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[1,2]

[1] University of Pennsylvania, Population Studies Center
[2] University of Pennsylvania, Socio-Spatial Climate Collaborative
[3] Greenspace Analytics, Inc.

Abstract The American Community Survey (ACS) is the largest household survey in the United States and indispensable for detailed analysis of specific places and populations. This paper introduces a technique to produce "small area" (e.g. census tract) estimates of any person- or household-level phenomenon that can be derived from ACS microdata variables. This is demonstrated by producing novel, tract-level estimates of 1) excess housing capacity, 2) prevalence of traditional living arrangements, and 3) household energy burden. We combine conventional spatial microsimulation techniques with binary-split decision trees to efficiently select local population margins from a large set of candidates. The result is place-specific microdata samples that are calibrated to match an information-rich set of known constraints (e.g. number of households by income group). A validation exercise indicates agreement between model output and known values (mean $R^2 = 0.78$). We conclude by discussing 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

Funding NG was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development Training Grant (T32-HD-007242-36A1). KU and DAC 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.

Corresponding Author Nick Graetz, Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, Email: ngraetz@sas.upenn.edu. Tel: 215-898-6441.

Electronic copy available at: https://ssrn.com/abstract=3574679

The American Community Survey (ACS) is the largest survey of U.S. households and provides extensive information on household demographics, finances, employment, health insurance, migration, ancestry, linguistics, housing conditions, and more. Given its uniquely large sample size, the ACS is indispensable for analyses that demand "high-resolution" data; e.g. examining patterns across neighborhoods, specific sub-populations, or both.

High-resolution research questions in the United States often go unanswered at smaller geographic levels because the necessary information is not tabulated or made public. However, the increasingly fractal nature of social and economic phenomena means that researchers and policy-makers are often interested in local estimation. And there is a need for more sophisticated data for social impacts, and especially social drivers, of ecological phenomena like climate change, at a time of growing concern about "eco-apartheid." (Cohen 2018; Wachsmuth, Cohen, and Angelo 2016). 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 solution has been through targeted, independent primary data collection projects in specific urban contexts (Sampson 2012). However, this decentralized, un-harmonized system of data collection and study of spatial phenomena does not provide a comprehensive national picture of how place increasingly structures social, health, and environmental stratification in the United States. Certainly, this system is missing huge populations that do not live in places like New York City or Chicago.

The ACS obtains survey responses from over 2 million housing units nationwide every year, but confidentiality concerns limit what data is made public. The Census Bureau publishes two kinds of data obtained by the ACS: 1) *Summary tables* report estimates (usually household or

Electronic copy available at: https://ssrn.com/abstract=3574679

person counts) for an array of common variables (e.g. income, age) and standard geographic entities. 2) The *Public Use Microdata Sample* (PUMS) contains de-identified microdata that provide the full range of questionnaire responses, but respondent location is not disclosed in detail. Summary tables provide greater *spatial resolution* at the expense of *attribute resolution*, while the PUMS provides the opposite. A consequence of this trade-off is a *de facto* narrowing of research questions (Sharkey and Faber 2014). In practice, analysis of ACS data is typically limited to questions that require *either* spatial or attribute resolution. The complexity of social life and policy, however, often demands both. More sophisticated types of analysis -- e.g. detailed cross-tabulations, correlation or regression analysis, third-party data fusion -- typically require (or benefit from) greater attribute resolution. The PUMS is amenable to these techniques, but it does not provide sufficient geographic detail to allow coincident analysis of spatial patterns. In principle, this limitation can be sidestepped by the statistical creation of microdata samples for individual "small areas", thereby providing a data product with *both* spatial and attribute resolution. Sakshaug and Raghunathan (2014) show how the Census Bureau might internally generate "synthetic" small area microdata samples that address confidentiality concerns (Sakshaug and Raghunathan 2014). However, this is an unlikely path forward given significant resource constraints facing the Census Bureau.

Alternatively, small area microdata samples can be created using existing, public ACS data in conjunction with spatial microsimulation techniques (Tanton 2014). Spatial microsimulation involves reweighting available microdata observations (e.g. PUMS) to create new, place-specific samples that replicate known, local population totals (e.g. the number of households with income below $10,000). The resulting small area samples allow for more complex analyses while maintaining high spatial resolution. Despite the centrality of the ACS to social science research

Electronic copy available at: https://ssrn.com/abstract=3574679

in the U.S., use of the ACS in spatial microsimulation applications is comparatively rare in the published literature (certainly compared to European counterparts). This paper combines conventional spatial microsimulation with decision trees to enable endogenous selection and "attribute binning" of population margin variables from a potentially large set of candidates. Binning of variable attributes (e.g. income categories) reduces the number of population totals that must be calibrated to, while preserving predictive information with respect to the "target variable" to be estimated for individual small areas. Importantly, the methodology and associated ACS-tailored code base allow for maximum flexibility: providing an automated process to estimate any potential ACS-derived variable for any small area across the United States.

We validate our methods by producing model estimates of four target variables for individual census tracts and comparing the results to published estimates from the Census Bureau. We then demonstrate three examples of real-world applications by producing (previously unpublished) tract-level estimates of excess housing capacity, the prevalence of traditional living arrangements, and household energy burden. We conclude with a discussion of outstanding methodological issues and how the technique can be extended to estimate target variables derived from non-ACS surveys of U.S. households or individuals.

Previous Research

Background on spatial microsimulation

Most research and policy questions requiring high spatial resolution ultimately seek to estimate a specific quantity for a specific place, a task generally known as "small area estimation" (SAE) (Rahman and Harding 2019). A number of techniques have been developed to address SAE,

Electronic copy available at: https://ssrn.com/abstract=3574679

generally classifiable as either *spatial microsimulation* (typically deterministic) or *statistical regression-based* approaches (probabilistic) (Whitworth et al. 2017).

As the impact of public policies, private investment, and residential segregation become increasingly fractal, researchers and policy-makers have increasingly recognized that modern spatial microsimulation (SM) can provide an important research tool (O'Donoghue, Morrissey, and Lennon 2014). SM techniques center on the ability to reweight or "calibrate" available microdata to match known small-area population totals ("small-area constraints") to produce place-specific microdata samples that are (to a degree) representative of the local population (Deville, Sarndal, and Sautory 1993). In other words, SM creates synthetic micro-populations for individual small areas that best match a specified set of known, population margins for the local population. Because SM produces microdata samples with maximal attribute resolution, it is possible to use individual-level covariates in predicting unobserved variables. This is in contrast to regression-based approaches, which generally regress a place-specific variable of interest (e.g. county-level smoking rates) on place-specific predictor variables (e.g. county-level mean income), and sometimes use the resulting model to predict outcomes at a higher spatial resolution (e.g. county-level) if predictor variables are available (Dwyer-Lindgren et al. 2014).

There are several important decisions that must be made in applying spatial microsimulation techniques (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,
4) and the validation of estimates.

Electronic copy available at: https://ssrn.com/abstract=3574679

To date, methodological research has largely focused on 2) in comparing the relative (dis)advantages of the three predominant calibration methods: combinatorial optimization (CO), generalized regression reweighting (GREG), and iterative proportional fitting (IPF) (Whitworth et al. 2017). For the purposes of our analysis, the choice of calibration technique is not of primary concern and we do not provide any comparison of methods (see Hermes and Poulsen 2012 for a comprehensive comparison) (Hermes and Poulsen 2012). Our methodological focus is 3), which has received far less attention; i.e. the sensitivity of estimates to choice of constraint variables and how one might optimally select these variables (Huang and Williamson 2001; Smith, Clarke, and Harland 2009).

Small-Area Estimation in the United States

Despite the substantive advantages of spatial microsimulation for SAE and the advent of faster, more efficient computational platforms, modern SM techniques are not applied broadly in the study of American communities. Applied research using SM has focused broadly on demography, poverty, health, transportation, public policy (Ballas et al. 2013; Rahman and Harding 2019); however, virtually all applied studies have focused on contexts outside of the United States. For example, synthetic microdata generated with an IPF approach has been used to calculate individual smoking rates across small areas in New Zealand and the UK (Smith, Pearce, and Harland 2011; Tomintz, Clarke, and Rigby 2008).

Present study

Our analysis contributes to the spatial microsimulation literature, and applied SAE more broadly in the United States, in three key ways. First, we demonstrate how modern SM techniques can be

Electronic copy available at: https://ssrn.com/abstract=3574679

applied nationally across the United States by using publicly-available data from the American Community Survey (ACS), the largest source of sociodemographic data available. There has been extremely limited research on applying SM methods to the ACS; Koh *et al.* (2015) and Level *et al.* (2014) applied SM to estimate smoking in New Bedford county (Massachusetts) and obesity in Wayne county (Michigan) (Koh, Grady, and Vojnovic 2015; Levy, Fabian, and Peters 2014). Second and related, 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. Third, we expand on SM methodology by introducing a novel decision tree-based method for optimally selecting and binning marginal constraint variables. We combine our tree-based selection method with a rigorous set of out-of-sample cross-validation tests to demonstrate how small-area estimates can be obtained across the United States using only publicly-available ACS datasets.

Spatial microsimulation: An illustrative example

Our focus is the use of spatial microsimulation for small area estimation (SAE). In this section, we use a simple example to illustrate the conventional approach and its challenges. The example is then used to illustrate the concepts underpinning our extension of the method to include binary-split decision trees.

It is useful to think of the SAE task as requiring two pieces of information about a specific locale: 1) the *frequency* of different population subgroups and 2) the *propensity* of those subgroups with respect to a "target variable" (outcome of interest) to be estimated. Consider a simple case: We wish to estimate the average years of schooling among adults in a town. Two variables -- occupation and wage level -- are observable for the population. The two necessary pieces of information are illustrated with matrices **A** and **B** (Table 1). **A** gives the number of

Electronic copy available at: https://ssrn.com/abstract=3574679

people in each subgroup (joint frequency); **B** gives the average years of schooling for each subgroup (propensity). In this case, the target variable represented in **B** is continuous in nature (years of schooling), but the logic that follows can be extended to binary outcomes (e.g. probability that a person has a college degree).

[Insert Table 1 here]

When **A** and **B** are known, the average years of schooling among the local population is a weighted mean given by the sum of the Hadamard product divided by the total population: $\Sigma(A \circ B) \div \Sigma(A)$. In this case, the result is 15.7 years.

However, most of the necessary data is unavailable in real-world cases. It is not unusual to only know the "margins" of **A** (i.e. the row and column sums). That is, we are ignorant of the joint frequency distribution of **A** and entirely ignorant of B. Table 2 shows the extent of what might be known in a real-world case, especially for small-area tabulated datasets.

[Insert Table 2 here]

In this situation, microdata observations sampled from a larger population (preferably inclusive of the small area) can be used to estimate the necessary-but-missing information. The microdata must include the variables in **A** ("margin variables") – or variables that can be manipulated to match them – along with any target variable(s). For example, we might have access to state- or national-level microdata resembling the following:

[Insert Table 3 here]

Electronic copy available at: https://ssrn.com/abstract=3574679

The general task of SAE is to use the available information in Tables 3 and 4 to "fill-in" **A** and wholly estimate **B** (usually implicitly). One way to do this is via microsimulation. Note that regression-based SAE techniques implicitly pursue the same goal by different means. For example, the multi-level mixed-effects regression and post-stratification (MRP) technique popular in political science is, in effect, a method for estimating **B** (Kastellec, Lax, and Phillips 2016). Similar mixed-effects models are frequently used in demography and epidemiology (Dwyer-Lindgren et al. 2014, 2016, 2017). However, these methods require that **A** is observed, limiting it to a specific subset of SAE applications. The sections that follow address the more general (and more difficult) case where only the margins of **A** are observed, such is the case with the tabulated ACS small-area datasets.

Given the information in Tables 2 and 3, any number of microsimulation calibration techniques (e.g. IPF, CO, GREG) can be employed. The goal of calibration is to find new microdata observation weights that produce a "small area sample" with aggregate margins close to the known margins in **A**. The calibrated sample effectively provides a guess as to the joint distribution of **A** (given the observed margins) and, further, an estimate of **B**. The latter is made possible by the joint observation of margin and target variables in the microdata.

Table 4 provides an example of frequency and propensity matrices – **A*** and **B***, respectively – that might be deduced from a successful calibration of the microdata (i.e. creation of a "local sample"). At this point, all of the data needed to calculate the small area estimate are available. Mean years of schooling in the town is estimated to be 16.4 years.

[Insert Table 4 here]

Electronic copy available at: https://ssrn.com/abstract=3574679

This simple example is instructive because it makes clear what spatial microsimulation seeks to do as well as a core challenge it faces. The values of **A*** and **B*** – and, hence, our final small area estimate – depend on the observation weights returned by the calibration step. However, there are potentially many "re-weightings" of the microdata that can replicate the known margins; we do not know which of these weightings provides an accurate SAE overall.

