Interpolating Population Distributions using Public-use Data: An Application to Income Segregation using American Community Survey Data

Matthew Simpson[‡]
SAS Institute
(to whom correspondence should be addressed)
Matt.Simpson@sas.com

Scott H. Holan
Department of Statistics, University of Missouri,
U.S. Census Bureau

Christopher K. Wikle
Department of Statistics, University of Missouri

and

Jonathan R. Bradley
Department of Statistics, Florida State University

November 15, 2021

[*]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, and under NSF grant SES-1853096. 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 those of the NSF or the U.S. Census Bureau.

[†]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.

[‡]The authors thank Noel Cressie for helpful discussion on earlier versions of this manuscript.

Abstract

Income segregation measures the extent to which households choose to live near other households with similar incomes. Sociologists theorize that income segregation can exacerbate the impacts of income inequality, and have developed indices to measure it at the metro area level, including the information theory index introduced in Reardon and Bischoff (2011), and the divergence index presented in Roberto (2015). To study their differences, we construct both indices using recent American Community Survey (ACS) estimates of features of the income distribution. Since the elimination of the decennial census long form, methods of computing these estimates must be updated to use ACS estimates and account for survey error. We propose a model-based method to interpolate estimates of features of the income distribution that accounts for this error. This method improves on previous approaches by allowing for the use of more types of estimates, and by providing uncertainty quantification. We apply this method to estimate U.S. census tract-level income distributions using ACS tabulations, and in turn use these to construct both income segregation indices. We find major differences between the two indices in the relative ranking of metro areas, as well as differences in how both indices correlate with the Gini index.

Keywords: Bayesian methods, Density estimation, Functional data, Income distribution, Pareto-linear procedure.

1 INTRODUCTION

Sociologists theorize that peer or neighborhood effects of income segregation can exacerbate the impacts of income inequality (Reardon and Bischoff, 2011, and references therein) – households are segregated by income to the extent that households with similar incomes live near each other. To study this, Reardon (2011) and Reardon and Bischoff (2011) develop the rank-order information theory index of income segregation in metro areas, which essentially compares the metro area income distribution to the census tract-level income distributions for each tract in the metro area. They then fit various regression models to the index using decennial census data. More recently, Roberto (2015) notes that the information theory index can give results that conflict with our intuitions about the meaning of income segregation, and suggests an alternative index based on the Kullback–Leibler (KL) divergence, called the divergence index.

Both of these indices require as inputs tract-level income distributions for each census tract within a given metro area. The Reardon and Bischoff (2011) analysis relied on decennial census data, which provides detailed distributional information at the tract level and has no associated survey error. Since the elimination of the decennial census long form, this information is no longer available, and instead data users must rely on American Community Survey (ACS) estimates with associated standard errors. The ACS provides fairly detailed information about tract-level income distributions in the form of *bin estimates*; i.e., estimates of the proportion or number of households in a given census tract 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.

Our goal is to use these and other estimates of features of the tract-level income distributions to estimate each tract-level income distribution. Then, in turn, we use these distributions to construct both income segregation indices and reproduce a portion of the Reardon and Bischoff (2011) analysis using both indices and more recent ACS data. Specifically, we assess the degree to which the Gini index, a measure of income inequality, predicts income segregation as measured by both indices at the household level, and at the household-

race level. The regressions we fit to study this relationship must account for the uncertainty in the ACS estimates we use as covariates, as well as in the estimated indices we use as responses. We fit error-in-variables (EIV) regressions to account for this uncertainty. We find that the two indices substantially disagree about the relative ranking of U.S. metro areas in terms of income segregation. Additionally, the correlation between income inequality as measured by the Gini index and income segregation can depend on which index is used, though this difference is less stark.

Our methodological contribution is to construct tract-level income distributions using only ACS estimates of features of those distributions. Many authors use a method called the "Pareto-linear procedure" (PRLN) to construct these distributions using bin estimates, typically as an intermediate step to obtain an estimate of the Gini index, e.g., Jargowsky (1996); Nielsen and Alderson (1997); Hipp (2007a,b); Moller et al. (2009); Hipp et al. (2013); Braithwaite (2015), among others. PRLN assumes that income is uniformly distributed within bins that include or are below the median, and Pareto distributed in bins above the median, with some exceptions to handle special cases. 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).

