

# Interpolating Distributions for Populations in Nested Geographies using Public-use Data with Application to the American Community Survey

Matthew Simpson,<sup>1,2</sup> Scott H. Holan,<sup>2,3</sup> Christopher K. Wikle,<sup>2</sup> and Jonathan R. Bradley<sup>4</sup>

## Abstract

Statistical agencies often publish multiple data products from the same survey. First, they produce aggregate estimates of various features of the distributions of several socio-demographic quantities of interest. Often these area-level estimates are tabulated at small geographies. Second, statistical agencies frequently produce weighted public-use microdata samples (PUMS) that provide detailed information of the entire distribution for the same socio-demographic variables. However, the public-use micro areas usually constitute relatively large geographies in order to protect against the identification of households or individuals included in the sample. These two data products represent a trade-off in official statistics: publicly available data products can either provide detailed spatial information or detailed distributional information, but not both. We propose a model-based method to combine these two data products to produce estimates of detailed features of a given variable at a high degree of spatial resolution. Our motivating example uses the disseminated tabulations and PUMS from the American Community Survey to estimate U.S. Census tract-level income distributions and statistics associated with these distributions.

**Keywords:** Bayesian methods, Functional data, Multi-scale model, Small area estimation, Spatial statistics.

---

<sup>1</sup>(to whom correspondence should be addressed) themattsimpson@gmail.com

<sup>2</sup>Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

<sup>3</sup>U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100

<sup>4</sup>Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306-4330

# 1 Introduction

The trade-off between spatial and distributional detail is ever-present in official statistics (e.g., U.S. Census Bureau, 2017a, Section 2). Statistical agencies cannot publish observations from a sample survey that are geocoded with the precise location of each household without risking disclosure of the surveyed individuals. Instead, they only release observations geocoded at a *coarse-scale geography*, typically called Public-Use Microdata Areas (PUMAs). At a *fine-scale geography* such as the county-level, statistical agencies compromise by only releasing estimates of specific features of the distribution of a given variable (e.g., U.S. Census Bureau, 2017b, Section 2). Often a data-user is interested in a feature of the distribution of some variable at a specific geography for which there is no published estimate. We propose a method to overcome this trade-off in order to obtain estimates of any feature of a variable’s distribution at a fine-scale geography by combining data products that have detailed spatial information but limited distributional information with data products that have limited spatial information but detailed distributional information. This problem commonly arises when a data-user is interested in income distributions, so our solution is motivated by estimating unobserved features of income distributions using data published by the U.S. Census Bureau from the American Community Survey (ACS), but the general structure of the problem arises from other variables and data products published by official statistical agencies.

Typically, statistical agencies make available many *bin estimates* of income; i.e., estimates of the proportion or number of households in a given areal unit with an income in a small number of income bins. For example, Table A in Appendix A of the Supplementary Materials contains 2015 ACS 5-year period bin estimates for several Census tracts in Boone County, MO. Sociologists and economists are interested in various measures of income inequality and segregation by income, and often develop ways to convert the bin estimates into estimates of

their desired measures (e.g. Nielsen and Alderson, 1997). The most commonly used measure of income inequality is the Gini coefficient (Yitzhaki, 1979). Estimates of Gini coefficients are available at a wide variety of geographies from the ACS, but the first ACS release was in 2005, and indeed they are often not available from other surveys or other statistical agencies. To remedy this, many authors use a method called the “Pareto-linear procedure” (PR-LN) to construct an estimate of the Gini coefficient, e.g., Jargowsky (1996); Nielsen and Alderson (1997); Hipp (2007a,b); Moller et al. (2009); Hipp et al. (2013); Braithwaite (2015). PR-LN assumes that income is uniformly distributed within bins which are below the median or include the median, and Pareto distributed in bins above the median, with some exceptions to handle special cases. This yields an estimate of the entire distribution of income and in turn can be used to estimate the Gini coefficient and other features of the income distribution. The methodology is well-established, and is effective for income distributions (Miller, 1966; Aigner and Goldberger, 1970; Kakwani and Podder, 1976; Spiers, 1977; Henson and Welniak, 1980; Welniak, 1988).

Estimates of many other measures of income inequality and segregation by income are not typically made publicly available by statistical agencies, and several methods are used to construct appropriate estimates using the bin estimates (Kennedy et al., 1996; Jargowsky, 1996; Mayer et al., 2001; Hardman and Ioannides, 2004; Watson, 2009; Reardon, 2011; Reardon and Bischoff, 2011). One such index is the Bourgignon index, which depends on the full distribution of wealth Ioannides and Seslen (2002). To obtain wealth distributions at a fine-scale geography, Ioannides and Seslen (2002) first estimates a regression model to predict wealth with various demographic characteristics using a dataset geocoded at a coarse-scale geography. Then they use the regression model to predict the wealth of families in a dataset geocoded at a fine-scale geography, but with no income or wealth information. This is similar in spirit to our approach, but we only have unit-level data at the coarse-scale geography (PUMA-level), and have areal-level data at the fine-scale geography.

A common thread to many of these approaches is that they only use information directly from a given areal unit to estimate the entire distribution in that areal unit. The class of models we develop uses publicly available microdata at a coarse-scale geography to “fill in the gaps” of Census tract-level household income distributions “between” the available tract-level estimates, e.g., those in Table A. Our methodology requires two key pieces. First, publicly available *source statistics*, i.e., statistical estimates produced at the fine-scale geography. Second, publicly available *source observations*, i.e., weighted survey samples available at a coarse-scale geography. Further, the fine-scale geography must be nested within the coarse-scale geography, and the source observations must be organized into strata where each observation in a given stratum has the same survey weight. The source statistics anchor the tract-level distributions, while the source observations provide distributional detail. This allows us to borrow strength across fine-scale areal units and additionally exploit any spatial patterns in the fine-scale estimates to build better models. While PR-LN only uses bin estimates, in principle we can use all available estimates of features of income distributions. Finally, our approach uses margins of error associated with the tract-level estimates quite naturally, which is critical for propagating uncertainty into any desired model-based estimates. None of the approaches above take into account this uncertainty, potentially biasing their estimates of Gini coefficients and other quantities.

The source statistics are functionals of the latent tract-level income distributions, so our model borrows elements from functional data analysis (FDA) — see e.g., Ramsay and Silverman (2005), and Ferraty and Vieu (2006) for overviews. However, our case differs from the usual FDA case because latent functions we are trying to estimate are probability distribution functions (PDFs), or equivalently any function which uniquely determines the latent probability distribution such as a cumulative distribution function (CDF) or quantile function. This puts constraints on the latent function that are not typical for FDA, and necessarily implies a different modeling strategy. We also allow for spatial dependence between

the tract-level parameters to control the tract-level income distributions, using a low-rank approach from Bradley et al. (2017) based on Obled-Creutin basis functions (Obled and Creutin, 1986), but see also Wikle and Cressie (1999), Cressie and Johannesson (2006), Cressie and Johannesson (2008), and Cressie and Wikle (2011, Section 4.1.4) for similar low-rank approaches. The literature on spatial functional data has primarily focused on geostatistical data or point processes, though some approaches to functional areal data exist — see Delicado et al. (2010) for a review of spatial FDA, Porter et al. (2014) for an application with areal referenced functional covariates, and Yang et al. (2015) for functional geostatistical responses with multi-dimensional functional covariates. Our work is also related to the approach to spatial quantile regression in Reich et al. (2011), but we do not have observations at the locations of interest; instead, we have estimates of a variety of distributional features.

The source observations are weighted survey samples, and there is a large literature discussing how to properly model and draw inference from these sort of data. One related thread of the literature discusses how inference is to proceed using data with associated survey weights, e.g., using likelihood based inference (Chambers et al., 2012), Bayesian analysis (Rubin, 1983; Little, 1991, 1993), calibrated Bayes (Little, 2012, 2015), and whether and how to post-stratify (Gelman, 2007; Si et al., 2015). The typical goal in this literature is to use weighted data from a sample survey to produce reliable estimates of population-level quantities for the population which was sampled. Our case is different since we want to perform inference on the quantities of certain sub-populations that make up the population from which the microdata was sampled. This literature does ultimately inform the class of models we construct, in particular Gelman (2007), Chambers et al. (2012), and Si et al. (2015), as we model the latent income distribution of each stratum separately.

The remainder of the paper is organized as follows. In Section 2 we describe the ACS and the challenges which arise when trying to construct a model of a latent population of households and their incomes. Section 3 describes how we overcome those challenges and

develops the class of models we propose. In Section 4 we demonstrate via a simulation study that a key assumption in our models does not prevent us from providing quality estimates of unobserved features of tract-level distributions. We fit the proposed models to ACS income data in Section 5 and compare model-based estimates to held-out direct estimates of various features of income distributions. Finally, in Section 6, we discuss our results and conclude.

## 2 American Community Survey and Model Motivation

The U.S. Census Bureau administers the ACS to produce a variety of annually released data products used by public and private institutions. There are two main types of data products. First, ACS estimates of various quantities are tabulated and published for several geographies, including Census tracts, counties, states, and national. Second, raw data files in the form of Public-Use Microdata Samples (PUMS) are released to the public. The PUMS are organized into PUMAs, and they contain a weighted sample of households and of residents living in each PUMA; more detailed location information about these residents and households is not available due to disclosure limitations. Each PUMA is designed to contain around 100,000 people, and Census tracts are nested within PUMAs.

The PUMS sample in a given PUMA for a given period is a subset of the full ACS sample for that same area and period, and the sample weights in the PUMS are not the same as the weights used to construct the ACS estimates (U.S. Census Bureau, 2017b). Both the ACS estimates and PUMS are currently published based on one and five years of samples, known as 1-year and 5-year period estimates and PUMS respectively, though areal units with less than 65,000 people only have published 5-year period estimates, and in previous years areal units with at least 20,000 people also had published 3-year period estimates (U.S. Census Bureau, 2014).

We will use the 5-year period estimates as our source statistics since they are produced

for all Census tracts with few exceptions, which necessitates using the 5-year PUMS as our source observations. A map of an example PUMA is contained in Figure A.1 in the Supplementary Materials. The ACS published the following 5-year tract-level income distribution period estimates: mean income, median income, Gini coefficient of income, the 20th, 40th, 60th, 80th, and 95th percentiles of income, and the proportion of households with incomes in 12 income bins defined by the following breaks: \$5,000, \$10,000, \$15,000, \$20,000, \$25,000, \$35,000, \$50,000, \$75,000, \$100,000, \$150,000, and \$200,000 (U.S. Census Bureau, 2017f,g,h,i,j). Each tract-level estimate also has a corresponding margin of error (MOE) so that estimate  $\pm$  MOE determines a 90% confidence interval, and MOE/1.645 is the standard error of the estimate.

For simplicity, consider only a single PUMA. Suppose the  $r$ th tract in that PUMA has population  $N_r$  and let  $\mathbf{Y}_r$  denote the latent  $N_r$ -dimensional vector of population incomes for tract  $r$ ,  $r = 1, 2, \dots, R$ , and let  $\mathbf{Y} = (\mathbf{Y}'_1, \mathbf{Y}'_2, \dots, \mathbf{Y}'_R)'$ . Our goal is to learn about  $\mathbf{Y}$  via the source statistics and source observations. Let  $q_{ur}$  denote the  $u$ th source statistic for tract  $r$  with standard error  $E_{ur}$ ,  $u = 1, 2, \dots, U$ , and let  $\mathbf{q}_r = (q_{1r}, q_{2r}, \dots, q_{Ur})'$ . The source observations are the PUMS, i.e., observations of household income geocoded at the PUMA level with attached survey weights. Each unique weight defines a unique stratum in the PUMS. The strata may depend on various demographic and geographic characteristics, and we expect households or individuals in different strata to be systematically different from each other. Let  $s = 1, 2, \dots, S$  index strata,  $\mathbf{Z}_s$  denote the  $n_s$ -dimensional vector of source observations for stratum  $s$ , and  $\mathbf{Z} = (\mathbf{Z}'_1, \mathbf{Z}'_2, \dots, \mathbf{Z}'_S)'$ .

The first challenge is that due to disclosure limitations there is no publicly available information on tract and stratum interactions. So we assume that the tract-level estimates and the stratum-level observations come from separate but dependent populations. While this assumption is not true, it may be benign. If the within-stratum distributions are similar to the within-tract distributions in the unobserved population, then the stratum-level ob-

servations should still provide detailed information about the tract-level distributions. This seems reasonable in the ACS case since we know that the PUMS strata are based in part on Census tracts (U.S. Census Bureau, 2017c), but for other surveys and other geographies it may not be tenable. In Appendix B, we empirically verify this requirement for household income from the ACS.

The second challenge is that the latent population is large. A single PUMA may contain about 60,000 or 70,000 households, and modeling the latent population would require fitting a model with that many latent variables. For an entire state, these numbers are in the millions. This is challenging since the computational cost of fitting the model will be at least  $\mathcal{O}(N)$  where  $N$  is the latent population size, and complex models make it more difficult. We avoid modeling latent populations altogether and let probability distributions take on the role of populations. This is similar to marginalizing out the latent process, which is common in spatial statistics (e.g., Wikle, 2010), but we directly specify the marginal model.

Let  $\pi_{\theta_r}$  denote the PDF for tract  $r$  and  $\pi_{\tilde{\theta}_s}$  denote the PDF for stratum  $s$  — stratum-level parameters will always have an overhead  $\sim$ . Then stratum-level observations are assumed to be drawn iid from the stratum-level distributions, i.e.,  $Z_{is} | \tilde{\theta}_s \stackrel{ind}{\sim} \pi_{\tilde{\theta}_s}$  for  $i = 1, 2, \dots, n_s$  and  $s = 1, 2, \dots, S$ . Source statistics are assumed to be centered on the corresponding functionals of the tract-level distributions. Let  $Q_u(\cdot)$  denote the  $u$ th functional so that  $q_{ur}$  is an estimate of  $Q_u(\pi_{\theta_r})$ . For example if  $q_{ur}$  is an estimate of mean income in tract  $r$ , it is centered on the mean of the probability distribution  $\pi_{\theta_r}$ . Typically a central limit theorem applies to these estimates, and there is no available information about error correlations, so we assume

$$q_{ur} | \theta_r \stackrel{ind}{\sim} N(Q_u(\pi_{\theta_r}), E_{ur}^2) \text{ for } u = 1, 2, \dots, U \text{ and } r = 1, 2, \dots, R,$$

where  $E_{ur}$  is the known standard error associated with  $q_{ur}$ , and  $N(m, s^2)$  denotes the normal distribution with mean  $m$  and variance  $s^2$ .

From the perspective of the data-user, household incomes within a stratum are exchangeable, which motivates our assumption that stratum-level observations are conditionally independent given the stratum-level parameters. To complete the model and build dependence between the model’s implicit tract-level and stratum-level populations, we assume  $\{\boldsymbol{\theta}_r : r = 1, 2, \dots, R\}$  and  $\{\tilde{\boldsymbol{\theta}}_s : s = 1, 2, \dots, S\}$  are conditionally independent given some higher level parameter  $\boldsymbol{\theta}$ . A graphical depiction of our framework is provided in Figure 1.

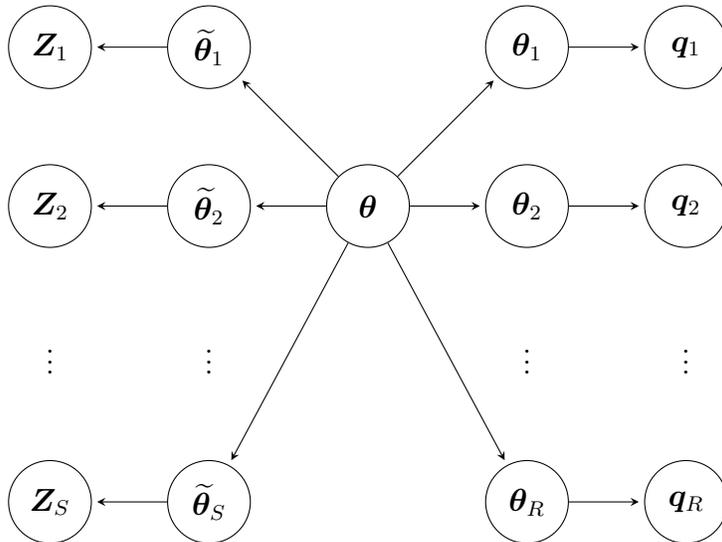


Figure 1: Graphical depiction of the modeling framework. Arrows represent conditional dependence relationships between nodes.

Essentially we have removed  $\mathbf{Y}$  from the model and, by doing so, have completely removed the  $\mathcal{O}(N)$  computational cost associated with it. An additional computational benefit to this modeling strategy is that  $Q_u(\pi_{\boldsymbol{\theta}_r})$  is likely to be differentiable in  $\boldsymbol{\theta}_r$  everywhere even when the corresponding function of  $\mathbf{Y}_r$  is not differentiable in  $\mathbf{Y}_r$  (e.g.,  $Q_u(\cdot)$  may be a quantile). Thus, standard Hamiltonian Monte Carlo (HMC) methods should work to fit the model (see e.g., Betancourt and Girolami, 2015, for an introduction). In the models we fit, the  $Q_u(\boldsymbol{\theta}_r)$ s are always differentiable everywhere and we use the software package **Stan** (Gelman et al., 2015; Stan Development Team, 2016) to do Markov chain Monte Carlo (MCMC) via HMC. The inferential goal is still to learn about  $\mathbf{Y}$  even though it is no longer explicitly included in

the model. We simulate it from the posterior predictive distribution after fitting the model. Let  $\boldsymbol{\theta}_r^{(i)}$  denote the  $i$ th posterior replicate of  $\boldsymbol{\theta}_r$ . Then, we simulate

$$Y_{ir}^{(j)} \stackrel{ind}{\sim} \pi_{\boldsymbol{\theta}_r^{(i)}} \text{ for } i = 1, 2, \dots, N_r, \text{ } r = 1, 2, \dots, R, \text{ and } j = 1, 2, \dots, N_{rep},$$

and can construct any functions of  $\mathbf{Y}$ , such as tract-level quantiles or moments.

Let  $[A|B]$  denote the conditional density of  $A$  given  $B$ . Then the modeling framework we propose can be concisely specified as follows:

$$q_{ur}|\boldsymbol{\theta}_r \stackrel{ind}{\sim} \text{N}(Q_u(\pi_{\boldsymbol{\theta}_r}), E_{ur}^2), \quad (\text{tract-level data model}) \quad (1)$$

for  $u = 1, 2, \dots, U$  and  $r = 1, 2, \dots, R$ ,

$$Z_{is}|\tilde{\boldsymbol{\theta}}_s \stackrel{ind}{\sim} \pi_{\tilde{\boldsymbol{\theta}}_s}, \quad (\text{stratum-level data model}) \quad (2)$$

for  $i = 1, 2, \dots, n_s$  and  $s = 1, 2, \dots, S$ ,

$$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_R|\boldsymbol{\theta} \sim [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_R|\boldsymbol{\theta}], \quad (\text{tract-level parameter model}) \quad (3)$$

$$\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_S|\boldsymbol{\theta} \sim [\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_S|\boldsymbol{\theta}], \quad (\text{stratum-level parameter model}) \quad (4)$$

$$\boldsymbol{\theta} \sim [\boldsymbol{\theta}]. \quad (\text{prior}) \quad (5)$$

Our *tract-level process model* is implicitly specified via Equations (1)-(5) as

$$Y_{ir} \stackrel{ind}{\sim} \pi_{\boldsymbol{\theta}_r}, \text{ for } i = 1, 2, \dots, N_r, \text{ and } r = 1, 2, \dots, R. \quad (6)$$

In Section 3 we specify the parametric family  $\pi_{\boldsymbol{\theta}}$  for this model.

### 3 Methods and Model Development

To fully specify the model in Section 2, we need several things: 1) analytical formulas for the  $Q_u(\pi_{\theta_r})$ s as functions of the  $\theta_r$ s, 2) a parametric family  $\pi_{\theta}$  indexed by the parameter  $\theta$ , 3) a tract-level parameter model, 4) a stratum-level parameter model, and 5) a prior on  $\theta$ . Each of these pieces are related to the others, but it is useful to start with the  $Q_u(\pi_{\theta_r})$ s.

#### 3.1 Source statistics and parametric families

The types of source statistics available for use in Equation (1) vary across applications, but in this subsection we will discuss in detail three typically available types of source statistics: moment estimates, quantile estimates, and bin estimates. Let  $\pi_{\theta_r}(y)$  denote the tract-level PDF of income for tract  $r$ ,  $\Pi_{\theta_r}(y)$  denote the corresponding CDF, and  $\Pi_{\theta_r}^{-1}(\tau)$  denote the corresponding quantile function. For ease of notation, we will often drop the subscripts and refer to these as  $\pi$ ,  $\Pi$ , and  $\Pi^{-1}$  when appropriate.

The source statistics put an important constraint on the choice of parametric family: the relevant functionals for the family must be easy to compute in order to fit the model efficiently. Most simple parametric families work, but this constraint rules out many more complex approaches. We consider two basic parametric families: normal and lognormal. We also consider finite mixtures of both parametric families, though we use an approximation to handle quantile estimates. Other more complex parametric families may be more appropriate for income distributions (e.g., the Pareto-lognormal; Hajargasht and Griffiths, 2013), but this makes estimation of our larger model much more difficult. Once the parametric family has been chosen and functions for computing each of the  $Q_u(\cdot)$ s specified, then (1) and (2) are fully specified.

Table 1 contains relevant information for constructing the most common moments (means and variances), quantiles, and bin proportions, as well as the Gini coefficient in the lognormal

Distribution ( $\pi$ ):	Normal	Lognormal
Parameters:	$\mu \in \mathbb{R},$ $\sigma > 0$	$\mu \in \mathbb{R},$ $\sigma > 0$
Mean:	$\mu$	$\exp(\mu + \sigma^2/2)$
Variance:	$\sigma^2$	$(e^{\sigma^2} - 1) \exp(2\mu + \sigma^2)$
$\Pi(y)$ :	$\Phi\left(\frac{y-\mu}{\sigma}\right)$	$\Phi\left(\frac{\log y - \mu}{\sigma}\right)$
$\Pi^{-1}(\tau)$ :	$\mu + \sigma\Phi^{-1}(\tau)$	$\exp\{\mu + \sigma\Phi^{-1}(\tau)\}$
Gini:	—	$2\Phi(\sigma/\sqrt{2}) - 1$

Table 1: Common parametric families to choose for  $\pi$  and relevant quantities for constructing common  $Q_u(\pi)$ s.  $\Pi(y)$  denotes the CDF of the family and  $\Pi^{-1}(\tau)$  denotes the quantile function. The function  $\Phi(\cdot)$  denotes the standard normal CDF.

case. The Gini coefficient is a measure of income inequality that ranges from 0, where each household has the same income, to 1, where one household has all of the income. Note that if  $q_{ur}$  is a bin estimate with bounds  $a < b$ , then  $Q_u(\pi_{\theta_r}) = \Pi_{\theta_r}(b) - \Pi_{\theta_r}(a)$ .

Each of the quantities in Table 1 are differentiable in the parameters, including each CDF and each quantile function, so we can use HMC to fit the model. For attaining additional flexibility, we use finite mixture models. Suppose we have a  $K$  component mixture with mixture probabilities in tract  $r$  denoted by  $\{\omega_{rk} : k = 1, 2, \dots, K\}$ . We will write e.g.  $Y \sim \sum_{k=1}^K \omega_k \mathbf{N}(m_k, s_k^2)$  to denote that  $Y$  is distributed according to a  $K$  component mixture of normals. Then, we can compute the mixture distribution means, variances, and bin proportions for that tract using the formulas in Table 2, based on the means, variances, and bin proportions of the base distribution, though we cannot write down similar formulas for the quantiles of finite mixtures. Similarly we cannot write down similar formulas for the Gini coefficient of finite mixtures in general, but the Gini coefficient for a  $K$ -component mixture of lognormals is (Young, 2011; Modalsli, 2011):

$$G = \sum_{i=1}^K \sum_{j=1}^K \frac{\omega_i \omega_j m_i}{m} \left\{ 2\Phi\left(\frac{\log m_i - \log m_j + \sigma_i^2/2 + \sigma_j^2/2}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right) - 1 \right\}, \quad (7)$$

where  $\mu_k$  and  $\sigma_k^2$  are the parameters of the  $k$ th component,  $\omega_k$  is the probability associated

Functional	Component-specific Value	Mixture Value
Mean:	$m_k = \mathbb{E}(Y h = k)$	$m = \sum_{k=1}^K \omega_k m_k$
Variance:	$V_k = \text{Var}(Y h = k)$	$V = \sum_{k=1}^K \omega_k \{(m_k - m)^2 + V_k\}$
Bin Proportions:	$B_k = P(a \leq Y \leq b h = k)$	$B = \sum_{k=1}^K \omega_k B_k$

Table 2: Formulas for computing mixture functionals using the component functionals and mixture probabilities. The variable  $h$  is the component indicator so that, e.g.,  $\mathbb{E}(Y|h = k)$  is the mean of the  $k$ th mixture component.

with component  $k$ ,  $m_k = e^{\mu_k + \sigma_k^2/2}$ , and  $m = \sum_{k=1}^K \omega_k m_k$ .

To handle quantile estimates when the quantile function is not available in closed form, we use the delta method to “invert” the quantile data model. Suppose  $q$  is an estimate of the  $\tau$ th quantile,  $\Pi^{-1}(\tau)$ , with standard error  $E$  — we omit explicit dependence on  $r$ ,  $u$ , or  $\boldsymbol{\theta}_r$  here. We originally assumed that  $q \sim \text{N}(\Pi^{-1}(\tau), E^2)$ . Using the delta method we have  $\tau \sim \text{N}(\Pi(q), E^2/\pi(q)^2)$ . We call  $\tau$  an *inverted quantile estimate*. Since  $\pi$  depends on the tract-level parameter,  $\boldsymbol{\theta}_r$ , both the mean and the variance of  $\tau$ ’s data model now depend on  $\boldsymbol{\theta}_r$ . This makes MCMC more difficult, so we approximate this data model by choosing a specific value for  $E/\pi(q)$ . To do this, we fit a simple model for each tract separately in order to initially estimate  $\pi_{\boldsymbol{\theta}}(q)$ , then use it to construct an estimate of  $E/\pi_{\boldsymbol{\theta}}(q)$ .