For example, Table 4 shows that **A*** accurately approximates the known margins in **A**. For practical purposes, we define "successful" calibration as one that produces local sample margins that are within some tolerance of the known margins (see below for more details). However, comparing the "true" **A** and **B** (Figures 1 and 2) with the simulated **A*** and **B*** make clear that successful calibration, while *necessary*, is not *sufficient* to guarantee reliable results. Indeed, the small area estimate derived from the calibrated sample (16.4 years) is quite different from the true value (15.7 years). This is unavoidable to an extent because it is rooted in the paucity of local information with which spatial microsimulation (or any other SAE technique) must grapple. The problem is especially pronounced when the microdata population differs from the local population in some fundamental way. Calibration techniques typically attempt to find the set of weights that satisfy the known margins while minimizing deviance from the initial weights. When the populations are substantially different (which may be difficult to determine), even well-calibrated samples can misrepresent **A** and **B**.

The primary remedy available to practitioners is to increase the number of margin variables (i.e. "constrained" variables) used in the calibration step. In our example, we might consider adding variables for gender and age. Additional variables reduce the number of potential weighting schemes that can satisfy the known population margins, effectively shrinking the

Electronic copy available at: https://ssrn.com/abstract=3574679

solution space and increasing confidence in the joint distribution (**A***) implied by any single (successfully) calibrated sample. In addition, while it is possible that "high-wage scientists" in the larger population could exhibit different educational outcomes than local "high-wage scientists", it is *less likely* that "high-wage, male scientists age 45-55" fundamentally differ between the two populations. Consequently, the addition of margin variables also increases confidence in **B***, provided that the additional margin variables are predictive of the target variable.

However, there are at least three practical challenges to using a larger number of margin variables in a spatial microsimulation exercise:

1) The choice of variables may be limited by data availability. This is not usually a significant problem when working with censuses and large household surveys like the ACS. As shown above, the ACS provides a rich set of potential margin variables.
2) The likelihood of successful calibration tends to decline with the number of margins (i.e. constraints) that must be replicated in the local sample. Adding margin constraints only increases confidence in the joint distribution (**A***) if the sample actually adheres to those constraints.
3) It may be unclear how to select from among candidate margin variables in the first place. There is an obvious preference for variables that are predictive of the target. However, given (2) and the likelihood of multicollinearity and interaction effects among candidate variables, the variable selection task is decidedly non-trivial. Worse still, adding variables that are *not* predictive of the target can actually reduce SAE quality, since it makes calibration more difficult without offering greater confidence in **B***.

Electronic copy available at: https://ssrn.com/abstract=3574679

The variable selection process -- encompassing challenges (2) and (3) -- is generally opaque in published spatial microsimulation studies. The selection process might be informed by analytical/correlational criteria, but more common is reliance on "expert judgement", convention, or an *ad hoc* process of trial and error (Huang and Williamson 2001; O'Donoghue et al. 2014; Smith et al. 2009). The issue of variable selection has received far less attention than that of calibration techniques. Yet, it is the interplay of the two that ultimately matters for SAE quality. To our knowledge, margin variables are always specified in advance of sample calibration and used uniformly across all spatial regions.

Attribute binning via decision tree

Consideration of challenges (2) and (3) leads us to two important observations: First, it is not the number of *variables* (e.g. occupation) that degrades sample calibration so much as the number of *individual margins* (Sales, Construction, etc.). Second, not all individual margins for a given variable are equally important for prediction of the target (i.e. income groups, levels of educational attainment).

The technique introduced below exploits these facts to provide a generic approach for variable selection in microsimulation studies. This is accomplished by reducing the attribute resolution of margin variables while seeking to minimize the loss of predictive ability with respect to the target. The goal is to "compress" the margins *data* (i.e. reduce their number) in an effort to increase the likelihood of successful calibration, while preserving useful *information* with respect to prediction. This is accomplished via selective "binning" of margin variable attributes.

Electronic copy available at: https://ssrn.com/abstract=3574679

To demonstrate this visually, recall our original example consisting of two variables (occupation and wage level) and seven individual margins. Looking at **B**, we notice that rows 1 and 2 are quite similar; that is, there is little difference in average years of schooling between individuals working in Sales and Construction. Looking at variation across the columns (wage level), we notice it is possible to break **B** into four distinct rectangles or "bins" (Table 5).

[Insert Table 5 here]

Figure 2 suggests that we might combine low-wage and medium-wage individuals into a single category and further distinguish between those employed in Sales and Construction, on the one hand, and those in employed in Management and Science on the other. Doing so reduces the number of margins that must be calibrated from seven to four.

A preferred binning strategy is one where the individuals (microdata observations) assigned to a bin exhibit little variation with respect to the target variable; this allows the number of margins to be reduced with minimal effect on SAE. From the perspective of SAE, we care less about distinguishing among individuals that exhibit little or no variation in outcomes. Imagine that high-wage workers in Sales or Construction *always* have 16 years of schooling. In that (extreme) case, we care only that the calibrated sample contain the correct total number of such individuals; further distinguishing among these individuals on the basis of occupation is irrelevant to the calculation of mean years of schooling for the population.

A binning strategy with these general characteristics can be deduced from a binary-split decision tree fitted to available microdata. A decision tree consists of a series of recursive, binary splits of available predictor variables, where successive "nodes" exhibit increasing uniformity

Electronic copy available at: https://ssrn.com/abstract=3574679

with respect to a response variable (Breiman 1993). They are a kind of predictive model with the added advantage (for our purposes) that analysis of the "split decisions" across a tree's nodes implies a strategy for binning the attributes of predictor variables.

Below is the decision tree associated with the binning strategy identified in Figure 6, fitted using the *rpart* package in the *R* language (Therneau and Atkinson 2019). The "leaf" nodes at the bottom show how mean years of schooling differs across the four bins.

[Insert Figure 1 here]

A decision tree "grown" without constraint will continue to split the predictor (i.e. margin) variables until each leaf node exhibits zero variation, typically resulting in no binning and no reduction in the number of margins. Varying the stopping criteria produces trees of varying complexity, with less complex trees generally yielding more aggressive binning strategies and fewer margins. Consequently, the extent to which the attribute resolution of the margin variables is compressed depends on the size and complexity of the associated decision tree. In general, our objective is to find the minimum amount of compression (i.e. maximum attribute resolution) that is needed to ensure successful calibration.

Iterative approach to "tree-binned" calibration

Identification of a potential binning strategy requires fitting a decision tree to microdata. However, we do not know if the initial microdata sample is representative of the local population; indeed, the assumption is that it is not. A binning strategy derived from unrepresentative microdata may result in inefficient attribute compression (i.e. information loss).

Electronic copy available at: https://ssrn.com/abstract=3574679

We address this by employing an iterative process of calibration and binning that attempts to evolve the observation weights toward a preferable solution.

The algorithm that follows is applied to each small area independently. It begins with an initial microdata sample (S) containing *n* households, associated observation weights W, and target variable Y drawn from a larger population as well as a set of *m* known, categorical candidate margins (M) for a specific small area (i.e. the data inputs are analogous to Figures 3 and 4). From this information it is possible to create a *n* x *m* "dummy" matrix (D) indicating whether a given observation is a member of a given margin. The basic "tree-binned" algorithm is:

<u>Part 1</u>

1a. Fit a decision of tree (T) predicting Y with complexity α using microdata S and observation weights W.

1b. Deduce a margins binning strategy (B) from the split decisions of T.

1c. Use B to generate "binned" versions of M and D (M1 and D1, respectively)

1d. Calibrate new weights (W1) using M1, D1, and initial weights W

1e. IF calibration error is less than tolerance ε

THEN proceed to Part 2

ELSE reduce α and repeat steps 1a-1d

<u>Part 2</u>

2a. Compute the target value ($\hat{Y}$) predicted by T for each observation in S

2b. Compute the small area estimate ($\hat{X}$) using $\hat{Y}$ and W1

2c. IF $\hat{X}$ is practically unchanged from previous iterations

Electronic copy available at: https://ssrn.com/abstract=3574679

THEN stop algorithm

ELSE set W = W1 and proceed to 1a

Part 1 is an iterative procedure to identify the maximum-complexity decision tree that results in successful calibration. Calibration is successful when the local sample margins are within some tolerance (ε) of the known margins. We make use of the fact that ACS summary table variables are accompanied by a standard error, allowing for a "soft" calibration step (1d) that uses the mean absolute Z-score across margins to assess calibration quality (Davies 2018). Calibration itself is performed using the "logit" generalized regression estimator as implemented in the *grake* package (Muller 2017). The algorithm as written above assumes that α is (slowly) reduced until this occurs. In practice, we employ a variation on the bisection search method to more efficiently identify a near-maximum α.

A binning strategy is *deduced from a decision tree* (1b) by analyzing the binary split decisions across the tree for each margin variable. The splits within a variable are not necessarily mutually exclusive. In Figure 3, imagine that the tree instead split Wages on the right-hand branch so as to group "Medium" and "High" wage individuals together. In this case, the tree's "deduced" binning strategy would only bin the Occupation variable; the Wages variable would retain maximum attribute resolution.

Part 2 calculates a local (small area) estimate ($\hat{X}$) for the target variable using the successfully calibrated weights (W1) and associated decision tree (T) from Part 1. The algorithm's overall stopping criterion is the convergence of $\hat{X}$ on a final estimate. That is, Part 1 is repeated using the latest calibration weights (W1) until calibration error is less than tolerance ε. This implies that

Electronic copy available at: https://ssrn.com/abstract=3574679

the local sample, decision tree, and binning strategy all become more realistic with each iteration as the weights evolve.

The predicted $\hat{Y}$ is T's "leaf node" value for each observation, analogous to the leaf node values exhibited in Figure 3. Use of $\hat{Y}$ instead of Y for the SAE calculation mitigates the possibility that the calibration step could unduly weight an unusual observation, biasing $\hat{X}$. In effect, we treat all observations assigned to the same leaf node as identical with respect to the target variable. The calibrated weights W1 seek to provide a best estimate of total node weight, but intra-node variation in weights is immaterial to the calculation of $\hat{X}$.

A final complication concerns the construction of S and D. In practice, we wish to use both household- and person-level margin variables (e.g. household size and an individual's race) as this maximizes the amount of available information. This necessitates that S consist of person-level observations nested within households, with household attributes replicated for members of the same household. And it requires that D (a household-level matrix) contain 0/1 for household-level columns and the applicable number of household members for person-level columns. This is similar to the strategy used by Bar-Gera et al. (2009) (also see Section 4.3.1 of Muller 2017) (Bar-Gera et al. 2009; Muller 2017). The advantage is that the code base can handle any potential margin variable or target variable, whether household- or person-level in nature.

Beyond the identification of binning strategies, decision trees offer two additional advantages for our purposes:

1) A tree's hierarchical nature provides implicit variable selection. The margin variables most influential in predicting the target variable are found in splits "higher up" the tree. As tree size is constrained (i.e. binning becomes more aggressive), the retained splits are

Electronic copy available at: https://ssrn.com/abstract=3574679

necessarily those involving the most influential variables. Consequently, employing a large set of candidate variables does not degrade SAE quality; the only penalty is computational.

2) Trees provide a measure of relative margin variable importance. The improvement in prediction attributable to each variable's splits are summed and compared across the tree, providing a measure of relative importance that can be used in the calibration process. Specifically, our measure of calibration quality is a weighted mean using variable-specific importance weights from the decision tree. This puts greater emphasis on accurately replicating the margins of highly-predictive/influential variables.

Data and validation examples

Although the methodology described above is applicable to any data source that meets the input requirements, we utilize the ACS exclusively for our application. The validation and demonstration outputs we describe below use *only* data sourced from the 2012-2016 (5-year) ACS. These data take two forms: categorical margin variables and microdata. Recall that the Census Bureau publishes two kinds of data obtained by the ACS:

1) *Summary tables* report estimates (usually household or person counts) and margins of error calculated by the Census Bureau. For example, Table B19001 reports the estimated number of households in each of 16 income categories ranging from "Less than $10,000" to "$200,000 or more". Tables are published for standard geographic entities, ranging from block groups and census tracts to counties and states.
2) The *Public Use Microdata Sample* (PUMS) contains de-identified microdata that provide the full range of responses collected from ACS questionnaires. For example, the PUMS

Electronic copy available at: https://ssrn.com/abstract=3574679

reports each respondent household's income as a numeric value (e.g. $45,500) rather than a categorical range. To protect anonymity, respondent location is disclosed only at the level of Public Use Microdata Areas. So-called PUMA's are relatively large geographic entities, each containing at least 100,000 residents.

The margin variables utilize by our model contain household or person counts, by block group, for specific categories and are sourced from ACS summary tables. They correspond to object M in the algorithm pseudocode. The Census Bureau publishes hundreds of summary tables, providing a large potential set of candidate variables. For this exercise, we constructed 21 candidate margins variables, summarized in Appendix Table 1. For example, the "education" margin variable is constructed from ACS summary table B15002. It provides the number of persons in each of 8 educational attainment categories (e.g. "HS graduate") that are ordinal in nature (i.e. the categories have a natural ordering). In principle, our technique places no upper limit on the number of candidate margin variables that can be considered; this is simply our attempt at a reasonably large set covering a range of socio-economic and dwelling characteristics. In general, it is beneficial to include any margin variable that might be predictive of the target. The tree-binning process determines whether (and how) to utilize the variable. Since the margin variables are observed for individual block groups (over 200,000 nationally), we can use these data to perform spatial microsimulation for any geographic unit ("small area" or otherwise) that is an aggregation of block groups.