However, PRLN suffers from several limitations, especially with respect to our problem. First, PRLN does not quantify uncertainty about the income distribution – it only provides a point estimate. Thus, confidence intervals and standard errors are not available for estimates of the Gini index or segregation indices based on PRLN. Second, PRLN is only able to use bin estimates. The ACS provides many other estimates of features of the income distribution including quantiles, income shares, and the Gini index. Taking these into account should result in more accurate estimates of the income distribution of interest. Third, PRLN does not take into account the standard error associated with the estimates that it does use. This is understandable given that PRLN was designed to be used with decennial census data. However, if data users ignore the standard errors in ACS data, their analyses will understate uncertainty (to the extent they quantify uncertainty at all) and potentially be biased.

We solve these three issues with PRLN by taking a latent density estimation approach

based on PRLN, which we call latent PRLN (L-PRLN). This approach is able to take into account multiple diverse types of estimates associated with a given distribution, and naturally accounts for the inherent uncertainty associated with the estimates used by the model. These estimates are estimates of 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), Ferraty and Vieu (2006), and Kokoszka and Reimherr (2017) for overviews. However, our case differs from the usual FDA case because the latent functions we are attempting to estimate are probability distribution functions (PDFs), or equivalently any function that 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. There are several small area estimation (SAE) approaches concerned with estimation of income and other related quantities (such as poverty or per capita household expenditures etc.) Nevertheless, these are typically unit-level models that directly use income (or other proxy variable) (e.g., see Battese et al., 1988; Elbers et al., 2003; Marchetti et al., 2012; Molina and Rao, 2010; Tarozzi and Deaton, 2009; Tzavidis et al., 2008, among others) or area-level models (e.g., see Fay III and Herriot, 1979) that use a direct estimator of income as the model inputs. In contrast, our model uses features of the income distribution as the model inputs rather than income (or another proxy variable) directly. Specifically, we can not fit the SAE models previously listed based on the information that encompasses our model inputs.

Similarly, our approach is also related to the literature on density estimation. The most popular approach is kernel density estimation (e.g. Scott, 2015), but this approach does not directly apply to our setting since we do not have observations drawn from the distribution of interest. Another approach is log splines (Kooperberg and Stone, 1992; Stone et al., 1994), which is subject to the same criticism for our problem. In essence, however, our model is fundamentally inspired by PRLN and can be motivated from that perspective. The choice of PRLN as a starting point for our model was in part based on computational convenience as well as its semiparametric specification. Nevertheless, other parametric alternatives could also be considered, e.g., see Singh and Maddala (1976) and Dagum (1977) among others.

However, purely parametric approaches may be less flexible accross a broad array of applications.

The remainder of the paper is organized as follows. In Section 2 we begin by describing the ACS and available estimates of features of the income distribution, then in Section 2.1 we describe PRLN, and use it to motivate L-PRLN in Section 2.2. In Section 3 we compare L-PRLN and PRLN in a pair of tests. First, in Section 3.1 we conduct a simulation study where we repeatedly sample from a fixed synthetic population and fit both models to each sample. Second, in Section 3.2 we fit both models to ACS data and compare model-based estimates to held-out direct estimates of various features of the income distributions. Next, in Section 4, we return to the income segregation index problem. Here we describe both indices, estimate both of them using ACS data, then use both in a partial reproduction of the analysis of Reardon and Bischoff (2011) using more recent ACS data. Finally, in Section 5, we discuss our results and conclude. Supplementary material includes several appendices referenced in the paper.

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, 2017a). Both the ACS estimates and PUMS are currently published based on one and five years of the survey, 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, in previous years areal units with at least 20,000 people also had published 3-year period estimates (U.S. Census Bureau, 2014).

At the PUMA level, the PUMS provides detailed distributional information about a wide variety of variables measured on households and individuals. At the tract level, however, only a set of specific estimates are available. Many variables only have basic summary statistics published, such as means. Some variables, such as household income or age of householder, have more detailed information available, though not necessarily the information a data user is interested in. In 2015 the ACS published the following 5-year tract-level income distribution period estimates: mean income, median income, Gini index of income, the 20th, 40th, 60th, 80th, and 95th percentiles of income, income shares of each quintile and the top 5% of the income distribution, 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, 2017d,e,f,g,h). 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.