The simple tract-level model we consider for this purpose for a given tract  $r$  is:

$$q_{ur}|\boldsymbol{\theta}_r \sim \text{N}(Q_u(\pi_{\boldsymbol{\theta}_r}), E_{ur}^2) \text{ for } u = 1, 2, \dots, U \text{ and}$$

$$\boldsymbol{\theta}_r \sim \text{prior},$$

where  $\pi_{\boldsymbol{\theta}_r}$  is either a normal or lognormal distribution, depending on the support of the data. These single-tract models can be fit very quickly, and provide a minimal loss of accuracy in estimates of the  $\pi_{\boldsymbol{\theta}_r}(q_{ur})$ s to the extent that the target variable is approximately (log)normally distributed within tracts, especially near the desired quantiles.

Given the estimates of the  $\pi_{\boldsymbol{\theta}_r}(q_{ur})$ s no matter their source, let  $V_{ur} = R_{ur}^2/\pi_{\boldsymbol{\theta}_r}(q_{ur})^2$ .

Then in each tract and for each inverted quantile estimate,  $\tau_{ur}$ , we find the parameters  $(\hat{a}_{ur}, \hat{b}_{ur})$  of the best fitting inverse gamma distribution to  $V_{ur}$ 's marginal posterior distribution using posterior replicates of  $V_{ur}$  from tract  $r$ 's single-tract model. These inverse gamma fits should be compared to the draws from the posterior distribution to ensure that the inverse gamma is a reasonable approximation, but in practice we find it is fairly close. If  $\tau_{ur}|V_{ur} \sim N(\Pi^{-1}(q_{ur}), V_{ur})$  with  $V_{ur} \sim \text{IG}(\hat{a}_{ur}, \hat{b}_{ur})$  where  $\text{IG}(a, b)$  denotes the inverse Gamma distribution with shape and scale parameters  $a$  and  $b$  respectively, then marginalizing out  $V_{ur}$  yields  $\tau_{ur} \sim T_{d_{ur}}(\Pi^{-1}(q_{ur}), \hat{D}_{ur})$ , where  $T_d(m, s)$  is the student's T distribution with degrees of freedom  $d$ , location parameter  $m$ , and scale parameter  $s$ , and where  $\hat{d}_{ur} = 2\hat{a}_{ur}$  and  $\hat{D}_{ur} = \sqrt{\hat{b}_{ur}/\hat{a}_{ur}}$ . This yields  $\hat{d}_{ur}$ s that are at least 15 for all tracts in each PUMA considered in Section 5 and typically much larger, so we further approximate these  $T$  distributions with

$$\tau_{ur}|\boldsymbol{\theta}_r \stackrel{\text{ind}}{\sim} N(\Pi_{\boldsymbol{\theta}_r}^{-1}(q_{ur}), \hat{E}_{ur}^2) \text{ for } r = 1, 2, \dots, R, \quad (8)$$

and for  $u$  such that  $q_{ur}$  is a quantile estimate,

where  $\hat{E}_{ur} = \hat{D}_{ur} \sqrt{\hat{d}_{ur}/(\hat{d}_{ur} - 2)}$  is the standard deviation of  $\tau_{ur}$ 's approximate  $T$  distribution. We use this as the data model for  $\tau_{ur}$  in all of the mixture models, i.e., Equation (8) replaces Equation (1) for all  $q_{ur}$ s that are quantile estimates.

### 3.2 Tract-level and stratum-level parameter models

Next we discuss how to specify Equations (3) and (4), the tract-level and stratum-level parameter models. To the extent possible, the tract-level model should mirror the stratum-level model since the stratum-level observations are supposed to “fill in the gaps” of the tract-level estimates. The details of these parameter models depend in large part on the specific parametric family chosen, and in this subsection we will focus on the normal and lognormal cases from Section 3.1.

### 3.2.1 iid base parameter models

First we consider what we call *iid base parameter models*, i.e., normal or lognormal distributions with parameters that are a priori iid across tracts or strata. In both cases,  $\boldsymbol{\theta}_r = (\mu_r, \sigma_r^2)$ , where  $\mu_r$  is the tract-level mean parameter and  $\sigma_r^2$  is the tract-level variance parameter, and similarly  $\tilde{\boldsymbol{\theta}}_s = (\tilde{\mu}_s, \tilde{\sigma}_s^2)$ . Then we assume that

$$\mu_r | \mu, \delta_\mu \stackrel{iid}{\sim} N(\mu, \delta_\mu^2), \quad \sigma_r^2 | \sigma^2, \delta_\sigma \stackrel{iid}{\sim} \text{IG}(\delta_\sigma^{-2}, \delta_\sigma^{-2} \sigma^2) \quad \text{for } r = 1, 2, \dots, R,$$

and

$$\tilde{\mu}_s | \mu, \tilde{\delta}_\mu \stackrel{iid}{\sim} N(\mu, \tilde{\delta}_\mu^2), \quad \tilde{\sigma}_s^2 | \sigma^2, \tilde{\delta}_\sigma \stackrel{iid}{\sim} \text{IG}(\tilde{\delta}_\sigma^{-2}, \tilde{\delta}_\sigma^{-2} \sigma^2) \quad \text{for } s = 1, 2, \dots, S,$$

where  $\boldsymbol{\theta} = (\mu, \sigma^2)$  is the top-level parameter. The parameters  $\delta_\sigma$  and  $\tilde{\delta}_\sigma$  are defined so that they are proportional to the standard deviations of  $\sigma_r^2$  and  $\tilde{\sigma}_s^2$ , respectively.

### 3.2.2 iid mixture parameter models

Next we develop *iid mixture parameter models*, i.e., finite mixtures of either normal or lognormal distributions with parameters that are iid across tracts or strata. Here we have an additional set of parameters to construct models for: the mixture weights. Suppose we have a  $K$  component mixture of normals or lognormals and let  $\mu_k$  and  $\sigma_k^2$  denote the parameters of the  $k$ th top-level normal or lognormal distribution with associated mixture weight  $\omega_k$ . Similarly, let  $\mu_{rk}$  and  $\sigma_{rk}^2$  denote the parameters of the  $k$ th tract-level normal or lognormal distribution for tract  $r$  with mixture weight  $\omega_{rk}$ , and let  $\tilde{\mu}_{sk}$  and  $\tilde{\sigma}_{sk}^2$  denote the parameters of the  $k$ th stratum-level normal or lognormal distribution for stratum  $s$  with mixture weight

$\tilde{\omega}_{sk}$ . Then the iid mixture model is:

$$\mu_{rk} | \mu_k, \delta_\mu \stackrel{iid}{\sim} \text{N}(\mu_k, \delta_\mu^2), \quad \sigma_{rk}^2 | \sigma_k^2, \delta_\sigma \stackrel{iid}{\sim} \text{IG}(\delta_\sigma^{-2}, \delta_\sigma^{-2} \sigma_k^2),$$

for  $r = 1, 2, \dots, R$  and  $k = 1, 2, \dots, K$ , and

$$\tilde{\mu}_{sk} | \mu_k, \tilde{\delta}_\mu \stackrel{iid}{\sim} \text{N}(\mu_k, \tilde{\delta}_\mu^2), \quad \tilde{\sigma}_{sk}^2 | \sigma_k^2, \tilde{\delta}_\sigma \stackrel{iid}{\sim} \text{IG}(\tilde{\delta}_\sigma^{-2}, \tilde{\delta}_\sigma^{-2} \sigma_k^2),$$

for  $s = 1, 2, \dots, S$  and  $k = 1, 2, \dots, K$ . To construct a model for the mixture weights we will work on the multivariate logit scale. Suppose that  $\omega_k = \exp(\xi_k) / \sum_{i=1}^K \exp(\xi_i)$  for  $k = 1, 2, \dots, K$ ,  $\omega_{rk} = \exp(\xi_{rk}) / \sum_{i=1}^K \exp(\xi_{ri})$  for  $k = 1, 2, \dots, K$  and  $r = 1, 2, \dots, R$ , and  $\tilde{\omega}_{sk} = \exp(\tilde{\xi}_{sk}) / \sum_{i=1}^K \exp(\tilde{\xi}_{si})$  for  $k = 1, 2, \dots, K$  and  $s = 1, 2, \dots, S$ . Then we assume

$$\begin{aligned} \xi_{rk} | \xi_k, \delta_\xi &\stackrel{iid}{\sim} \text{N}(\xi_k, \delta_\xi^2) \text{ for } r = 1, 2, \dots, R \text{ and } k = 1, 2, \dots, K, \\ \tilde{\xi}_{sk} | \xi_k, \tilde{\delta}_\xi &\stackrel{iid}{\sim} \text{N}(\xi_k, \tilde{\delta}_\xi^2) \text{ for } s = 1, 2, \dots, S \text{ and } k = 1, 2, \dots, K. \end{aligned}$$

This model is easy to reparameterize in order to speed up MCMC — see Appendix C.

### 3.2.3 Spatial base tract-level parameter models

The iid parameter models discussed above are straightforward to write down and implement, but often constructing plausible dependence structures will improve the parameter models. This is difficult in the stratum-level parameter model because we have no information to tell us about which strata are likely to be dependent. However, we typically can assume that tracts nearby each other have similar tract-level parameters. We also can typically expect some degree of nonstationarity — tracts in similar regions of different cities are more likely to have similar tract-level parameters, for example.

We construct spatially correlated tract-level parameter models using a method from

Bradley et al. (2017) to construct what we call a truncated integrated Karhunen-Loève expansion (TIKL expansion) using Obled-Creutin basis functions (Obled and Creutin, 1986). See Appendix D for a complete description and justification of this process. In short, let  $\boldsymbol{\psi}_r$  denote a known set of  $M$  basis functions evaluated at tract  $r$ . Then in the *spatial base tract-level parameter models* we assume that

$$\begin{aligned} \mu_r | \mu, \boldsymbol{\eta}_\mu, \delta_\mu &\stackrel{iid}{\sim} \text{N}(\mu + \boldsymbol{\psi}'_r \boldsymbol{\eta}_\mu, \delta_\mu^2) \quad \text{and} \\ \sigma_r^2 | \sigma^2, \boldsymbol{\eta}_\sigma, \delta_\sigma &\stackrel{iid}{\sim} \text{IG}(\delta_\sigma^{-2}, \delta_\sigma^{-2} \sigma^2 \exp\{\boldsymbol{\psi}'_r \boldsymbol{\eta}_\sigma\}) \quad \text{for } r = 1, 2, \dots, R, \text{ with} \\ \boldsymbol{\eta} | \delta_{\eta_\mu}, \delta_{\eta_\sigma} &\sim \text{Cauchy}(\mathbf{0}_{2M}, \boldsymbol{\Sigma}_\eta), \end{aligned}$$

and

$$\boldsymbol{\Sigma}_\eta = \begin{bmatrix} \delta_{\eta_\mu}^2 \mathbf{I}_M & \mathbf{0}_{M \times M} \\ \mathbf{0}_{M \times M} & \delta_{\eta_\sigma}^2 \mathbf{I}_M \end{bmatrix},$$

where  $\boldsymbol{\eta} = (\boldsymbol{\eta}'_\mu, \boldsymbol{\eta}'_\sigma)'$ ,  $\mathbf{0}_{2M}$  is an  $2M$ -dimensional vector of zeroes,  $\mathbf{0}_{M \times M}$  is an  $M \times M$  matrix of zeroes, and  $\mathbf{I}_M$  is the  $M \times M$  identity matrix. In this model  $\boldsymbol{\psi}'_r \boldsymbol{\eta}_\mu$  denotes the spatial random effect for  $\mu$  in tract  $r$ , while  $\boldsymbol{\psi}'_r \boldsymbol{\eta}_\sigma$  represents the spatial random effect for  $\sigma^2$  in tract  $r$ , and together they control the spatial dependence between the tract-level distributions of the target variable.

We use this low-dimensional parameterization of  $\boldsymbol{\Sigma}_\eta$  in order to reduce the size of the parameter space — with  $\boldsymbol{\Sigma}_\eta = \text{diag}(\delta_{\eta_i}^2 : i = 1, 2, \dots, 2M)$  we found that the model estimated each of the first  $M$   $\delta_{\eta_i}$ s to be essentially the same, and likewise for the second  $M$   $\delta_{\eta_i}$ s. Typically, Obled-Creutin basis functions imply that the random effect vector is uncorrelated in the univariate case, so this should not be too constraining. In practice we have found that assuming a fat-tailed distribution on  $\boldsymbol{\eta}$  makes MCMC via HMC easier and does not impact our results, hence the Cauchy assumption. We use the same basis functions for both  $\mu_r$  and

$\sigma_r^2$ , which significantly simplifies the basis function selection problem.

### 3.2.4 Spatial mixture tract-level parameter models

In *spatial mixture tract-level parameter models*, adding spatial dependence in this manner can quickly cause over-parameterization. To mitigate this we only allow the mean and variance parameters to have spatial dependence. With either a mixture of normals or of lognormals for tract  $r = 1, 2, \dots, R$  we assume:

$$\begin{aligned}\mu_{rk} | \mu_k, \boldsymbol{\eta}_\mu, \delta_\mu &\stackrel{iid}{\sim} \text{N}(\mu_k + \boldsymbol{\psi}'_r \boldsymbol{\eta}_{\mu_k}, \delta_\mu^2), \\ \sigma_{rk}^2 | \sigma_k^2, \delta_\sigma &\stackrel{iid}{\sim} \text{IG}(\delta_\sigma^{-2}, \delta_\sigma^{-2} \sigma_k^2 \exp\{\boldsymbol{\psi}'_r \boldsymbol{\eta}_{\sigma_k}\}), \\ \xi_{rk} | \xi_k, \delta_\xi &\stackrel{iid}{\sim} \text{N}(\xi_k, \delta_\xi^2),\end{aligned}$$

for  $k = 1, 2, \dots, K$  with

$$\boldsymbol{\eta} | \delta_{\eta_\mu}, \delta_{\eta_\sigma} \sim \text{Cauchy}(\mathbf{0}_{2KM}, \boldsymbol{\Sigma}_\eta),$$

and

$$\boldsymbol{\Sigma}_\eta = \begin{bmatrix} \delta_{\eta_\mu}^2 \mathbf{I}_{KM} & \mathbf{0}_{KM \times kM} \\ \mathbf{0}_{KM \times KM} & \delta_{\eta_\sigma}^2 \mathbf{I}_{KM} \end{bmatrix},$$

where  $\boldsymbol{\eta} = (\boldsymbol{\eta}'_{\mu_1}, \dots, \boldsymbol{\eta}'_{\mu_K}, \boldsymbol{\eta}'_{\sigma_1}, \dots, \boldsymbol{\eta}'_{\sigma_K})'$ . This structure allows for spatially dependent mixture models on each of the tract-level distributions, but guards against over-parameterization.

## 3.3 Priors, standardized data, and identification

To aid in the choice of priors for Equation (5) and to make MCMC easier, we standardize the data used to fit the models. The type of standardization used depends on the parametric

family used in the model. We consider mixtures of normals or lognormals, though both single-component models constitute special cases. In both cases we standardize based on the PUMS observations, but this also impacts the tract-level data models.

The vector of source observations is denoted by  $\mathbf{Z}$ . Let  $\overline{Z}$  denote its mean and  $S_Z$  denote its standard deviation. Then in the normal model we standardize the source observations in the natural way to create  $\mathbf{Z}^* = (\mathbf{Z} - \overline{Z})/S_Z$ . If  $q_{ur}$  is a quantile or mean estimate it is standardized in the same way:  $q_{ur}^* = (q_{ur} - \overline{Z})/S_Z$ . If  $q_{ur}$  is a standard deviation estimate the standardized version is  $q_{ur}^* = q_{ur}/S_Z$  while if it is a variance estimate the standardized version is  $q_{ur}^* = q_{ur}/S_Z^2$ . In the case of bin estimates the estimates do not need to be standardized, but the bounds defining the bins do need to be standardized. If  $b$  is a bound for one of the bin estimates, then the standardized version is  $b^* = (b - \overline{Z})/S_Z$ .

In the lognormal case we standardize on the log scale. Let  $\log \mathbf{Z}$  denote the vector of log source observations,  $\overline{\log Z}$  denote its mean, and  $S_{\log Z}$  denote its standard deviation. Then instead of assuming  $Z_{is} \stackrel{ind}{\sim} \sum_{k=1}^K \tilde{\omega}_{sk} \text{LN}(\tilde{\mu}_{sk}, \tilde{\sigma}_{sk}^2)$ , where  $\text{LN}(m, s^2)$  denotes the lognormal distribution, we assume  $\log Z_{js}^* \stackrel{ind}{\sim} \sum_{k=1}^K \tilde{\omega}_{sk} \text{N}(\tilde{\mu}_{sk}, \tilde{\sigma}_{sk}^2)$  where  $\log \mathbf{Z}^* = (\log \mathbf{Z} - \overline{\log Z})/S_{\log Z}$ . This changes the bounds of the bin estimates in an analogous way to the normal case:  $b^* = (\log b - \overline{\log Z})/S_{\log Z}$ . For other estimates it is typically easier to change the functional rather than the estimate by transforming the parameters that the functional depends on. Let  $\mu_{rk}^* = \mu_{rk} S_{\log Z} + \overline{\log Z}$  and  $\sigma_{rk}^* = \sigma_{rk} S_{\log Z}$ . Then the functionals can be computed with these parameters and the original mixture probabilities using the formulas in Tables 1 and 2.

We specify priors for both classes of models in the same way given these reparameterizations. We need priors for the  $\mu_{ks}$ , the  $\sigma_{ks}$ , the  $\xi_{ks}$ ,  $\delta_\mu$ ,  $\delta_\sigma$ ,  $\delta_\xi$ ,  $\tilde{\delta}_\mu$ ,  $\tilde{\delta}_\sigma$ ,  $\tilde{\delta}_\xi$ ,  $\delta_{\eta_\mu}$ , and  $\delta_{\eta_\sigma}$ . Our default prior assumes all of these parameters are mutually independent with  $\mu_k \sim \text{N}(0, 1)$  for  $k = 1, 2, \dots, K$ ,  $\sigma_k \sim \text{N}^+(1, 1)$  for  $k = 1, 2, \dots, K$ ,  $\xi_k \sim \text{N}(0, 1)$  for  $k = 1, 2, \dots, K$ ,  $\delta_\mu \sim \text{N}^+(0, 1)$ ,  $\delta_\sigma \sim \text{N}^+(0, 1)$ ,  $\delta_\xi \sim \text{N}^+(0, 0.5^2)$ ,  $\tilde{\delta}_\mu \sim \text{N}^+(0, 1)$ ,  $\tilde{\delta}_\sigma \sim \text{N}^+(0, 1)$ ,  $\tilde{\delta}_\xi \sim \text{N}^+(0, 0.5^2)$ ,  $\delta_{\eta_\mu} \sim \text{N}^+(0, 0.1^2)$ , and  $\delta_{\eta_\sigma} \sim \text{N}^+(0, 0.1^2)$ . Here  $\text{N}^+(m, s^2)$  denotes the nor-

mal distribution with mean  $m$  and variance  $s^2$  truncated from below at zero. In Appendix E we discuss this prior, the robustness of our results to our chosen priors, and simulate from the prior predictive distribution to demonstrate that the priors are not too constraining in general and in particular on the amount of spatial dependence allowed in the model.

## 4 Design-Model Hybrid Simulation Study

A key assumption in our models is that there are two separate but dependent populations, one divided spatially into tracts and the other divided into strata based on unobserved variables that potentially depend on space. In reality there is only one population, though we argued in Section 3 that this mismatch between model and reality is likely not a problem. To demonstrate this, we construct a synthetic population over the Boone County PUMA and its Census tracts, then repeatedly sample from it and create synthetic tract-level ACS estimates and a synthetic PUMS. Using these synthetic data products, we fit several of the models discussed above and evaluate them based on predictions of various features of the tract-level distributions of the target variable.

The population is generated to have the same number of households per tract as the 2014 ACS 5-year period estimates of household population for the Boone County, MO PUMA. We also divide the population into the same 106 strata that exist in the 2014 Boone County 5-year PUMS. The population of each stratum is assumed to be approximately proportional to  $n_s w_s$  where  $n_s$  is the sample size of stratum  $s$  in the PUMS, and  $w_s$  is the survey weight associated with stratum  $s$ . To fully specify the population we need to know number of households in each tract/stratum combination, though in reality this is unknown. Nevertheless, we know that the PUMS strata are based in part on Census tracts (U.S. Census Bureau, 2017c), so in our synthetic population we assign the households in a given stratum to a small number of tracts, typically only one or two, so that tract and stratum assignments are closely related.

Next, an income is generated for each household using a two-component mixture of lognormals with parameters that depend on both their tract and stratum. We do not fully describe how the synthetic population is generated here; instead, see the Supplementary Material for the R code (R Core Team, 2017) used to generate the incomes. The resulting tract-level distributions are mixtures of lognormals. Depending on how many strata are included within a given tract, the income distribution may be a two or three component mixture of lognormals, though most are approximated well by two components. Figure A.1 in the Supplementary Materials contains maps of the true tract-level means, medians, and standard deviations of income for the synthetic population.

Holding the population fixed, we repeatedly sample from it using a stratified random sampled based on the strata defined by the 2014 PUMS. Similar to the real ACS, approximately 10% of the population is sampled without replacement, and the sample size of each stratum is proportional to its sample size in the PUMS. We create two synthetic data products using this sample: ACS estimates and PUMS observations. The synthetic PUMS observations are created by sub-sampling approximately 38.5% of the sample in each stratum, which again mimics the real PUMS. The synthetic ACS estimates are created using the full sample and associated weights in each tract, and the associated standard errors are created using successive difference replication (Judkins, 1990; Fay and Train, 1995), the method used in the ACS (U.S. Census Bureau, 2017d,e). We construct bin estimates, median estimates, and mean estimates in order to fit the models. We use the same 12 bin estimates that are available in the ACS, defined by the following breaks: \$5,000, \$10,000, \$15,000, \$20,000, \$25,000, \$35,000, \$50,000, \$75,000, \$100,000, \$150,000, and \$200,000. We also construct Gini coefficient estimates and several other quantile estimates as hold-out estimates so that we can compare them to model-based estimates of the same quantities.

We fit several models using the generated data products: the iid base model (Base IID), the spatial base model with two basis functions (Base TIKL-2), the iid mixture model with

either two or three mixture components (Mix-2 IID and Mix-3 IID), and the spatial mixture model with two or three mixture components and either two or four basis functions (Mix-2 TIKL-2, Mix-2 TIKL-4, Mix-3 TIKL-2, and Mix-3 TIKL-4 respectively). In each model we used the priors discussed in Section 3.3. Before fitting the mixture models, we fit the single-tract models discussed at the end of Section 3.1 in order to approximate the data model for the median estimates. Each model was fit using `Rstan` (Stan Development Team, 2016) to do MCMC via HMC with four chains, and after a warm-up of 2,000 iterations per chain for tuning and burn-in, a further 2,000 iterations per chain were kept as draws from each model’s posterior distribution. See Appendix C for more details on MCMC implementation and diagnostics. Additionally, we fit the PR-LN procedure on the synthetic bin estimates. We execute this sampling and model fitting process on the same synthetic population 1008 times in parallel — due to the number of cores per sever node we had access to, 1008 iterations took the same amount of time as 1000 iterations.

Table 3 contains the difference in RMSE between the model-based estimates (or PR-LN) and the corresponding hold-out estimates as a percentage of the RMSE of the hold-out estimates, averaged over each tract. Table F.1 in Appendix F of the Supplementary Materials contains results for additional percentile estimates. Each model-based estimate is the posterior median of the relevant estimand from the posterior predictive distribution of the process model, i.e., Equation (6), and each RMSE is the RMSE of a given hold-out or model-based estimated compared to the corresponding value in the population. So a value of  $-10$  in the table indicates that the model-based estimates have an RMSE which are 10% smaller than the RMSE of the design-based estimates, on average across tracts. In short, we do not have any difficulty predicting unobserved features of tract-level distributions with the mixture models despite the fact that the model assumes that there are two populations instead of one. Treating the data as coming from two separate but dependent populations instead of a single population simplifies modeling and computation without preventing the

model from successfully predicting unobserved features of tract-level distributions.

	Base Models			Mix-2 Models			Mix-3 Models		
	PR-LN	IID	TIKL-2	IID	TIKL-2	TIKL-4	IID	TIKL-2	TIKL-4
10th	<b>-1.94</b>	284.03	311.44	2.55	10.42	<b>0.15</b>	0.97	4.34	<b>-7.59*</b>
20th	-4.02	170.79	188.47	0.06	-3.05	<b>-8.82</b>	-0.07	<b>-5.73</b>	<b>-13.16*</b>
30th	-2.79	46.76	54.57	1.49	<b>-7.39</b>	-7.28	2.14	<b>-8.60</b>	<b>-11.01*</b>
40th	<b>-4.64</b>	34.05	34.60	-3.02	0.88	1.55	-3.97	<b>-4.05</b>	<b>-7.08*</b>
50th	<b>-1.58</b>	144.29	142.12	-1.56	3.76	2.76	<b>-2.27</b>	2.67	<b>-2.64*</b>
60th	<b>-3.34*</b>	235.49	232.60	<b>-1.00</b>	4.73	3.34	<b>-1.67</b>	3.61	-0.82
70th	<b>-4.29</b>	236.34	233.81	<b>-3.94</b>	13.09	12.34	<b>-4.73*</b>	4.80	0.31
80th	<b>-0.83</b>	131.29	130.24	<b>-10.35</b>	24.34	25.70	<b>-11.49*</b>	5.98	5.81
90th	<b>2.25</b>	84.31	82.28	<b>-12.01</b>	28.69	35.31	<b>-15.16*</b>	8.28	42.27
Gini	<b>10.64</b>	408.49	402.15	<b>6.27</b>	14.46	12.70	<b>4.27*</b>	13.98	14.47

Table 3: Difference between RMSE of model-based estimates and RMSE of hold-out estimates as a percentage of RMSE of hold-out estimates for the fixed population, averaged over each tract. The three best performing models for each estimate (in each row) are bold and best performing model is additionally starred. Based on 1008 repeated samples from the synthetic population.

Neither of the base models do as well as the mixture models, with average RMSEs (averaged over tracts) typically at least 50% worse than the RMSEs of the hold-out estimates, and often 100% or even 300% worse depending on the quantity being estimated. This is not surprising since we designed our synthetic population to have income distributed according to a mixture of lognormals within each tract. Notably, there is little difference between the iid and spatial base models. The mixture models with two components do much better than the base models, with average RMSEs which are typically no more than 30% worse than the hold-out estimates, and often better. The additional flexibility afforded by the second mixture component is indispensable. In a small range of the lower percentiles of the distribution — the 10th through 35th percentiles — Mix-2 TIKL-4 significantly improves on the Mix-2 IID in terms of average RMSE, though for every other estimate the Mix-2 IID is comparable and often is superior, particularly in the upper percentiles. The Mix-2 IID model is never more than 7% worse than the hold-out estimates in terms of average RMSE, and is typically as good or better.