The microdata inputs come from the Census Bureau PUMS files, which consist of de-identified microdata containing the full range of responses collected from ACS questionnaires. The geographic location of each PUMS observation is identified by a specific Public Use

Electronic copy available at: https://ssrn.com/abstract=3574679

Microdata Area (PUMA). The raw PUMS is processed to create a person-level dataset exhibiting concordance with the margin variables in Table 1. An example is shown in Appendix Table 2, which reports microdata for three households in Philadelphia across four of the margin variables. Note that Appendix Table 2 corresponds to object S in the algorithm pseudocode with the exception that S is assumed to include a column with a "target variable" of interest (Y). Importantly, Y can be any variable (continuous or binary) that can be defined for either household or person PUMS records. Since the number of raw variables in the PUMS is very large, there is significant flexibility with respect to construction of the target variable.

Validation strategy

A challenge of microsimulation -- and SAE 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 to "known" small area estimates derived information in ACS summary tables. We selected four target variables to use for validation: mean years of schooling, percent of the population with public health insurance, percent of the population that is non-Hispanic white, and mean hours worked per week (Appendix Table 3). These include both continuous (numerical) and discrete (binary) variables; the small area estimate is a mean value for the former; a population proportion for the latter.

A validation exercise should test model performance using assumptions and data inputs similar to those in likely for legitimately "unknown" target variables. Consequently, we exclude candidate margin variables that might give a model artificially-high predictive ability ("Excluded variables" column of Appendix Table 3). For example, model estimates of "Mean years schooling" are nearly perfect when the "education" margin variable is utilized (as expected, since

Electronic copy available at: https://ssrn.com/abstract=3574679

they are derived from the same source data). But this is not indicative of model performance for real-world cases where the target variable is not necessarily highly-correlated with a margin variable.

Two study areas are used for validation: 1) Gwinnett County, GA, a suburban county northeast of Atlanta; 2) Philadelphia County, PA, a highly-diverse county encompassing Philadelphia and surrounding inner suburbs. Tree-binned, spatial microsimulation model results are compared to summary table values at census tract level. There are 113 populated census tracts in Gwinnett County and 377 in Philadelphia County; the median tract population is 7,000 and 4,000, respectively.

This use of census tracts (which are quite small; only block groups offer higher spatial resolution) and the exclusion of highly-correlated margin variables means that our validation exercise provides a stiff test of our model. SAE quality generally improves as the geographic unit of analysis and number of candidate margin variables increase in size. Consequently, our tract-level validation results should be interpreted as lower-bounds on model performance for most real-world applications. Descriptions of all validation test statistics can be found in Appendix Section 1.

Validation results

Overall, the validation exercise suggests that the tree-binned spatial microsimulation approach generates estimates in broad agreement with known values (Appendix Figures 1-8). The mean model value-added is 0.72 across the eight case studies (range 0.42 to 0.94), which is encouraging given the stiff test imposed by our validation strategy. One advantage of a tree-

Electronic copy available at: https://ssrn.com/abstract=3574679

based approach is ability to calculate relative variable importance weights. We discussed above how this information is used to weight the calibration process, but it also has post-analysis diagnostic value. Figure 2 summarizes variable importance for the case of mean years schooling in Gwinnett County. The boxplots show the variation in variable importance across census tracts for each margin variable.

[Insert Figure 2 here]

Two features of Figure 2 are particularly notable. First, there is considerable variance in relative importance for any given variable. This is due to the tree-binning algorithm performing tract-specific selection and binning of the candidates; i.e. which margin variables matter (and how much they matter) varies across space. This differs significantly from the conventional approach of selecting a fixed set of variables for the entire study area in advance of sample calibration. Second, the ranking of the variables does not necessarily conform to "expert judgement". We suspect that few practitioners would be inclined to include "yrbuilt", "health_insurance", "hhsize", or "commute_time" in a spatial microsimulation model of educational attainment. Yet all of these exhibit strong importance.

Out-of-sample demonstration

We provide three examples of tree-binned spatial microsimulation being used to estimate unobserved phenomena at high spatial resolution. Our demonstrations make use of the considerable household- and person-level detail in the PUMS to construct unique target variables that are not available in small-area summary tables. In short, any variable that can be calculated for either household or person PUMS records is eligible for small-area estimation.

Electronic copy available at: https://ssrn.com/abstract=3574679

Example 1: Excess housing capacity

Our "excess housing capacity" variable is constructed by analyzing the age, sex, and relationship of a household's members to determine the number of bedrooms needed to accommodate everyone. We define the necessary minimum using the UK government's bedroom entitlement rules for public housing assistance (Entitledto 2019). The rules are:

- Couples receive their own a bedroom
- Non-coupled individuals age 16+ receive their own bedroom
- Two children age 0-9 can share a bedroom whatever their sex
- Two children age 0-15 can share a bedroom if they are the same sex

The detailed nature of the PUMS allows us to apply these rules to every household in the sample. The difference between the actual and necessary number of bedrooms is the "excess" (possibly negative). Our model is then used to predict the mean excess number of bedrooms for each census tract. Figure 3 shows the results for Gwinnett and Philadelphia counties.

[Insert Figure 3 here]

The results identify specific places (orange-yellow) where we estimate there is a glut of empty/unnecessary bedrooms. The model suggests that the planned community of Peachtree Corners, GA (northwest corner of Gwinnett County) has, on average, more than two excess bedrooms per household -- an incredible feat of superfluity. Conversely, the lowest-valued tracts (dark blue) are associated with negative excess capacity; the number of bedrooms available to households in these areas is, on average, lower than the necessary minimum as defined above.

Electronic copy available at: https://ssrn.com/abstract=3574679

Notably, Peachtree Corners is directly adjacent to the area around Norcross, GA where we estimate there is the greatest "bedroom crunch" in the county.

Example 2: Prevalence of "traditional" household structure

Next, we estimate the percent of children living in households with a "traditional" family structure. The detail of the PUMS records again allows us to create a target variable that is highly-specific. We define "traditional" families as those with the following characteristics:

- A heterosexual married couple with (only) biological or adopted children
- Possibly including the parents/in-laws of the married couple (i.e. grandma and grandpa)
- Only one of the couple is in the labor force (i.e. single-earner household)
- All adults have been married just once in their lifetime

By classifying each PUMS household as "traditional" or not along these lines, we can then create a binary variable assigned to each *child* in the sample (1 if in a traditional household; 0 otherwise) that serves as the target variable. We do not claim that this a universal or even preferable definition of "traditional" household structure. Rather, the point is that such structure can be defined in *any way* allowed by PUMS household- and person-record variables. Figure 4 shows the census-tract model estimates.

[Insert Figure 4 here]

The results show that a distinct minority (< 30%) of children in either county live in "traditional" households, as defined above. In large parts of West and North Philadelphia, our estimates suggest there are practically *no* children (< 5%) living in such households. The spatial patterns

Electronic copy available at: https://ssrn.com/abstract=3574679

indicate that socio-economics alone cannot explain the phenomenon. For example, parts of northwest and northeast Philadelphia exhibit similar prevalence of traditional household structure, despite the former having higher education levels (see Appendix Figure 2).

Example 3: Household energy burden

Finally, we estimate a measure of the typical "energy burden" (Drehobl and Ross 2016) experienced by households in each census tract. The energy burden is defined as the ratio of total energy expenditures (electricity, natural gas, and heating oil) to household income. The PUMS contains self-reported expenditures for either the most recent month (electricity and gas) or year (heating oil). Given the temporal/seasonal variability and likelihood of reporting error, this measure is inherently noisy across households. We address this by exploiting our ability to estimate the *median* energy burden across households, which is more robust to outliers than the mean. Figure 5 shows the census-tract model estimates.

[Insert Figure 5 here]

The results suggest that energy burden is considerably more problematic for households in parts of Philadelphia compared to those in Gwinnett County. Some research suggests that a burden of 6% is the upper threshold for "affordable" energy (Fisher, Sheehan and Colton 2013). Large areas of north and west Philadelphia exhibit median energy burdens in excess of 6%, implying that a majority of households in these areas are in some state of "energy insecurity" (Hernandez and Bird 2010; Hernandez 2013). While we have chosen to display the median estimate here, the inherent flexibility of our technique also allows for estimation of the percent of households above or below a given threshold.

Electronic copy available at: https://ssrn.com/abstract=3574679

Discussion

In this study, we demonstrate 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, the American Community Survey. Using the examples of Gwinnett County, GA, (a suburban county northeast of Atlanta) and Philadelphia County, PA (a highly-diverse county encompassing Philadelphia and surrounding inner suburbs), we show how a tree-based spatial microsimulation approach can accurately predict unobserved ACS summaries at the census-tract level by comparing to published Census tables. We then apply this technique to estimate more complex cross-tabulated summaries that are not available in public Census tables, including timely local environmental indicators characterizing high-consumption neighborhoods (mean excess bedrooms) and high-burden neighborhoods (energy expenditure as a percent of household income). Particularly in Philadelphia, we observe extreme residential segregation along these lines. Immense energy burden is very concentrated in the communities of North Philadelphia, while wealthier households in northeastern suburban neighborhoods are much more likely to contain excess bedrooms, accounting for family size. Researchers often want to analyze these types of policy-relevant indicators that are high spatial resolution *and* high attribute resolution, but in the publicly-available Census data are forced to choose one at the expense of the other. Here we choose to demonstrate how we can estimate three such indicators using only publicly-available data, but our estimation framework can be easily applied to any combination of variables collected in the ACS.

Future directions

Electronic copy available at: https://ssrn.com/abstract=3574679

We noted above 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. For example, "excess housing capacity" is a function of the number, sex, age, and relation of household members according to a set of rules. However, it is possible to 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 small 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. This quantity becomes the target metric for subsequent small areas estimates using the technique described here. In this way, the application of our technique – and the range of target variable 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, its spatial contours have taken on a central research importance (Chetty et al. 2018). The increasingly fractal spatial dimensions of social life in the

Electronic copy available at: https://ssrn.com/abstract=3574679

United States requires investment in 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 geo-coded.

Electronic copy available at: https://ssrn.com/abstract=3574679

References

Ballas, Dimitris, Graham Clarke, Stephen Hynes, John Lennon, Karyn Morrissey, and Cathal O'Donoghue. 2013. "A Review of Microsimulation for Policy Analysis." Pp. 35–54 in.

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."

Bazuin, Joshua Theodore and James Curtis Fraser. 2013. "How the ACS Gets It Wrong: The Story of the American Community Survey and a Small, Inner City Neighborhood." *Applied Geography* 45:292–302.

Breiman, Leo. 1993. *Classification and Regression Trees*. 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.

Chetty, Raj, Harvard University, Nber N. John Friedman, Nathaniel Hendren, John Abowd, Peter Bergman, David Deming, Edward Glaeser, David Grusky, Lawrence Katz, Enrico Moretti, Robert Sampson, Caroline Dockes, Michael Droste, Benjamin Goldman, Jack Hoyle, Federico Gonzalez Rodriguez, Jamie Gracie, Matthew Jacob, Martin Koenen, Sarah Merchant, Donato Onorato, Kamelia Stavreva, Wilbur Townsend, and Joseph Winkelmann. 2018. *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility* *.

Cohen, Daniel Aldana. 2018. "Water Crisis and Eco-Apartheid in São Paulo: Beyond Naive Optimism About Climate-Linked Disasters." *International Journal of Urban and Regional Research*, November.

Davies, Gareth. 2018. "A Thesis Submitted for the Degree of Doctor of Philosophy Examination of Approaches to Calibration in Survey Sampling." (March).

Deville, Jean-Claude, Carl-Erik Sarndal, and Olivier Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88(423):1013.

Drehobl, Ariel and Lauren Ross. 2016. *Lifting the High Energy Burden in America's Largest Cities: How Energy Efficiency Can Improve Low Income and Underserved Communities*.

Dwyer-Lindgren, Laura, Amelia Bertozzi-Villa, Rebecca W. Stubbs, Chloe Morozoff, Michael J. Kutz, Chantal Huynh, Ryan M. Barber, Katya A. Shackelford, Johan P. Mackenbach, Frank J. van Lenthe, Abraham D. Flaxman, Mohsen Naghavi, Ali H. Mokdad, and Christopher J. L. Murray. 2016. "US County-Level Trends in Mortality Rates for Major Causes of Death, 1980-2014." *JAMA* 316(22):2385.

Dwyer-Lindgren, Laura, Amelia Bertozzi-Villa, Rebecca W. Stubbs, Chloe Morozoff, Johan P. Mackenbach, Frank J. van Lenthe, Ali H. Mokdad, and Christopher J. L. Murray. 2017. "Inequalities in Life Expectancy Among US Counties, 1980 to 2014." *JAMA Internal Medicine* 177(7):1003.

Dwyer-Lindgren, Laura, Ali H. Mokdad, Tanja Srebotnjak, Abraham D. Flaxman, Gillian M. Hansen, and Christopher JL Murray. 2014. "Cigarette Smoking Prevalence in US Counties: 1996-2012." *Population Health Metrics* 12(1):5.

Electronic copy available at: https://ssrn.com/abstract=3574679

Entitledto. 2019. “Your Bedroom Entitlement.” Retrieved June 19, 2019 (https://www.entitledto.co.uk/help/Calculating-Your-Bedroom-Entitlement).