2.1 The Pareto-linear procedure

The fundamental problem is to estimate a density π using estimates of various features of that density. PRLN does this by only using the bin estimates. Let $k = 1, 2, \ldots, K$ index bins, and let $\kappa_1 = 0 < \kappa_2 < \cdots < \kappa_K < \kappa_{K+1} = \infty$, denote the bin boundaries, which we will refer to as knots. Then the PRLN density is given by

$$\pi(x) = \sum_{k=1}^{K} p_k f_k(x). \qquad \text{(PRLN density)} \tag{1}$$

where p_k is the probability associated with bin k, and f_k is the probability density within bin k, with support $(\kappa_k, \kappa_{k+1}]$, except in the uppermost bin where f_K has the support (κ_K, ∞). Let k^* denote the index of largest knot below the median according to the bin estimates. Then the PRLN density defines the f_ks via

$$\begin{aligned} f_k(x) &= \frac{1}{\kappa_{k+1} - \kappa_k} \times \mathbb{1}(\kappa_k < x \leq \kappa_{k+1}) && \text{if } k \leq k^* \ , \\ &= \frac{\alpha_k \kappa_k^{\alpha_k} x^{-\alpha_k - 1}}{1 - \left(\frac{\kappa_k}{\kappa_{k+1}}\right)^{\alpha_k}} \times \mathbb{1}(\kappa_k < x \leq \kappa_{k+1}) && \text{if } k^* < k < K \ , \\ &= \alpha_k \kappa_k^{\alpha_K} x^{-\alpha_k - 1} \times \mathbb{1}(\kappa_K < x) && \text{if } k = K \ . \end{aligned} \tag{2}$$

The unknown parameters of the model, which need to be estimated, are the knot probabilities, $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$, as well as the Pareto parameters, $\boldsymbol{\alpha} = (\alpha_{k^*+1}, \alpha_{k^*+2}, \ldots, \alpha_K)$.

PRLN estimates the p_ks with the associated bin estimates, which we will denote by b_k for $k = 1, 2, \ldots, K$. Then PRLN estimates the Pareto parameters as follows. Let $B_k = \sum_{i=k}^{K} b_i$ for $k = 1, 2, \ldots, K$. The initial PRLN estimate for α_k is given by

$$\widehat{\alpha}_k = \log(B_k / B_{k-1}) / \log(\kappa_k / \kappa_{k-1}).$$

If $\widehat{\alpha}_k \leq 1$, then in the truncated Pareto bins, PRLN reverts to a uniform distribution. In the uppermost bin, which is untruncated Pareto distributed, PRLN instead tries to use $\widehat{\alpha}_{K-1}$, i.e. the $\widehat{\alpha}$ from the bin just below it, as long as that bin was truncated Pareto distributed. If $\widehat{\alpha}_{K-1} \leq 1$, then it tries to use $\widehat{\alpha}_{K-2}$, and so on, until it reaches the last Pareto distributed bin. If it runs out of Pareto bins in this manner, then PRLN assumes that the uppermost bin is a point mass at the lower bound.

The PRLN density a judicious choice because there is not much information about the income distribution between the boundaries of the bins defining the bin estimates. This makes it difficult to estimate a large number of p_ks, or a larger number of parameters associated with the f_ks. The chosen knots help to minimize the number of p_ks as much as possible, and by assuming uniform distributions within the lower bins, PRLN further reduces the number of parameters to estimate. Additionally, since income distributions are known to have approximately Pareto right tails the Pareto bins are likely to fit well.

2.2 L-PRLN: A semiparametric latent density model

Despite its effectiveness, PRLN suffers from three major flaws for our purposes. It cannot quantify uncertainty and only provides point estimates, it cannot take into account all available estimates of features of the income distribution, and it does not take into account the standard error associated with the ACS estimates. Our key innovation to solve these problems is to treat the density as latent, and the published estimates as estimations of functionals of that density with some associated error.

Let $u = 1, 2, \ldots, U$ index the available published estimates, e.g. from the ACS, let q_u denote the estimate and S_u its standard error, and let $Q_u(\cdot)$ denote the functional that takes a probability distribution and returns the value of the estimand for that distribution. For example, if q_u is an estimate of the mean, $Q_u(\pi) = \mathrm{E}_\pi[X]$. Typically a central limit theorem applies for the estimates, so we assume

$$q_u|\pi, S_u \overset{ind}{\sim} \mathrm{N}(Q_u(\pi), S_u^2) \qquad \text{(data model)} \tag{3}$$

for $u = 1, 2, \ldots, U$. The estimate errors are correlated, but these correlations are not available in the ACS, and in general are rarely publicly available. When they are available, (3) can be modified appropriately to take into account the full error covariance matrix.