In the three-component mixture models we again see that adding spatial dependence only appears to help estimate a small number of quantiles of the tract-level distributions, and otherwise Mix-3 IID model is comparable and often significantly better than the Mix-3 TIKL-2 model in terms of average RMSE. The Mix-3 IID models significantly improve on their Mix-2 counterparts for many of the estimates considered, particularly in the upper quantiles of the distribution. Similarly in the lower quantiles of the tract-level distributions, Mix-3 TIKL-2 significantly improves on Mix-2 TIKL-2 and sometimes even Mix-2 TIKL-4.

The best performing model is Mix-3 IID, unless the the 10th through 40th percentiles are key quantities of interest. Mix-3 TIKL-4 does the best in that range of the distribution and works well through most of the rest of the distribution, with the exception of the far right tail and as a result the Gini coefficient. All of the mixture models provide reasonably estimates of most of the unobserved features of tract-level income distributions we considered, typically with the only major exceptions being in one tail of the distribution or the other. The PR-LN procedure is worse than our best model for any given estimand, except for a few quantiles near the middle of the distribution where PR-LN is the best. It performs fairly well overall, but is significantly worse than our best performing models on several quantiles while never offering a compelling advantage on other quantiles. A practical advantage of PR-LN is that it is easy to compute due to its simplicity.

## **5 Multi-Source ACS Estimation for Unpublished Tabulations**

We use our modeling framework to estimate U.S. Census tract-level income distributions using 2015 ACS 5-year period estimates of features of tract-level income distributions and the 2015 5-year PUMS. A few households in the PUMS have zero or negative income in both PUMS files; so to fit the lognormal models we use a small offset to make all of the

observations strictly positive. This offset must be taken into account in all of the tract-level estimate data models and posterior predictive distributions. Let  $Z_{\text{offset}} = \min\{\min\{\mathbf{Z}\} - 1, 0\}$  denote the offset, which is defined to be negative or potentially zero. Then the tract-level estimate data model, stratum-level observation data model, and tract-level latent process model (predictive distribution) become, respectively:

$$\begin{aligned} Z_{is} | \tilde{\boldsymbol{\theta}}_s &\stackrel{\text{ind}}{\sim} Z_{\text{offset}} + \pi_{\tilde{\boldsymbol{\theta}}_s} && \text{for } i = 1, 2, \dots, n_s \text{ and } s = 1, 2, \dots, S, \\ q_{ur} | \boldsymbol{\theta}_r &\stackrel{\text{ind}}{\sim} N(Q_u(Z_{\text{offset}} + \pi_{\boldsymbol{\theta}_r}), E_{ur}^2) && \text{for } u = 1, 2, \dots, U \text{ and } r = 1, 2, \dots, R, \end{aligned}$$

and

$$Y_{it} | \boldsymbol{\theta}_r \stackrel{\text{ind}}{\sim} Z_{\text{offset}} + \pi_{\boldsymbol{\theta}_r} \quad \text{for } i = 1, 2, \dots, N_r \text{ and } r = 1, 2, \dots, R.$$

Computing  $Q_u(Z_{\text{offset}} + \pi_{\boldsymbol{\theta}_r})$  depends on  $Q_u(\cdot)$  in general, but is usually a simple function of  $Q_u(\pi_{\boldsymbol{\theta}_r})$ . See Appendix G for details. We also attempted modeling  $Z_{\text{offset}}$  as an unknown parameter, but this made fitting the models significantly more difficult for little benefit (results not reported here).

We fit the models to four separate PUMAs using 2015 5-year ACS estimates and PUMS: PUMA 821 in Colorado (a wealthy rural PUMA south of Denver), PUMA 3502 in Illinois (a wealthy PUMA in the northern portion of Chicago), PUMA 600 in Missouri (Boone County, MO, a college town and rural outlying areas), and PUMA 600 in Montana (a sparsely populated rural PUMA). Figure A.1 in the Supplementary Materials contains maps of each PUMA and each of their Census tracts, shaded according to the 2015 ACS 5-year period estimate of median household income. We fit the following models to data from each PUMA: the iid base model (Base IID), the spatial base model with two basis functions (Base TIKL-2), the iid mixture model with two or three mixture components (Mix-2 IID and Mix-3), and

the spatial mixture model with either two or three mixture components and either two or four basis functions (Mix-2 TIKL-2, Mix-2 TIKL-4, Mix-3 TIKL-2, Mix-3 TIKL-4). There are two exceptions to this: in the IL and MT PUMAs, MCMC is very difficult for the Mix-3 TIKL-2 and Mix-3 TIKL-4 models, so we omit them. We used the priors discussed in Section 3.3 for all models. To fit each model we used `Rstan` (Stan Development Team, 2016) to do MCMC via HMC with four chains, a warm-up of 4,000 iterations per chain for tuning and burn-in, and a further 4,000 iterations per chain were kept as draws from each model’s posterior distribution. See Appendix C for more detail on MCMC implementation and diagnostics.

Table 4 compares the hold-out estimates to model-based estimates of the same quantities in the MO PUMA, while Tables F.2, F.3, and F.4 in Appendix F of the Supplementary Materials do the same for the IL, MO, and MT PUMAs, respectively. We will focus on the MO PUMA, but the others are similar. The model-based estimates are posterior medians of the relevant estimands from the posterior predictive distribution, and we report the absolute percentage difference between the model-based estimate and the hold-out estimate, averaged across all tracts in the PUMA. In general, each model gets reasonably close to the hold-out estimates. Our best models produce estimates of these features which are within 10% of the hold-out estimates on average, indicating that the models are working well at predicting unobserved features of tract-level income distributions. Notably, the base lognormal models do reasonably well in comparison to the more complicated mixture models, while in the simulation study in Section 4 they did not. Adding spatial dependence to the base model again has little impact on the results — fundamentally the tract-level parameters in the base lognormal model are so constrained by the tract-level estimates there is not much room for dependence to alter the tract-level parameters.

The additional flexibility granted by using a mixture model does increase the ability of the models to faithfully reproduce the held-out 20th percentile estimates, and moving

Model	20th	40th	60th	80th	95th	Gini
PR-LN	<b>2.96*</b>	<b>2.80*</b>	<b>3.35*</b>	<b>3.83*</b>	<b>5.14*</b>	<b>4.30*</b>
Base IID	26.27	6.42	7.59	6.62	<b>6.70</b>	9.10
Base TIKL-2	26.08	6.30	7.57	6.62	<b>6.67</b>	9.04
Mix-2 IID	12.50	6.56	<b>5.70</b>	<b>5.15</b>	7.06	<b>6.27</b>
Mix-2 TIKL-2	10.16	4.95	7.18	8.55	15.26	9.49
Mix-2 TIKL-4	10.32	5.16	6.94	8.46	15.43	9.83
Mix-3 IID	10.07	5.65	<b>5.52</b>	<b>5.24</b>	6.96	<b>5.69</b>
Mix-3 TIKL-2	<b>6.78</b>	<b>4.69</b>	6.46	6.89	12.85	8.78
Mix-3 TIKL-4	<b>6.78</b>	<b>4.74</b>	6.34	6.96	13.08	8.84

Table 4: Absolute percentage difference between the model-based estimates and the hold-out ACS estimates averaged across all tracts for each model and hold-out estimate type for the MO PUMA. The three best performing models for each estimate (in each column) are bold, and the best performing model is additionally starred. 4 of the 29 tracts are omitted from the calculation of the 95th percentile column because they did not have a published estimate.

up to three mixture components helps a bit more. Adding spatial dependence to the two-component and three-component mixture models helps them capture the lower percentiles of income, especially the 20th percentile, and additional basis functions tends to improve these estimates further. For the upper percentiles of the income distribution as well as the Gini coefficient, adding spatial dependence to the mixture models tends to hurt their estimates, consistent with the results from the simulation study in Section 4. The best performing models are again the three-component mixture models, though which of the two is better depends on how important the lower quantiles of tract-level income distributions are to inferential goals relative to the upper quantiles.

## 6 Discussion

In both the simulation study in Section 4 and income example in Section 5, we found that adding spatial dependence to the mixture models tends to improve the model-based estimates of percentiles in the lower half of the distribution, but hurts the estimates of percentiles in the upper half of the distribution. This result is somewhat surprising. Since the tract-level

distributions were designed to be a mixture of lognormals in the simulation study, this cannot be explained by a mismatch between the true tract-level distributions and parametric family used in the model. We performed an exploratory analysis of several of the source statistics and the held out direct estimates using the Moran’s I test of spatial association with a two-sided alternative using a binary weight matrix (Banerjee et al., 2015, Section 4.1). Table A.2 in Appendix A of the Supplementary Materials contains the resulting p-values. There is virtually no spatial dependence in the CO estimates, some dependence in only a few of the MT estimates, in most of the IL estimates, and in all of the MO estimates. But whether the test showed significant spatial dependence for the estimates of a given percentile in a given PUMA appears to have no relationship with whether or not the model-based estimates for that percentile were improved by adding spatial dependence.

More important than spatial dependence in the estimates is how spread out the bin estimates are throughout a given PUMA’s income distribution. Table A.3 in Appendix A of the Supplementary Materials displays the number of the bin estimates which are strictly below each of the held out percentile estimates for each PUMA considered in Section 5. For wealthy PUMAs, such as the CO and IL PUMAs, half or more of the bin estimates are below the 40th percentile. This means that there is little information in the bin estimates to learn about percentiles outside of the lower portion of the distribution. Perhaps because of this paucity of information, the spatial models do well for a larger range of the distribution in these two PUMAs. The MT PUMA is near as wealthy as those two PUMAs, but the bins are spread out a bit more and the spatial models do not help quite as much in the upper quantiles.

In Section 4 our best models tend to do better than PR-LN when compared to the truth, but in Section 5 they almost always do worse than PR-LN when compared to held out direct estimates. This may be because the assumptions of PR-LN are a better approximation to U.S. tract-level income distributions than they are to the income distributions

in the simulation study, but it also may be because the model-based estimates are further from the hold-out estimates to the extent that they are closer to the corresponding unobserved population values. In any case, our framework improves on PR-LN in several ways. First, it allows the data-user to use a wide variety of estimates of features tract-level income distributions, such as common moment and quantile estimates, rather than only using bin estimates. Further, it takes into account the coarse-scale observations that are often available, i.e. the PUMS, allowing in principle for better interpolation within bins. Second, our framework takes into account the standard error of the estimates, and further uncertainty quantification of model-based estimates is straightforward in the Bayesian paradigm. Third, our framework includes the ability to exploit spatial dependence in the underlying income distributions through low-rank spatially dependent tract-level parameter models via a TIKL expansion. This does come at a computational cost, but also tends to improve estimates of the lower quantiles of the income distribution. Finally, our framework is more general than PR-LN in the sense that it applies to variables other than income. For example, it can be applied to other dollar-denominated unit-level variables, unit-level counts that could be modeled continuously such as Age, or unit-level proportions, though each of these may require a different choice of base distribution. The methodology can also be applied to similar variables from other statistical agencies so long as observations from a coarse-scale geography and estimates of features of distributions from a fine-scale geography exist. To this end, we included formulas in Section 3.1 for mixtures of normals in addition to mixtures of lognormals, and include `Stan` code for mixtures of normals in the Supplementary Material.

To the extent that PR-LN outperforms our model-based estimates on income data, we can remedy this through a more judicious choice of base distribution, though there is work to be done to ensure that we can perform trustworthy MCMC with alternative base distributions at a reasonable computational cost. Our results also suggest that we could improve the model somehow allowing for spatial dependence in the lower portion of the distribution but

not the upper portion, but in mixture models it is unclear how to accomplish this since each parameter of the mixture distribution impacts every quantile.

In the ACS/PUMS case, some issues do perfectly generalize to other settings. The definition of the PUMAs may change during that 5-year period, complicating the use of the PUMS. We deliberately chose PUMAs which were unchanged during the 5-year period, but data-users may not have this luxury. In principle the modeling framework can be adjusted to accommodate this, though we have not explored any alternatives. We have experimented with scaling up our models to multiple PUMAs, and we find that there is little benefit to modeling several PUMAs jointly while the additional computational costs are substantial. Indeed, MCMC was already fairly difficult in the larger PUMAs we considered, both in terms of number of tracts and number of PUMS observations, such as the IL and MT PUMAs.

As a final note, the source statistics are an important piece of information that help pin down the tract-level distributions — without good source statistics that cover a wide variety of features of the distribution of the target variable, no model is likely to produce quality estimates of the tract-level distributions. A large suite of bin or quantile estimates is enough to pin them down in most cases, but every tract-level estimate helps. In Section 5 we held out several quantile estimates and Gini coefficient estimates in order to compare them to model-based estimates of the same quantities, but in any real-world application of this methodology they should be included in the final model. Including additional quantiles is straightforward, and we included a line in Table 1 with the Gini coefficient for the lognormal distribution, while Equation (7) contains the Gini coefficient for a finite mixture of lognormal distributions.

## 7 Acknowledgements

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. This article is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the NSF or the U.S. Census Bureau. The authors thank Noel Cressie for helpful discussion.

The computation for this work was performed on the high performance computing infrastructure provided by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia, MO.

## 8 Supplementary Material

**Online Appendix:** Includes several appendices adding relevant detail to the paper.

**Appendix A: Exploratory tables and figures.** Includes various tables and figures referenced throughout the paper that are useful, but not necessary, for understanding the data and results in this paper.

**Appendix B: Comparing tract and stratum-level distributions.** Shows how to compare tract-level and stratum-level distributions in order to check a key model assumption.

**Appendix C: MCMC and reparameterization.** Includes details on how MCMC was performed, including how the model was reparameterized to facilitate MCMC.

**Appendix D: Truncated integrated Karhunen-Loève expansions.** Includes details on the TIKL expansion used in the spatial models.

**Appendix E: Priors and posterior robustness.** Includes information on the sensitivity of our results to our priors.

**Appendix F: Additional results.** Includes additional tables of results for Sections 4 and 5, referenced in the paper.

**Appendix G: Functionals for shifted distributions.** Includes formulas for the functionals corresponding to source statistics under the shifted lognormal and shifted mixture of lognormal distributions used in Section 5.

## A Exploratory Tables and Figures

This appendix contains several tables and figures that useful for understanding the data that were referenced in the main text

Tract	Bins									
	<10	≥10 <15	≥15 <25	≥25 <35	≥35 <50	≥50 <75	≥75 <100	≥100 <150	≥150 <200	≥200
2	9.8	9.3	25.8	13.7	20.4	14.3	4.0	2.8	0.0	0.0
3	31.9	16.0	21.1	12.4	3.3	6.8	4.1	1.9	1.1	1.4
5	46.6	8.3	19.5	6.4	10.3	3.8	1.7	0.9	2.5	0.0
6	7.2	3.2	4.4	3.6	16.1	17.3	14.2	23.0	5.8	5.4
7	10.5	10.8	15.3	15.7	16.6	18.9	9.1	2.7	0.4	0.0
9	17.6	10.3	21.5	14.6	18.4	10.4	4.9	2.2	0.0	0.0

Table A.1: Bin estimates for selected tracts in PUMA 600 (Boone County) in MO. All estimates are 2015 ACS 5-year period estimates, and come from ACS Table S1901. Each bin estimate is the percentage of households in that tract with an income within a set of bounds, including the lower bound but excluding the upper bound. Both bounds are denominated in \$1,000. The ACS tables also include an associated margin of error for each estimate (not displayed here).

Boone County, MO; Median Income

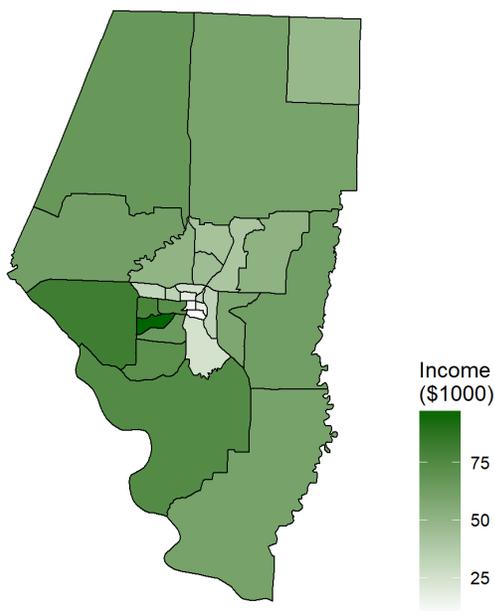


Figure A.1: An example PUMA with nested tracts: PUMA 600 (Boone County) in MO. Tracts are shaded according to 2015 ACS 5-year estimates of median household income.

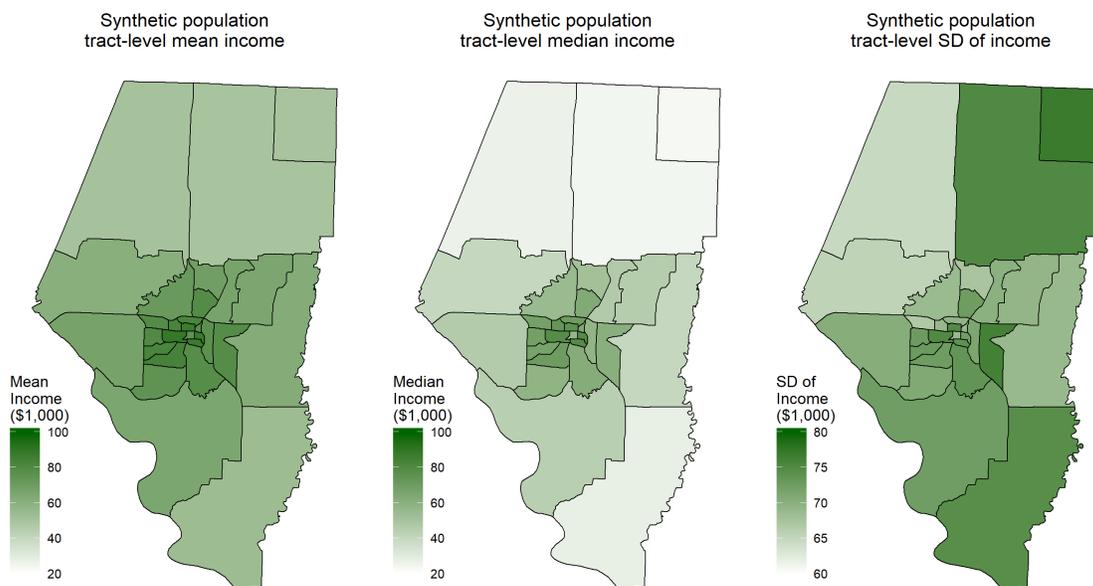


Figure A.1: True tract-level means, medians, and standard deviations of income for the synthetic population. The first two exhibit a noticeable inside-out spatial pattern, while the third is a bit different but still appears to have spatial dependence.

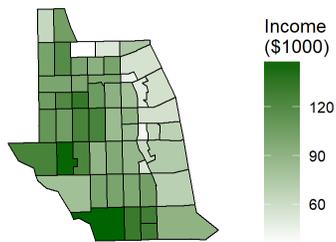
Estimate	CO	IL	MO	MT
Mean	0.35	0.00	0.00	0.03
Median	0.88	0.00	0.00	0.07
< 5	0.40	0.00	0.00	0.53
5-10	0.51	0.73	0.00	0.37
10-15	0.54	0.79	0.00	0.72
15-20	0.07	0.52	0.00	0.88
20-25	0.82	0.63	0.00	0.27
25-35	0.62	0.00	0.09	0.03
35-50	0.47	0.00	0.00	0.71
50-75	0.92	0.06	0.00	0.95
75-100	0.14	0.16	0.00	0.80
100-150	0.01	0.12	0.00	0.12
$\geq 150$	0.78	0.00	0.00	0.55
20th %tile	0.78	0.00	0.00	0.06
40th %tile	0.82	0.00	0.00	0.03
60th %tile	0.84	0.00	0.00	0.16
80th %tile	0.80	0.00	0.00	0.04
Gini	0.91	0.33	0.02	0.55

Table A.2: P-values of Moran’s I tests for spatial dependence among each estimate type in each PUMA used in Section 5, using the binary weight matrix. Other choices of the weight matrix did not materially affect this analysis. All bin estimates are denominated in \$1,000, i.e., 5-10 denotes the bin including incomes of at least \$5,000 but below \$10,000.

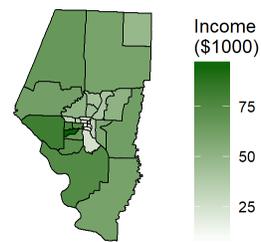
	20th	40th	60th	80th
CO	7	8	9	10
IL	5	7	9	10
MO	2	5	7	8
MT	4	6	7	8

Table A.3: Number of bin estimates strictly below each percentile, of 12 total, for each PUMA used in Section 5.

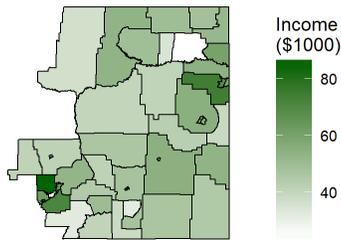
PUMA 3502, IL; Median Income



Boone County, MO; Median Income



PUMA 600, MT; Median Income



PUMA 821, CO; Median Income

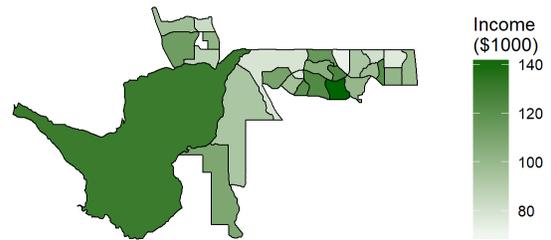


Figure A.1: Maps of each PUMA used in this section with each of the Census tracts. Each tract within each PUMA is shaded according to the 2015 ACS 5-year period estimate of median household income.

## B Comparing Tract and Stratum-Level Distributions

In Section 3 we argued that a key requirement for the model to be successful is that tract-level distributions look similar to stratum-level distributions. In this appendix we construct visual diagnostics to evaluate this assumption for 2014 and 2015 ACS data as well as for both the fixed and even populations we constructed in the simulation study in Section 4. To do this we construct stratum-level density histograms with bins equal to the bins from the observed tract-level bin estimates, and compare these to tract-level plots of density estimates constructed from the bin estimates. The bins in the ACS data are defined by the following breaks: \$5,000, \$10,000, \$15,000, \$20,000, \$25,000, \$35,000, \$50,000, \$75,000, \$100,000, \$150,000, and \$200,000, defining 12 income categories. For the purposes of these plots, we will assume that the bottom category is  $[\$0, \$5,000)$  and the top category is  $[\$200,000, \$250,000]$ . Any observed incomes below \$0 or above \$250,000 in the PUMS are set to \$0 and \$250,000 respectively for the purposes of these plots. The tract-level density estimates are computed by dividing each bin estimate by the width of their respective bins.

Figure B.0 contains the tract-level density plots for the 2015 5-year ACS estimates for the CO PUMA and Figure B.0 contains the tract-level density histograms for the 5-year PUMS data from the CO PUMA. Similarly, Figures B.0 and B.0 contain the plots for the IL PUMA, Figures B.0 and B.0 for the MO PUMA, and Figures B.0 and B.0 for the MT PUMA. For the most part the shapes of the tract-level and stratum-level distributions look pretty similar with the major exception that several tracts in the MO have much of their mass in the lower income categories while none of the MO strata look similar. This pattern occurs for the some of the other PUMAs as well, but the difference is not nearly as stark. This suggests that the model will have trouble with those tracts, and is probably a reason why the model-based estimates have more trouble in the lower quantiles of the income distributions

in Section 5. This problem often implies that the tract-level parameters for those tracts are weakly identified which, in turn, implies that MCMC is a bit more difficult — requiring a higher target Metropolis acceptance rate to avoid divergent transitions (see Section C).

This difference also implies that priors on tract-level parameters cannot be arbitrarily tight in order to allow the stratum-level side of the model to identify the tract-level parameters when the complexity of the model is high relative to amount of tract-level data. For example tight priors would be desirable when trying to fit a mixture model with many mixture components. But a tight prior will not work well because when the observed data suggests that a given tract’s distribution has a shape which is substantially different from all stratum-level distributions, the model may not be flexible enough to accommodate this difference. Despite this difference between some of the tract-level and stratum-level distributions, we saw in Section 5 that the model does a good job of reproducing the hold-out estimates.