Hermes, Kerstin and Michael Poulsen. 2012. “A Review of Current Methods to Generate Synthetic Spatial Microdata Using Reweighting and Future Directions.” *Computers, Environment and Urban Systems* 36(4):281–90.

Huang, Zengyi and Paul Williamson. 2001. *A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small Area Microdata Contents*.

Kastellec, Jonathan P., Jeffrey R. Lax, and Justin Phillips. 2016. “Estimating State Public Opinion With Multi-Level Regression and Poststratification Using R.” *Unpublished Manuscript* 13.

Koh, Keumseok, Sue C. Grady, and Igor Vojnovic. 2015. “Using Simulated Data to Investigate the Spatial Patterns of Obesity Prevalence at the Census Tract Level in Metropolitan Detroit.” *Applied Geography* 62:19–28.

Levy, Jonathan I., Maria Patricia Fabian, and Junenette L. Peters. 2014. “Community-Wide Health Risk Assessment Using Geographically Resolved Demographic Data: A Synthetic Population Approach.” *PLoS ONE* 9(1).

Morley, S. K., T. V. Brito, and D. T. Welling. 2018. “Measures of Model Performance Based On the Log Accuracy Ratio.” *Space Weather* 16(1):69–88.

Muller, Kirill. 2017. “A Generalized Approach to Population Synthesis.” *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.

Rahman, Azizur and Ann Harding. 2019. *Small Area Estimation and Microsimulation Modelling*.

Sakshaug, Joseph W. and Trivellore Raghunathan. 2014. “Generating Synthetic Microdata to Estimate Small Area Statistics in the American Community Survey.” *Statistics in Transition* 15(3):341–68.

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

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

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.

Electronic copy available at: https://ssrn.com/abstract=3574679

Spielman, Seth E. and David C. Folch. 2015. “Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization” edited by A. R. Hernandez Montoya. *PLOS ONE* 10(2):e0115626.

Spielman, Seth E., David Folch, and Nicholas Nagle. 2014. “Patterns and Causes of Uncertainty in the American Community Survey.” *Applied Geography* 46:147–157.

Tanton, Robert. 2014. “A Review Os Spatial Microsimulation Methods.” *International Journal of Microsimulation* 7:4–25.

Therneau, Terry and Beth Atkinson. 2019. “Rpart: Recursive Partitioning and Regression Trees.”

Tofallis, Chris. 2015. “A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation.” *Journal of the Operational Research Society* 66(8):1352–62.

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.” 40(3):341–53.

Wachsmuth, David, Daniel Aldana Cohen, and Hillary Angelo. 2016. “Expand the Frontiers of Urban Sustainability.” *Nature* 536(7618):391–93.

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.

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.”

Bazuin, Joshua Theodore and James Curtis Fraser. 2013. “How the ACS Gets It Wrong: The Story of the American Community Survey and a Small, Inner City Neighborhood.” *Applied Geography* 45:292–302.

Breiman, Leo. 1993. *Classification and Regression Trees*. Chapman & Hall.

Chetty, Raj, Harvard University, Nber N. John Friedman, Nathaniel Hendren, John Abowd, Peter Bergman, David Deming, Edward Glaeser, David Grusky, Lawrence Katz, Enrico Moretti, Robert Sampson, Caroline Dockes, Michael Droste, Benjamin Goldman, Jack Hoyle, Federico Gonzalez Rodriguez, Jamie Gracie, Matthew Jacob, Martin Koenen, Sarah Merchant, Donato Onorato, Kamelia Stavreva, Wilbur Townsend, and Joseph Winkelmann. 2018. *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility* *.

Cohen, Daniel Aldana. 2018. “Water Crisis and Eco-Apartheid in São Paulo: Beyond Naive Optimism About Climate-Linked Disasters.” *International Journal of Urban and Regional Research*, November.

Davies, Gareth. 2018. “A Thesis Submitted for the Degree of Doctor of Philosophy Examination of Approaches to Calibration in Survey Sampling.” (March).

Deville, Jean-Claude, Carl-Erik Sarndal, and Olivier Sautory. 1993. “Generalized Raking Procedures in Survey Sampling.” *Journal of the American Statistical Association* 88(423):1013.

Drehobl, Ariel and Lauren Ross. 2016. *Lifting the High Energy Burden in America’s Largest*

Electronic copy available at: https://ssrn.com/abstract=3574679

Cities: How Energy Efficiency Can Improve Low Income and Underserved Communities.

Dwyer-Lindgren, Laura, Amelia Bertozzi-Villa, Rebecca W. Stubbs, Chloe Morozoff, Michael J. Kutz, Chantal Huynh, Ryan M. Barber, Katya A. Shackelford, Johan P. Mackenbach, Frank J. van Lenthe, Abraham D. Flaxman, Mohsen Naghavi, Ali H. Mokdad, and Christopher J. L. Murray. 2016. "US County-Level Trends in Mortality Rates for Major Causes of Death, 1980-2014." *JAMA* 316(22):2385.

Dwyer-Lindgren, Laura, Amelia Bertozzi-Villa, Rebecca W. Stubbs, Chloe Morozoff, Johan P. Mackenbach, Frank J. van Lenthe, Ali H. Mokdad, and Christopher J. L. Murray. 2017. "Inequalities in Life Expectancy Among US Counties, 1980 to 2014." *JAMA Internal Medicine* 177(7):1003.

Dwyer-Lindgren, Laura, Ali H. Mokdad, Tanja Srebotnjak, Abraham D. Flaxman, Gillian M. Hansen, and Christopher JL Murray. 2014. "Cigarette Smoking Prevalence in US Counties: 1996-2012." *Population Health Metrics* 12(1):5.

Entitledto. 2019. "Your Bedroom Entitlement." Retrieved June 19, 2019 (https://www.entitledto.co.uk/help/Calculating-Your-Bedroom-Entitlement).

Hermes, Kerstin and Michael Poulsen. 2012. "A Review of Current Methods to Generate Synthetic Spatial Microdata Using Reweighting and Future Directions." *Computers, Environment and Urban Systems* 36(4):281–90.

Huang, Zengyi and Paul Williamson. 2001. *A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small Area Microdata Contents*.

Kastellec, Jonathan P., Jeffrey R. Lax, and Justin Phillips. 2016. "Estimating State Public Opinion With Multi-Level Regression and Poststratification Using R." *Unpublished Manuscript* 13.

Koh, Keumseok, Sue C. Grady, and Igor Vojnovic. 2015. "Using Simulated Data to Investigate the Spatial Patterns of Obesity Prevalence at the Census Tract Level in Metropolitan Detroit." *Applied Geography* 62:19–28.

Levy, Jonathan I., Maria Patricia Fabian, and Junenette L. Peters. 2014. "Community-Wide Health Risk Assessment Using Geographically Resolved Demographic Data: A Synthetic Population Approach." *PLoS ONE* 9(1).

Morley, S. K., T. V. Brito, and D. T. Welling. 2018. "Measures of Model Performance Based On the Log Accuracy Ratio." *Space Weather* 16(1):69–88.

Muller, Kirill. 2017. "A Generalized Approach to Population Synthesis." *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.

Rahman, Azizur and Ann Harding. 2019. *Small Area Estimation and Microsimulation Modelling*.

Sakshaug, Joseph W. and Trivellore Raghunathan. 2014. "Generating Synthetic Microdata to Estimate Small Area Statistics in the American Community Survey." *Statistics in Transition* 15(3):341–68.

Electronic copy available at: https://ssrn.com/abstract=3574679

Sampson, Robert J. 2012. *Great American City : Chicago and the Enduring Neighborhood Effect*. The 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.”

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.

Spielman, Seth E. and David C. Folch. 2015. “Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization” edited by A. R. Hernandez Montoya. *PLOS ONE* 10(2):e0115626.

Spielman, Seth E., David Folch, and Nicholas Nagle. 2014. “Patterns and Causes of Uncertainty in the American Community Survey.” *Applied Geography* 46:147–157.

Tanton, Robert. 2014. “A Review Os Spatial Microsimulation Methods.” *International Journal of Microsimulation* 7:4–25.

Therneau, Terry and Beth Atkinson. 2019. “Rpart: Recursive Partitioning and Regression Trees.”

Tofallis, Chris. 2015. “A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation.” *Journal of the Operational Research Society* 66(8):1352–62.

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.” 40(3):341–53.

Wachsmuth, David, Daniel Aldana Cohen, and Hillary Angelo. 2016. “Expand the Frontiers of Urban Sustainability.” *Nature* 536(7618):391–93.

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.

Electronic copy available at: https://ssrn.com/abstract=3574679

Tables and Figures

Table 1. The top panel shows frequency matrix “A” (number of individuals, by occupation and wage level) and the bottom panel shows propensity matrix “B” (average years of schooling, by occupation and wage level).

	Low wages	Medium wages	High wages
Sales	25	40	15
Construction	35	15	10
Management	15	25	30
Science	10	15	25

	Low wages	Medium wages	High wages
Sales	13	14	16
Construction	12	13	16
Management	16	17	19
Science	17	18	20

Electronic copy available at: https://ssrn.com/abstract=3574679

Table 2. Typical extent of knowledge for frequency matrix.

	Low wages	Medium wages	High wages	**TOTAL**
Sales	?	?	?	**80**
Construction	?	?	?	**60**
Management	?	?	?	**70**
Science	?	?	?	**50**
TOTAL	**85**	**95**	**80**	**260**

Electronic copy available at: https://ssrn.com/abstract=3574679

Table 3. Example of individual microdata observations sampled from larger population.

Observation ID	**Observation weight**	**Wage level**	**Occupation**	**Years schooling**
1	15	Medium	Management	16
2	10	High	Construction	14
3	16	High	Science	21
...	...	...	...	...
N	12	Low	Sales	12

Electronic copy available at: https://ssrn.com/abstract=3574679

Table 4. **A*** (frequency) and **B*** (propensity), based on a calibrated local sample.

A*	Low wages	Medium wages	High wages	**TOTAL**
Sales	27	28	24	**79**
Construction	14	32	15	**61**
Management	27	19	24	**70**
Science	17	16	17	**50**
TOTAL	**85**	**95**	**80**	**260**

B*	Low wages	Medium wages	High wages
Sales	14	15	17
Construction	12	14	16
Management	17	16	21
Science	16	19	21

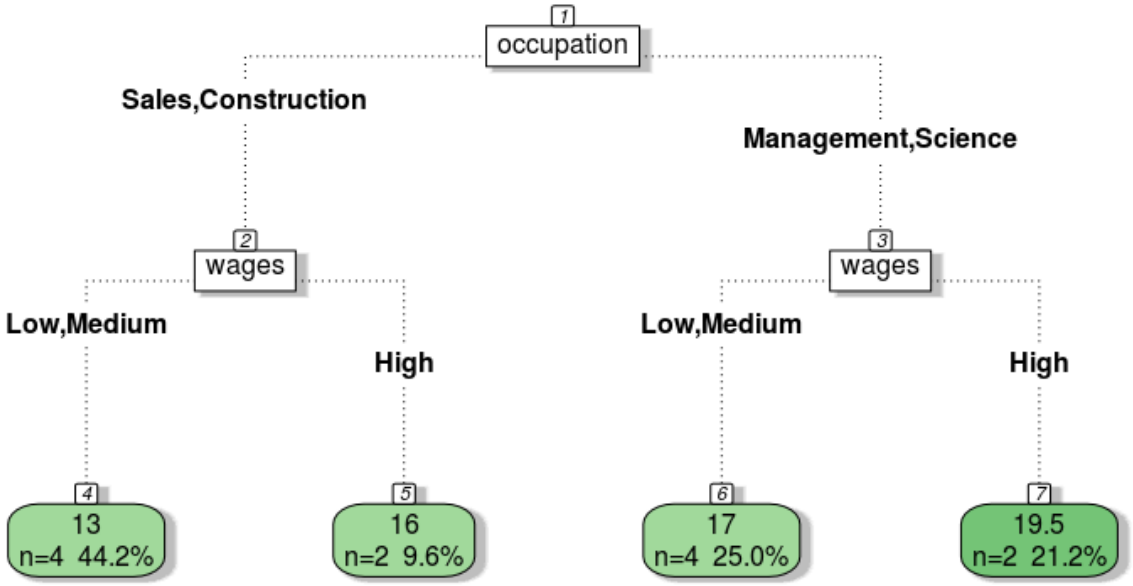
Electronic copy available at: https://ssrn.com/abstract=3574679

Table 5. A possible margin binning strategy visualized using the propensity matrix.