Next, we need a model for π. In theory, the class of densities used by log spline density estimation (Stone et al., 1994) or kernel density estimation (Scott, 2015) could be used here, but a fundamental constraint is that we need to be able to compute $Q_u(\pi)$ quickly for many different Q_us, including, for example, the mean of the density. Instead we use the PRLN density defined in (1) and (2). As long as $\alpha_K > 1$, then the CDF, quantile function, mean, income shares, and Gini coefficient of the PRLN density are all available in closed form. If $\alpha_K > 2$ then additionally the variance is available in closed form. See Appendix B for formulas for each of these functionals.

By treating the PRLN density as latent, we are able to solve all three limitations of PRLN. We easily take into account the standard errors and propagate that uncertainty into our estimates of the latent population and any distributional features of interest. Additionally, we are able to take into account a much wider variety of available estimates of features of

the income distribution.

2.3 Estimation and interpolation

To construct estimates of any feature of the distribution of interest, including interpolating between the end points of the bins, we will use the Bayesian posterior predictive distribution for the latent population in the area of interest. This allows us to construct a posterior distribution for any distributional feature of interest, so long as it can be easily computed for a finite population and we can easily simulate from π conditional on its parameters. Additionally, it allows us to partially take into account the fact that the latent population is finite. We can even relax the finite population requirement so long as the distributional feature can be easily computed as a function of the model parameters.

We first must sample from the posterior of $\boldsymbol{\theta}$ to be able to sample from the posterior predictive distribution. We do this using the No-U-Turn Sampler (NUTS; Hoffman and Gelman, 2014), a variant of Hamiltonian Monte Carlo (HMC; Neal, 2011). One reason for this choice is that conditional conjugacy in a Gibbs sampler is hopeless due to the form of the Q_us. Additionally, NUTS tends to be more robust and efficient than other MCMC options even when conjugacy relationships are available (Betancourt and Girolami, 2015). We use the software package `Stan` (Gelman et al., 2015; Stan Development Team, 2016) to perform NUTS. NUTS requires the log-posterior and thus log-likelihood be available in closed form, up to an additive constant. The log-likelihood is implied by equation (3) with the Q_us defined in Appendix B.

To construct the posterior predictive distribution of the latent population, let N denote an estimate of the population of the area of interest, e.g. from the ACS. Let $i = 1, 2 \ldots, N$ index the latent population, let Y_i denote the ith latent income, and let $\boldsymbol{\theta}$ denote the full vector of unknown parameters. Then for each posterior sample $\boldsymbol{\theta}^{(m)}$, $m = 1, 2, \ldots, M$ we generate the latent population via

$$Y_i^{(m)} | \boldsymbol{\theta}^{(m)} \overset{iid}{\sim} \pi_{\boldsymbol{\theta}^{(m)}} \qquad \text{(posterior predictive distribution)} \tag{4}$$

for $i = 1, 2, \ldots, N$ and $m = 1, 2, \ldots, M$. This is easily performed in a two step process.

First, generate the bin the observation belongs to using $(p_1^{(m)}, \ldots, p_K^{(m)})$ where p_k denotes the probability of bin k. Then conditional on bin k being chosen, $Y_i^{(m)}$ is generated from the density within that bin, f_k, conditional on $\boldsymbol{\theta}$, or more precisely the elements of $\boldsymbol{\theta}$ that determine f_k. Then the posterior distribution of any feature of the latent distribution of income can be obtained as a function of $\boldsymbol{Y}^{(m)} = (Y_1^{(m)}, Y_2^{(m)}, \ldots, Y_N^{(m)})$ for $m = 1, 2, \ldots, M$.

In principle, the standard error of N can be taken into account by treating the true size of the population as an unknown, denoted by η, with estimate N and standard error H. Then for each draw from the MCMC sampler, a new value of η can be drawn via

$$\eta^{(m)} \overset{iid}{\sim} \mathrm{N}(N, H^2).$$

Subsequently, $Y_i^{(m)}$ can be drawn via (4) for $i = 1, 2, \ldots, \eta^{(m)}$. We do not use this approach here and, instead, treat the tract-level population estimates as the truth since it is unlikely to have a major impact on the results, but in cases where the population estimates are near zero and their standard errors are large, it may be worthwhile.