### 2015 5-year ACS Tract-level Density Estimates for the CO PUMA

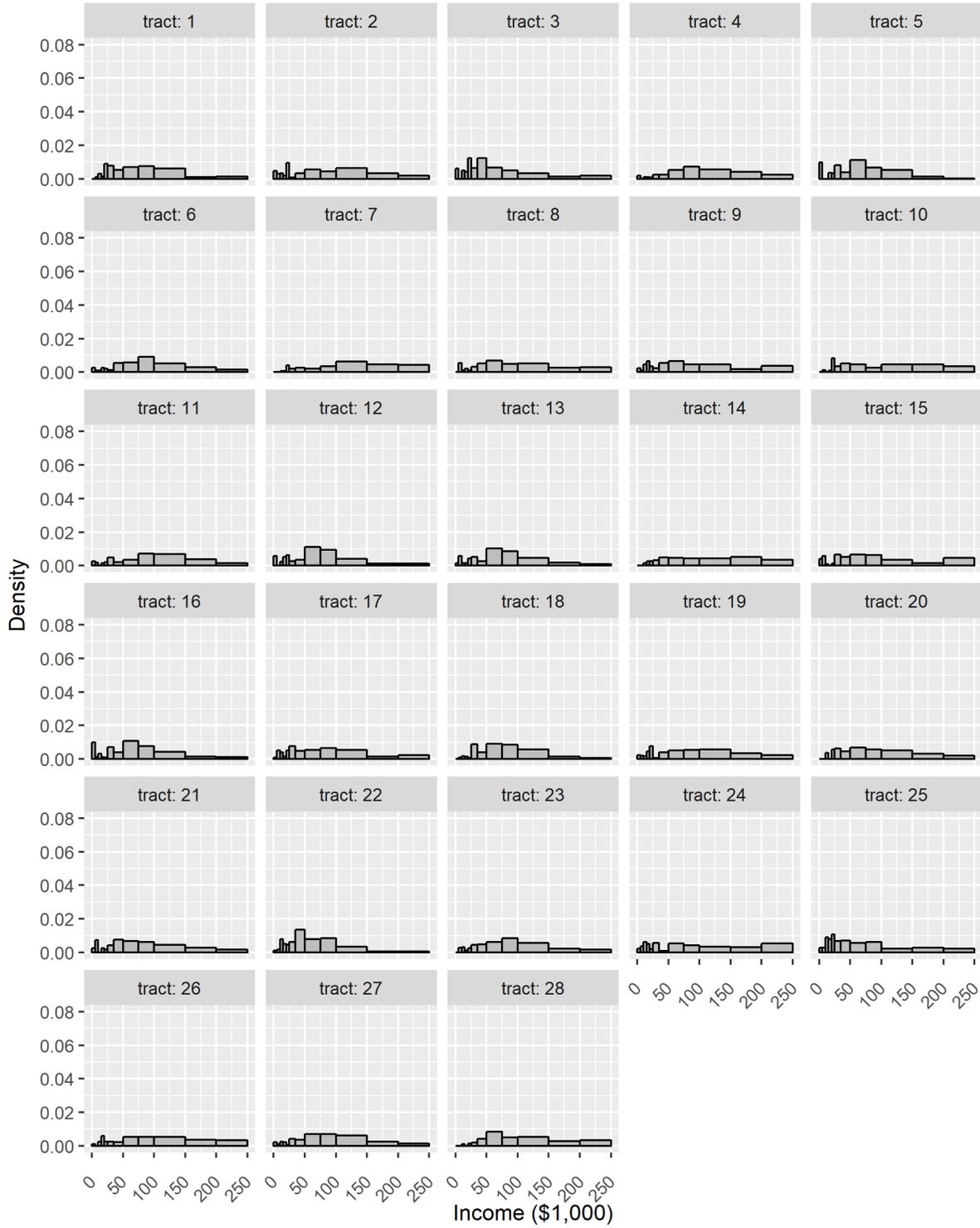


Figure B.0: 2015 5-year ACS density estimates for PUMA 821, CO.

### 2015 5-year PUMS Stratum-level Density Histograms for the CO PUMA

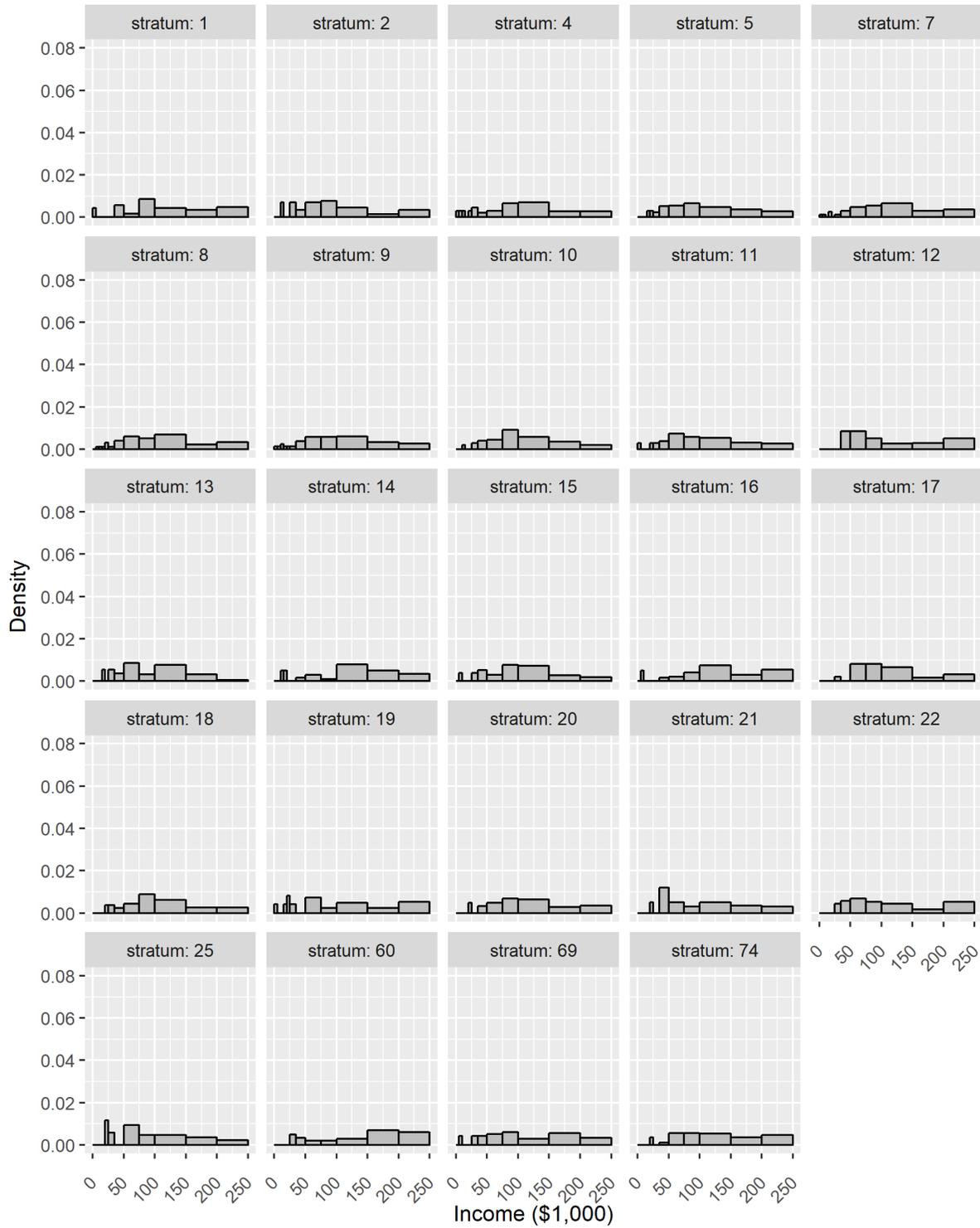


Figure B.0: 2015 5-year observed PUMS densities for strata with at least 17 observations in PUMA 821, CO.

### 2015 5-year ACS Tract-level Density Estimates for the IL PUMA

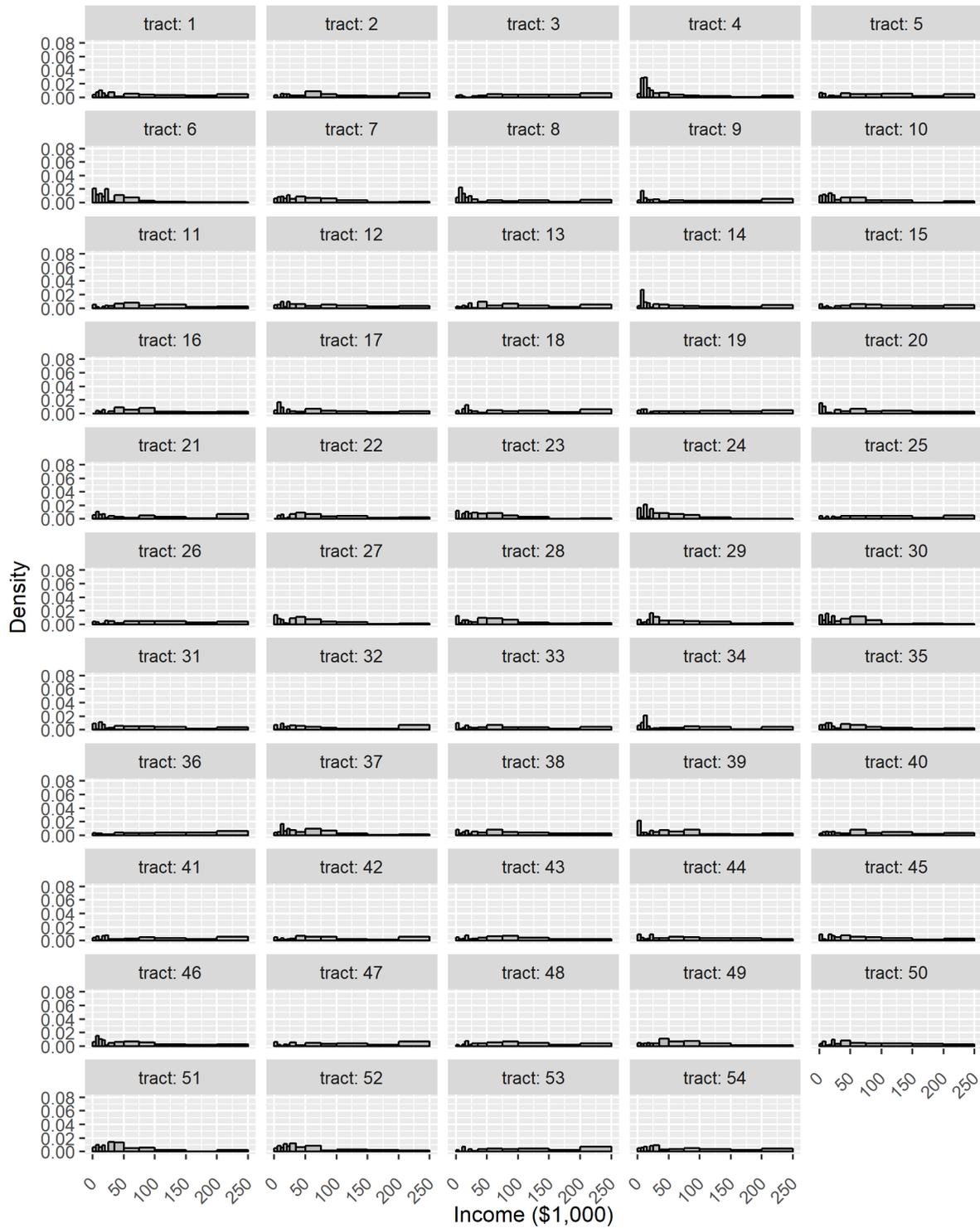


Figure B.0: 2015 5-year ACS density estimates for PUMA 3502, IL.

2015 5-year PUMS Stratum-level Density Histograms for the IL PUMA

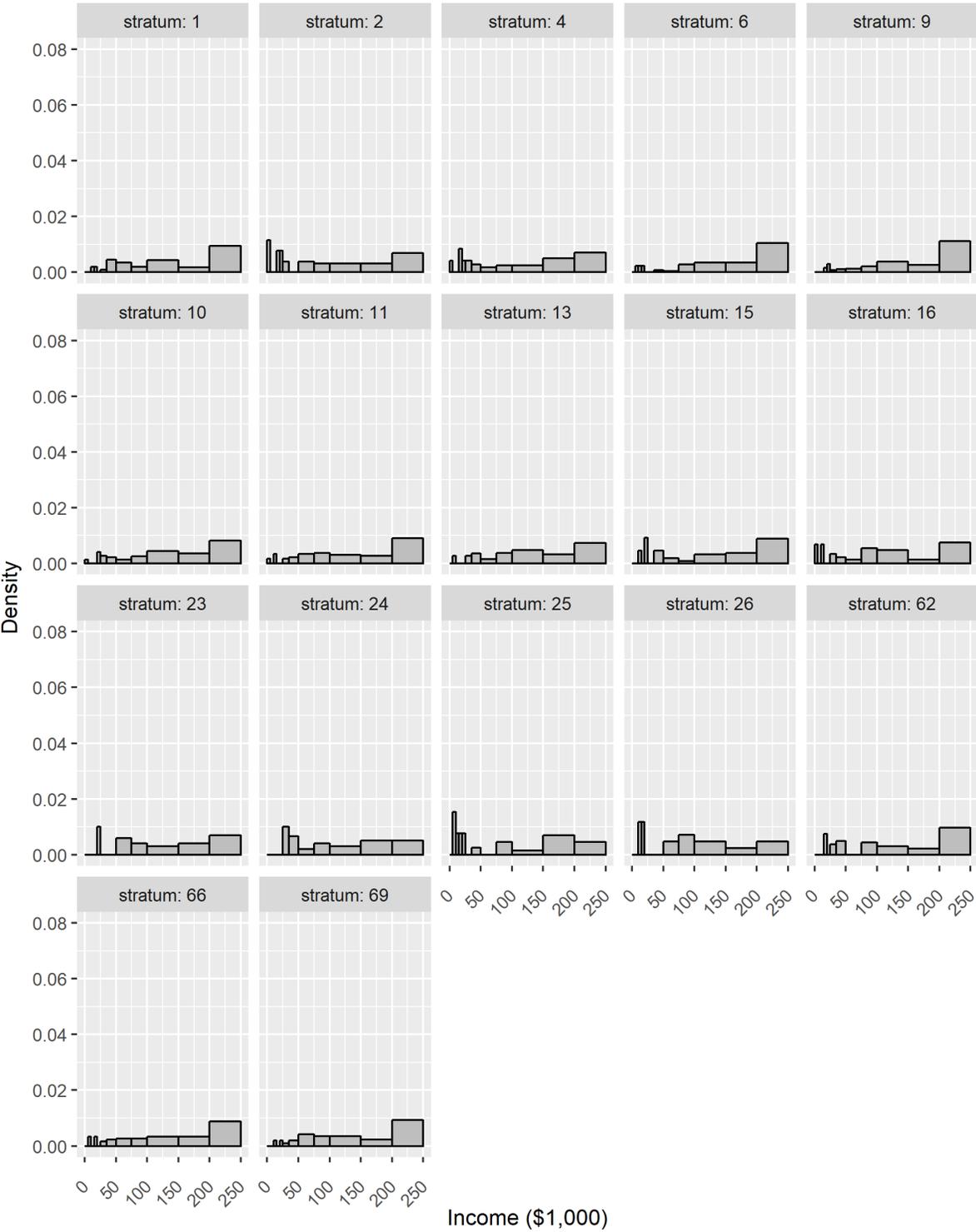


Figure B.0: 2015 5-year observed PUMS densities for strata with at least 17 observations in PUMA 3502, IL.

### 2015 5-year ACS Tract-level Density Estimates for the MO PUMA

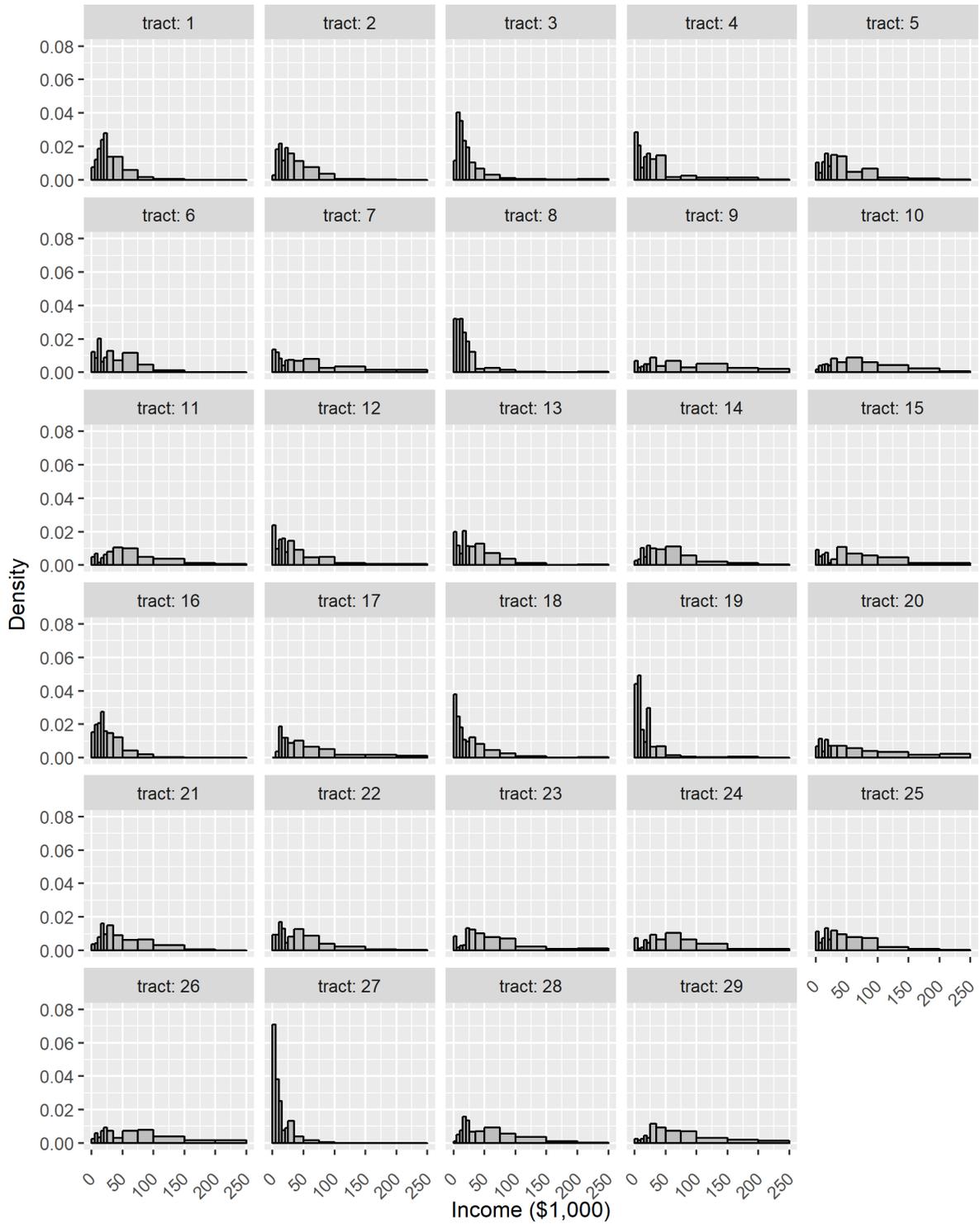


Figure B.0: 2015 5-year ACS density estimates for PUMA 600, MO (Boone County).

### 2015 5-year PUMS Stratum-level Density Histograms for the MO PUMA

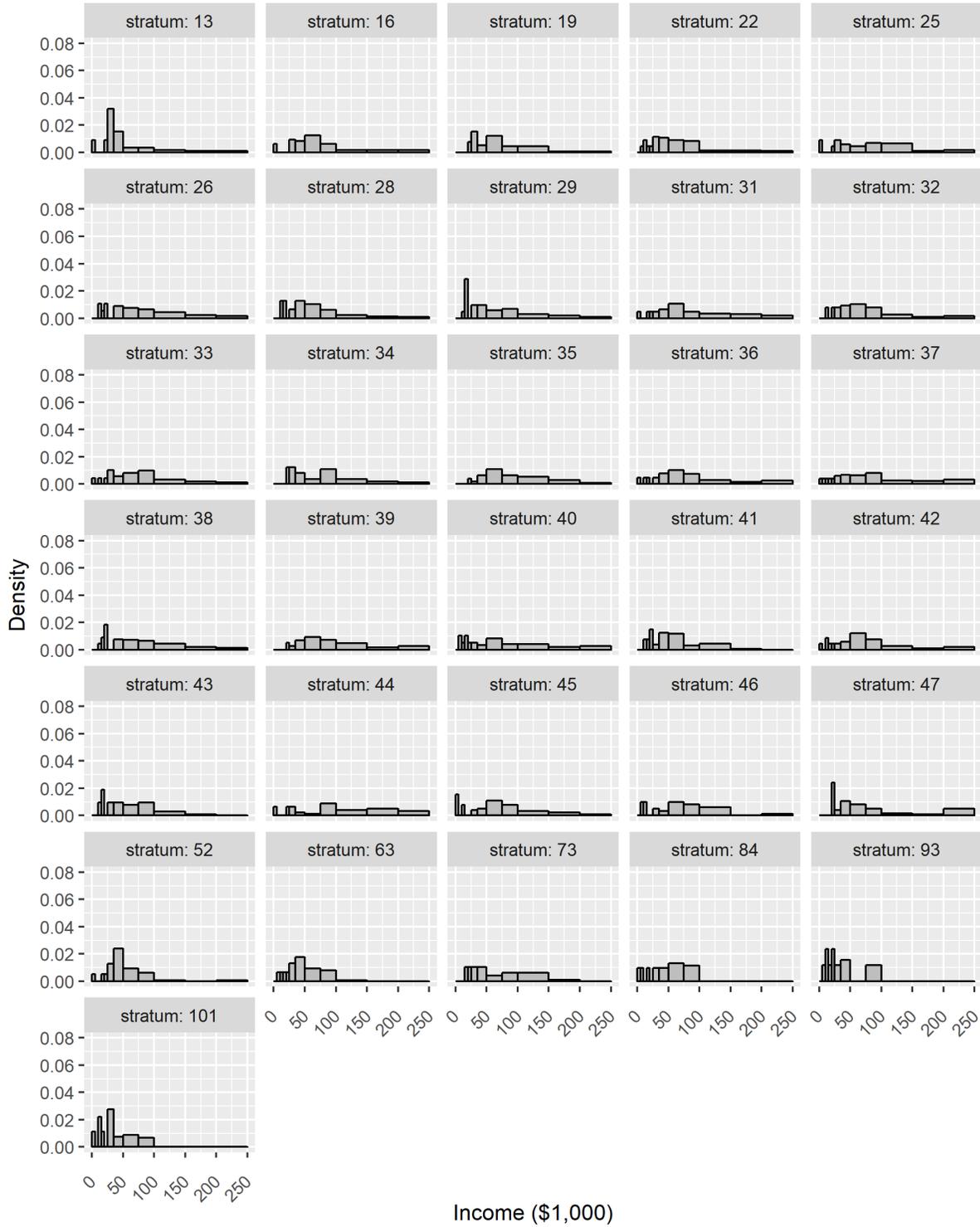


Figure B.0: 2015 5-year observed PUMS densities for strata with at least 17 observations in PUMA 600, MO (Boone County).

### 2015 5-year ACS Tract-level Density Estimates for the MT PUMA

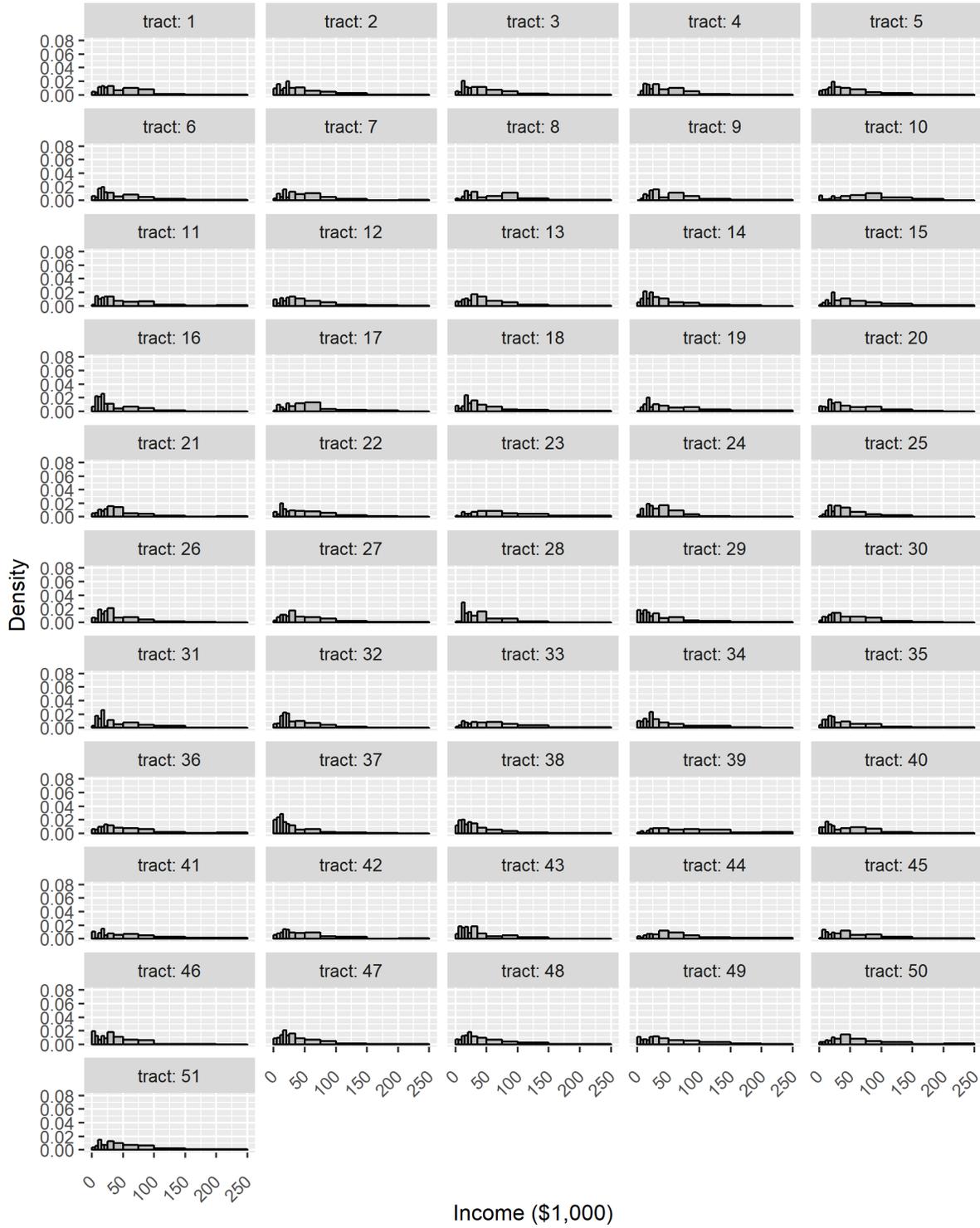


Figure B.0: 2015 5-year ACS density estimates for PUMA 600, MT.