	Low wages	Medium wages	High wages
Sales	13	14	16
Construction	12	13	16
Management	16	17	19
Science	17	18	20

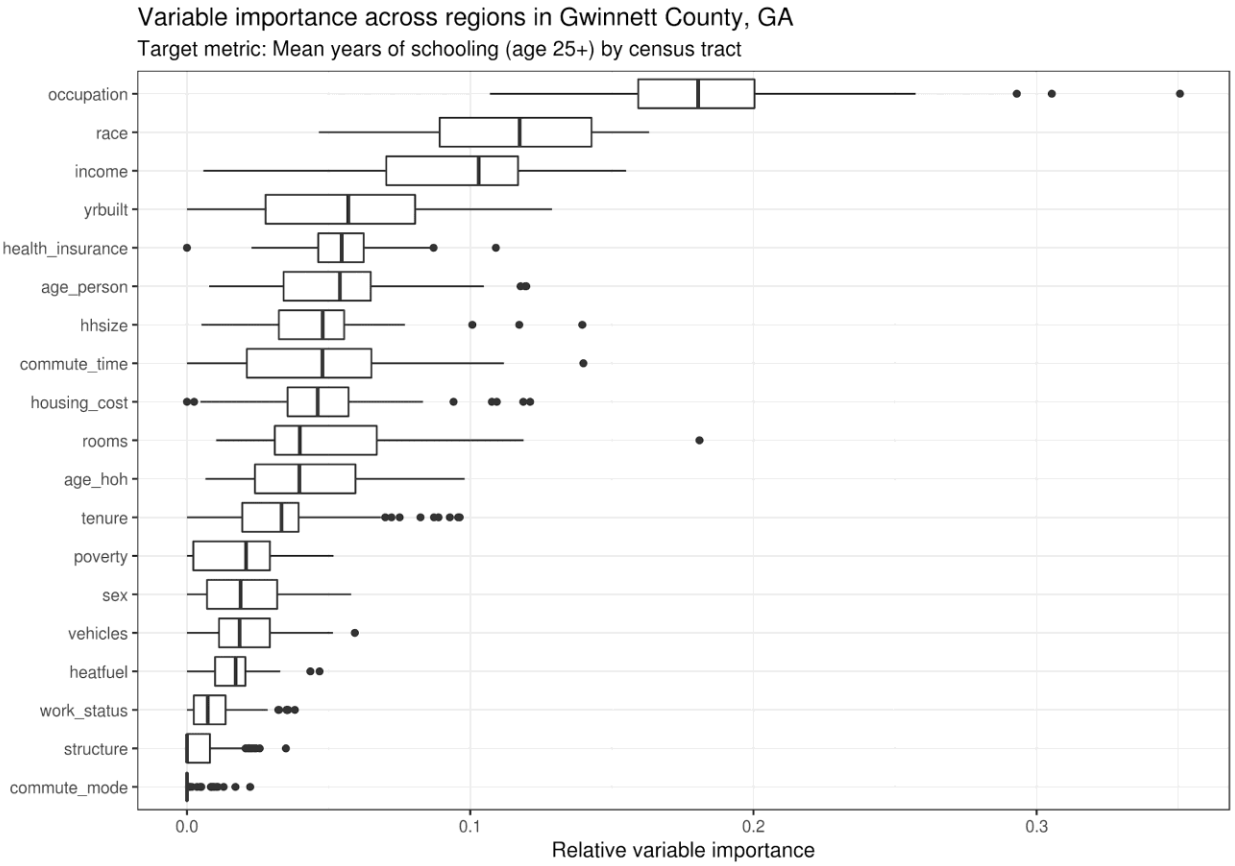
Electronic copy available at: https://ssrn.com/abstract=3574679

Figure 1. Example of decision tree used to deduce a binning strategy.



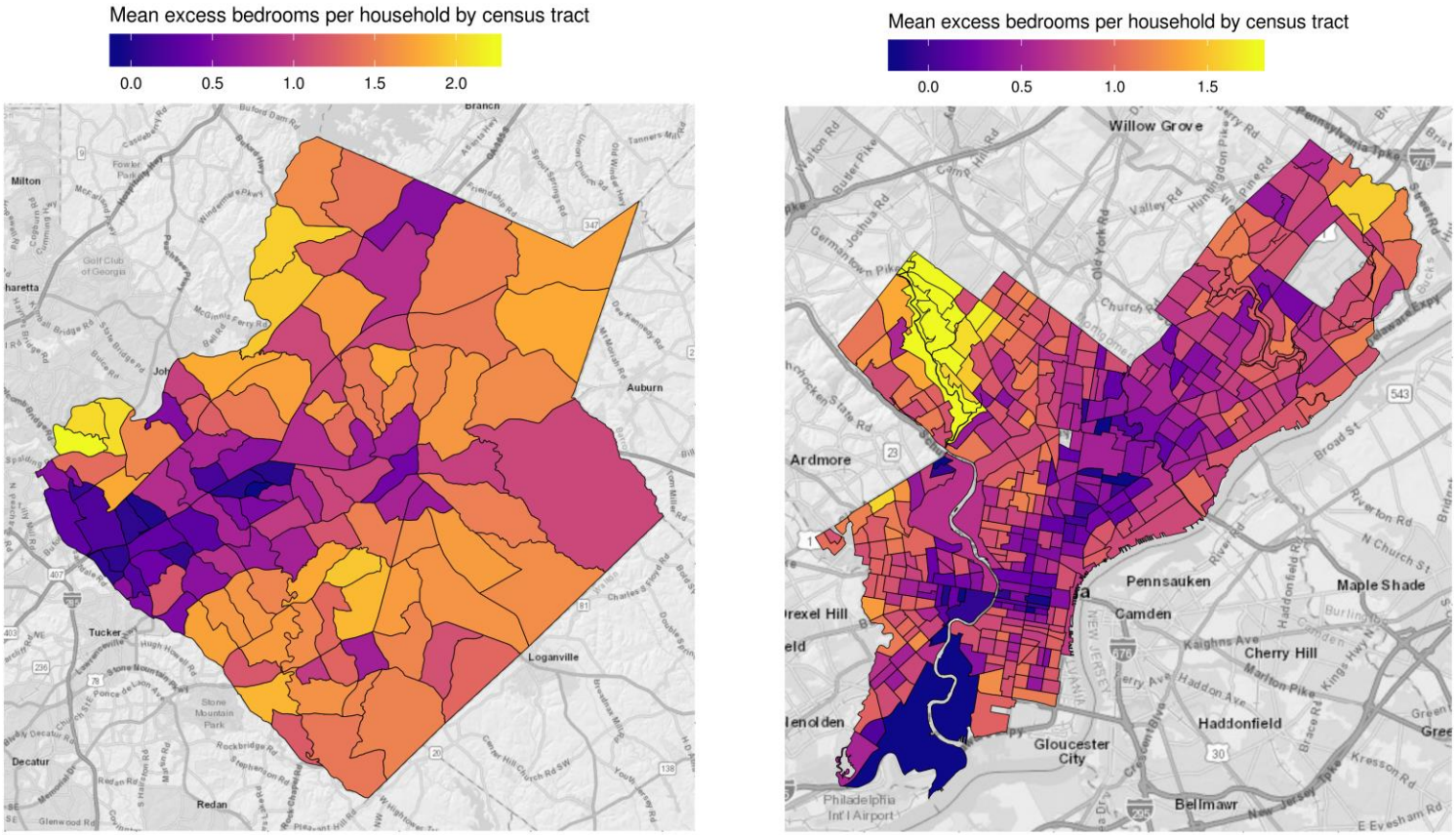
Electronic copy available at: https://ssrn.com/abstract=3574679

Figure 2. The distribution across Census tracts of variable importance scores in tree-based binning. Boxes represent the interquartile range, whiskers represent the 2.5 and 97.5 percentiles, and the black line represents the median.



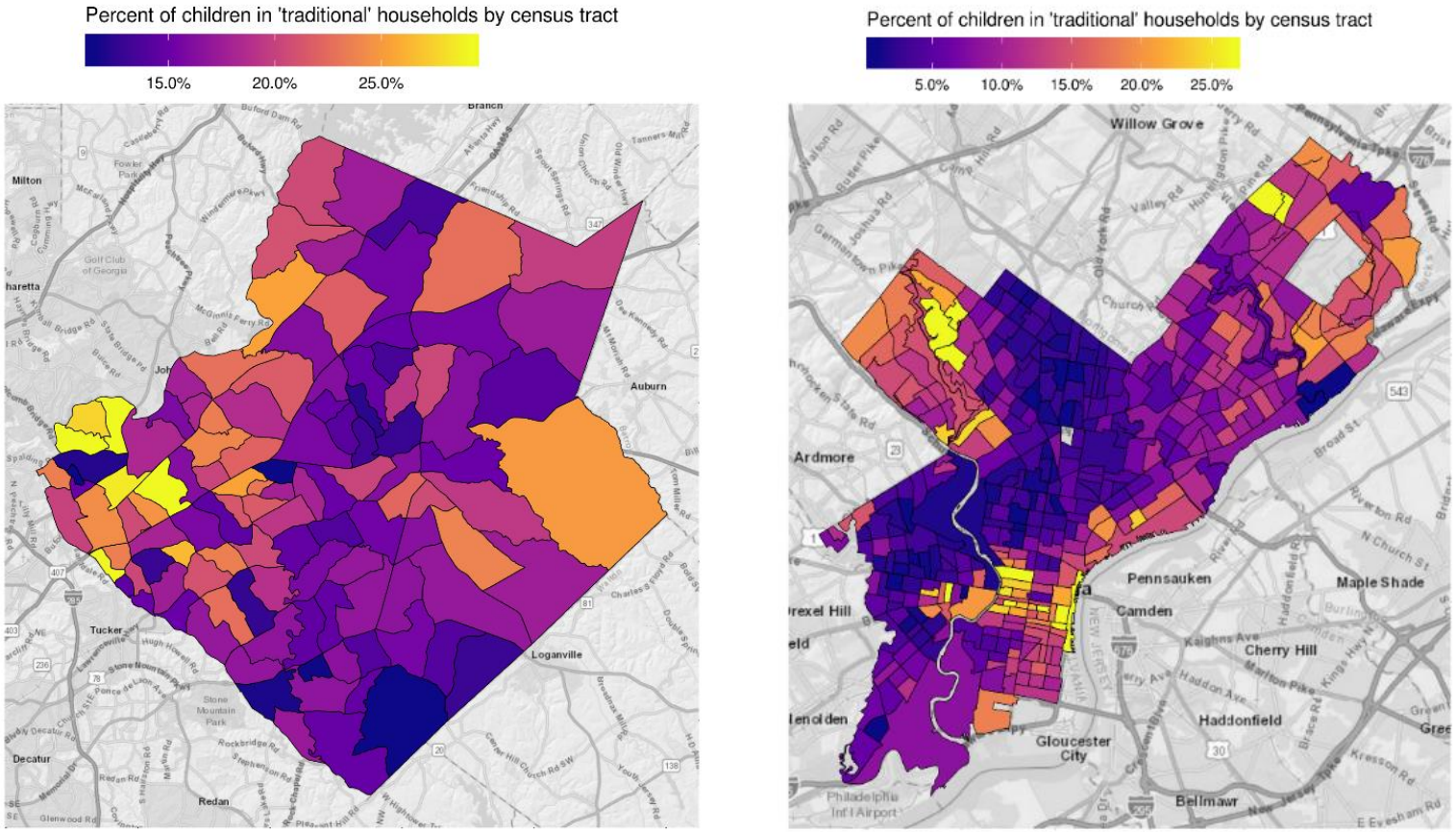
Electronic copy available at: https://ssrn.com/abstract=3574679

Figure 3. Small area estimates of excess housing capacity in Gwinnett (left) and Philadelphia (right) counties



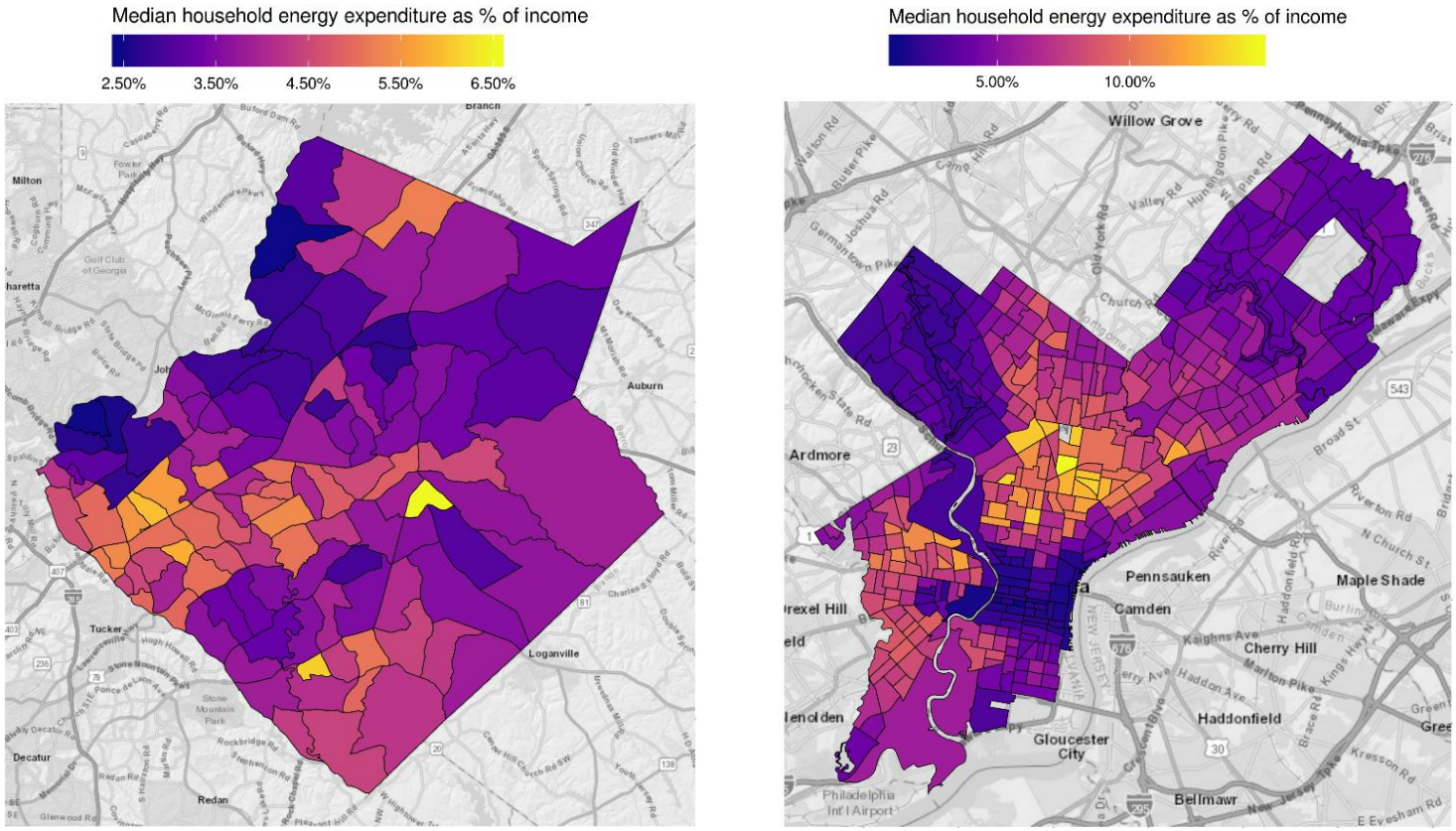
Electronic copy available at: https://ssrn.com/abstract=3574679

Figure 4. Small area estimates of traditional household structure in Gwinnett (left) and Philadelphia (right) counties.



