2) the proportion of people

in a census-tract population living below the federal poverty threshold and in households that spend greater than 50 percent of monthly income on rent or owner costs.

BACKGROUND

History of Microsimulation

Many of the foundational ideas in computational "microsimulation" were developed in economics. In 1957, Orcutt (1957) published "A New Type of Socio-economic System" in the *Review of Economics and Statistics*, reprinted in 2007 in the *International Journal of Microsimulation* (Orcutt 2007). Orcutt described how synthetic microdata can be used to produce projections of complex social, labor, and demographic systems from widely available tabulated data sets, although in this early work, spatial resolution was not yet a concern. The work of Orcutt and colleagues may have been the beginning of microsimulation studies, but they built on older techniques for the sociological and demographic analysis of population change (Zagheni 2015); for example, the synthetic cohort life table approach central to the field of demography can be considered a microsimulation. Indeed, the procedure described by Orcutt (1957) involves a demographic model of marriage, divorce, fertility, and mortality rates. These foundations of microsimulation have now extended into various fields, including contemporary statistics (multiple imputation for incomplete data) and demography (multistate life tables) (Rubin 1987; Schafer 1997). In reviewing Orcutt et al. (1961) for the *American Sociological Review* and discussing whether the methods might be useful in sociological research, Wager (1962) wrote that the "uses discussed often requires major commitments of funds, specialized talent, and electronic equipment." As we will discuss, these concerns about microsimulation continue to be cited today, although they are far weaker obstacles than at the time of Wager's review.

The extension of these techniques to SM first appeared in the field of geography, in which they continue to be developed. Wilson and Pownall (1976) introduced synthetic reconstruction techniques to spatial analysis (e.g., iterative proportional fitting) and later extended the technique in the areas of urban analysis and transportation research (Beckman, Baggerly, and McKay 1996; Birkin and Clarke 1988; Wilson and Pownall 1976). Methodological innovations since Wilson and Pownall have focused on developing more efficient and accurate calibration algorithms to the chosen population constraints, such as combinatorial optimization (Huang and Williamson 2002; Williamson, Birkin, and Rees 1998) and deterministic reweighting (Ballas et al. 2007). Rahman and others have demonstrated how an SM approach coupled with robust validation can avoid many of the pitfalls of statistical approaches to small-area estimation that are common in epidemiology and other fields (Das et al. 2019; Rahman 2017; Rahman and Harding 2016; Rahman et al. 2013).

Challenges in SM

Despite the substantive advantages of SM for small-area estimation and the advent of faster, more efficient computational platforms, modern techniques are not applied

broadly in sociological research on U.S. communities. Applied research using SM has focused broadly on poverty, health, transportation, and public policy (O'Donoghue, Morrissey, and Lennon 2014; Rahman and Harding 2016; Sakshaug and Raghunathan 2014). However, with the exception of Sakshaug and Raghunathan (2014), virtually all applied studies focus on contexts outside the United States. For example, synthetic microdata generated with an iterative proportional fitting approach has been used to calculate smoking rates across small areas in New Zealand and the United Kingdom (Smith, Pearce, and Harland 2011; Tomintz, Clarke, and Rigby 2008).

In describing progress and persistent gaps in the development and application of SM techniques, we must note the important decisions that must be made in any microsimulation study (O'Donoghue et al. 2014): (1) the data sources and spatial scope, (2) the data creation and calibration methodology, (3) which variables to use as population constraints, and (4) the validation of estimates. To date, the methodological research has largely focused on data creation and calibration in comparing the relative (dis)advantages of the three predominant calibration methods: combinatorial optimization, generalized regression (GREG) reweighting, and iterative proportional fitting (Whitworth et al. 2017).

In contrast, there is little discussion in the current microsimulation literature on a rigorous method for selecting population constraints. The selection process is generally opaque in published SM studies (Huang and Williamson 2002; Smith, Clarke, and Harland 2009). It might be informed by correlational criteria or possibly stepwise regression, but reliance on "expert judgement," convention, or an *ad hoc* process of trial and error is more common (Huang and Williamson 2002; O'Donoghue et al. 2014; Smith et al. 2009). The issue of variable selection has received far less attention than that of calibration techniques. Given the likelihood of multicollinearity and interaction effects among candidate variables, the constraint selection task is decidedly nontrivial. Worse still, including constraint variables that do not enhance predictive ability could reduce the quality of outputs by making it difficult or impossible for the SM model to calibrate the microdata weights to the selected population margins.

Voas and Williamson (2000) offered the following considerations in discussing their choice of population constraints for an SM using UK census tables:

> While these tables cover a reasonable cross-section of census topics, it is quite possible that a better set of eight tables could be chosen. The choice of constraints is guided by three main considerations: 1) computer resources, as every additional table included will increase the number of iterations required to achieve a given level of fit, 2) the perceived importance of the topics, and 3) the extent of correlation with other variables, since fit on unconstrained tables will be affected by how far those counts have been determined by the constraints. (p. 351)

Computational limitations have historically been a tremendous barrier to constraint selection and quantity—and indeed the adoption of microsimulation techniques more broadly (Orcutt 1957; Wager 1962). But the other points raised by Voas and Williamson (2000) relate to the theoretical and empirical relevance of possible constraint variables, and results are sensitive to selection. Two decades later, Whitworth

(forthcoming) reiterates these same points in framing constraint selection as an arbitrary choice that must be conceptually justified:

> Firstly, the researcher must decide which data attributes to optimize against; a choice which should be driven by the research question. In certain cases, for instance, having an accurate representation of age, gender, and marital status may be critically important; in others, gender may be unimportant but income and educational status may be critical. After determining which attributes to optimize against, the researcher then builds constraints from the macro tables and maps them between the micro and macro data. (p. 15)

Sakshaug and Raghunathan (2014) identified the same constraint selection problem. In their SM analysis of the National Health Interview Survey, the authors selected variables on the basis of their "common usage" and recode categorical variables to binary "to ease computation." The authors' validation and robustness tests do not include varying the selection of constraints.

The additional question of whether and how to collapse, aggregate, or otherwise "bin" individual constraint variable margins is, to our knowledge, entirely unaddressed in the SM literature, for example, whether and how to "bin" a 16-category ordinal income variable to reduce the number of margins to be calibrated. SM practitioners do regularly bin constraint variables, but the process is *ad hoc*, almost always undocumented, and without theoretical rationale.

Present Study

Our study has two connected aims. First, our broader aim is to translate the development and advantages of contemporary SM from the fields of geography, economics, and urban planning to sociological research. We draw on topics for which geographic context has become increasingly important in the United States, including racialized stratification in labor and housing. Toward this end, our methodological aim involves combining a generalized reweighting approach to SM with high-quality data from the ACS. We extend previous research using this technique to include a rigorous treatment to the problem of selecting population constraints using decision trees, validating against existing techniques with a variety of out-of-sample statistics. We develop a generalized SM tool that can be applied broadly to examine a vast array of local research and policy questions using the ACS, and we demonstrate several use cases related to geographic and racial/ethnic stratification in labor, housing, and family.