### 2015 5-year PUMS Stratum-level Density Histograms for the MT PUMA

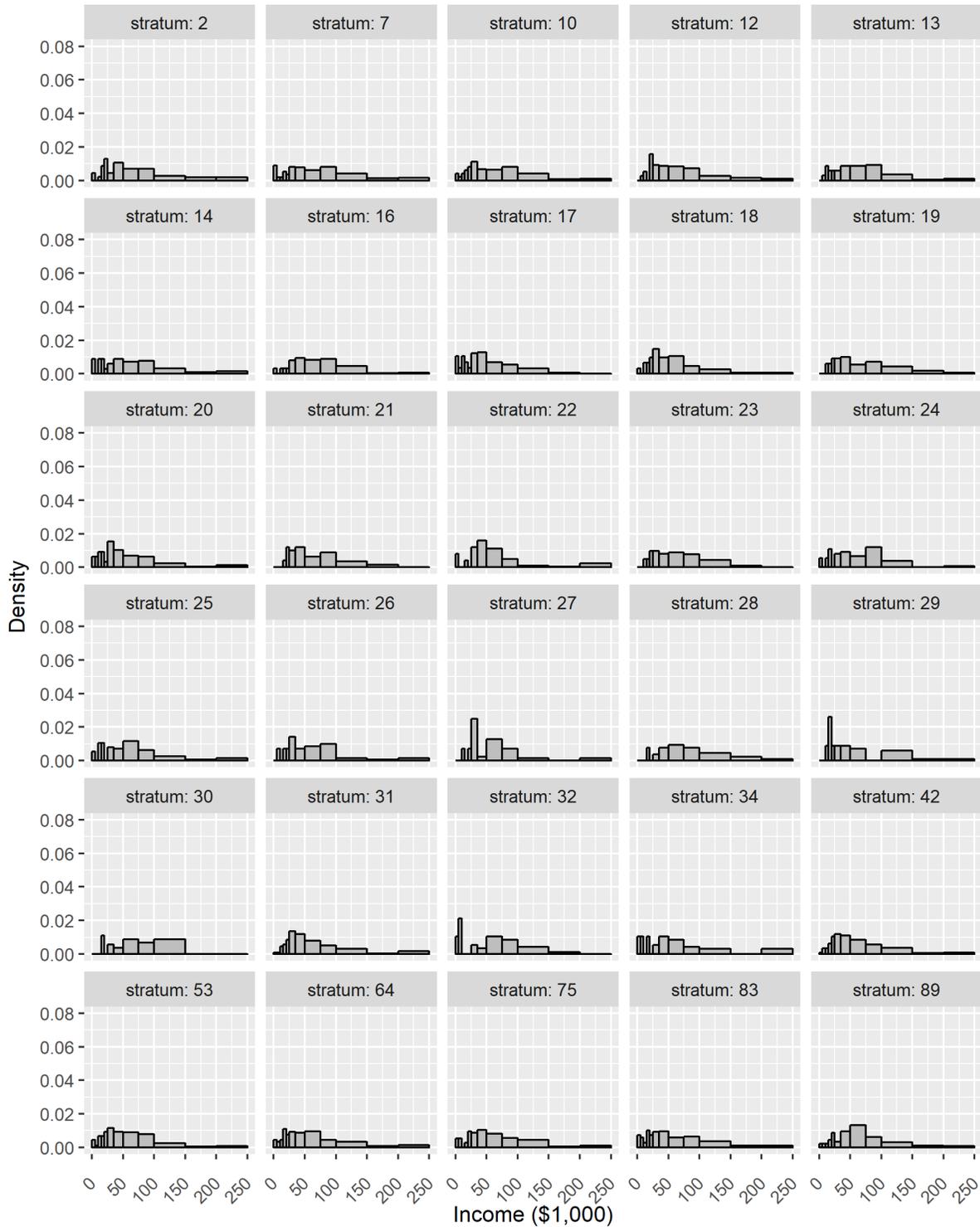


Figure B.0: 2015 5-year observed PUMS densities for strata with at least 17 observations in PUMA 600, MT.

## C MCMC and Reparameterization

Sampling from the posterior distribution via MCMC in our models is somewhat delicate, and here we describe some of the issues that arise and how to overcome them. We use HMC instead of a Gibbs sampler partially because the structure of the model will necessitate Metropolis steps for all of the parameters of the tract-level distributions, but also because HMC is robust in large, hierarchical models (Betancourt and Girolami, 2015). One common issue in HMC is “divergent transitions,” i.e. the gradient of the log posterior is numerically infinite. This biases the sampler and is also an indication that the Markov chain is not geometrically ergodic, which is necessary for a Markov chain central limit theorem and thus to be able to compute MCMC standard errors of e.g. posterior means (Betancourt, 2017). This is caused by complex geometries in the posterior density, and can be solved by reparameterizing the model. We discuss how to do this in Appendix C.1. Mixture models pose several problems for MCMC related to identification and label switching, which we discuss in Appendix C.2. Finally, Appendix C.3 discusses the `Stan` settings we use when fitting the model, and strategies for getting trustworthy MCMC for our models.

### C.1 Reparameterization

In MCMC algorithms based on both Gibbs samplers and on HMC, the parameterization of the target distribution can profoundly affect the behavior of the Markov chain. Essentially, the problem is that under some parameterizations the geometry of the target distribution is complex and as a result most standard MCMC algorithms will have trouble fully exploring certain regions with high mass in the target distribution (Betancourt and Girolami, 2015). A large literature exists describing centered parameterization (CPs) and non-centered parameterizations (NCPs) in the context of Gibbs samplers and data augmentation algorithms (see e.g. Van Dyk and Meng, 2001), and much of it applies to HMC as well

— at least to Euclidean HMC (Betancourt and Girolami, 2015). For example, the simple hierarchical model  $y|\theta \sim N(\theta, 1)$ ,  $\theta|\mu, \sigma \sim N(\mu, \sigma^2)$  is written in terms of a CP. The model  $y|\varepsilon, \mu, \sigma \sim N(\mu + \sigma\varepsilon, 1)$ ,  $\varepsilon \sim N(0, 1)$  is an equivalent NCP, and the posterior of the CP can be recovered from the posterior of the NCP and vice-versa via the one-to-one transformation  $\theta(\varepsilon) = \mu + \sigma\varepsilon$ .

For some parameters in any given model the CP is ideal for producing Markov chains which converge quickly and mix efficiently, while for others the NCP is ideal. A priori it is not obvious how to tell the difference between the two. Through trial and error we found a good parameterization of our models that results in HMC algorithms (using `Stan`) which explore the posterior distribution quickly and efficiently, though there may be other ways to improve further. The models as described in Section 3 are written entirely in terms CPs, so we only describe the parameters which require NCPs and how to construct them from their CPs. We have found that using NCPs for all tract-level and stratum-level parameters significantly eases MCMC. In the mixture models with tract-level spatial dependence we use the following reparameterizations (the other models are special cases):

$$\begin{aligned}
\varepsilon_{\mu_{rk}}(\mu_{rk}) &= (\mu_{rk} - \mu_k - \boldsymbol{\psi}'_r \boldsymbol{\eta}_{\mu_k}) / \delta_\mu && \text{for } k = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, R, \\
\varepsilon_{\sigma_{rk}}(\sigma_{rk}) &= (\log \sigma_{rk} - \log \sigma_k - \boldsymbol{\psi}'_r \boldsymbol{\eta}_{\sigma_k}) / \delta_\sigma && \text{for } k = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, R, \\
\varepsilon_{\xi_{rk}}(\xi_{rk}) &= (\xi_{rk} - \xi_k) / \delta_\xi && \text{for } k = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, R, \\
\tilde{\varepsilon}_{\mu_{sk}}(\tilde{\mu}_{sk}) &= (\tilde{\mu}_{sk} - \mu_k) / \tilde{\delta}_\mu && \text{for } k = 1, 2, \dots, K \text{ and } s = 1, 2, \dots, S, \\
\tilde{\varepsilon}_{\sigma_{sk}}(\tilde{\sigma}_{sk}) &= (\log \tilde{\sigma}_{sk} - \log \sigma_k) / \tilde{\delta}_\sigma && \text{for } k = 1, 2, \dots, K \text{ and } s = 1, 2, \dots, S, \\
\tilde{\varepsilon}_{\xi_{sk}}(\tilde{\xi}_{sk}) &= (\tilde{\xi}_{sk} - \xi_k) / \tilde{\delta}_\xi && \text{for } k = 1, 2, \dots, K \text{ and } s = 1, 2, \dots, S.
\end{aligned}$$

Under this transformation, we have

$$\begin{aligned}
\varepsilon_{\mu_{rk}} &\stackrel{iid}{\sim} N(0, 1) && \text{for } k = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, R, \\
\varepsilon_{\sigma_{rk}} &\stackrel{iid}{\sim} N(0, 1) && \text{for } k = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, R, \\
\varepsilon_{\xi_{rk}} &\stackrel{iid}{\sim} N(0, 1) && \text{for } k = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, R, \\
\tilde{\varepsilon}_{\mu_{sk}} &\stackrel{iid}{\sim} N(0, 1) && \text{for } k = 1, 2, \dots, K \text{ and } s = 1, 2, \dots, S, \\
\tilde{\varepsilon}_{\sigma_{sk}} &\stackrel{iid}{\sim} N(0, 1) && \text{for } k = 1, 2, \dots, K \text{ and } s = 1, 2, \dots, S, \\
\tilde{\varepsilon}_{\xi_{sk}} &\stackrel{iid}{\sim} N(0, 1) && \text{for } k = 1, 2, \dots, K \text{ and } s = 1, 2, \dots, S.
\end{aligned}$$

These parameterizations are used in all of the `.Stan` model files included in the Supplementary Material.

## C.2 Label switching and identification

Once reparameterized, there is still one more MCMC issue which must be handled — many parameters are unidentified due to label switching for mixture models (Stephens, 2000; Celeux et al., 2000; Jasra et al., 2005), and since we model  $K$  mixture probabilities with  $K$   $\xi_k$ s. This is only a problem for using MCMC diagnostics to assess the quality of our Markov chains since we do not want to do inference on the parameters of the mixture components — our objects of interest are the tract-level PDFs, but not necessarily the parameters defining those PDFs. MCMC diagnostics on these parameters may be misleading, however, because label switching causes them to jump to and from different modes. We look at diagnostics only on functions of parameters that we know are not unidentifiable due to these issues. For example functionals of the tract-level, stratum-level, and top-level distributions such as moments, quantiles, and CDF evaluations. Specifically, we look at the functionals of the tract-level distributions corresponding to the source statistics used in the model, and also all

of the  $\delta$  parameters (hierarchical standard deviation parameters) since they are fully identified. The main diagnostic we use is the Gelman-Rubin diagnostic using multiple chains (aka Split-R diagnostic Gelman and Rubin, 1992; Gelman et al., 2014) as computed by `Stan`, which can diagnose convergence and also failures of geometric ergodicity.

### C.3 `Stan` settings and MCMC strategies

In order to fit our models, we often have to deviate from standard settings in `Stan` (Stan Development Team, 2017). The three main settings we change and their default values are the target Metropolis acceptance rate, `adapt_delta= 0.8`, the maximum treedepth, `max_treedepth= 11`, and the initialization range, `init_r= 2`. See Stan Development Team (2017) and Stan Development Team (2016) for an explanation of all of these parameters, though we will briefly describe them here.

HMC essentially is a Metropolis-Hastings algorithm with a very complex proposal which is tuned to hit a target Metropolis acceptance rate. A higher acceptance rate leads to less dependence in the chain and other desirable properties, but it comes at a computational cost. Often, increasing this rate can remove a small number of divergent transitions, particularly after the best parameterization of the model has been chosen. In our larger models, we set this parameter to 0.99, but 0.9 is typically a good choice the smaller models, particularly with less than three mixture components or without spatial dependence.

When developing this proposal, `Stan` builds a binary tree until it finds a proposal that is “good enough” in some sense. Large trees result in a very high computational cost per iteration, so the algorithm uses a cap on the depth of the tree to mitigate this. When cap is binding for a high percentage of iterations, however, the algorithm can perform poorly and exhibit slow mixing or even stuck chains. This happens occasionally in the larger models, so we increase the tree depth to accommodate it.

To fit any model, **Stan** first transforms each parameter to the unconstrained real line. Then to initialize each chain, by default, **Stan** draws a random value from the  $U(-\text{init\_r}, \text{init\_r})$  distribution for each unconstrained parameter, where  $U(a, b)$  denotes the uniform distribution over the interval  $(a, b)$ , though the initialization can also be manually specified. Then during the warmup (burn-in) period, **Stan** adapts several parameters of the HMC algorithm to the target distribution. The initialization of a chain can sometimes drastically impact how the adaptation turns out for that chain, and thus its behavior. In particular, if a chain is initialized very far away from the “typical set” (the bulk of the target distribution’s probability mass, Betancourt, 2017), the algorithm can become adapted to a range outside of the typical set and then perform poorly when the chain is actually in the typical set. Decreasing the initialization range can help solve this at the risk of losing the robustness associated with starting each chain in a drastically different region of the parameter space. Increasing the number of iterations dedicated to adaptation can also help — this is controlled by the `warmup` parameter.

In Section 4 we set `max_treedepth= 12`, and set `init_r= 0.5`. We set `adapt_delta` depending on the model, but it ranges from 0.95 in the base lognormal models to 0.99 in the larger spatial mixture models. We fit 4 chains for each model, each for 4,000 iterations, 2,000 of which are used for warm-up (for burn-in and adaptation). Since we fit eight models each to 1008 datasets, we have to keep the number of posterior draws to a manageable size, we cannot tweak the MCMC for each of models fit, and some of the models have MCMC problems some of the time during the simulation study. In Section 5 we tailor the settings specifically to each model until we are able to get the model to fit reasonably well, though they are similar to the settings used in Section 4. The settings for each model for each PUMA can be seen in the R code in Supplementary Material. In some cases the MCMC could still be improved, particularly in the larger mixture models with spatially dependent tract-level parameters.

In other scenarios the settings we use may not be enough to achieve trustworthy MCMC. If a large number of divergent transitions are a problem, reparameterizing some of the parameters in terms of a CP instead of the NCP we use may help, see Appendix C.1. Using tighter priors can also help, though at the risk of making the results more dependent on the prior in an undesirable way. Sometimes changing the model can also help. In previous versions of our model we assumed that the tract-level and stratum-level standard deviation parameters were lognormally distributed instead of assuming the variances were inverse gamma distributed. This made MCMC easier in some cases but caused problems in others — we include the code for these models in the Supplementary Material, though we only use the inverse gamma models in the paper.

Finally, often MCMC and computational problems are an indication that there is a mismatch between the model and data, perhaps because the model is too large and complex. This is often a problem for larger versions of models, either with too many mixture components, or with too many spatial basis functions, and it is unlikely that any amount of tweaking the MCMC settings will solve it. In particular, in the PUMAs that exhibit little or no spatial dependence in tract-level source statistics, it is often hard to fit models with many (e.g. three or more) mixture components and spatial dependence. For example in the CO PUMA the results for three component spatially dependent models should be interpreted with this in mind, and in the IL and MT PUMAs we were only able to fit the three component mixture models with iid tract-level parameters. In practice this should not be an issue for users of this methodology since they should only use models with as many mixture components or spatial basis functions as are necessary — we deliberately tried to fit models that were more complex than necessary for completeness.

## D Truncated Integrated Karhunen-Loève Expansions

In this section we describe the truncated integrated Karhunen-Loève (TIKL) expansions with Obled-Creutin (OC) basis functions (Obled and Creutin, 1986) used in Section 3.2 to create spatial dependence between the tract-level parameters of the tract-level probability distributions. The TIKL expansion comes from Bradley et al. (2017), though we name it here, and can be defined in terms of the standard Karhunen-Loève (KL) expansion, e.g. Obled and Creutin (1986) and Cressie and Wikle (2011, page 165).

### D.1 Standard Karhunen-Loève expansion

The standard KL expansion is given by Theorem 1, and is a consequence of Mercer’s Theorem and its generalizations (Ferreira and Menegatto, 2009, Theorem 1.1).

**Theorem 1 (Karhunen-Loève)** *Let  $Y(\mathbf{u})$  be a zero-mean square-integrable stochastic process defined on  $\mathcal{D} \subseteq \mathbb{R}^D$  with valid covariance function  $C(\mathbf{u}, \mathbf{v})$  for  $\mathbf{u}, \mathbf{v} \in \mathcal{D}$ . Then there exists a sequence of mean-zero, uncorrelated random variables  $\{\alpha_k : k = 1, 2, \dots\}$  with corresponding variances (eigenvalues)  $\{\lambda_k : k = 1, 2, \dots\}$ ,  $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$  and a sequence of orthonormal eigenfunctions,  $\{\phi_k(\mathbf{u}) : k = 1, 2, \dots\}$  such that*

$$\lambda_k \phi_k(\mathbf{u}) = \int_{\mathcal{D}} C(\mathbf{u}, \mathbf{v}) \phi_k(\mathbf{v}) d\mathbf{v} \text{ for } k = 1, 2, \dots, \quad (9)$$

$$C(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\mathbf{u}) \phi_k(\mathbf{v}), \quad (10)$$

$$\alpha_k = \int_{\mathcal{D}} Y(\mathbf{u}) \phi_k(\mathbf{u}) d\mathbf{u} \text{ for } k = 1, 2, \dots, \quad (11)$$

$$Y(\mathbf{u}) = \sum_{k=1}^{\infty} \alpha_k \phi_k(\mathbf{u}). \quad (12)$$

*Additionally if  $Y(\mathbf{u})$  is Gaussian process, then  $\alpha_k \stackrel{ind}{\sim} N(0, \lambda_k)$ .*

Equation (12) is the KL expansion of  $Y(\mathbf{u})$ , while equation (9) is the Fredholm integral equation that allows for the computation of  $\lambda_k$  and  $\phi_k(\mathbf{u})$  from  $C(\mathbf{u}, \mathbf{v})$ .

For a geostatistical process  $Y(\mathbf{u})$  with unknown covariance function, we can use this result to flexibly model  $Y(\mathbf{u})$  via the truncated KL (TKL) expansion

$$Y(\mathbf{u}) \approx \widehat{Y}(\mathbf{u}) = \sum_{k=1}^K \phi_k(\mathbf{u})\alpha_k = \boldsymbol{\phi}(\mathbf{u})'\boldsymbol{\alpha}.$$

Often this is combined this with a fine scale variation term,  $\varepsilon(\mathbf{u})$ , which represents the error associated with truncating the KL expansion. This yields another form of the TKL expansion:

$$\widehat{Y}(\mathbf{u}) = \sum_{k=1}^K \phi_j(\mathbf{u})\alpha_k + \varepsilon(\mathbf{u}) = \boldsymbol{\phi}(\mathbf{u})'\boldsymbol{\alpha} + \varepsilon(\mathbf{u})$$

as a typical class of models based on the truncated KL expansion. In order to complete this model we need to specify the known eigenfunctions,  $\boldsymbol{\phi}(\mathbf{u})$ . The OC eigenfunctions can be constructed from any set of spatial generating basis functions (GBFs), denoted by  $\{g_k(\mathbf{u}), k = 1, 2, \dots, K\}$ , and an appropriate weighting matrix  $\mathbf{F}$  with entries  $F_{ij}, i, j = 1, 2, \dots, K$ . The OC eigenfunctions are then  $\boldsymbol{\phi}^{OC}(\mathbf{u}) = \mathbf{F}'\mathbf{g}(\mathbf{u})$ . In order to ensure these eigenfunctions yield a valid KL expansion, we additionally require that  $\mathbf{F}\mathbf{W}\mathbf{F}' = \mathbf{I}_K$  where  $\mathbf{I}_K$  is the  $K \times K$  identity matrix and  $\mathbf{W}$  is a  $K \times K$  matrix with elements  $W_{ij} = \int_{\mathcal{D}} g_i(\mathbf{u})g_j(\mathbf{u})d\mathbf{u}$ . This condition is satisfied by  $\mathbf{F} = \mathbf{R}^{-1}\mathbf{G}$  where  $\mathbf{R}$  is any matrix such that  $\mathbf{R}'\mathbf{R} = \mathbf{W}$  and  $\mathbf{G}$  is a  $K \times K$  orthogonal real matrix. This allows us the rewrite  $\widehat{Y}(\mathbf{u})$  as

$$\widehat{Y}(\mathbf{u}) = \boldsymbol{\phi}(\mathbf{u})'\boldsymbol{\alpha} + \varepsilon(\mathbf{u}) = \mathbf{g}'(\mathbf{u})\mathbf{R}^{-1}\mathbf{G}\boldsymbol{\alpha} + \varepsilon(\mathbf{u}) = \boldsymbol{\psi}'(\mathbf{u})\boldsymbol{\eta} + \varepsilon(\mathbf{u})$$

where  $\boldsymbol{\psi}'(\mathbf{u}) = \mathbf{g}'(\mathbf{u})\mathbf{R}^{-1}$ , and  $\boldsymbol{\eta} = \mathbf{G}\boldsymbol{\alpha}$ . If we assume that  $\boldsymbol{\alpha}$  is Gaussian, i.e.  $\boldsymbol{\alpha} \sim$

$N(\mathbf{0}_K, \mathbf{\Lambda})$ , then  $\boldsymbol{\eta} \sim N(\mathbf{0}_K, \mathbf{G}\mathbf{\Lambda}\mathbf{G}')$  where  $\mathbf{0}_K$  is an  $K$ -dimensional vector of zeroes and  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_K)$ . Both  $\mathbf{G}$  and  $\mathbf{\Lambda}$  are unknown, so it is convenient to reparameterize in terms of  $\boldsymbol{\Sigma} = \mathbf{G}\mathbf{\Lambda}\mathbf{G}'$ . So we can model a sequence of mean-zero point-level random variables  $\{Y_i \equiv Y(\mathbf{u}_i) : i = 1, 2, \dots, n\}$  as

$$Y_i = \boldsymbol{\psi}'_i \boldsymbol{\eta} + \varepsilon_i \text{ for } i = 1, 2, \dots, n \quad (13)$$

where  $\boldsymbol{\psi}_i = \boldsymbol{\psi}(\mathbf{u}_i)$ ,  $\varepsilon_i \equiv \varepsilon(\mathbf{u}_i)$ ,  $\boldsymbol{\eta} \sim N(\mathbf{0}_r, \boldsymbol{\Sigma})$ , and  $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ . The only missing piece is a specification of the GBFs, but nearly any spatial basis functions can be used. We will use the bisquare basis functions defined by

$$g_k(\mathbf{u}) = \begin{cases} [1 - (||\mathbf{u} - \mathbf{c}_k||/w)^2]^2 & \text{if } ||\mathbf{u} - \mathbf{c}_k|| \leq w, \\ 0 & \text{otherwise; } \mathbf{u} \in \mathcal{D} \end{cases}$$

where  $\mathbf{c}_k \in \mathcal{D}$  is a knot point and  $w$  is 1.5 times the smallest distance between any two knots. The knots could be equally spaced, or placed according to a space-filling design. Given a choice of GBFs,  $\{g_k(\mathbf{u}) : k = 1, 2, \dots, K\}$ , the matrix  $\mathbf{W}$  can be approximated via Monte Carlo simulation so that  $\mathbf{R}^{-1}$  can be approximated. Suppose  $\mathbf{u}_t \stackrel{iid}{\sim} U(\mathcal{D})$  for  $t = 1, 2, \dots, N_W$  where  $U(\mathcal{D})$  denotes the uniform distribution over the set  $\mathcal{D}$ . Then we can approximate each element of  $\mathbf{W}$  with

$$\widehat{W}_{ij} \equiv \frac{|\mathcal{D}|}{N_W} \sum_{t=1}^{N_W} g_i(\mathbf{u}_t) g_j(\mathbf{u}_t) \rightarrow W_{ij} \equiv |\mathcal{D}| \int_{\mathcal{D}} g_i(\mathbf{u}) g_j(\mathbf{u}) \frac{1}{|\mathcal{D}|} d\mathbf{u}$$

almost surely as  $N_W \rightarrow \infty$  via the law of large numbers.

## D.2 Integrating a TKL expansion

Let  $A \subset \mathcal{D}$  denote some areal unit on which we have data or want to do inference. Then we can define the mean of  $Y(\mathbf{u})$  over  $A$  as  $Y(A) = \int_A Y(\mathbf{u})d\mathbf{u}/|A|$ . With areal-level data, we can take advantage of the integrated KL (IKL) expansion to provide a decomposition for  $Y(A)$  (Bradley et al., 2017):

**Theorem 2 (Integrated Karhunen-Loève)** *Under the conditions of Theorem 1, for any  $A \subset \mathcal{D}$*

$$Y(A) = \sum_{k=1}^{\infty} \phi_k(A)\alpha_k \quad (14)$$

where  $\phi_k(A) = \int_A \phi_k(\mathbf{u})d\mathbf{u}$ , and for any  $A \subset \mathcal{D}$ ,  $B \subset \mathcal{D}$  we have

$$\text{cov}\{Y(A), Y(B)\} = \sum_{k=1}^{\infty} \lambda_k \phi_k(A)\phi_k(B). \quad (15)$$

This theorem allows us to model areal-level data similarly to how the standard KL expansion is used to model point-level data in Appendix D.1. Specifically we employ a truncated IKL (TIKL) expansion so that

$$\widehat{Y}(A) = \boldsymbol{\phi}(A)' \boldsymbol{\alpha} + \varepsilon(A) = \boldsymbol{\psi}'(A) \boldsymbol{\eta} + \varepsilon(A) \quad (16)$$

where  $\psi_k(A) = \int_A \psi_k(\mathbf{u})d\mathbf{u}/|A|$ ,  $\varepsilon(A) = \int_A \varepsilon(\mathbf{u})d\mathbf{u}/|A|$  and both  $\psi_k(\mathbf{u})$  and  $\varepsilon(\mathbf{u})$  were defined in Appendix D.1. For a sequence of areal units  $\{A_i \subset \mathcal{D} : i = 1, 2, \dots, n\}$  let  $\psi_{ik} \equiv \psi_k(A_i)$  and assume that  $\varepsilon_i \equiv \varepsilon(A_i) \stackrel{iid}{\sim} N(0, \sigma^2)$  and  $\boldsymbol{\eta} \sim N(\mathbf{0}_K, \boldsymbol{\Sigma})$  so that

$$Y_i \equiv Y(A_i) = \boldsymbol{\psi}'_i \boldsymbol{\eta} + \varepsilon_i \quad (17)$$

for  $i = 1, 2, \dots, n$ . Monte Carlo simulation can be used to approximate  $\psi_k(A)$  for any given  $A$  much like  $\mathbf{W}$  was approximated. Let  $\mathbf{u}_t \stackrel{iid}{\sim} U(A)$  for  $t = 1, 2, \dots, N_A$ , then

$$\widehat{g}_k(A) \equiv \frac{1}{N_A} \sum_{t=1}^{N_A} g(\mathbf{u}_t) \rightarrow g_k(A) \equiv \int_A g_k(\mathbf{u}) \frac{1}{|A|} d\mathbf{u}$$

almost surely as  $N_A \rightarrow \infty$  via the law of large numbers. Then  $\widehat{\boldsymbol{\psi}}'(A) = \widehat{\mathbf{g}}'(A) \widehat{\mathbf{R}}^{-1}$  approximates  $\boldsymbol{\psi}'(A)$ .