Electronic copy available at: https://ssrn.com/abstract=3574679

Figure 5. Small area estimated of median household energy burden in Gwinnett (left) and Philadelphia (right) counties.



Electronic copy available at: https://ssrn.com/abstract=3574679

Appendix

Electronic copy available at: https://ssrn.com/abstract=3574679

Appendix Table 1. Summary of candidate margin variables.

	Variable	Level	Description	Example category	No. of categories	Ordinal?	ACS table(s)
1	age_hoh	Household	Age of householder	Householder 25 to 34 years	9	Yes	B25007
2	age_person	Person	Age of person	Under 10 years	16	Yes	B01001
3	commute_mode	Person	Commute mode	Personal vehicle or taxi	4	No	B08301
4	commute_time	Person	Commute time	30 to 34 minutes	12	Yes	B08303
5	education	Person	Educational attainment	HS graduate	8	Yes	B15002
6	health_insurance	Person	Type of health insurance	Employer-based	5	No	B27010
7	heatfuel	Household	Heating fuel	Natural gas	6	No	B25040
8	hhsize	Household	Household size	1-person household	7	Yes	B25009
9	housing_cost	Household	Housing cost as % of income	50.0 percent or more	10	Yes	B25070, B25091
10	income	Household	Household income	Less than $10,000	16	Yes	B19001
11	occupation	Person	Occupation class	Science, education, and health care	7	No	C24010
12	poverty	Person	Income relative to poverty line	2.00 and over	7	Yes	C17002
13	race	Person	Race or ethnicity	Black	5	No	B02001, B03002
14	rooms	Household	Number of rooms	6 rooms	9	Yes	B25017
15	sex	Person	Sex	Female	2	No	B01001
16	structure	Household	Dwelling type	Single-family	4	No	B25024
17	tenure	Household	Housing tenure	Renter occupied	3	No	B25003, B25091
18	type	Household	Housing unit type	Occupied	3	No	B25002, B09019
19	vehicles	Household	Number of vehicles	1 vehicle available	6	Yes	B25044
20	work_status	Person	Work status past 12 months	Did not work	3	No	B23027
21	yrbuilt	Household	Year dwelling built	Built 1939 or earlier	10	Yes	B25034

Electronic copy available at: https://ssrn.com/abstract=3574679

Appendix Table 2. Example of processed PUMS data

	Household ID	Person ID	Household weight	income	hhsize	race	health_insurance
1	1	1	19	Less than $10,000	2-person household	Black	Medicaid and means-tested public
2	1	2	19	Less than $10,000	2-person household	Black	None
3	2	1	30	$75,000 to $99,999	4-person household	Asian	Employer-based
4	2	2	30	$75,000 to $99,999	4-person household	White	Employer-based
5	2	3	30	$75,000 to $99,999	4-person household	White	Employer-based
6	2	4	30	$75,000 to $99,999	4-person household	White	Employer-based
7	3	1	54	$60,000 to $74,999	6-person household	Other	None
8	3	2	54	$60,000 to $74,999	6-person household	Other	Medicaid and means-tested public
9	3	3	54	$60,000 to $74,999	6-person household	Other	Medicaid and means-tested public
10	3	4	54	$60,000 to $74,999	6-person household	Other	Medicaid and means-tested public
11	3	5	54	$60,000 to $74,999	6-person household	Other	Medicaid and means-tested public
12	3	6	54	$60,000 to $74,999	6-person household	Other	None

Appendix Table 3. Summary of target variables used in validation exercise

Validation target variable	**Variable type**	**ACS table(s)**	**Excluded variable(s)**
Mean years schooling (age 25+)	Continuous	B15002	education
Percent of population with public health insurance	Discrete	B27003	health_insurance
Percent of population that is non-Hispanic white	Discrete	B02001	race
Mean hours worked per week (age 16-64)	Continuous	B23018, B23027	N/A

Electronic copy available at: https://ssrn.com/abstract=3574679

Appendix Section 1. Description of validation tests

We report three different error measures for comparing model estimates of a given target variable ($\hat{X}$) to "known" values derived from summary tables (x). The first is the canonical coefficient of determination ($R^2{}_{pred}$):

$$R^2{}_{pred} = 1 - \frac{\Sigma(x - \hat{x})^2}{\Sigma\left(x - \underline{x}\right)^2}$$

The second is "symmetric accuracy" (ζ), a measure of percent error that does not suffer from the drawbacks of the common mean absolute percentage error (Morley, Brito, and Welling 2018; Tofallis 2015):

$$\zeta = exp\left(log\left(\frac{\hat{x}}{x}\right)\right) - 1$$

The third is simply absolute error (π):

$$\pi = |\hat{x} - x|$$

For continuous target variables we report $R^2{}_{pred}$ and ζ (median and mean of latter). For discrete variables, where the estimate is necessarily a population proportion [0, 1], we report $R^2{}_{pred}$ and π (median and mean of latter).

Finally, we also report a measure of model "value-added". In the absence of sample calibration, a "naive" small area estimate ($\hat{X}_{naive}$) is simply the target variable mean using the initial microdata sample and observation weights; i.e. the sample estimate using default PUMS observation weights prior to any calibration (analogous to a "null model"). Our technique "adds value" only to the extent that it outperforms $\hat{X}_{naive}$. Comparison of $\hat{X}_{naive}$ to observed values (x)

Electronic copy available at: https://ssrn.com/abstract=3574679

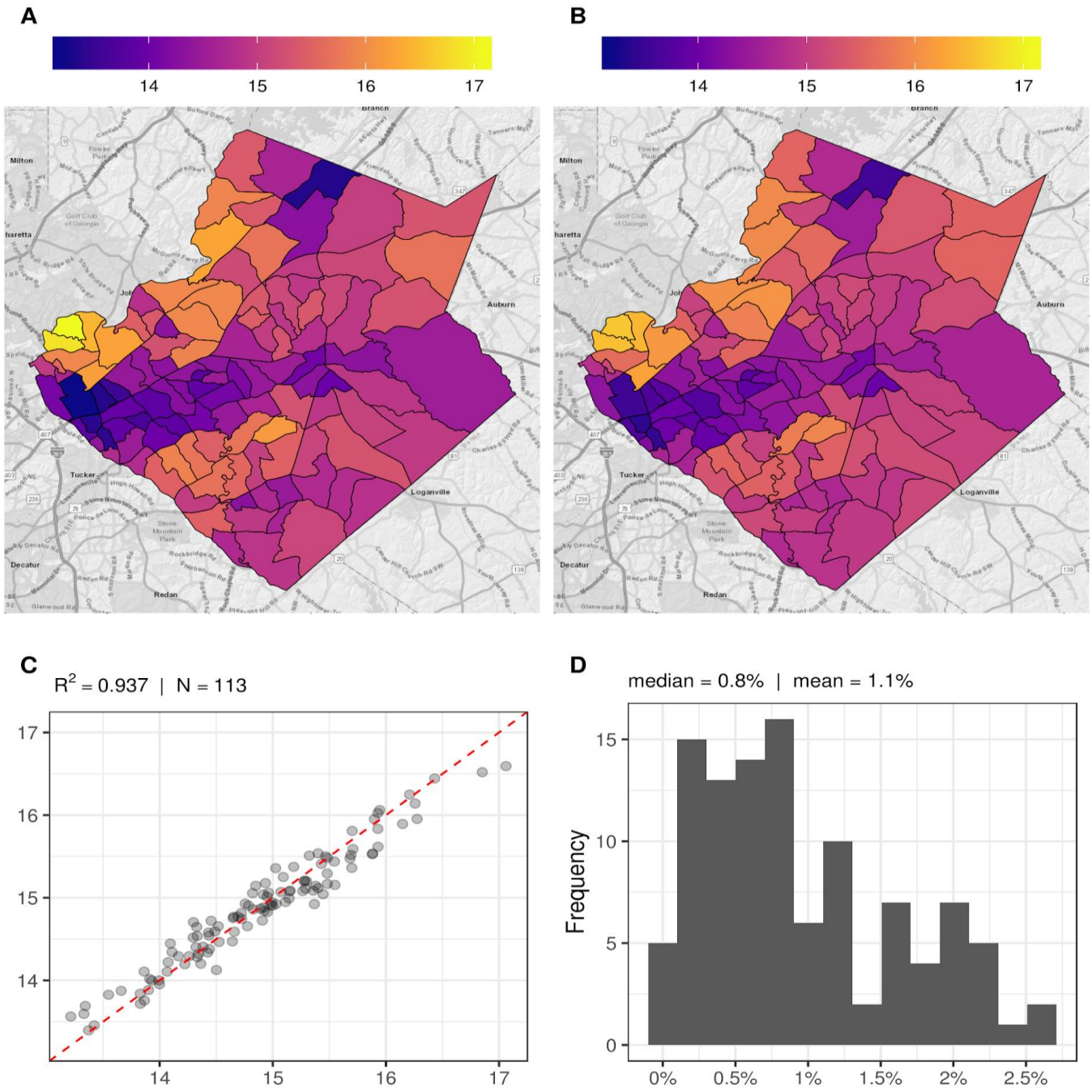
yields $R^2{}_{naive}$, which is analogous to the $R^2{}_{pred}$ calculated using $\hat{X}$. We define the model value-added (V) as:

$$V = \frac{\left(R^2{}_{pred} - R^2{}_{naive}\right)}{\left(1 - R^2{}_{naive}\right)}$$

Note that V has a maximum value of 1 (perfect prediction) and is negative if $R^2{}_{pred}$ is less than $R^2{}_{naive}$ (i.e. the model is worse than naive prediction).