METHODS

TBSM

The technique introduced here uses decision trees to inform both the selection and binning of candidate constraint variables within an SM model. We call our method "tree-binned spatial microsimulation." We motivate the method using a toy example and then provide details of our actual implementation using ACS data. Our focus, and that of nearly all studies in the SM literature, is the use of SM for small-area estimation:

Table 1. Examples of Variables Used in Spatial Microsimulation Study

Constraint Variable	Number of Categories	Examples
Age (years)	14 (ordinal)	18–25, 26–30, . . . , ≥ 85
Education (years)	7 (ordinal)	0–3, 4–6, . . . , ≥ 16
Occupation	4 (nominal)	A, B, C, D

the estimation of a "target variable" or quantity or outcome of interest that is observable in the microdata but unknown for the small area.

Imagine we want to estimate per capita income (target variable) in a specific town (small area). We have microdata sampled from the national adult population that allow us to observe individuals' income, age, years of education, and occupation. Income is a continuous variable. The other three are discrete candidate constraint variables whose categories align with known local population totals (e.g., number of people by age group) taken from a separate data source such as a census. This is the most common data structure for SM studies (see Table 1).

We begin by fitting a decision tree to the microdata with income as the response (dependent) variable and the three candidate constraints as predictor (independent) variables. A decision tree consists of a series of recursive, binary splits of available predictor variables, where successive "nodes" exhibit increasing uniformity with respect to a response variable (Breiman 1993). In our example, the fitted tree (Figure 1) assigns each of the microdata individuals to one of five groups ("terminal nodes") for which we report the mean income. The tree is built by greedily selecting the split that maximizes the "purity" (minimizes the internal variance) of the resulting nodes. Consequently, decision trees reveal a natural ranking of candidate constraint variables in terms of their ability to explain the target variable; splits that occur earlier/higher in the tree are more influential in terms of explaining the variance of the target variable; subsequent splits contribute less explanatory power. The hierarchical nature of decision trees means they implicitly capture "interaction effects" between predictors because selected splits are conditional on splits higher in the tree. We also choose to use a single decision tree rather than a tree ensemble (e.g., random forest) because our primary aim is not optimization of prediction accuracy but, rather, to derive a single, suitable binning strategy from a tree's split decisions. The multiple trees resulting from an ensemble is ambiguous in this respect.

Note that in Figure 1, the bottom two terminal nodes created by a split on occupation exhibit a relatively small difference in mean income ($45,000 versus $40,000). We could "prune" the tree to remove the occupation split, and the resulting tree would be less complex but nearly as good at explaining variation in individuals' income. In practice, decision tree algorithms often use k-fold cross-validation to decide how to optimally prune the tree. The cross-validation step calculates the mean and standard error of the out-of-sample model error (across k folds) for trees of varying complexity. We use the rpart package in the R language with $k = 30$ (Therneau and Atkinson 2019). We follow a common rule of thumb in decision tree analysis (the "1-S.E. rule")

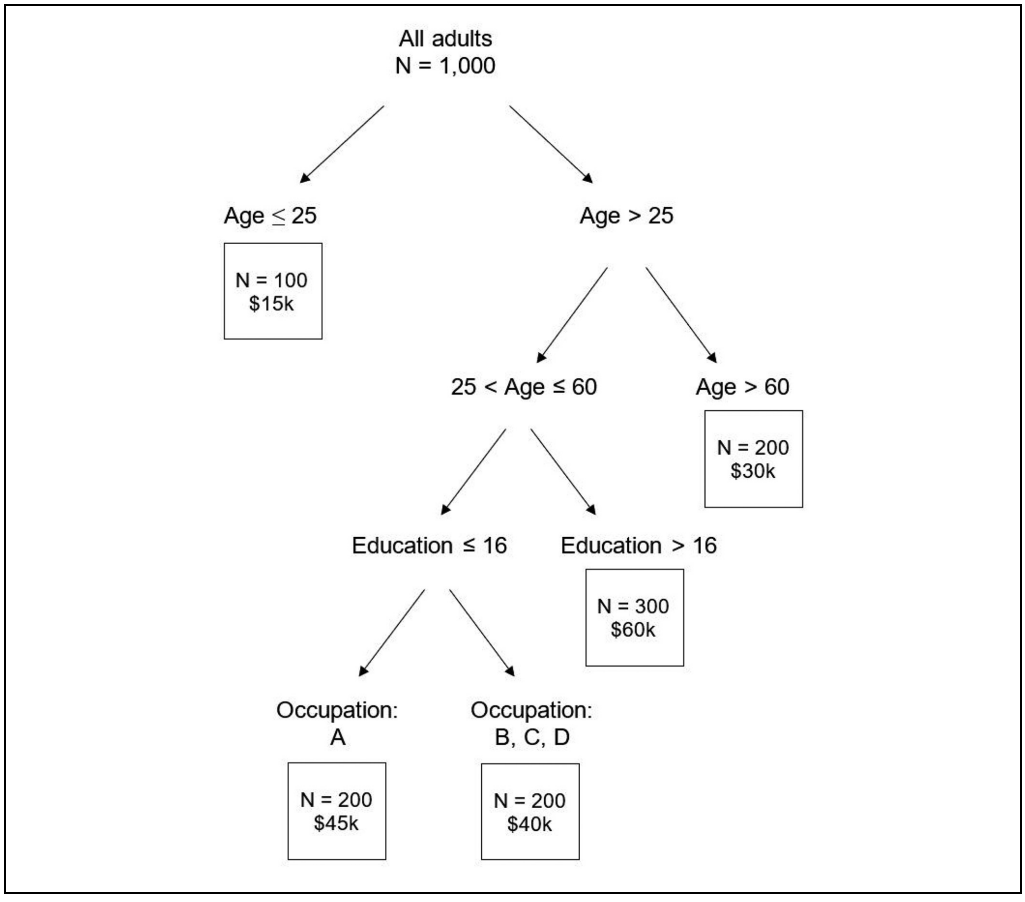

Figure 1. Example decision tree fit to microdata with income as the target variable.

by selecting the least complex tree that exhibits mean out-of-sample error within 1 standard error of the minimum (Wood 2017). Simulation studies show that this rule of thumb generally does a good job of preventing the tree from fitting to noise in the training data (Therneau and Atkinson 2019). A pruned tree need not contain all the candidate constraint variables present in the microdata. Consequently, a decision tree provides explicit selection of constraint variables from among the full candidate set.