### D.3 Applying the TIKL expansion to tract-level distributions

We defined the TIKL above, culminating in the statement of Equation (17). This construction assumes that we observe the areal-level variables we are attempting to model, but in the spatial models we constructed in Section 3.2 we need to apply the TIKL to latent tract-level parameters, e.g.  $\mu_r$  and  $\sigma_r^2$  for tract  $r$  in the base spatial model. To do this, we construct separate independent TIKL expansions for the  $\mu_r$ s and the  $\sigma_r^2$ s, but using the same basis functions for both. Let  $\boldsymbol{\psi}_r$  denote the set of  $M$  OC basis functions integrated over tract  $r$ , defined in Appendices D.1 and D.2. We rescale the OC basis functions by their standard deviations to make it easier to choose priors for  $\delta_{\eta\mu}$  and  $\delta_{\eta\sigma}$  that allow for spatial dependence between the tract-level parameters, but does not enforce it. In Appendix E we explore the degree to which our chosen priors combined with this rescaling allows for varying degrees of spatial dependence in the source statistics by simulating from the priors and the model.

Let  $\boldsymbol{\psi}$  denote the  $R \times M$  matrix with  $r$ th row  $\boldsymbol{\psi}_r$ . Then let  $\boldsymbol{\Psi}$  be the  $R \times M$  matrix formed by taking each column of  $\boldsymbol{\psi}$  and dividing each element of that column by the standard deviation of the column, i.e.  $\psi_{rm} = \psi_{rm}/\text{SD}(\psi_{1m}, \psi_{2m}, \dots, \psi_{Rm})$ . Let  $\boldsymbol{\psi}'_r$  denote the  $r$ th row

of  $\Psi$ . Then from Section 3.2 in the spatial base tract-level parameter models we assumed

$$\begin{aligned} \mu_r | \mu, \boldsymbol{\eta}_\mu, \delta_\mu &\stackrel{iid}{\sim} \text{N}(\mu + \boldsymbol{\psi}'_r \boldsymbol{\eta}_\mu, \delta_\mu^2) \quad \text{and} \\ \log \sigma_r | \log \sigma, \boldsymbol{\eta}_\sigma, \delta_\sigma &\stackrel{iid}{\sim} \text{N}(\log \sigma + \boldsymbol{\psi}'_r \boldsymbol{\eta}_\sigma, \delta_\sigma^2) \quad \text{for } r = 1, 2, \dots, R, \text{ with} \\ \boldsymbol{\eta} | \delta_{\eta_\mu}, \delta_{\eta_\sigma} &\sim \text{Cauchy} \left( \mathbf{0}_{2M}, \begin{bmatrix} \delta_{\eta_\mu}^2 \mathbf{I}_M & \mathbf{0}_{M \times M} \\ \mathbf{0}_{M \times M} & \delta_{\eta_\sigma}^2 \mathbf{I}_M \end{bmatrix} \right). \end{aligned}$$

The scale matrix of  $\boldsymbol{\eta}$  is diagonal both because univariate TIKL expansions tend to create uncorrelated  $\boldsymbol{\eta}$ s and because we assumed that the TIKL expansions on  $\mu_r$  and  $\sigma_r^2$  are independent for simplicity. We mentioned in Section 3.2 that we use a Cauchy distribution instead of a normal because it turns out to make MCMC significantly easier without materially impacting our results. Applying the TIKL expansion to the spatially dependent mixture models is analogous — the same  $\boldsymbol{\psi}_r$  is used for both the mean and standard deviation parameter of each mixture component.

## E Priors And Posterior Robustness

Here we investigate how informative our priors, including using the prior predictive distribution. In Section 3.3 we specified the following priors for all of our models:  $\mu_k \sim \text{N}(0, 1^2)$  for  $k = 1, 2, \dots, K$ ,  $\sigma_k \sim \text{N}^+(1, 1^2)$  for  $k = 1, 2, \dots, K$ ,  $\xi_k \sim \text{N}(0, 1^2)$  for  $k = 1, 2, \dots, K$ ,  $\delta_\mu \sim \text{N}^+(0, 1^2)$ ,  $\delta_\sigma \sim \text{N}^+(0, 1^2)$ ,  $\delta_\xi \sim \text{N}^+(0, 0.5^2)$ ,  $\tilde{\delta}_\mu \sim \text{N}^+(0, 1^2)$ ,  $\tilde{\delta}_\sigma \sim \text{N}^+(0, 1^2)$ ,  $\tilde{\delta}_\xi \sim \text{N}^+(0, 0.5^2)$ ,  $\delta_{\eta_\mu} \sim \text{N}^+(0, 0.1^2)$ , and  $\delta_{\eta_\sigma} \sim \text{N}^+(0, 0.1^2)$ . Here  $\text{N}^+(m, s^2)$  is the normal distribution with mean  $m$  and standard deviation  $s$  truncated from below at zero.

Because of the standardization of the data, these priors are readily interpretable in terms of standard deviations of the original potentially log-scaled source observations. The priors on the  $\mu_k$ s imply that a priori we expect each  $\mu_k$  to be within 1-2 standard deviations of the

mean of the source observations, so it is rather weak relative to the scale of the data. In the 2-component mixture model with the prior specified as above, there is about a 65% chance a priori that  $\omega_{rk}$  is within 0.1 of  $\omega_k$ , and an 85% chance it is within 0.2, though these numbers are higher to the extent that  $\omega_k$  is farther away from 0.5. Further, this prior implies that  $\omega_k$  has about a 97% chance of being between 0.1 and 0.9 a priori, and an 83% chance of being between 0.2 and 0.8. These priors are somewhat tight on the  $\omega_{ks}$ ,  $\omega_{rk}s$ , and  $\tilde{\omega}_{sk}s$  in order to help identify the tract-level mixture probabilities using the stratum-level observations, but flexible enough that any given tract can have mixture probabilities which are significantly different from the top-level mixture probabilities. MCMC is somewhat delicate in these models, though we have developed a strategy to make it easier. See Appendix C for details.

To investigate the level of spatial dependence implied by our priors, we simulate five draws from the prior predictive distribution of several of the bin estimates in the Mix-3 TIKL-4 model, and compare maps of the prior predictive bin estimates to maps of the estimates. We look at four bin estimates for each PUMA considered in Section 5: the \$10,000 to \$15,000 bin, the \$35,000 to \$50,000 bin, the \$75,000 to \$100,000 bin, and the \$150,000 to \$200,000 bin. Figures E.0—E.0 contain the maps for the CO PUMA, Figures E.0—E.0 contain the maps for the IL PUMA, Figures E.0—E.0 contain the maps for the MO PUMA, and Figures E.0—E.0 contain the maps for the MT PUMA. In each figure, the top left map is of the observed estimates while the other five maps are simulated from the prior predictive distribution. These plots suggest that our priors can easily accommodate the sort of spatial dependence we see in the bin estimates, and in general can accommodate a wide range of values for the bin estimates.

Further, we experimented with a variety of priors that were significantly looser on all parameters except  $\delta_{\eta\mu}$  and  $\delta_{\eta\sigma}$ , and our results were largely unchanged (not reported here), though the prior predictive distributions were significantly more variable. These looser priors made MCMC more difficult, and increasing the standard deviation of  $\delta_{\eta\mu}$  or  $\delta_{\eta\sigma}$ , made it

nearly impossible. So while computational concerns practically necessitate the regularization imposed by our somewhat tight priors, the prior predictive exercise in this section and our experience with looser priors suggest that our results are not being driven by the priors.

Prior predictive simulations for Colorado PUMA  
Proportion of households with income between \$5,000 and \$10,000

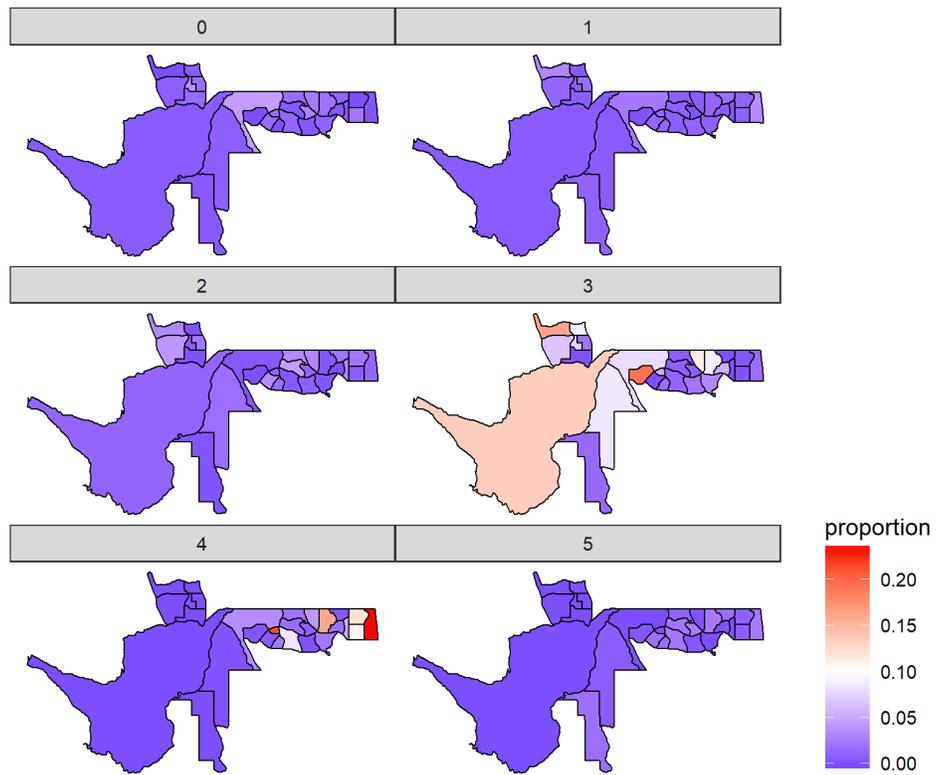


Figure E.0: Maps of bin estimates for the \$10,000 to \$15,000 bin in PUMA 821, CO. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Colorado PUMA  
Proportion of households with income between \$5,000 and \$10,000

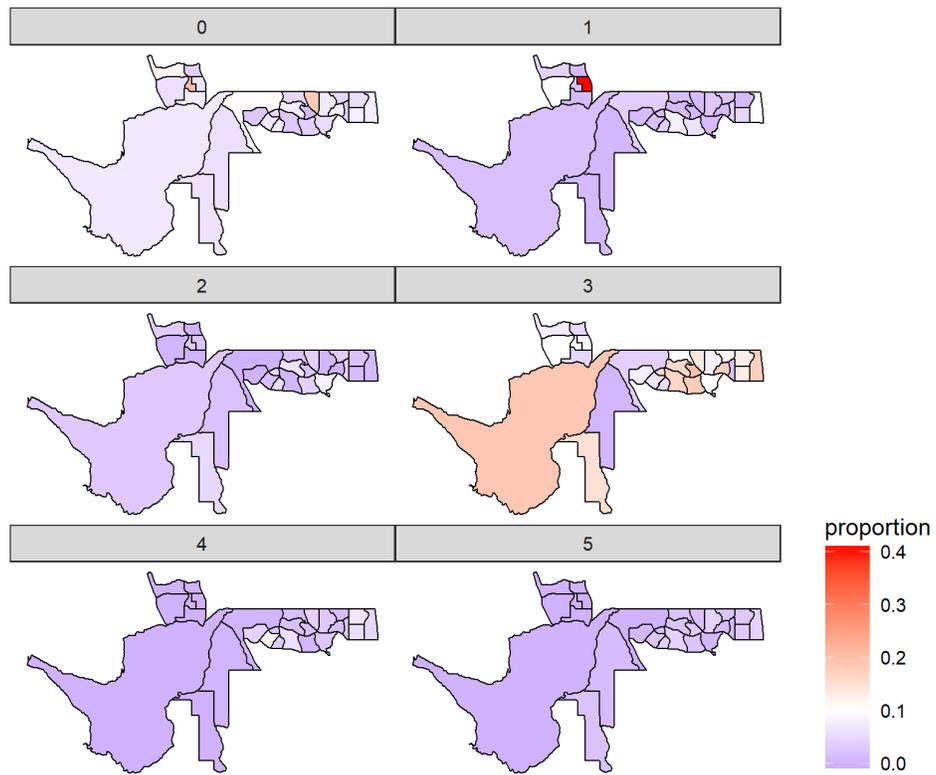


Figure E.0: Maps of bin estimates for the \$35,000 to \$50,000 bin in PUMA 821, CO. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Colorado PUMA  
Proportion of households with income between \$5,000 and \$10,000

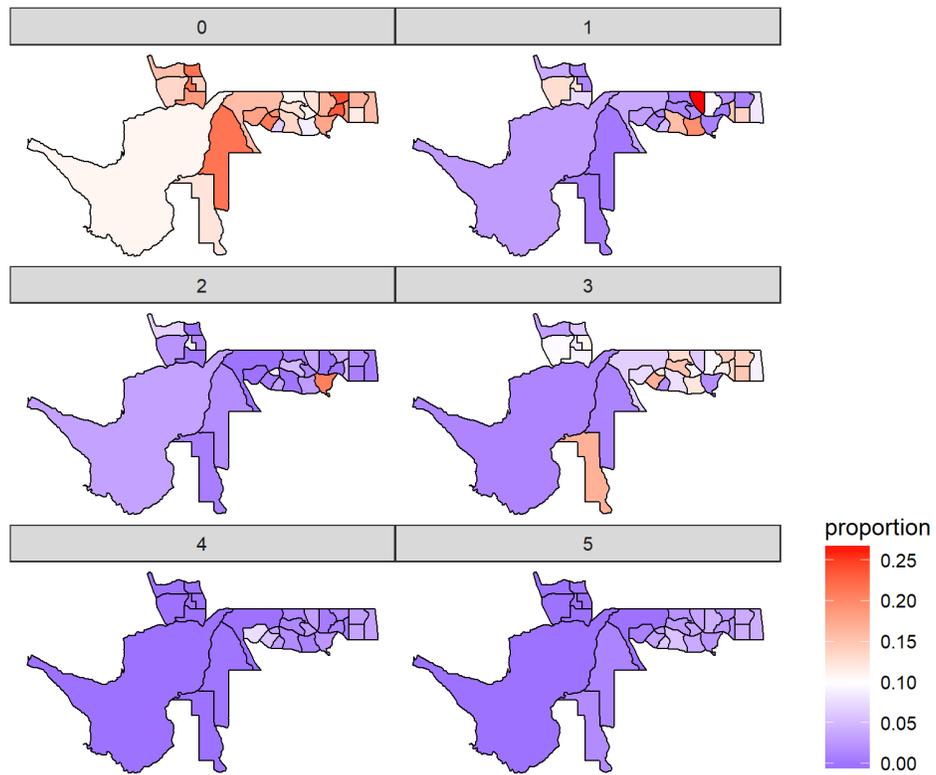


Figure E.0: Maps of bin estimates for the \$75,000 to \$100,000 bin in PUMA 821, CO. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Colorado PUMA  
Proportion of households with income between \$5,000 and \$10,000

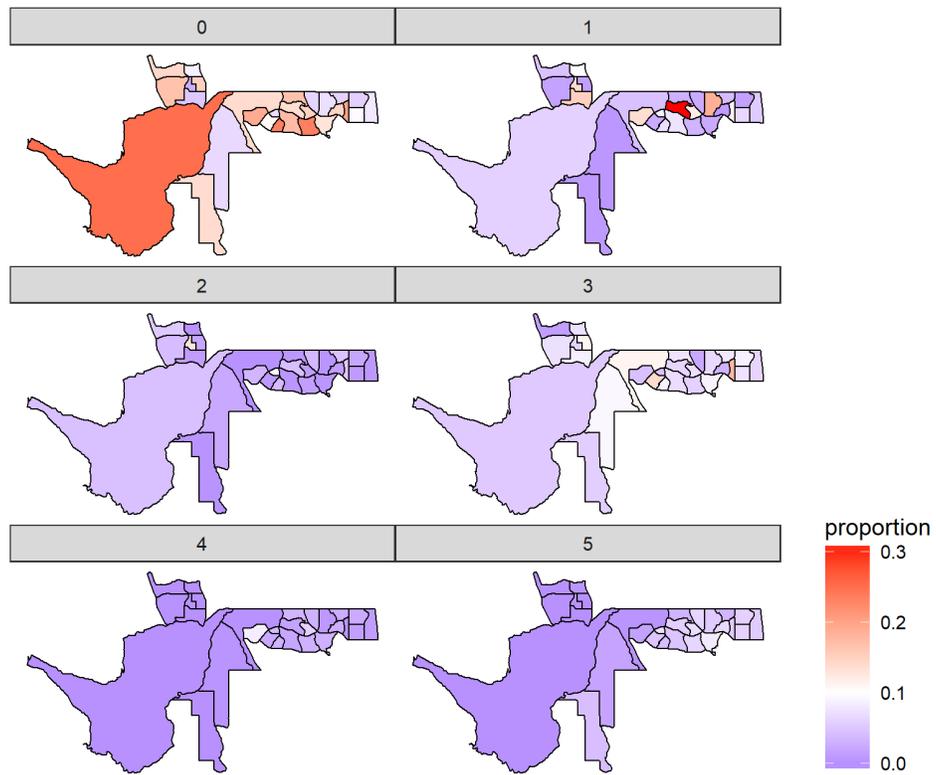


Figure E.0: Maps of bin estimates for the \$150,000 to \$200,000 bin in PUMA 821, CO. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Illinois PUMA  
Proportion of households with income between \$5,000 and \$10,000

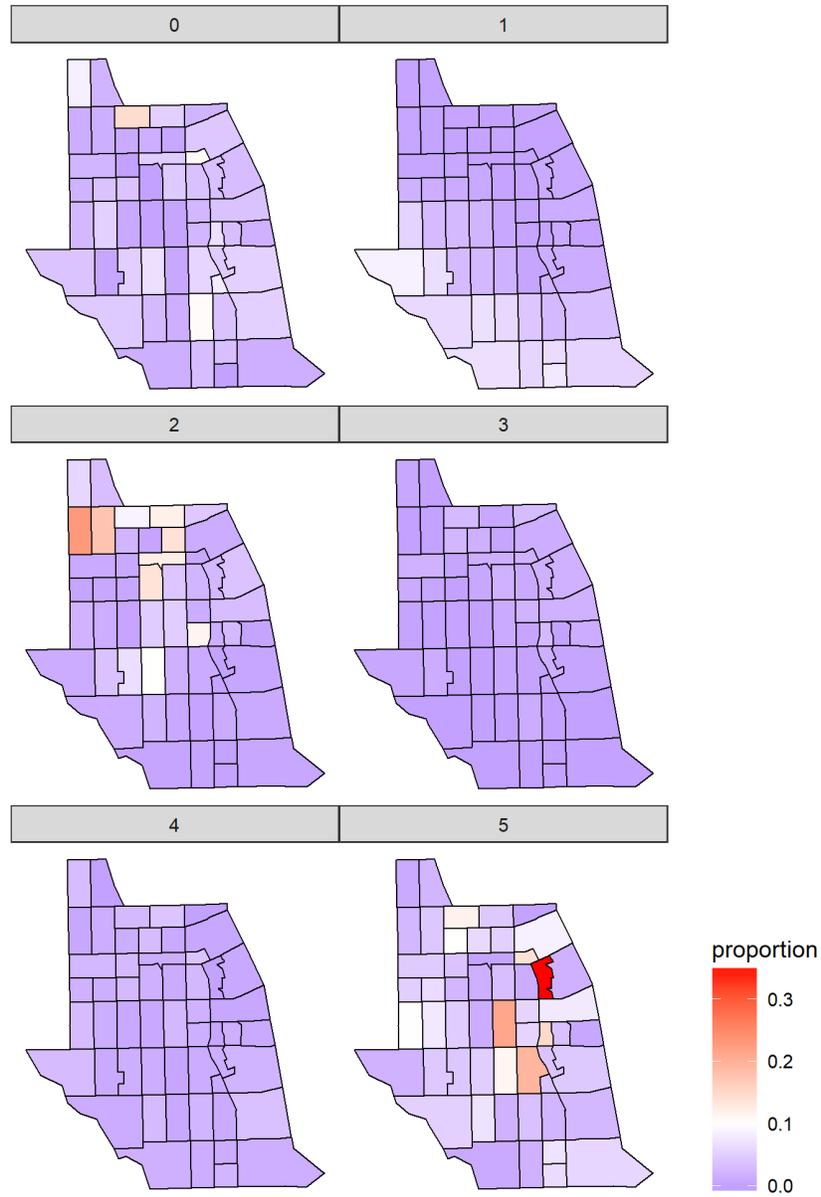


Figure E.0: Maps of bin estimates for the \$10,000 to \$15,000 bin in PUMA 3502, IL. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Illinois PUMA  
Proportion of households with income between \$5,000 and \$10,000

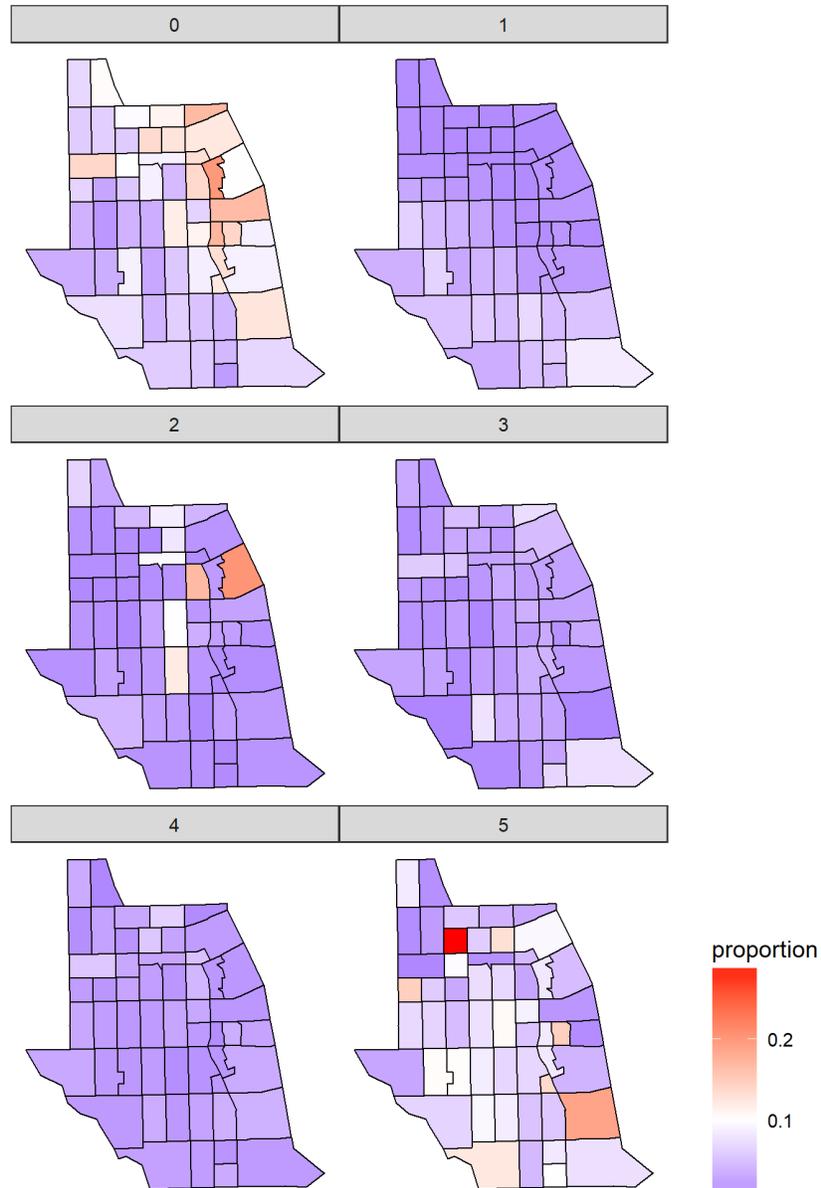


Figure E.0: Maps of bin estimates for the \$35,000 to \$50,000 bin in PUMA 3502, IL. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Illinois PUMA  
Proportion of households with income between \$5,000 and \$10,000

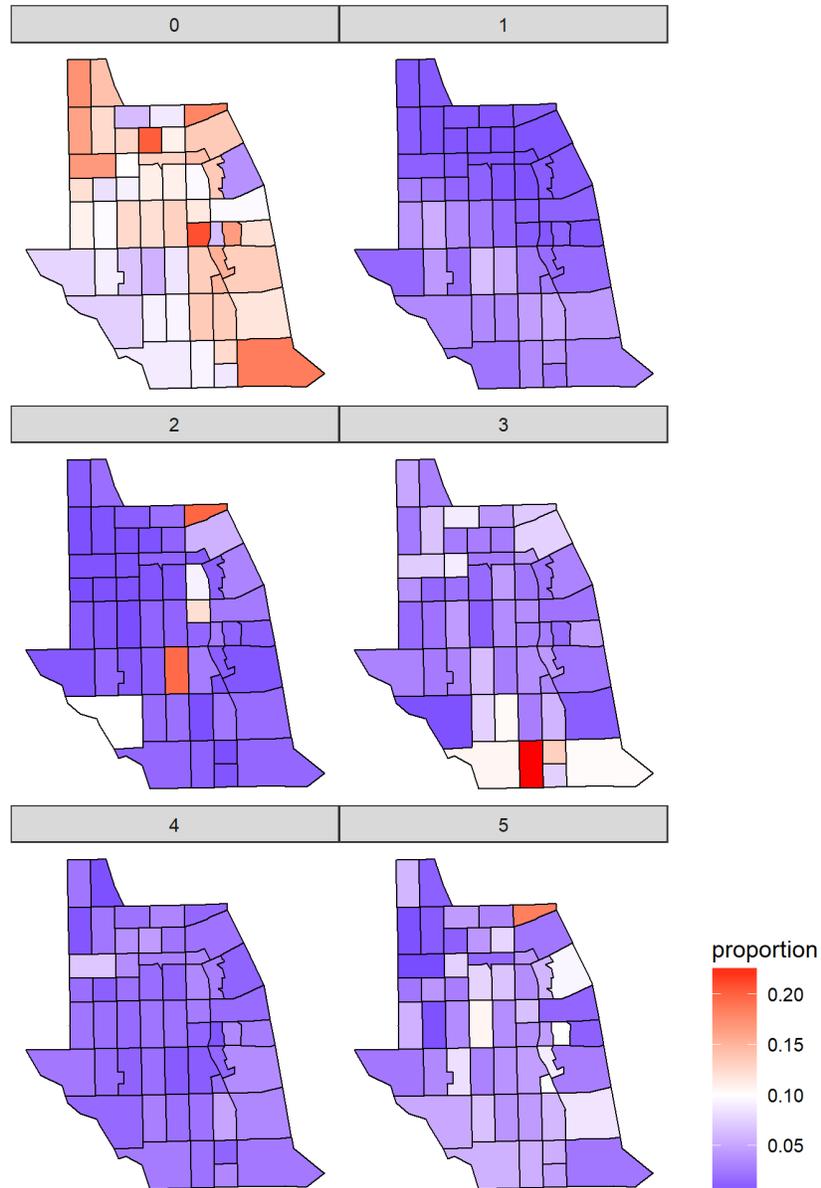


Figure E.0: Maps of bin estimates for the \$75,000 to \$100,000 bin in PUMA 3502, IL. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Illinois PUMA  
Proportion of households with income between \$5,000 and \$10,000