Electronic copy available at: https://ssrn.com/abstract=3574679

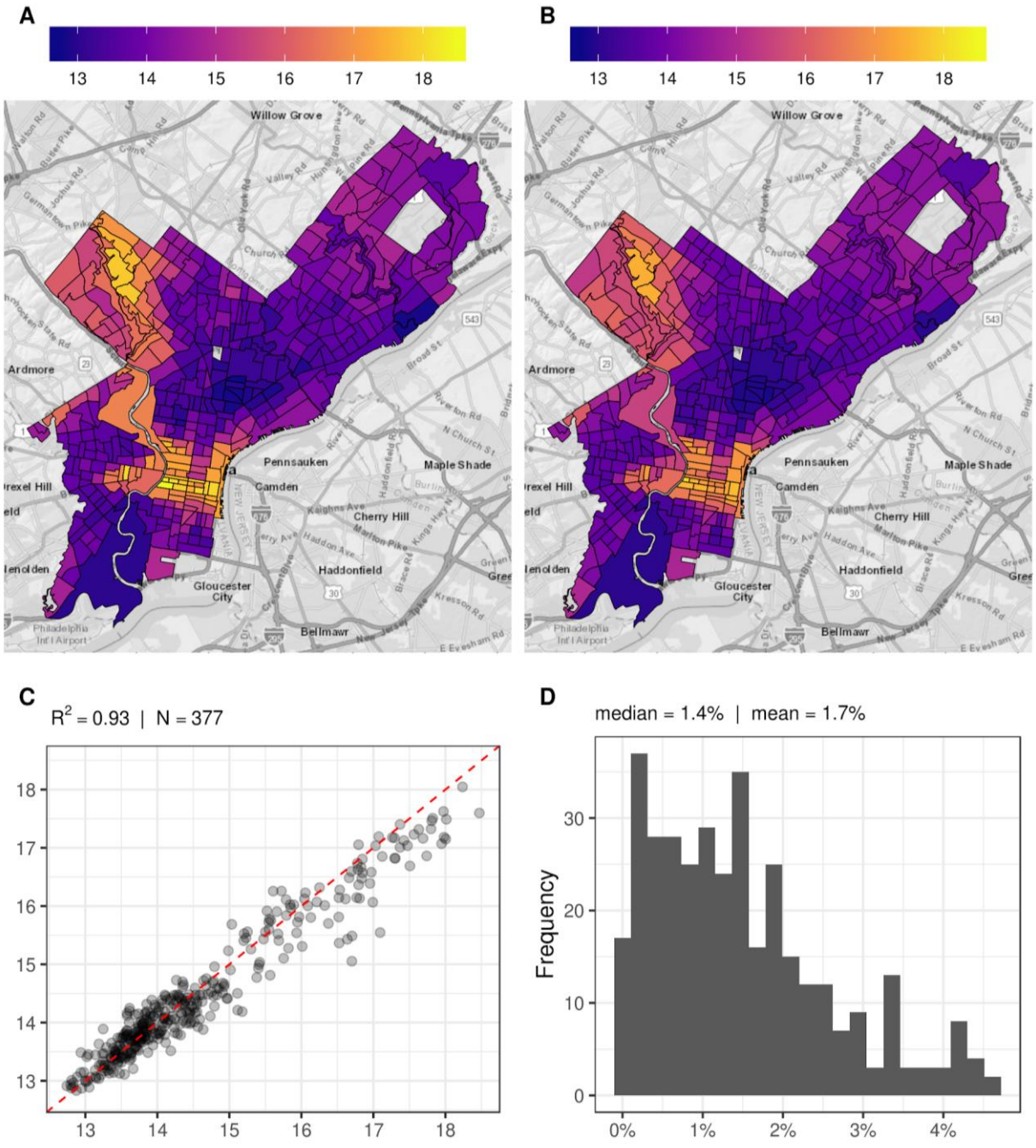
Appendix Figure 1. Mean years of schooling (age 25+) by census tract in Gwinnett County, GA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.927.

Electronic copy available at: https://ssrn.com/abstract=3574679

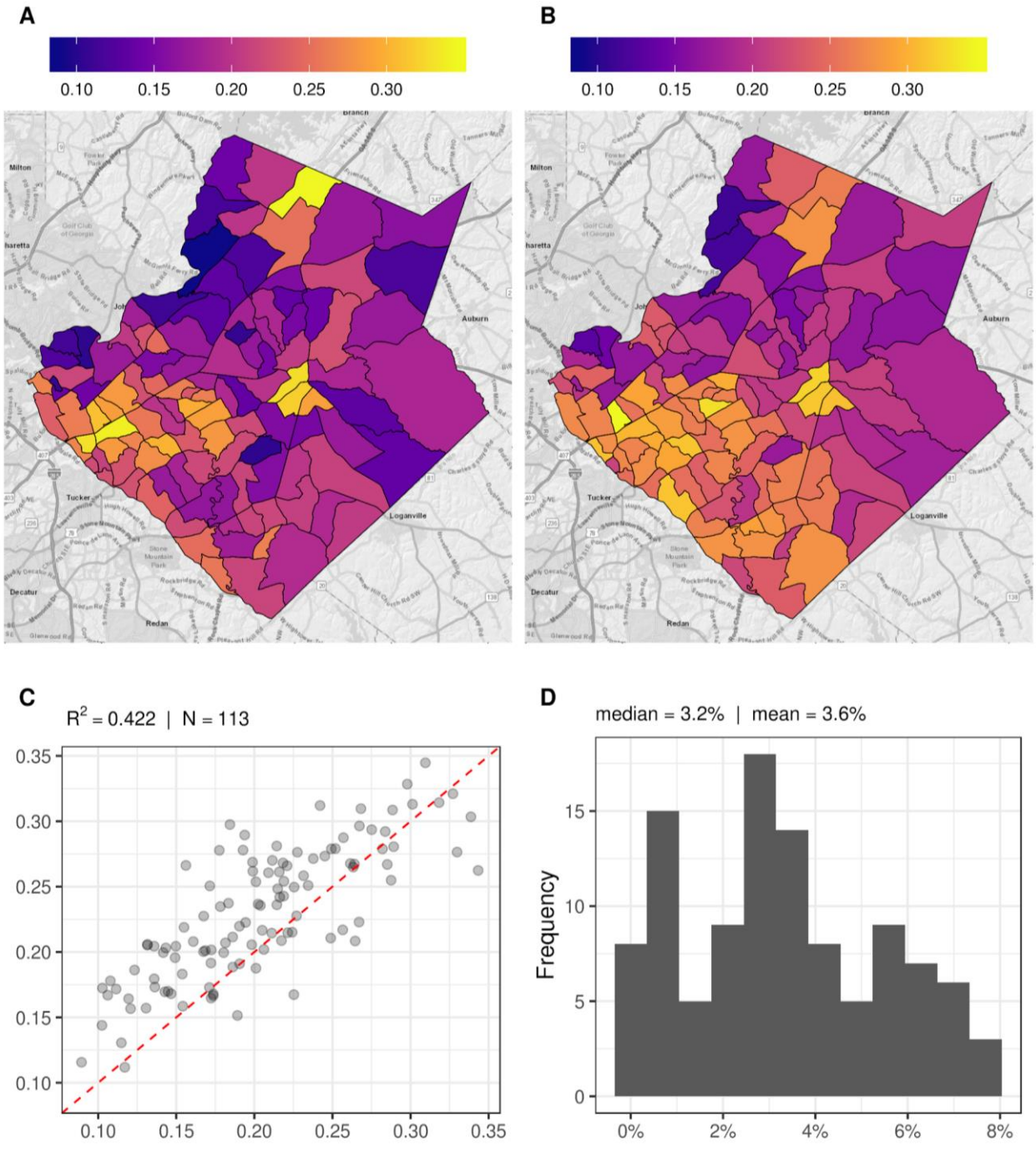
Appendix Figure 2. Mean years of schooling (age 25+) by census tract in Philadelphia County, PA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.828.

Electronic copy available at: https://ssrn.com/abstract=3574679

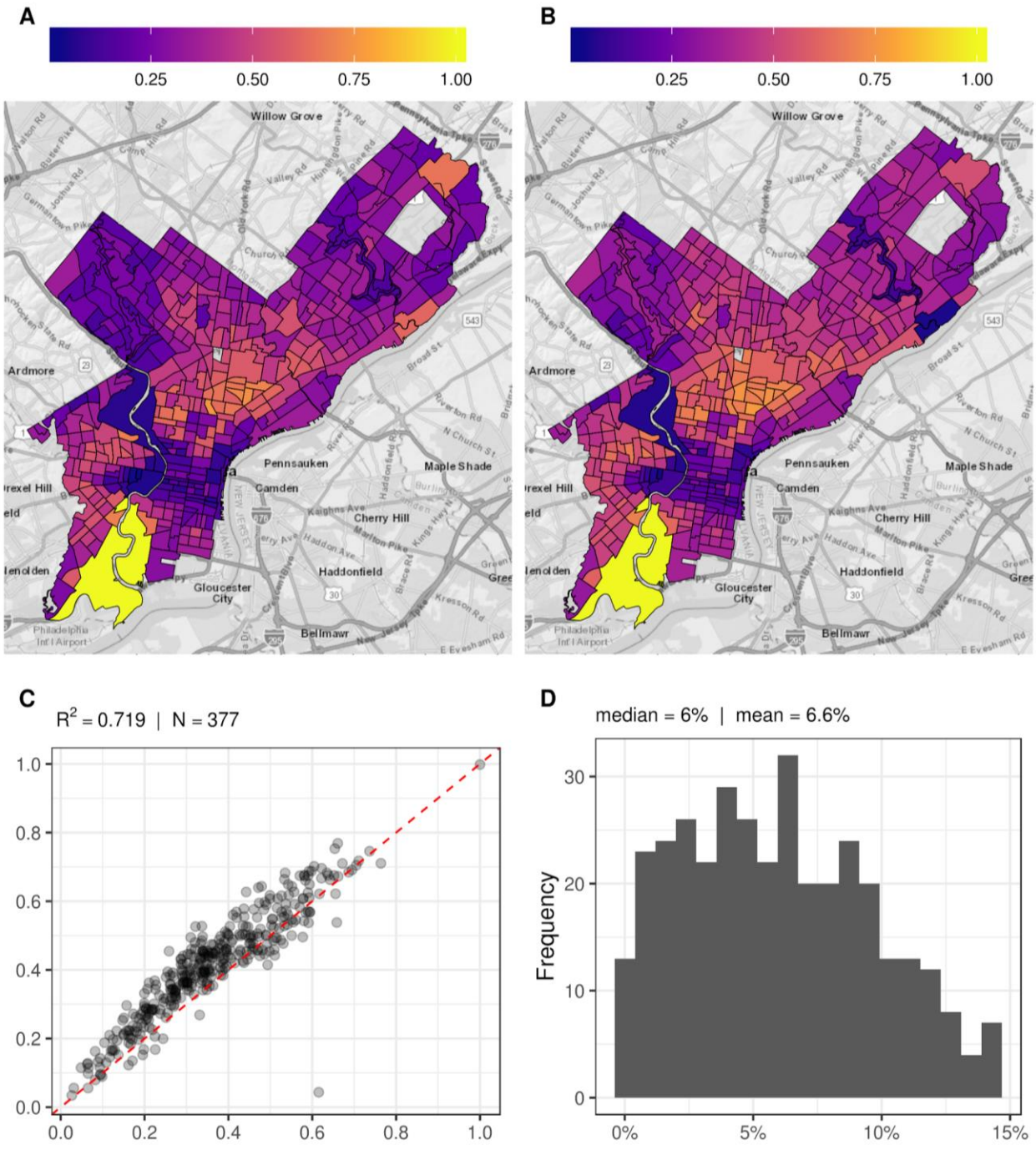
Appendix Figure 3. Percent of population with public health insurance by census tract in Gwinnett County, GA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.422.

Electronic copy available at: https://ssrn.com/abstract=3574679

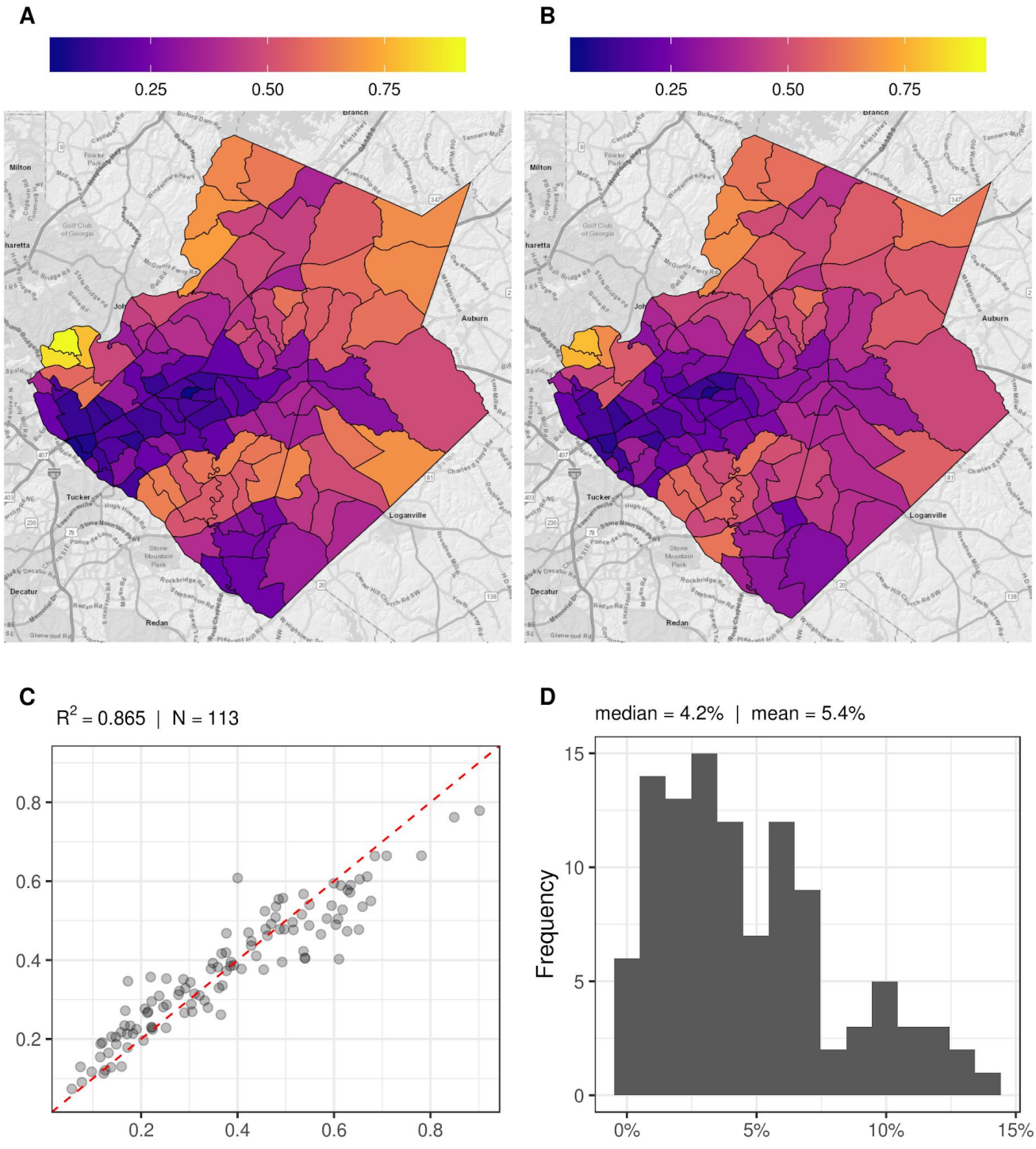
Appendix Figure 4. Percent of population with public health insurance by census tract in Philadelphia County, PA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.629.