The decision tree's split points also reveal a "binning strategy" for the constraint variables. In our example, age is split at 25 and 60 years old. Years of education is split at 14, and occupation, which is nominal rather than ordinal in nature, has a split that groups occupations B, C, and D into one node and occupation A into another. Because split points are selected to minimize node impurity, the "bins" defined by the split points reflect a preferred way of grouping individual constraint categories. The constraint variable binning strategy "deduced" from Figure 1 is shown in Table 2. In this case, the total number of categories (population constraints) defined by the candidate constraints is reduced from 25 originally (14 + 7 + 4) to just 7 after binning.

Table 2. Example of a Constraint Variable Binning Strategy Deduced from the Decision Tree Fit in Figure 1

Constraint Variable	Binned Categories
Age (years)	Three bins: 18–25, 26–60, $\geq 60+$
Education (years)	Two bins: 0–15, ≥ 16
Occupation	Two bins: [A], [B, C, D]

Note that a constraint variable's split points across the entire tree need not be mutually exclusive (as is the case in Figure 1). In the case of more complex trees, the deduced binning strategy maintains as much category resolution as is necessary to respect all of a variable's unique splits across the tree.

Subsequent calibration of the microdata uses these binned category definitions, reweighting the microdata to mimic the local (binned) population constraints. For example, if the decision tree reveals that distinguishing between 60- to 75-year-olds and those ≥ 75 years old is not particularly important for predicting the target variable, then there is little reason to calibrate the local sample to those individual population constraints. It makes more sense to calibrate to the binned population constraint (the total number of people ≥ 60 years old in this case). It is not the number of constraint variables *per se* that make sample calibration difficult as much as the number of individual population totals that must be replicated. TBSM effectively reduces the number of individual constraints and increases the likelihood of successfully replicating the selected population totals.

Importantly, the use of cross-validation to select appropriate tree complexity helps guard against overfitting to noise in the microdata. An SM model using a large number of constraints may "calibrate successfully" (i.e., replicate totals), but this does not necessarily lead to reliable small-area estimation. In general, TBSM seeks to identify and use the smallest number of population constraints consistent with strong out-of-sample prediction of the target variable. In addition to selecting and binning candidate constraint variables in a theoretically defensible way, decision trees have a number of secondary but useful features for the purposes of SM models. First, the target variable can be continuous, binary, or multinomial in nature, and computation time is similar across cases. Conversely, common "off the shelf" variable selection techniques (e.g., stepwise regression) are typically not amenable to multinomial target variables and tend to be more computationally expensive in the binary (logistic) case. Second, decision trees are quick to compute and scale well with both the number of observations and the number of predictors. This is a particular strength in our application using ACS data (described below), which includes a relatively large number of candidate constraint variables. Third, decision trees provide an overall measure of relative variable importance, calculated by summing the improvement in prediction attributable to each variable's splits across the tree.

Although the methodology described above is applicable to standard SM data inputs, we use the ACS exclusively for our applied TBSM model. The validation and

demonstration outputs described in the following sections use only data sourced from the 2012 to 2016 (five-year) ACS. The data inputs take two forms: (1) microdata lacking geographic specificity and (2) block-group-level population counts for individual constraint variable categories (e.g., the number of households with income less than $10,000). The latter are extracted from U.S. Census Bureau "summary tables," which provide a large potential set of candidate constraint variables. We constructed 21 candidates for use within our model (see Appendix Table 1 in the online supplement), intended to cover a range of socioeconomic and dwelling characteristics. The ACS microdata are processed to create person-level microdata exhibiting concordance with the constraint variables in Appendix Table 1 (see Section 1 of the online supplement). In principle, the TBSM technique places no upper limit on the number of candidate constraint variables. We use only univariate candidate constraints for simplicity, but multivariate or cross-tabulated constraints (e.g., age groups by sex) can also be used. In general, it is beneficial to include any candidate constraint variable that might be predictive of the target. Because these variables are observed for individual block groups (more than 200,000 nationally), we can perform SM for any geographic unit (small area) that is an aggregation of block groups (e.g., census tract, county).

The ACS data provide both household- and person-level constraint variables (e.g., an individual's race and the number of people in that individual's household). To use all this information within a single framework, the data inputs to the tree-fitting and calibration steps consist of person-level observations nested within households, with household attributes replicated for members of the same household. This is similar to the strategy used by Bar-Gera et al. (2009) (see also Section 4.3.1 in Muller 2017). The advantage is that the code base can handle any potential constraint or target variable, whether household or person level in nature. See Section 1 of the online supplement for more information on available ACS data.

For our applied model using ACS data, calibration is performed using the GREG estimator (Tanton et al. 2011). GREG calibration is iterative and continues until the (reweighted) sample population totals are within some tolerance of the local population totals. We leverage the facts that (1) local population totals from ACS summary tables are accompanied by a standard error, and (2) the decision tree provides importance weights for each constraint variable. Consequently, calibration quality is measured by the importance-weighted mean absolute z score across margins, and GREG iterations are terminated when this quantity falls below 0.125 (this tolerance is arbitrary, but sensitivity testing demonstrated this threshold to be appropriate). This allows the calibration step to put greater weight on the replication of population constraints with low uncertainty and those associated with constraint variables that are more predictive of the target variable.

Validation Strategy

A challenge of microsimulation, and small-area estimation in particular, is validation of model output, given that the techniques are typically used to estimate unobserved phenomena. In our case, we can compare model output with known small-area

Table 3. Review of Unique Variables and Constraints Used in Applied Microsimulation Studies

Reference	Variables	Constraints
Smith et al. (2011)	4	—
Ballas et al. (2007)	6	18
Campbell and Ballas (2013)	8	—
Lovelace, Ballas, and Watson (2014)	5	40
Anderson (2007)	7	26
Smith et al. (2011)	4	21
Tomintz et al. (2008)	4	18
Morrissey and O'Donoghue (2011)	5	—
Ifesemen, Bestwick-Stevenson, and Edwards (2019)	7	29
Smith et al. (2011)	4	—
Ballas et al. (2007)	6	18
Campbell and Ballas (2013)	8	—
Lovelace et al. (2014)	5	40
Anderson (2007)	7	26
Smith et al. (2011)	4	21
Tomintz et al. (2008)	4	18
Morrissey and O'Donoghue (2011)	5	—
Ifesemen et al. (2019)	7	29

estimates derived from information in ACS summary tables. We selected four target variables to use for validation: mean years of schooling, percentage of the population with public health insurance, mean hours worked per week, and mean annual household income (see Appendix Table 3 in the online supplement). These include both continuous (numerical) and discrete (binary) variables; the small-area estimate is a mean value for the former and a population proportion for the latter. A validation exercise should test model performance using assumptions and data inputs similar to those for legitimately "unknown" target variables. Consequently, we exclude candidate margin variables that might give a model artificially high predictive ability (the "excluded variables" column of Appendix Table 3). For example, model estimates of "mean years schooling" are nearly perfect when the "education" margin variable is used (as expected, because they are derived from the same source data). But this is not indicative of model performance for real-world cases in which the target variable is not necessarily highly correlated with a margin variable.