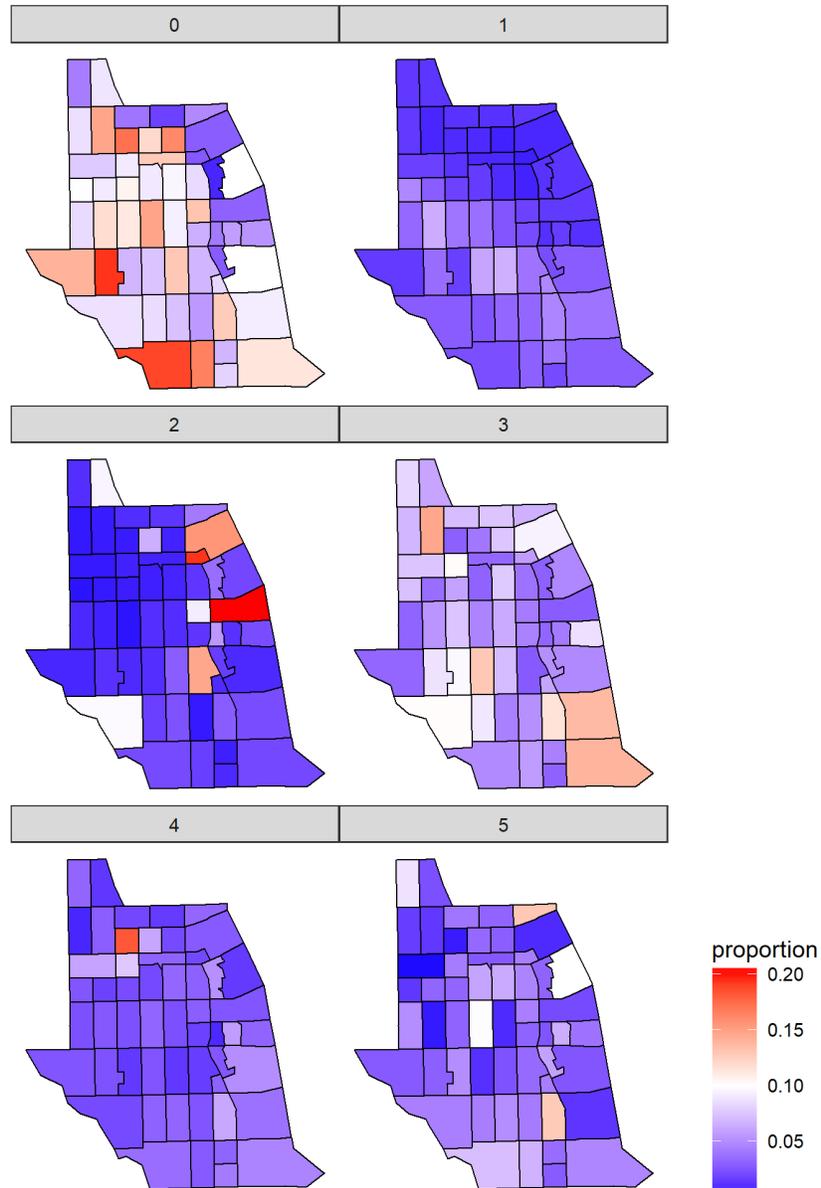


Figure E.0: Maps of bin estimates for the \$150,000 to \$200,000 bin in PUMA 3502, IL. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Boone County PUMA  
 Proportion of households with income between \$5,000 and \$10,000

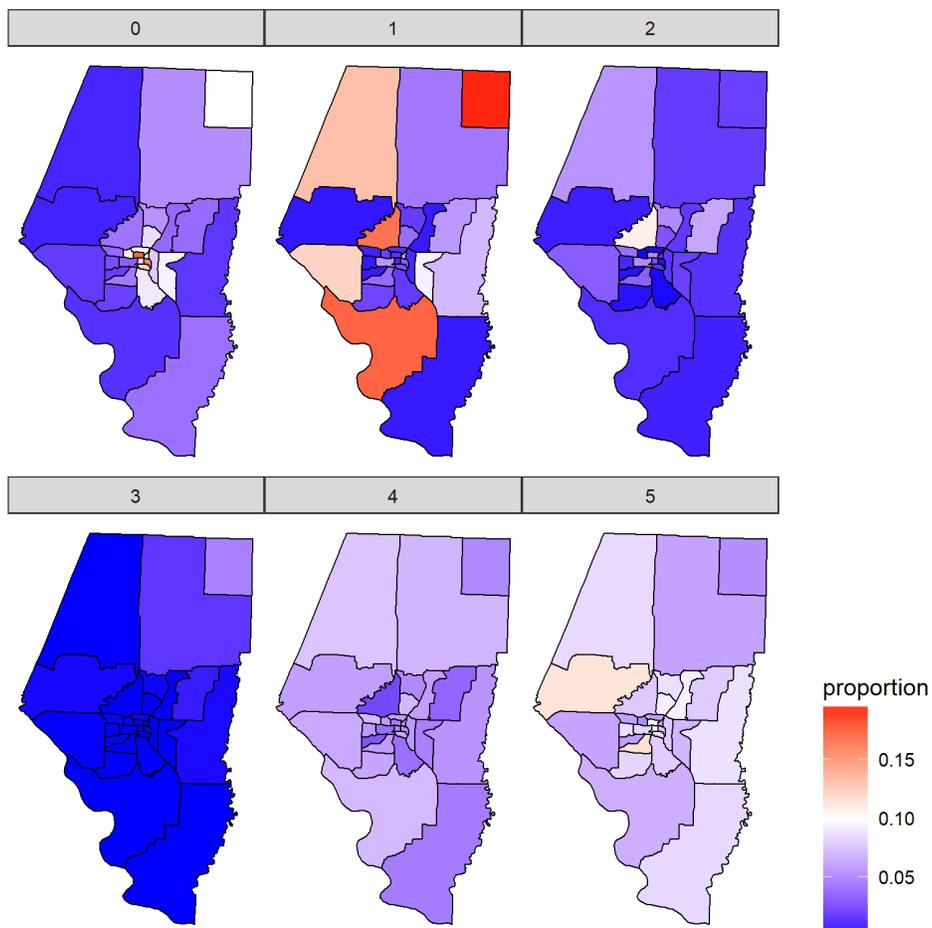


Figure E.0: Maps of bin estimates for the \$10,000 to \$15,000 bin in PUMA 600, MO (Boone County). The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Boone County PUMA  
 Proportion of households with income between \$5,000 and \$10,000

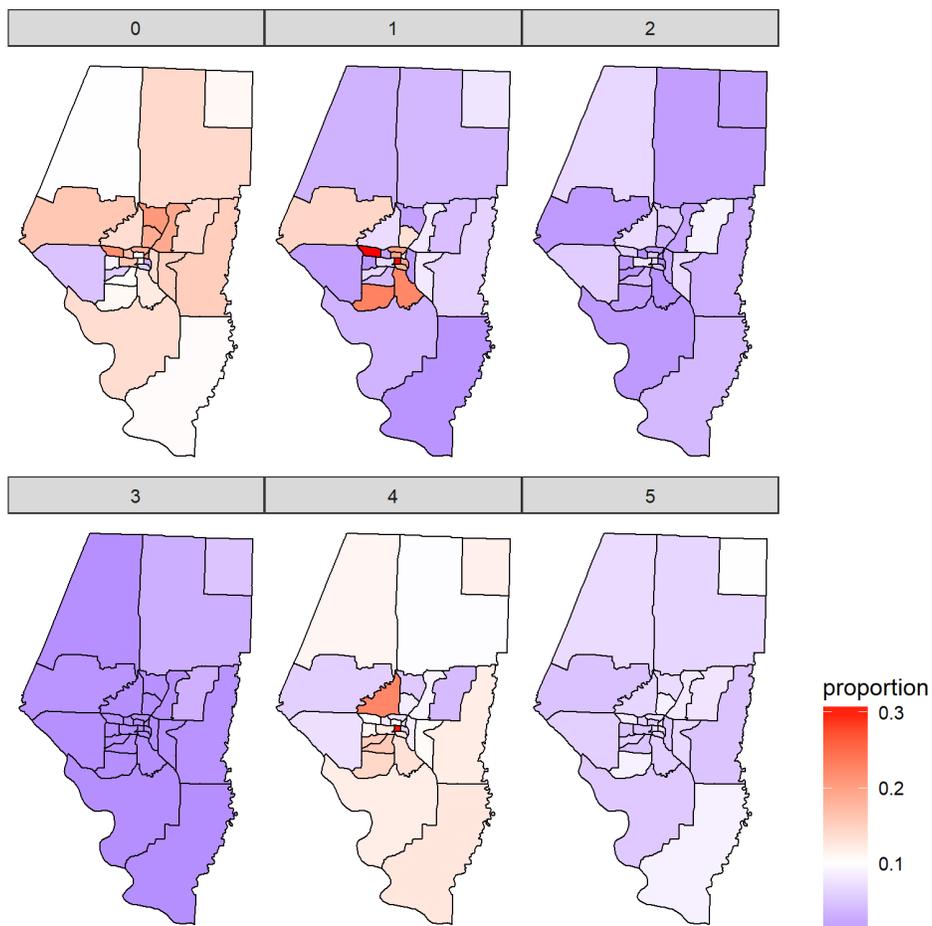


Figure E.0: Maps of bin estimates for the \$35,000 to \$50,000 bin in PUMA 600, MO (Boone County). The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Boone County PUMA  
 Proportion of households with income between \$5,000 and \$10,000

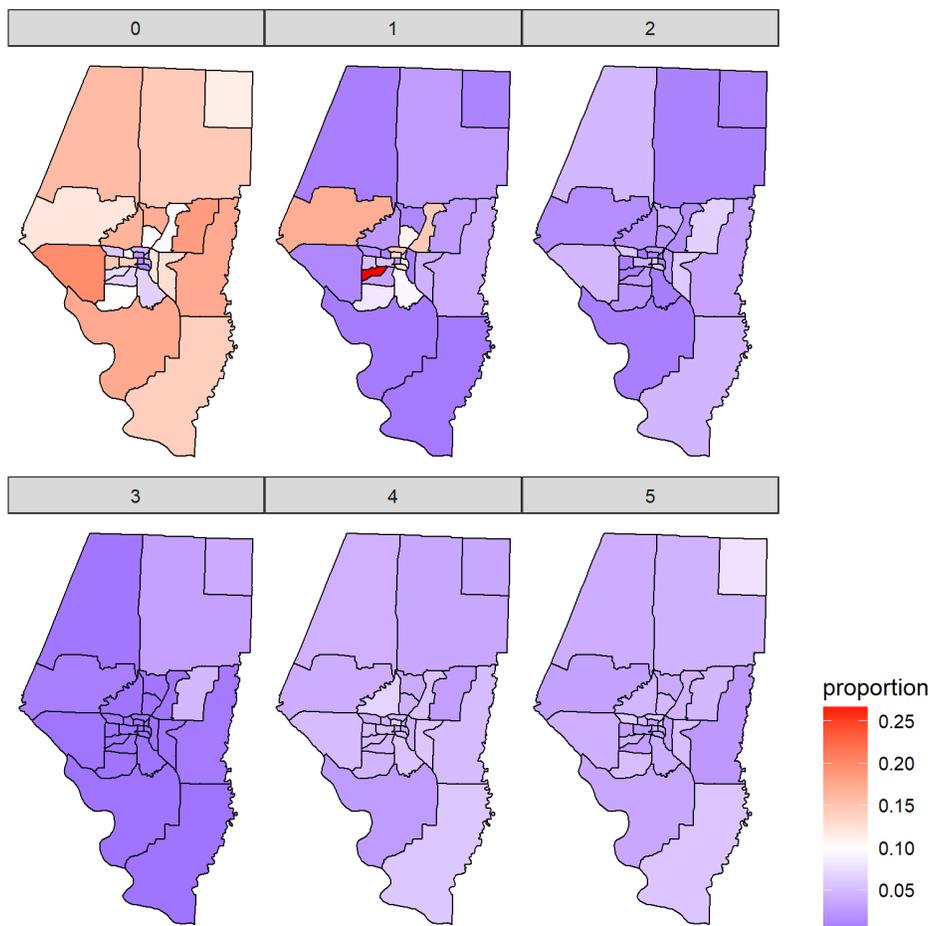


Figure E.0: Maps of bin estimates for the \$75,000 to \$100,000 bin in PUMA 600, MO (Boone County). The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Boone County PUMA  
 Proportion of households with income between \$5,000 and \$10,000

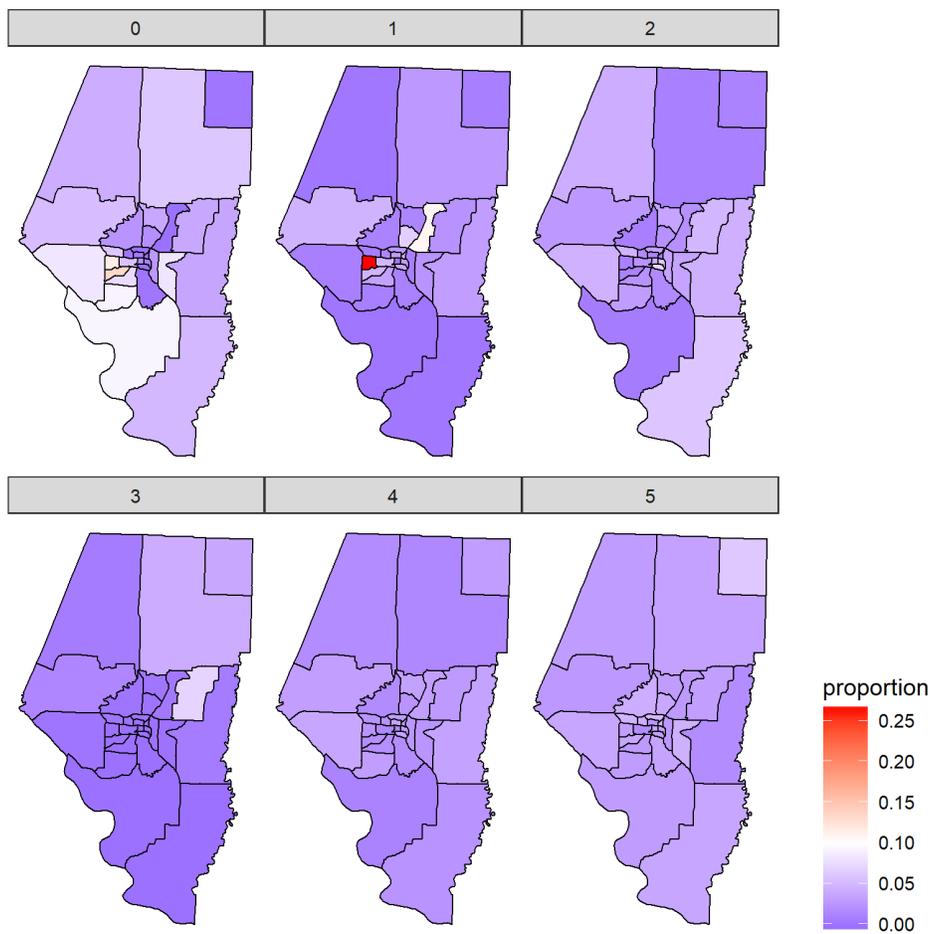


Figure E.0: Maps of bin estimates for the \$150,000 to \$200,000 bin in PUMA 600, MO (Boone County). The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Montana PUMA  
 Proportion of households with income between \$5,000 and \$10,000

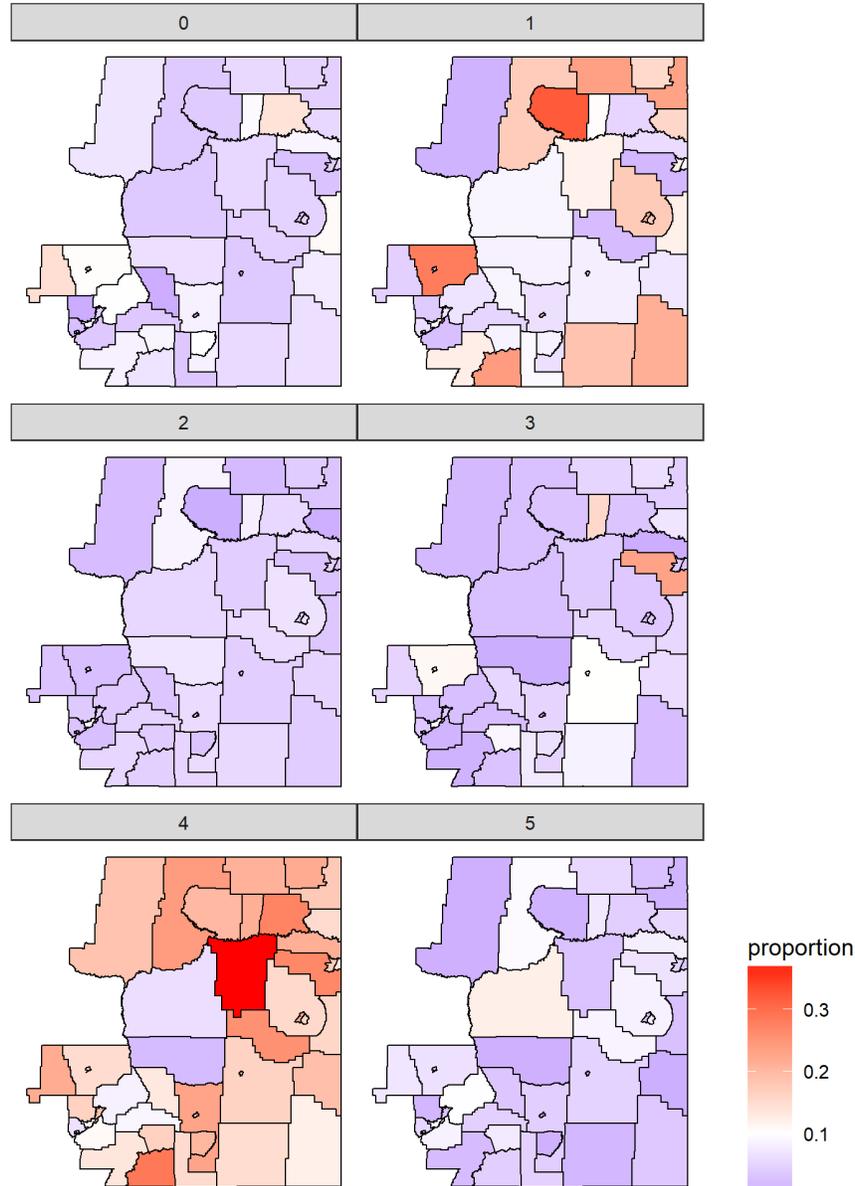


Figure E.0: Maps of bin estimates for the \$10,000 to \$15,000 bin in PUMA 600, MT. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Montana PUMA  
Proportion of households with income between \$5,000 and \$10,000

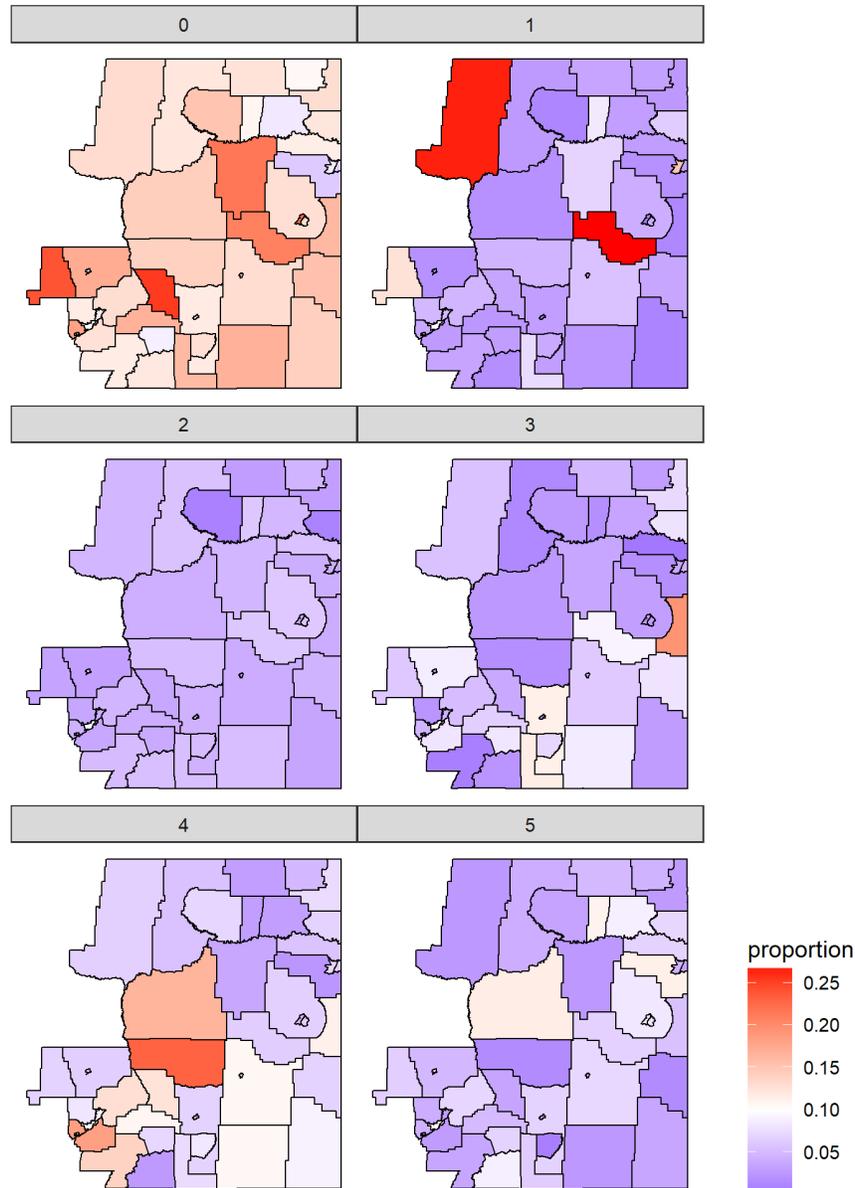


Figure E.0: Maps of bin estimates for the \$35,000 to \$50,000 bin in PUMA 600, MT. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Montana PUMA  
Proportion of households with income between \$5,000 and \$10,000

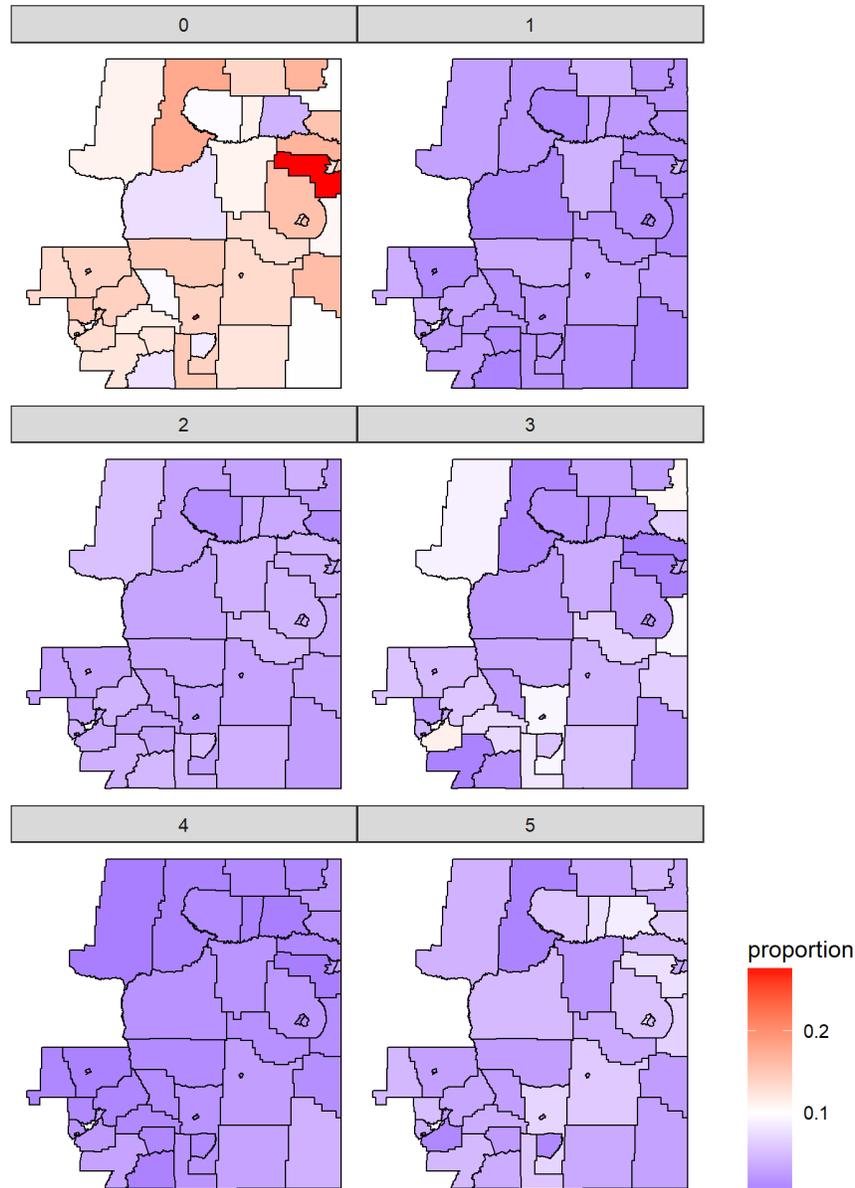


Figure E.0: Maps of bin estimates for the \$75,000 to \$100,000 bin in PUMA 600, MT. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

Prior predictive simulations for Montana PUMA  
Proportion of households with income between \$5,000 and \$10,000

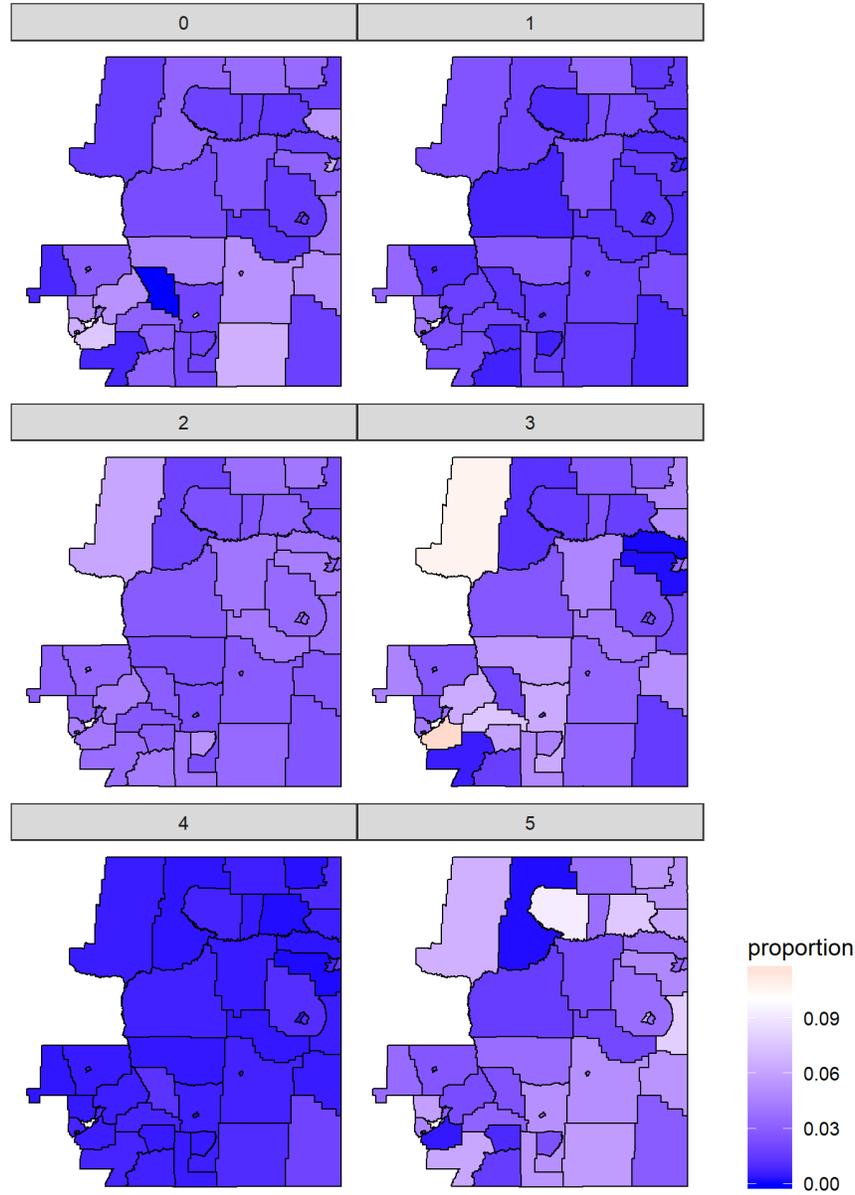


Figure E.0: Maps of bin estimates for the \$150,000 to \$200,000 bin in PUMA 600, MT. The top left map is the observed estimates, the rest are simulated from the prior predictive distribution.