Electronic copy available at: https://ssrn.com/abstract=3574679

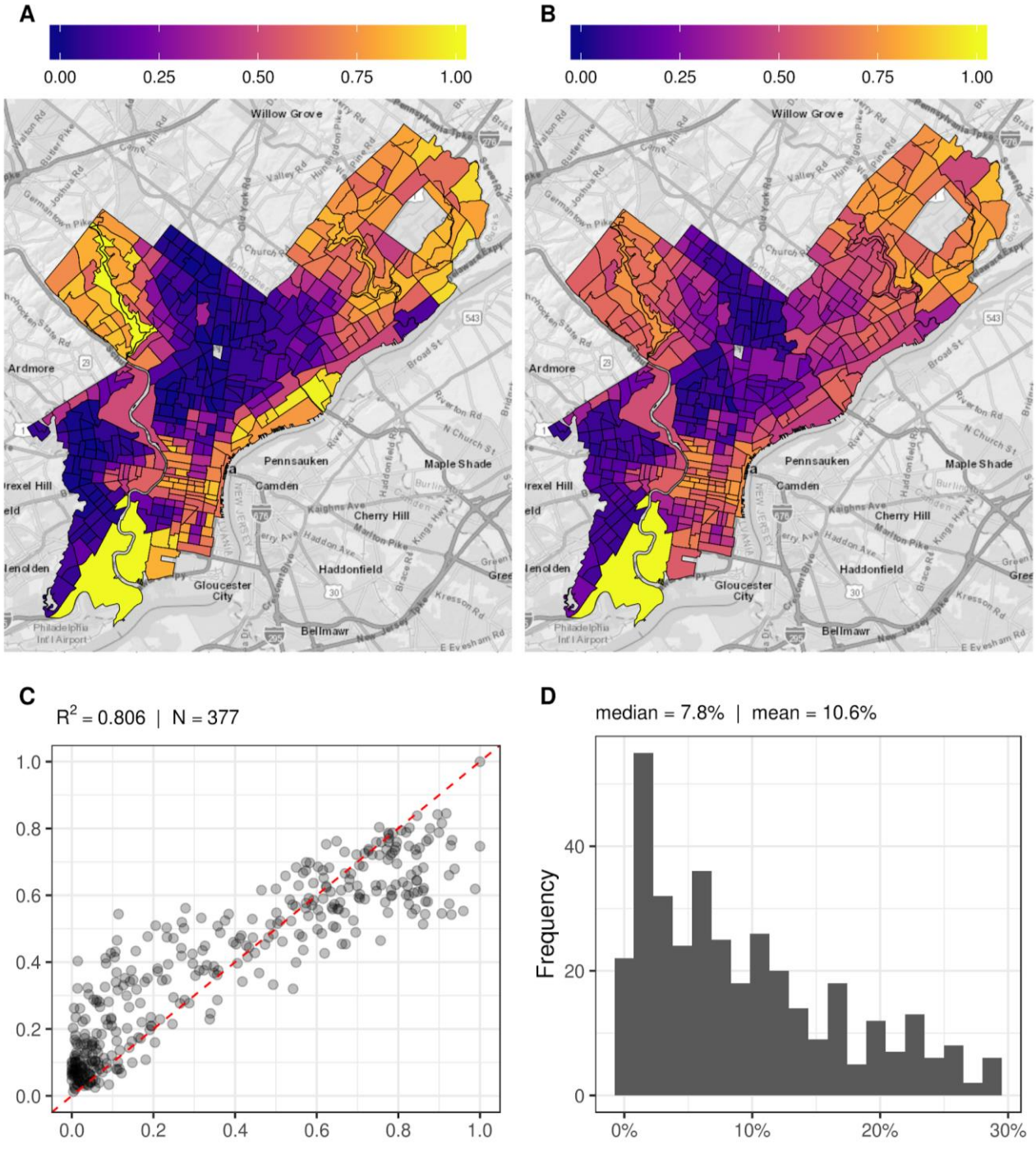
Appendix Figure 5. Percent non-Hispanic white by census tract in Gwinnett County, GA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.819.

Electronic copy available at: https://ssrn.com/abstract=3574679

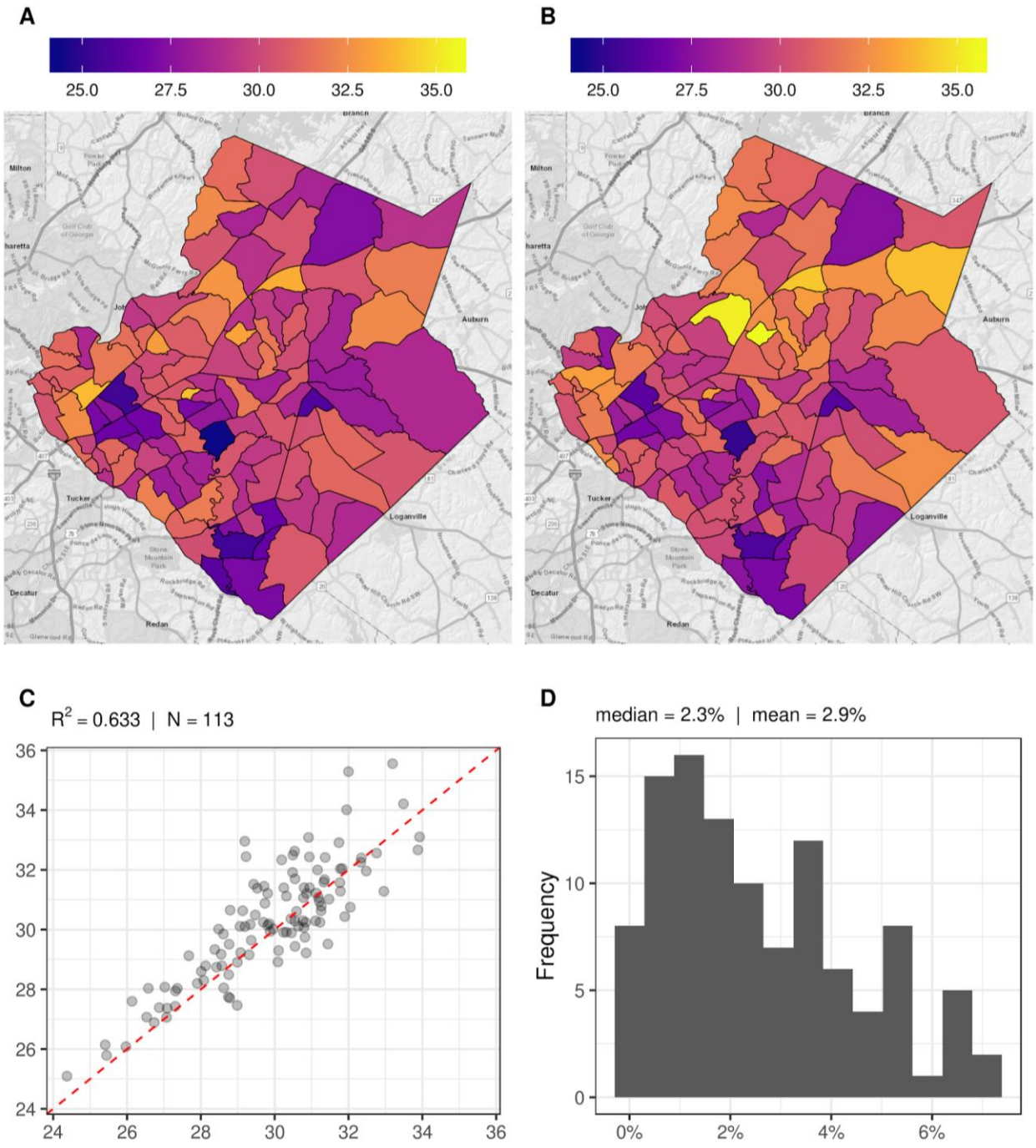
Appendix Figure 6. Percent of population with public health insurance by census tract in Philadelphia County, PA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.627.

Electronic copy available at: https://ssrn.com/abstract=3574679

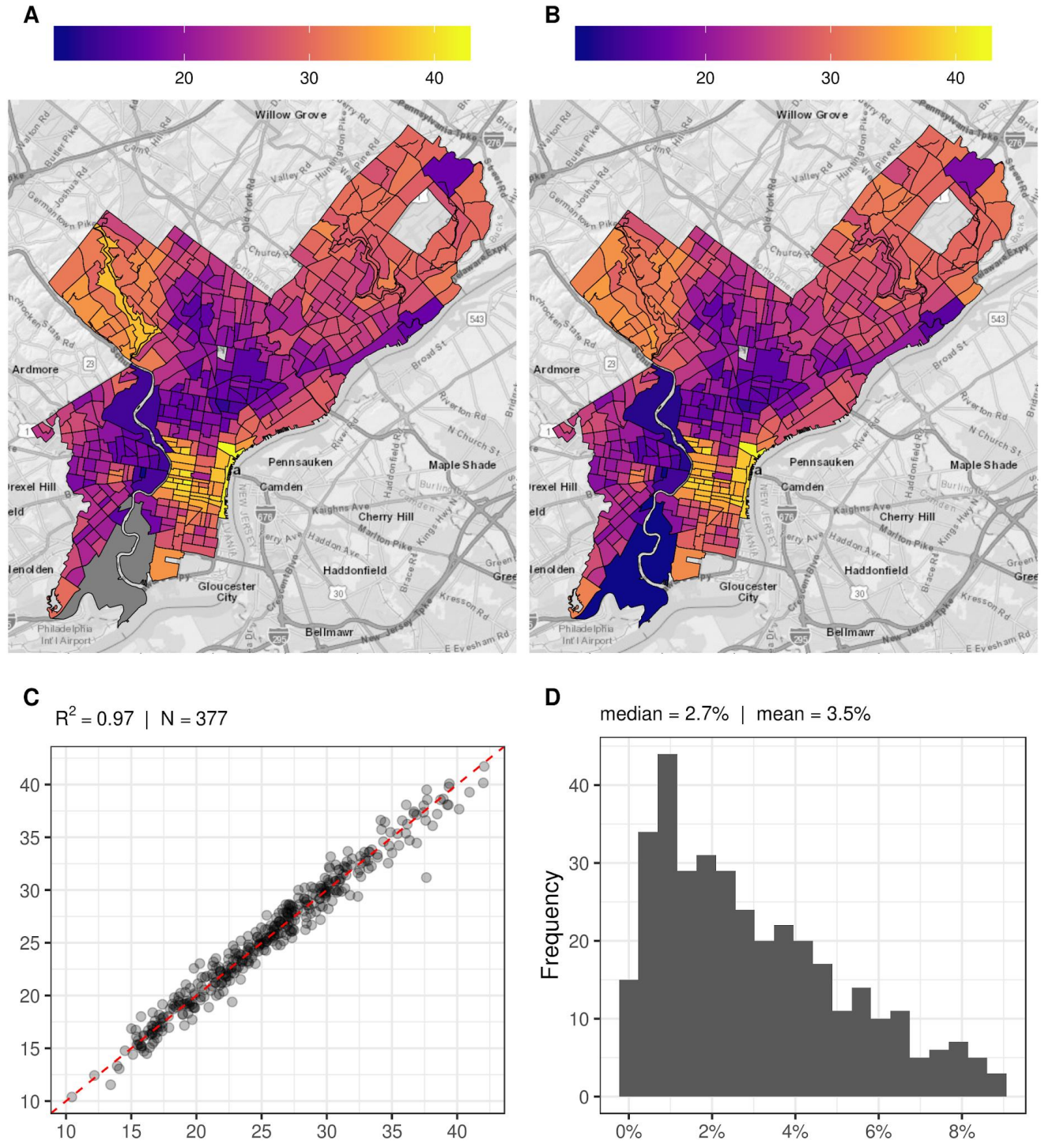
Appendix Figure 7. Mean hours worked by census tract in Gwinnett County, GA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.601.

Electronic copy available at: https://ssrn.com/abstract=3574679

Appendix Figure 8. Mean hours worked by census tract in Philadelphia County, PA.



Panel A shows estimates from the Census Bureau summary table B15002 (2012-2016 ACS) while Panel B shows estimates from the spatial microsimulation model (excluding education as a candidate constraint variable). Panel C illustrates the concordance between Census estimates (x-axis) and model estimates (y-axis). Panel D shows the distribution of symmetric accuracy. The model value-added relative to naïve estimates is 0.627.

Electronic copy available at: https://ssrn.com/abstract=3574679