To assess the value added of our approach, we construct a "baseline" SM model for comparison with the TBSM model. The baseline model is an attempt at a plausible "standard" approach to constraint selection. Reviewing applied SM studies (see Table 3), we find that researchers typically select about five constraint variables, with a total of about 30 individual population constraints. The baseline model uses the same data as the TBSM model but selects up to five constraint variables via stepwise regression (optimal model selected via the Akaike information criterion) and then bins categories to produce approximately 30 total population constraints. The latter step uses the binr package to produce roughly evenly distributed bins for each constraint variable (Izrailev 2016). Constraint selection is considerably more *ad hoc* among practitioners

than the baseline model implies; stepwise regression is not universally used, and there is effectively no documentation of how researchers choose to (or not to) bin constraints. However, we doubt the baseline model is significantly underperforming (and may well outperform) what one would expect from SM practitioners asked to construct a model using the same data. We focus on out-of-sample predictive validity to demonstrate how our TBSM technique performs against the baseline model. Validation is critically important in ensuring robust estimates and demonstrating that SM can be used to make valid inferences at the small-area level. Any new proposed technique can be validated by calculating out-of-sample fit statistics comparing small-area outputs from synthetic microdata to a known small-area estimate that was not used in the SM (Ballas et al. 2007; Rahman and Harding 2016; Voas and Williamson 2000) and demonstrating that the proposed technique outperforms existing techniques using the same fit statistics.

We first estimate average years of education in the five-county New York City metropolitan area and compare with known estimates. Second, we estimate the four indicators above across six study areas spread across the urban-rural continuum. We use the urban-rural codes developed by the U.S. Department of Agriculture Economic Research Service, which were modified and made available by the National Center for Health Statistics (USDA ERS 2015). We chose two large metropolitan counties (Philadelphia, Pennsylvania, and San Francisco, California), two fringe metropolitan counties (Westchester, New York, and Essex, Massachusetts), and two small metropolitan counties (Lee, Alabama, and Bibb, Georgia). TBSM model results are compared with summary table values at the census-tract level. Descriptions of all validation test statistics can be found in Section 2 of the online supplement.

Last, we provide two examples of TBSM to estimate unobserved phenomena at high spatial resolution. Our demonstrations use the considerable household- and person-level detail in the Public-Use Microdata Sample (PUMS) to construct unique target variables that are not available in small-area summary tables. In short, any variable that can be calculated from either household or person PUMS records is eligible for small-area estimation across the United States. Here we focus on two such variables within Philadelphia County: (1) the proportion of the census-tract population (2012–2016) who have children and work in essential occupation, stratified by race/ethnicity, and (2) the proportion of the census-tract population (2012–2016) living below the federal poverty threshold and in households that spend more than 50 percent of monthly income on rent or owner costs, stratified by race/ethnicity.

RESULTS

Validation Results

Overall, the validation exercise suggests the TBSM approach generates estimates in broad agreement with known values (see Table 4). Figure 2 presents out-of-sample predictions of tract-level average years of education across all of New York City using a conventional approach and our tree-based approach. Our method provides a high relative likelihood (0.926), low absolute percentage error (0.023), and high variance

Table 4. Fit Statistics Comparing the Baseline Microsimulation Method with the TBSM Method

Place	Metric	MAPE (Baseline)	MAPE (TBSM)	R^2 (Baseline)	R^2 (TBSM)
Bibb, GA	Mean hours worked per week (age 1664 years)	.11	.11	.79	.80
Bibb, GA	Mean household income	.28	.38	.75	.55
Bibb, GA	Mean years of schooling (age $\geq$ 25 years)	.04	.02	.47	.86
Bibb, GA	Percentage of population with public health insurance	.25	.26	.64	.62
Essex, MA	Mean hours worked per week (age 16–64 years)	.05	.05	.71	.67
Essex, MA	Mean household income	.17	.18	.84	.82
Essex, MA	Mean years of schooling (age $\geq$ 25 years)	.06	.02	.22	.85
Essex, MA	Percentage of population with public health insurance	.43	.25	.43	.80
Lee, AL	Mean hours worked per week (age 1664 years)	.07	.05	.75	.81
Lee, AL	Mean household income	.17	.21	.75	.58
Lee, AL	Mean years of schooling (age $\geq$ 25 years)	.04	.03	.38	.56
Lee, AL	Percentage of population with public health insurance	.32	.35	.66	.63
New York City	Mean years of schooling (age $\geq$ 25 years)	.04	.02	.28	.83
Philadelphia, PA	Mean hours worked per week (age 16–64 years)	.05	.05	.92	.94
Philadelphia, PA	Mean household income	.21	.23	.80	.79
Philadelphia, PA	Mean years of schooling (age $\geq$ 25 years)	.05	.03	.14	.84
Philadelphia, PA	Percentage of population with public health insurance	.36	.35	.53	.51
San Francisco, CA	Mean hours worked per week (age 16–64 years)	.07	.05	.70	.88
San Francisco, CA	Mean household income	.34	.32	.47	.52
San Francisco, CA	Mean years of schooling (age $\geq$ 25 years)	.06	.02	.00	.88
San Francisco, CA	Percentage of population with public health insurance	.60	.70	.79	.77
Westchester, NY	Mean hours worked per week (age 16–64 years)	.10	.11	.70	.72
Westchester, NY	Mean household income	.24	.23	.84	.85
Westchester, NY	Mean years of schooling (age $\geq$ 25 years)	.02	.02	.88	.89
Westchester, NY	Percentage of population with public health insurance	2.27	2.11	.54	.55

Note: MAPE = mean absolute percentage error; TBSM = tree-binned spatial microsimulation.