## F Additional Results

This appendix contains addition tables of results from Sections 4 and 5.

	Base Models			Mix-2 Models			Mix-3 Models		
	PR-LN	IID	TIKL-2	IID	TIKL-2	TIKL-4	IID	TIKL-2	TIKL-4
5th	2.44	301.76	331.47	<b>-0.06</b>	17.02	6.16	<b>-0.78</b>	9.62	<b>-2.68*</b>
10th	<b>-1.94</b>	284.03	311.44	2.55	10.42	<b>0.15</b>	0.97	4.34	<b>-7.59*</b>
15th	-1.86	232.61	255.15	-0.68	1.81	<b>-6.70</b>	-1.59	<b>-2.59</b>	<b>-12.35*</b>
20th	-4.02	170.79	188.47	0.06	-3.05	<b>-8.82</b>	-0.07	<b>-5.73</b>	<b>-13.16*</b>
25th	-1.96	103.44	116.09	1.48	-7.13	<b>-9.49</b>	2.07	<b>-8.21</b>	<b>-12.91*</b>
30th	-2.79	46.76	54.57	1.49	<b>-7.39</b>	-7.28	2.14	<b>-8.60</b>	<b>-11.01*</b>
35th	-3.12	21.00	24.65	0.40	<b>-3.55</b>	-3.06	0.21	<b>-6.61</b>	<b>-8.92*</b>
40th	<b>-4.64</b>	34.05	34.60	-3.02	0.88	1.55	-3.97	<b>-4.05</b>	<b>-7.08*</b>
45th	<b>-3.76</b>	84.63	83.32	-3.17	5.23	5.19	<b>-4.18*</b>	1.06	<b>-3.76</b>
50th	<b>-1.58</b>	144.29	142.12	-1.56	3.76	2.76	<b>-2.27</b>	2.67	<b>-2.64*</b>
55th	<b>-2.56*</b>	199.88	197.19	-0.79	4.22	2.52	<b>-1.32</b>	4.18	<b>-0.98</b>
60th	<b>-3.34*</b>	235.49	232.60	<b>-1.00</b>	4.73	3.34	<b>-1.67</b>	3.61	-0.82
65th	<b>-5.45*</b>	243.49	240.66	<b>-2.51</b>	7.21	6.07	<b>-3.36</b>	3.21	-1.15
70th	<b>-4.29</b>	236.34	233.81	<b>-3.94</b>	13.09	12.34	<b>-4.73*</b>	4.80	0.31
75th	<b>-3.11</b>	190.48	188.60	<b>-8.98</b>	14.39	14.56	<b>-9.59*</b>	1.69	-1.92
80th	<b>-0.83</b>	131.29	130.24	<b>-10.35</b>	24.34	25.70	<b>-11.49*</b>	5.98	5.81
85th	<b>-1.81</b>	58.81	58.38	<b>-9.61</b>	29.13	32.20	<b>-11.47</b>	8.20	17.16
90th	<b>2.25</b>	84.31	82.28	<b>-12.01</b>	28.69	35.31	<b>-15.16*</b>	8.28	42.27
95th	54.25	295.90	289.72	<b>-13.30</b>	24.53	59.71	<b>-17.76*</b>	<b>10.18</b>	132.01
Gini	<b>10.64</b>	408.49	402.15	<b>6.27</b>	14.46	12.70	<b>4.27*</b>	13.98	14.47

Table F.1: Difference between RMSE of model-based estimates and RMSE of hold-out estimates as a percentage of RMSE of hold-out estimates for the fixed population, averaged over each tract. The three best performing models for each estimate (in each row) are bold and best performing model is additionally starred. Based on 1008 repeated samples from the synthetic population.

Model	20th	40th	60th	80th	95th	Gini
PR-LN	<b>3.44*</b>	<b>2.36*</b>	<b>1.81*</b>	<b>2.78*</b>	<b>3.38*</b>	<b>4.60*</b>
Base IID	14.83	5.64	3.56	4.53	7.89	12.88
Base TIKL-2	14.80	5.66	3.54	4.44	7.81	13.02
Mix-2 IID	13.18	4.88	3.11	3.96	<b>6.40</b>	<b>9.33</b>
Mix-2 TIKL-2	11.28	4.33	2.39	5.39	7.87	15.04
Mix-2 TIKL-4	11.17	4.40	<b>2.42</b>	5.33	7.62	14.96
Mix-3 IID	11.14	4.82	2.67	<b>3.11</b>	<b>4.48</b>	<b>8.72</b>
Mix-3 TIKL-2	<b>8.97</b>	<b>3.88</b>	2.53	3.63	8.27	14.06
Mix-3 TIKL-4	<b>9.59</b>	<b>3.88</b>	<b>2.39</b>	<b>3.48</b>	8.01	14.60

Table F.2: Absolute percentage difference between the model-based estimates and the hold-out ACS estimates averaged across all tracts for each model and hold-out estimate type for PUMA 821 in CO. The three best performing models for each estimate (in each column) are bold, and the best performing model is additionally starred. 14 of the 28 tracts are omitted in the calculation of the 95th percentile column because they did not have a published estimate.

Model	20th	40th	60th	80th	95th	Gini
PR-LN	<b>3.12*</b>	<b>2.57*</b>	<b>3.13*</b>	<b>4.39*</b>	<b>16.00</b>	<b>13.12</b>
Base IID	40.32	13.49	6.73	6.38	18.06	14.61
Base TIKL-2	40.29	13.43	6.64	6.24	18.14	14.74
Mix-2 IID	22.98	10.26	6.03	<b>5.80</b>	<b>13.85</b>	<b>9.11</b>
Mix-2 TIKL-2	<b>16.76</b>	<b>7.70</b>	<b>5.96</b>	8.28	21.22	14.46
Mix-2 TIKL-4	<b>18.22</b>	7.85	6.19	8.34	21.46	14.61
Mix-3 IID	23.47	<b>7.80</b>	<b>4.87</b>	<b>5.51</b>	<b>12.41*</b>	<b>8.53*</b>

Table F.3: Absolute percentage difference between the model-based estimates and the hold-out ACS estimates averaged across all tracts for each model and hold-out estimate type for PUMA 3502 in IL. The three best performing models for each estimate (in each column) are bold, and the best performing model is additionally starred. 47 of the 54 tracts are omitted in the calculation of the 95th percentile column because they did not have a published estimate. Additionally, one tract is omitted from the 20th, 40th, and 60th percentile columns, and 9 tracts are omitted from the 80th percentile column for the same reason.

Model	20th	40th	60th	80th	95th	Gini
PR-LN	<b>2.48*</b>	<b>2.76*</b>	<b>2.54*</b>	<b>2.40*</b>	<b>4.24*</b>	<b>3.39*</b>
Base IID	15.40	5.50	6.42	<b>6.40</b>	9.76	7.19
Base TIKL-2	15.06	5.42	6.48	<b>6.40</b>	9.68	7.15
Mix-2 IID	13.99	5.78	6.04	6.52	<b>8.76</b>	<b>5.94</b>
Mix-2 TIKL-2	<b>8.41</b>	<b>4.82</b>	<b>4.72</b>	7.92	11.35	11.06
Mix-2 TIKL-4	<b>8.36</b>	<b>4.79</b>	4.77	7.92	11.35	11.09
Mix-3 IID	9.72	5.52	<b>3.46</b>	<b>3.50</b>	<b>5.47</b>	<b>4.83</b>

Table F.4: Absolute percentage difference between the model-based estimates and the hold-out ACS estimates averaged across all tracts for each model and hold-out estimate type for PUMA 600 in MT. The three best performing models for each estimate (in each column) are bold, and the best performing model is additionally starred. 3 of the 51 tracts are omitted in the calculation of the 95th percentile column because they did not have a published estimate.

## G Functionals for Shifted Distributions

In Section 5 we used a shifted version of the lognormal model for the tract-level and stratum-level distributions, defined in terms of the non-positive offset  $Z_{\text{offset}}$ . If  $Y$  has a mixture of shifted lognormal distributions, then  $Y - Z_{\text{offset}} \sim \sum_{k=1}^K \omega_k \text{LN}(\mu_k, \sigma_k^2)$ . We briefly mentioned that this offset forces us to redefine each of the functionals in the tract-level data model of the larger model. Here we provide formulas for each of these functionals. Throughout this section we will let  $X$  denote a random variable with the “unshifted” probability distribution, and  $Y$  denote a random variable with the “shifted” probability distribution, i.e.  $Y = X + Z_{\text{offset}}$ . The formulas for the functionals of the unshifted distributions are found in Section 3.1. This section provides formulas that convert the results of those formulas into the corresponding functional of the same distribution shifted by  $Z_{\text{offset}}$ . These formulas apply no matter what the unshifted probability distribution is, but the context of this paper is it always either a lognormal or mixture of lognormals.

### G.1 Moments

Let  $m_X$  denote the mean of the unshifted distribution, and let  $v_X$  and  $s_X$  denote the variance and standard deviation respectively. Then for the shifted distribution  $m_Y = m_X + Z_{\text{offset}}$ ,  $v_Y = v_X$ , and  $s_Y = s_X$  from the basic properties of means and variances.

### G.2 Quantiles

Suppose that  $\Pi_X^{-1}(\tau)$  is the quantile function of the unshifted distribution. Then the quantile function of the shifted distribution is  $\Pi_Y^{-1}(\tau) = \Pi_X^{-1}(\tau) + Z_{\text{offset}}$  from the basic properties of quantiles.

### G.3 Bin estimates

Suppose  $B_X = P(a \leq X \leq b) = \Pi_X(b) - \Pi_X(a)$  is the probability mass between  $a$  and  $b$  in the unshifted distribution, where  $\Pi_X(\cdot)$  is the CDF of the unshifted distribution. Then the CDF of the shifted distribution is  $\Pi_Y(y) = \Pi_X(y - Z_{\text{offset}})$  so that  $B_Y = \Pi_X(b - Z_{\text{offset}}) - \Pi_X(a - Z_{\text{offset}})$ .

### G.4 Gini coefficient

The definition of the Gini coefficient for a random variable  $X$  with continuous CDF  $\Pi_X(x)$  in terms of its Lorenz curve  $L_X(\tau)$  is

$$G_X = 1 - 2 \int_0^1 L_X(\tau) d\tau$$

while  $L_X(\tau)$  is the Lorenz curve defined by

$$L_X(\tau) = \frac{1}{m_X} \int_0^\tau \Pi_X^{-1}(u) du,$$

where  $\Pi_X^{-1}(\cdot)$  is the quantile function for  $X$ , and  $m_X$  is the mean of  $X$ . Let  $Y = X + Z_{\text{offset}}$ . Then from the definition of the Lorenz curve and for  $Z_{\text{offset}} \neq m_X$

$$L_Y(\tau) = \frac{Z_{\text{offset}}}{Z_{\text{offset}} + m_X} \tau + \frac{m_X}{Z_{\text{offset}} + m_X} L_X(\tau),$$

which implies

$$G_Y = \frac{m_X}{Z_{\text{offset}} + m_X} G_X - \frac{Z_{\text{offset}}}{m_X + Z_{\text{offset}}}.$$

The Gini coefficient is only restricted to be non-negative when the underlying distribution has only positive support. So it is possible for  $G_Y$  to be negative since we defined  $Z_{\text{offset}}$  to be non-positive in Section 5, though this is unlikely so long as  $Z_{\text{offset}}$  is close to zero.

## References

- Aigner, D. J. and Goldberger, A. S. (1970). “Estimation of Pareto’s law from grouped observations.” *Journal of the American Statistical Association*, 65, 330, 712–723.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. London, UK: Chapman and Hall.
- Betancourt, M. (2017). “A Conceptual Introduction to Hamiltonian Monte Carlo.” *arXiv Preprint arXiv:1701.02434*.
- Betancourt, M. and Girolami, M. (2015). “Hamiltonian Monte Carlo for hierarchical models.” *Current Trends in Bayesian Methodology With Applications*, 79, 30.
- Bradley, J. R., Wikle, C. K., and Holan, S. H. (2017). “Regionalization of multiscale spatial processes using a criterion for spatial aggregation error.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 3, 815–832.
- Braithwaite, J. (2015). “Sexual violence in the backlands: Toward a macro-level understanding of rural sex crimes.” *Sexual Abuse*, 27, 5, 496–523.
- Celeux, G., Hurn, M., and Robert, C. P. (2000). “Computational and inferential difficulties with mixture posterior distributions.” *Journal of the American Statistical Association*, 95, 451, 957–970.
- Chambers, R. L., Steel, D. G., Wang, S., and Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. Boca Raton, FL: CRC Press.

- Cressie, N. and Johannesson, G. (2008). “Fixed rank kriging for very large spatial data sets.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 1, 209–226.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.
- Cressie, N. A. and Johannesson, G. (2006). “Spatial prediction for massive datasets.” In *Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, 1–11.
- Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). “Statistics for spatial functional data: some recent contributions.” *Environmetrics*, 21, 3-4, 224–239.
- Fay, R. E. and Train, G. (1995). “Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties.” In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA*, 154–159. Taylor & Francis.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.
- Ferreira, J. and Menegatto, V. (2009). “Eigenvalues of integral operators defined by smooth positive definite kernels.” *Integral Equations and Operator Theory*, 64, 1, 61–81.
- Gelman, A. (2007). “Struggles with survey weighting and regression modeling.” *Statistical Science*, 22, 2, 153–164.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., and Rubin, D. B. (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: CRC Press.

- Gelman, A., Lee, D., and Guo, J. (2015). “Stan: A probabilistic programming language for Bayesian inference and optimization.” *Journal of Educational and Behavioral Statistics*, 40, 5, 530–543.
- Gelman, A. and Rubin, D. B. (1992). “Inference from iterative simulation using multiple sequences.” *Statistical Science*, 7, 4, 457–472.
- Hajargasht, G. and Griffiths, W. E. (2013). “Pareto–lognormal distributions: Inequality, poverty, and estimation from grouped income data.” *Economic Modelling*, 33, 593–604.
- Hardman, A. and Ioannides, Y. M. (2004). “Neighbors’ income distribution: economic segregation and mixing in US urban neighborhoods.” *Journal of Housing Economics*, 13, 4, 368–382.
- Henson, M. F. and Welniak, E. (1980). “Money income of families and persons in the United States: 1978.” Series P 60, No. 123. US Government Printing Office.
- Hipp, J. R. (2007a). “Block, tract, and levels of aggregation: Neighborhood structure and crime and disorder as a case in point.” *American Sociological Review*, 72, 5, 659–680.
- (2007b). “Income inequality, race, and place: Does the distribution of race and class within neighborhoods affect crime rates?” *Criminology*, 45, 3, 665–697.
- Hipp, J. R., Butts, C. T., Acton, R., Nagle, N. N., and Boessen, A. (2013). “Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime?” *Social Networks*, 35, 4, 614–625.
- Ioannides, Y. M. and Seslen, T. N. (2002). “Neighborhood wealth distributions.” *Economics Letters*, 76, 3, 357–367.
- Jargowsky, P. A. (1996). “Take the money and run: Economic segregation in US metropolitan areas.” *American Sociological Review*, 61, 6, 984–998.

- Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). “Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling.” *Statistical Science*, 20, 1, 50–67.
- Judkins, D. R. (1990). “Fay’s method for variance estimation.” *Journal of Official Statistics*, 6, 3, 223.
- Kakwani, N. C. and Podder, N. (1976). “Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations.” *Econometrica: Journal of the Econometric Society*, 44, 1, 137–148.
- Kennedy, B. P., Kawachi, I., and Prothrow-Stith, D. (1996). “Income distribution and mortality: cross sectional ecological study of the Robin Hood index in the United States.” *The BMJ*, 312, 7037, 1004–1007.
- Little, R. J. (1991). “Inference with survey weights.” *Journal of Official Statistics*, 7, 4, 405.
- (1993). “Post-stratification: a modeler’s perspective.” *Journal of the American Statistical Association*, 88, 423, 1001–1012.
- (2012). “Calibrated Bayes, an alternative inferential paradigm for official statistics.” *Journal of Official Statistics*, 28, 3, 309.
- (2015). “Calibrated Bayes, an inferential paradigm for official statistics in the era of big data.” *Statistical Journal of the IAOS*, 31, 4, 555–563.
- Mayer, S. E. et al. (2001). “How the growth in income inequality increased economic segregation.” Tech. rep., Northwestern University/University of Chicago Joint Center for Poverty Research.
- Miller, H. P. (1966). “Income Distribution in the United States. A 1960 Census Monograph.” US Government Printing Office.

- Modalsli, J. H. (2011). “Inequality and growth in the very long run: inferring inequality from data on social groups.” Tech. rep., Department of Economics, University of Oslo.
- Moller, S., Alderson, A. S., and Nielsen, F. (2009). “Changing patterns of income inequality in US counties, 1970–2000.” *American Journal of Sociology*, 114, 4, 1037–1101.
- Nielsen, F. and Alderson, A. S. (1997). “The Kuznets curve and the great U-turn: income inequality in US counties, 1970 to 1990.” *American Sociological Review*, 62, 1, 12–33.
- Obled, C. and Creutin, J. (1986). “Some developments in the use of empirical orthogonal functions for mapping meteorological fields.” *Journal of Climate and Applied Meteorology*, 25, 9, 1189–1204.
- Porter, A. T., Holan, S. H., Wikle, C. K., and Cressie, N. (2014). “Spatial Fay–Herriot models for small area estimation with functional covariates.” *Spatial Statistics*, 10, 27–42.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. New York: Springer.
- Reardon, S. F. (2011). “Measures of income segregation.” *Unpublished Working Paper. Stanford Center for Education Policy Analysis*.
- Reardon, S. F. and Bischoff, K. (2011). “Income inequality and income segregation.” *American Journal of Sociology*, 116, 4, 1092–1153.
- Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). “Bayesian spatial quantile regression.” *Journal of the American Statistical Association*, 106, 493, 6–20.

- Rubin, D. B. (1983). “Comment on ”An evaluation of model-dependent and probability-sampling inferences in sample surveys,” by MH Hansen, WG Madow and BJ Tepping.” *Journal of the American Statistical Association*, 78, 384, 803–805.
- Si, Y., Pillai, N. S., and Gelman, A. (2015). “Bayesian nonparametric weighted sampling inference.” *Bayesian Analysis*, 10, 3, 605–625.
- Spiers, E. F. (1977). “Estimation of Summary Measures of Income Size Distribution from Grouped Data.” In *Proceedings of the Social Statistics Section—American Statistical Association*, 252–77.
- Stan Development Team (2016). “RStan: the R interface to Stan.” R package version 2.14.1.
- (2017). *Stan Modeling Language Users Guide and Reference Manual*.
- Stephens, M. (2000). “Dealing with label switching in mixture models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 4, 795–809.
- U.S. Census Bureau (2014). “American Community Survey Design and Methodology Report — Chapter 14: Data Dissemination.” [https://www2.census.gov/programs-surveys/acs/methodology/design\\_and\\_methodology/acs\\_design\\_methodology\\_ch14\\_2014.pdf](https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_ch14_2014.pdf).
- (2017a). “2011-2015 PUMS Accuracy of the Data.” [https://www2.census.gov/programs-surveys/acs/tech\\_docs/pums/ACS2011\\_2015\\_PUMS\\_README.pdf](https://www2.census.gov/programs-surveys/acs/tech_docs/pums/ACS2011_2015_PUMS_README.pdf).
- (2017b). “American Community Survey 2011-2015 ACS 5-year PUMS files ReadMe.” [https://www2.census.gov/programs-surveys/acs/tech\\_docs/pums/accuracy/2011\\_2015AccuracyPUMS.pdf](https://www2.census.gov/programs-surveys/acs/tech_docs/pums/accuracy/2011_2015AccuracyPUMS.pdf).

- (2017c). “American Community Survey Multiyear Accuracy of the Data (5-year 2011-2015).” [https://www2.census.gov/programs-surveys/acs/tech\\_docs/accuracy/MultiyearACSAccuracyofData2015.pdf](https://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/MultiyearACSAccuracyofData2015.pdf).
  - (2017d). “Documentation for the 2011-2015 Variance Replicate Estimates Tables.” [https://www2.census.gov/programs-surveys/acs/replicate\\_estimates/2015/documentation/5-year/2011\\_2015\\_Variance\\_Replicate\\_Tables\\_Documentation.pdf](https://www2.census.gov/programs-surveys/acs/replicate_estimates/2015/documentation/5-year/2011_2015_Variance_Replicate_Tables_Documentation.pdf).
  - (2017e). “Estimating ASEC Variances with Replicate Weights.” [http://thedataweb.rm.census.gov/pub/cps/march/Use\\_of\\_the\\_Public\\_Use\\_Replicate\\_Weight\\_File\\_final\\_PR.doc](http://thedataweb.rm.census.gov/pub/cps/march/Use_of_the_Public_Use_Replicate_Weight_File_final_PR.doc).
  - (2017f). “Five-year Public Use Microdata Sample, 2011 – 2015 American Community Survey.” <https://factfinder.census.gov/>.
  - (2017g). “Table B19080: Household Income Quintile Upper Limits, 2011 – 2015 American Community Survey.” <https://factfinder.census.gov/>.
  - (2017h). “Table B19083: Gini Index of Income Inequality, 2011 – 2015 American Community Survey.” <https://factfinder.census.gov/>.
  - (2017i). “Table S1901: Income in the Past 12 Months, 2011 – 2015 American Community Survey.” <https://factfinder.census.gov/>.
  - (2017j). “Table S2503: Financial Characteristics, 2011 – 2015 American Community Survey.” <https://factfinder.census.gov/>.
- Van Dyk, D. A. and Meng, X.-L. (2001). “The art of data augmentation.” *Journal of Computational and Graphical Statistics*, 10, 1, 1–50.

- Watson, T. (2009). “Inequality and the measurement of residential segregation by income in American neighborhoods.” *Review of Income and Wealth*, 55, 3, 820–844.
- Welniak, E. (1988). “Calculating indexes of income concentration (Gini’s) from grouped data: An empirical study.” Internal Memorandum, Income Statistics Branch, US Census Bureau.
- Wikle, C. K. (2010). “Hierarchical modeling with spatial data.” In *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, 89–106. Boca Raton, FL: CRC Press.
- Wikle, C. K. and Cressie, N. (1999). “A dimension-reduced approach to space-time Kalman filtering.” *Biometrika*, 86, 4, 815–829.
- Yang, W.-H., Wikle, C. K., Holan, S. H., Myers, D. B., and Sudduth, K. A. (2015). “Bayesian analysis of spatially-dependent functional responses with spatially-dependent multi-dimensional functional predictors.” *Statistica Sinica*, 25, 1, 205–223.
- Yitzhaki, S. (1979). “Relative deprivation and the Gini coefficient.” *The Quarterly Journal of Economics*, 93, 2, 321–324.
- Young, A. (2011). “The Gini coefficient for a mixture of ln-normal populations.”