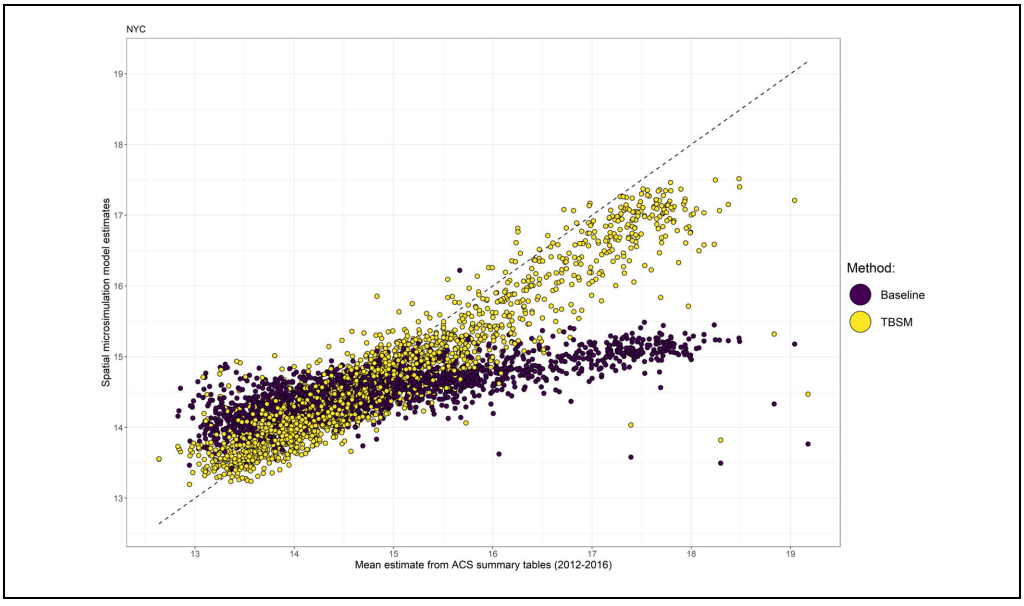

Figure 2. Out-of-sample validation results comparing observed estimates of tract-level mean years of educational attainment in the five counties of New York City with (1) the tree-based spatial microsimulation approach and (2) a conventional approach using stepwise regression to select constraints.
Note: ACS = American Community Survey; TBSM = tree-binned spatial microsimulation.

explained (R^2 = 0.832). In contrast, the stepwise selection method performed much worse on all three fit statistics. An important feature of this validation exercise is demonstrating the advantages of the tree-based method for avoiding shrinkage to the global average, a common feature of most small-area estimation techniques. These patterns are largely replicated when we extend this validation test to a wider set of urban-rural counties and target variables (see Figures 1–5 in the online supplement). The tree-based method consistently performs as well or better than the conventional approach.

Example 1: COVID-19 and the Racialized Intersection of Labor and Childcare

We present two examples using our method to predict target variables that are observed in the PUMS but unavailable at the tract level. First, we estimate the proportion of the census-tract population (2012–2016) who have children and work in essential occupations, stratified by race/ethnicity (Figure 3). Examining local variation in this target variable is important to contemporary sociological research for several reasons. Building on the debates around "essential" workers during the COVID-19 pandemic, we adopt the American Civil Liberties Union's (2020) coding of census occupational categories into an aggregate "essential worker" category. From a relational class structure perspective, this corresponds roughly to what is colloquially known as the working class, namely, occupations that are relatively low wage and tend

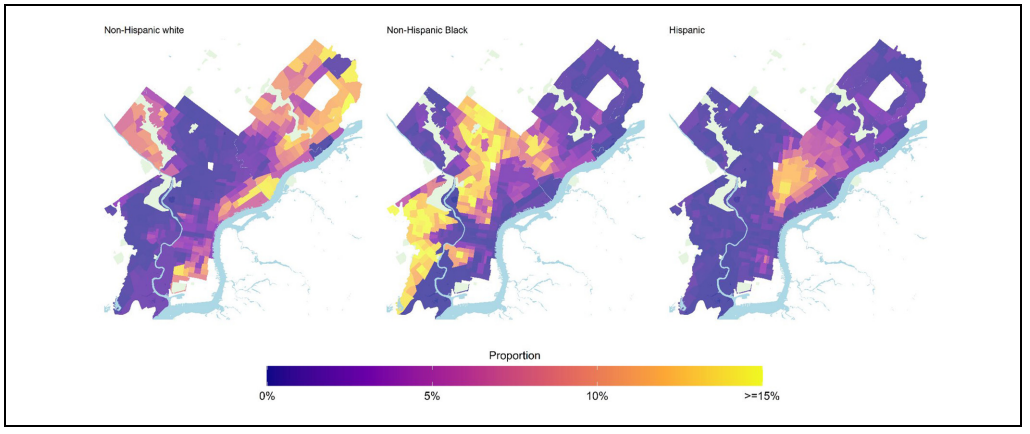

Figure 3. Small-area estimates of the proportion of the census-tract population (2012–2016) who have children and work in essential occupations, stratified by race/ethnicity.

to require in-person, rather than virtual, laboring conditions. Scholars across sociology and public health have discussed the racialized distribution of occupational risk for infection and mortality during the COVID-19 pandemic, but these analyses draw on national data or data from specific sites (Laster Pirtle 2020; McClure et al. 2020). Other studies have more broadly examined increasing racial segregation and labor fragmentation at higher levels using ACS data, but they have not been able to look at the overlaps of such variables at a high spatial resolution (Lichter, Parisi, and Taquino 2012). Many studies have examined the evolving family complexity in the United States using ACS data at a national or state level (Bloome 2017; Maralani 2013), but these studies increasingly take a longitudinal perspective to examine evolving cohabitation patterns and socioeconomic implications, including racialized contours (Carlson and Corcoran 2001; Williams, Simon, and Cardwell 2019). Still, empirical research is needed on the geographic distribution of these patterns. In terms of the COVID-19 pandemic, essential workers experience much higher risk that intersects with the lack of social safety nets related to unemployment and childcare in the United States. It is important to consider where parents living with children are exposed to the compound risks for infection at work, unemployment (with associated loss of employer-provided insurance, which may cover dependents), and increased need for childcare. As noted earlier, these types of compound risks are highly racialized, especially as they intersect with entrenched systems of urban racial segregation.

Our small-area estimates in Figure 3 illustrate the stark geographic concentration of this compound risk in Philadelphia, stratified by race/ethnicity. Essential workers living with children in Philadelphia are overwhelmingly non-Hispanic Black, and extremely segregated from similar populations racialized as non-Hispanic white or Hispanic. During the COVID-19 pandemic, these estimates could be used to guide policy around the availability of testing sites and the delivery of social safety net programs to neighborhoods most at risk for infection, loss of employment, and unmet childcare needs. From a theoretical perspective, sociologists can use these local data

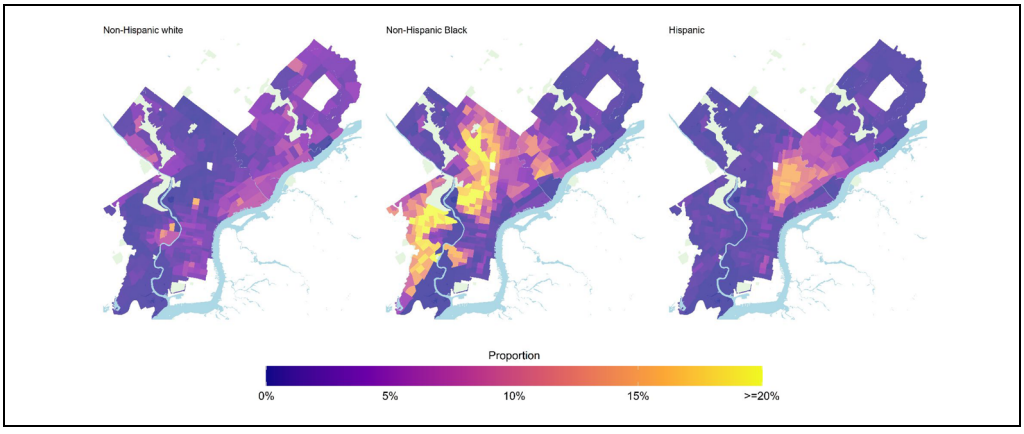

Figure 4. Small-area estimates of the proportion of the census-tract population (2012–2016) living below the federal poverty threshold and in households that spend more than 50 percent of monthly income on rent or owner costs, stratified by race/ethnicity.

across multiple ACS years to understand patterns of change within and between neighborhoods, comparing across geographies in the United States and reflecting on nuance that is often lost in discourse around national or state-level statistics.

Example 2: The Unequal Burden of Housing

Second, we estimate the proportion of the census-tract population (2012–2016) living below the federal poverty threshold and in households that spend more than 50 percent of monthly income on rent or owner costs, stratified by race/ethnicity (Figure 4). This level is often characterized as extreme housing cost, and households in this category are at high risk for eviction, particularly very low income households. As the cost of housing has exponentially increased across the United States, sociological research has sought to understand its patterns, determinants, and consequences, especially as it intersects with racial segregation and racialized housing markets in urban areas (Faber 2020; Howell and Korver-Glenn 2020; Sewell 2016). Contemporary research has applied relational perspectives to the intersection of poverty and rental markets. For example, Desmond and Wilmers (2019) use national data to ask whether the poor pay more for housing, demonstrating that landlords often derive greater profits from raising rents in poorer neighborhoods where property values and tax burdens are lower. In terms of the costs of home ownership, often seen as a tool for social mobility and building wealth, Taylor (2019) examined how, following the Housing and Urban Development Act of 1968, predatory subprime housing markets led to disproportionately high housing costs for Black homeowners. Scholars have drawn connections between this history of exploitative housing practice and policy and the contemporary COVID-19 pandemic; individuals living with extreme housing costs are far less equipped to absorb shocks related to unemployment or acute health events, and the

lifting of eviction moratoriums during the pandemic has contributed to surges in homelessness, infection, and mortality (Benfer et al. 2021).

Figure 4 illustrates the geographic concentration of extreme housing costs across Philadelphia. Relatively few individuals racialized as non-Hispanic white live in households with incomes below the poverty threshold and spending more than 50 percent of monthly income on rent or owner costs. However, rates are very high for individuals racialized as non-Hispanic Black, representing more than 20 percent of the population in many neighborhoods, particularly in the west and north sections of the city.

DISCUSSION

In this study, we introduced TBSM, a technique that uses decision trees to automatically select and bin population constraints for use within SM models. We demonstrated that our implementation of TBSM using public data from the ACS provides a reliable, scalable small-area estimation strategy that leverages the full information contained in the largest survey of social, economic, and demographic data in the United States. We tested the model across a diverse set of counties using an out-of-sample validation strategy, indicating that TBSM produces census tract–level estimates that are more accurate than estimates produced by a standard SM approach. We then applied this technique to estimate more complex cross-tabulated summaries that are not available in public census tables, including local indicators related to the racialized distribution of risk in labor and housing. These indicators are often studied at a higher geographic level by sociologists, and granular nuance is important and timely for considering the local causes and consequences of the inequitable spread of COVID-19 throughout the country. Particularly in our example of Philadelphia, we saw extreme racialized segregation along these lines. Researchers often want to analyze these types of policy-relevant indicators that have high spatial resolution and high attribute resolution, but with the publicly available census data, they are forced to choose one at the expense of the other. Here we demonstrated how we can estimate two such indicators using only publicly available data, but our estimation framework can be easily applied to any combination of variables collected in the ACS microdata.

Our approach is not without important limitations. We have improved on several key obstacles identified in the implementation of SM techniques more broadly in sociological study (issues related to computation and data synthesis, selection of constraint variables and appropriate binning), but our validation results suggest that researchers should exercise caution in applying this method to very rural counties and perhaps consider combining data across nearby counties (or combining similar census tracts). Still, validation tests, particularly in metropolitan areas and suburbs, show how TBSM can significantly improve on the performance of more conventional (and arbitrary) methods of constraint selection, leading to more accurate small-area estimates of unobserved indicators. This will be important for applying sociological theories as evidence mounts of very local patterns of extreme segregation in U.S. cities, which

until now have been examined mainly using national data sets or disparate city-specific efforts.

Future Directions

We noted earlier that the target variable can be any variable (continuous or binary) that can be defined for either household or person PUMS records. That is, the target variable must be a function of the "raw" PUMS variables (of which there are many). The examples presented here are straightforward, constructing the target variable as a fairly simple combination of other variables. However, one could construct a target variable defined by a more complex combination of the PUMS variables. This opens the possibility of using other (non-ACS) surveys to create the target-defining function. For example, the ACS is of no direct use if we wish to estimate household gasoline consumption; that information is not solicited by the ACS questionnaire. The National Household Travel Survey (NHTS), on the other hand, does report respondent gasoline consumption along with a set of household-level characteristics. However, as a much smaller survey, the NHTS cannot provide reliable estimates for small areas. If there is sufficient overlap between NHTS household characteristics and those in the PUMS, one can fit a model to the NHTS to estimate gasoline consumption for PUMS household records (Ummel 2016). This quantity becomes the target metric for subsequent small-area estimates using the technique described here. In this way, the application of our technique—and the range of target variables eligible for small-area estimation—can be greatly expanded.

CONCLUSIONS

Among sociologists, demographers, economists, and other scholars studying the persistence and widening of inequality in the United States, spatial contours have taken on central importance. The increasingly fractal spatial dimensions of social life in the United States require rigorous small-area estimation strategies to answer high-dimensional, policy-relevant research questions at a local level while maintaining confidentiality in the underlying data. These local estimates based on publicly available census tables can inform a research agenda focused on spatial equity and open up a variety of possibilities for linkage with other public and private data sources that are increasingly geocoded.

Acknowledgments

We would like to thank Xi Song for comments on previous drafts.

Funding

N.G. and D.A.C. were supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development Training Grant (T32-HD-007242-36A1), as well as funding from the University of Pennsylvania's Socio-Spatial Climate Collaborative, Population Studies Center, Kleinman Center for Policy Research, and Office of the Vice Provost for Research. K.U. and D.A.C. were supported through a Quartet Pilot Research Award and funded by the Eunice Shriver Kennedy National Institute of Child

Health and Development (Population Research Infrastructure Program) Population Studies Center NICHD P2C (HD044964) at the University of Pennsylvania. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Pennsylvania or National Institutes of Health.

ORCID iD

Nick Graetz https://orcid.org/0000-0002-4362-2059

References

American Civil Liberties Union. 2020. "Data Show COVID-19 Is Hitting Essential Workers and People of Color Hardest." ACLU Massachusetts. Retrieved November 8, 2021. https://www.aclum.org/en/publications/data-show-covid-19-hitting-essential-workers-and-people-color-hardest.

Anderson, Ben. 2007. "Creating Small Area Income Estimates for England: Spatial Microsimulation Modelling." Chimera Working Paper No. 2007-07. Essex, UK: University of Essex.

Ballas, Dimitris, Graham Clarke, Danny Dorling, and David Rossiter. 2007. "Using SimBritain to Model the Geographical Impact of National Government Policies." *Geographical Analysis* 39(1):44–77.

Bar-Gera, Hillel, Karthik Charan Konduri, Bhargava Sana, Xin Ye, and Ram M. Pendyala. 2009. "Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods." Paper presented at the 88th annual meeting of the Transportation Research Board, Washington, DC.

Beckman, Richard J., Keith A. Baggerly, and Michael D. McKay. 1996. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice* 30(6):415–29.

Benfer, Emily A., D. Vlahov, Marissa Y. Long, Evan Walker-Wells, J. L. Poteenger, Gregg Gonsalves, and Danya E. Keene. 2021. "Eviction, Health Inequity, and the Spread of COVID-19: Housing Policy as a Primary Pandemic Mitigation Strategy." *Journal of Urban Health* 98(1):1–12.

Birkin, M., and M. Clarke. 1988. "Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples." *Environment & Planning A* 20(12):1645–71.

Bloome, Deirdre. 2017. "Childhood Family Structure and Intergenerational Income Mobility in the United States." *Demography* 54(2):541–69.

Breiman, Leo. 1993. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall.

Cagney, Kathleen A., Christopher R. Browning, James Iveniuk, and Ned English. 2014. "The Onset of Depression during the Great Recession: Foreclosure and Older Adult Mental Health." *American Journal of Public Health* 104(3):498–505.

Campbell, Malcolm, and Dimitris Ballas. 2013. "A Spatial Microsimulation Approach to Economic Policy Analysis in Scotland." *Regional Science Policy and Practice* 5(3):263–88.

Carlson, Marcia J., and Mary E. Corcoran. 2001. "Family Structure and Children's Behavioral and Cognitive Outcomes." *Journal of Marriage and Family* 63(3):779–92.

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." NBER Working Paper No. w25147. Cambridge, MA: National Bureau of Economic Research.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The Association between Income and Life Expectancy in the United States, 2001–2014." *JAMA* 315(16):1750–66.

Das, Sumonkanti, Azizur Rahman, Ashraf Ahamed, and Sabbir Tahmidur Rahman. 2019. "Multi-level Models Can Benefit from Minimizing Higher-Order Variations: An Illustration Using Child Malnutrition Data." *Journal of Statistical Computation and Simulation* 89(6):1090–1110.

Desmond, Matthew, and Nathan Wilmers. 2019. "Do the Poor Pay More for Housing? Exploitation, Profit, and Risk in Rental Markets." *American Journal of Sociology* 124(4):1090–1124.

Faber, Jacob W. 2020. "We Built This: Consequences of New Deal Era Intervention in America's Racial Geography." *American Sociological Review* 85(5):739–75.

Howell, Junia, and Elizabeth Korver-Glenn. 2021. "The Increasing Effect of Neighborhood Racial Composition on Housing Values, 1980–2015." *Social Problems* 68(4):1051–71.

Huang, Zengyi, and Paul Williamson. 2002. "A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small Area Microdata Contents." Working paper, Department of Geography, University of Liverpool.

Ifesemen, Onosi Sylvia, Thomas Bestwick-Stevenson, and Kimberly L. Edwards. 2019. "Spatial Microsimulation of Osteoarthritis Prevalence at the Small Area Level in England—Constraint Selection for a 2-Stage Microsimulation Process." *International Journal of Microsimulation* 12(2): 37–51.

Izrailev, Sergei. 2016. "binr: Cut Numeric Values into Evenly Distributed Groups (Bins)." R package version 1.1. Retrieved November 8, 2021. https://rdrr.io/cran/binr/man/bins.html.

Laster Pirtle, Whitney N. 2020. "Racial Capitalism: A Fundamental Cause of Novel Coronavirus (COVID-19) Pandemic Inequities in the United States." *Health Education and Behavior* 47(4): 504–508.

Lichter, Daniel T., Domenico Parisi, and Michael C. Taquino. 2012. "The Geography of Exclusion." *Social Problems* 59(3):364–88.

Lovelace, Robin, Dimitris Ballas, and Matt Watson. 2014. "A Spatial Microsimulation Approach for the Analysis of Commuter Patterns: From Individual to Regional Levels." *Journal of Transport Geography* 34:282–96.

Maralani, Vida. 2013. "The Demography of Social Mobility: Black-White Differences in the Process of Educational Reproduction." *American Journal of Sociology* 118(6):1509–58.

McClure, Elizabeth S., Pavithra Vasudevan, Zinzi Bailey, Snehal Patel, and Whitney R. Robinson. 2020. "Racial Capitalism within Public Health: How Occupational Settings Drive Covid-19 Disparities." *American Journal of Epidemiology* 189(11):1244–53.

Morrissey, Karyn, and Cathal O'Donoghue. 2011. "The Spatial Distribution of Labour Force Participation & Market Earnings at the Sub-national Level in Ireland." *Review of Economic Analysis* 3(1):80–101.

Muller, Kirill. 2017. "A Generalized Approach to Population Synthesis." Publication No. 23514. Doctoral dissertation, ETH Zurich, ETH Zurich Research Collection.

O'Donoghue, Cathal, Karyn Morrissey, and John Lennon. 2014. "Spatial Microsimulation Modelling: A Review of Applications and Methodological Choices." *Microsimulation Association International Journal of Microsimulation* 7(1):26–75.

Orcutt, Guy H. 1957. "A New Type of Socio-economic System." *Review of Economics & Statistics* 39(2): 116–23.

Orcutt, Guy H. 2007. "A New Type of Socio-economic System." *International Journal of Microsimulation* 1(1):3–9.

Orcutt, Guy H., Martin Greenberger, John Korbel, and Alice M. Rivlin. 1961. *Microanalysis of Socioeconomic Systems: A Simulation Study*. New York: Harper.

Rahman, Azizur. 2017. "Estimating Small Area Health-Related Characteristics of Populations: A Methodological Review." *Geospatial Health* 12(1).

Rahman, Azizur, and Ann Harding. 2016. *Small Area Estimation and Microsimulation Modeling*. Boca Raton, FL: CRC Press.

Rahman, Azizur, Ann Harding, Robert Tanton, and Shuangzhe Liu. 2013. "Simulating the Characteristics of Populations at the Small Area Level: New Validation Techniques for a Spatial Microsimulation Model in Australia." *Computational Statistics and Data Analysis* 57(1):149–65.

Rubin, Donald. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Sakshaug, Joseph W., and Trivellore E. Raghunathan. 2014. "Generating Synthetic Data to Produce Public-Use Microdata for Small Geographic Areas Based on Complex Sample Survey Data with Application to the National Health Interview Survey." *Journal of Applied Statistics* 41(10):2103–22.

Sampson, Robert J. 2008. "Moving to Inequality: Neighborhood Effects and Experiments Meet Social Structure." *American Journal of Sociology* 114(1):189–231.

Sampson, Robert J. 2012. *Great American City: Chicago and the Enduring Neighborhood Effect*. Chicago: University of Chicago Press.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: CRC Press.

Sewell, Abigail A. 2016. "The Racism-Race Reification Process." *Sociology of Race and Ethnicity* 2(4): 402–32.

Sharkey, Patrick. 2013. *Stuck in Place: Urban Neighborhoods and the End of Progress toward Racial Equality*. Chicago: University of Chicago Press.

Sharkey, Patrick, and Jacob Faber. 2014. "Where, When, Why, and for Whom Do Residential Contexts Matter? Moving Away from the Dichotomous Understanding of Neighborhood Effects." *Annual Review of Sociology* 40(1):559–79.

Smith, Dianna M., Graham P. Clarke, and Kirk Harland. 2009. "Improving the Synthetic Data Generation Process in Spatial Microsimulation Models." *Environment and Planning A: Economy and Space* 41(5): 1251–68.

Smith, Dianna M., Jamie R. Pearce, and Kirk Harland. 2011. "Can a Deterministic Spatial Microsimulation Model Provide Reliable Small-Area Estimates of Health Behaviours? An Example of Smoking Prevalence in New Zealand." *Health and Place* 17(2):618–24.

Tanton, Robert, Yogi Vidyattama, Binod Nepal, and Justine McNamara. 2011. "Small Area Estimation Using a Reweighting Algorithm." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(4):931–51.

Taylor, Keeanga-Yamahtta. 2019. *Race for Profit: Black Homeownership and the End of the Urban Crisis*. Chapel Hill: University of North Carolina Press.

Therneau, Terry, and Beth Atkinson. 2019. "rpart: Recursive Partitioning and Regression Trees." R package version 4.1-15. Retrieved November 8, 2021. https://cran.r-project.org/web/packages/rpart/index.html.

Tomintz, Melanie N., Graham P. Clarke, and Janette E. Rigby. 2008. "The Geography of Smoking in Leeds: Estimating Individual Smoking Rates and the Implications for the Location of Stop Smoking Services." *Area* 40(3):341–53.

Ummel, Kevin. 2016. "Impact of CCL's Proposed Carbon Fee and Dividend Policy: A High-Resolution Analysis of the Financial Effect on U.S. Households." Working Paper. International Institute for Applied Systems Analysis. Retrieved November 8, 2021. https://citizensclimatelobby.org/wp-content/uploads/2016/02/Household-Impact-Study-Ummel.pdf.

USDA ERS (U.S. Department of Agriculture Economic Research Service). 2015. "County Typology Codes." Retrieved November 8, 2021. https://www.ers.usda.gov/data-products/county-typology-codes/.

Voas, David, and Paul Williamson. 2000. "An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata." *International Journal of Population Geography* 6(5):349–66.

Wager, L. Wesley. 1962. "Reviewed Work(s): *Microanalysis of Socioeconomic Systems: A Simulation Study*, by Guy H. Orcutt, Martin Greenberger, John Korbel and Alice M. Rivlin." *American Sociological Review* 27(3):15–16.

Whitworth, Alex. Forthcoming. "SynthACS: Spatial MicroSimulation Modeling with Synthetic American Community Survey Data." *Journal of Statistical Software*.

Whitworth, A., E. Carter, D. Ballas, and G. Moon. 2017. "Estimating Uncertainty in Spatial Microsimulation Approaches to Small Area Estimation: A New Approach to Solving an Old Problem." *Computers, Environment and Urban Systems* 63:50–57.

Williams, Deadric T., Laura Simon, and Marissa Cardwell. 2019. "Black Intimacies Matter: The Role of Family Status, Gender, and Cumulative Risk on Relationship Quality among Black Parents." *Journal of African American Studies* 23(1–2):1–17.

Williamson, P., M. Birkin, and P. H. Rees. 1998. "The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records." *Environment and Planning A: Economy and Space* 30(5):785–816.

Wilson, A. G., and C. E. Pownall. 1976. "A New Representation of the Urban System for Modelling and for the Study of Micro-level Interdependence." *Area* 8(4):246–54.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Boca Raton, FL: CRC Press.

Zagheni, Emilio. 2015. "Microsimulation in Demographic Research." Pp. 9780–85 in *International Encyclopedia of the Social and Behavioral Sciences*, 2nd ed., Vol. 15, edited by J. D. Wright. Oxford, UK: Elsevier.

Zhang, Xingyou, James B. Holt, Hua Lu, Anne G. Wheaton, Earl S. Ford, Kurt J. Greenlund, and Janet B. Croft. 2014. "Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System." *American Journal of Epidemiology* 179(8):1025–33.

Author Biographies

Nick Graetz is a postdoctoral research associate in the Department of Sociology at Princeton University. His research focuses on using quantitative methods to answer relational questions about the role of place in producing racialized inequality across the life-course, including the long-run consequences of racist housing policies on outcomes related to household wealth and population health. His methodological interests include spatial statistics, causal inference, and mediation/decomposition analysis.

Kevin Ummel is a data scientist and environmental economist whose expertise includes spatial microsimulation, machine learning and predictive modeling, and geospatial analysis. He previously worked on the Decent Living Energy Project as a research scholar at the International Institute for Applied Systems Analysis. As a consultant to the Citizens' Climate Lobby, he authored a national Household Impact Study to quantify the effect of carbon tax policy on individual households and created a carbon fee and dividend calculator to disseminate results to the public. As a senior associate at the Center for Global Development, he published on high-resolution carbon footprinting and spatiotemporal optimization of renewable energy systems and cocreated the CARMA global power plant database.

Daniel Aldana Cohen is an assistant professor of sociology at the University of California, Berkeley, where he directs the Socio-Spatial Climate Collaborative, or (SC)2. He also codirects the Climate and Community Project. He is a CIFAR Azrieli Global Scholar for 2021 to 2023. In 2018 and 2019, he was a member of the Institute for Advanced Study in Princeton, New Jersey. He works on the politics of climate change, investigating the intersections of climate change, housing, political economy, social movements, and inequalities of race and class in the United States and Brazil. As director of (SC)2, he is leading qualitative and quantitative research projects on whole-community climate mapping, green political economy, and eco-apartheid.