2.4 Inverted quantile estimates

To be able to use gradient based estimation methods such as NUTS, we use the delta method to "invert" the quantile data model. Suppose q is an estimate of the τth quantile, $\Pi^{-1}(\tau)$, with standard error S. We originally assumed that $q \sim \mathrm{N}(\Pi^{-1}(\tau), S^2)$. Using the delta method we obtain (5) as the data model for the corresponding *inverted quantile estimate*, τ,

$$\tau | \pi, q, S \sim \mathrm{N}\left(\Pi(q), \left[\frac{S}{\pi(q)}\right]^2\right). \tag{5}$$

Since π depends on several unknown parameters, NUTS is more difficult because it creates hard to eliminate divergences (see e.g. Betancourt and Girolami, 2015). So we plug in an estimate of $\pi(q)$ using a modification of the original PRLN. See Section 2.1 for a description of PRLN's estimation process. We modify PRLN in two ways. First, our initial estimate of α_k is

$$\widehat{\alpha}_k = \log\left(\frac{B_k + 0.0001}{B_{k-1} + 0.0001}\right) \Big/ \log(\kappa_k / \kappa_{k-1})$$

to prevent bins estimates of zero from causing problems. Second, instead of using a point mass as a last resort in the uppermost bin, we instead use $\widehat{\alpha}_K = 1.0001$. This is more realistic, and should result in a more accurate standard error, e.g., for $\tau = 0.95$. Note that for quantiles which are in bins that are uniform distributed, our plug-in estimate is $\widehat{\pi}(q) = b_{k^*}/(\kappa_{k^*+1} - \kappa_{k^*})$ where k^* is the index of the closest knot from below to q, and b_{k^*} is the corresponding bin estimate.

2.5 Priors

To complete the model, we need to choose priors for the p_ks and the α_ks. An extremely "uninformative" prior for $\boldsymbol{p}$ can cause problems for MCMC, so we opt for a weakly informative prior. Note also that the bins are not designed so that we would expect them to be equally probable *a priori*. Thus, we center $\boldsymbol{p}$ on the ACS 5-year period bin estimates for the entire United States, from the same year as the tract-level estimates, using a Dirichlet prior. Let $\boldsymbol{g}$ denote the country-level estimates, and let t denote a scale hyperparameter, then we assume

$$\boldsymbol{p} \sim \text{Dirichlet}(\boldsymbol{g}/t).$$

The value of t encodes the level of prior certainty that $\boldsymbol{g}$ is the true value of $\boldsymbol{p}$. A value of $t \geq 1$ is ideal since we do not necessarily expect $\boldsymbol{g}$ to be close to $\boldsymbol{p}$ with a high degree of certainty, but this must be balanced against computational considerations. When an element of $\boldsymbol{p}$ is close to zero in the posterior, this can cause problems for NUTS. See Section 5 for a discussion of this issue. As a result, we use a value of $t = 1/10$ which regularizes $\boldsymbol{p}$ away from zero.

For the α_ks, we restrict the prior mass to be above one so that the untruncated Pareto distribution in the rightmost bin has a well defined mean. Note that $\alpha_k > 2$ is necessary to ensure a well defined variance if a user wants to include estimates of the second moment of the income distribution in the data model. Nevertheless, we assume that the α_ks are iid truncated normal distributed as

$$\alpha_k \overset{iid}{\sim} \text{N}(2, 1^2)\mathbb{1}(\alpha_k > 1).$$

In practice we have found using PRLN that the tail bins tend to have estimated α_ks between around one and three, with smaller values in bins further in the right tail. In general, there is not much information in the data to learn the Pareto parameters, so this prior provides some useful regularization to help with model estimation.

3 EVALUATION OF L-PRLN

As mentioned above, the goal of the L-PRLN methodology as presented here is to provide estimates of income segregation indices and their uncertainty for ACS data, taking into account the survey errors and multiple data sources. Although the standard PRLN methodology cannot account for multiple and uncertain data sources, and only provides point estimates, it is useful to see how the L-PRLN approach compares to PRLN when considering only point estimates of income distributions. Thus, we consider two specific cases. First, in Section 3.1, we design a simulation study using a synthetic population generated over the Boone County, MO PUMA (Public-use Micro Area) and its census tracts. We repeatedly sample from this population and create synthetic tract-level ACS estimates, which we use to fit both PRLN and L-PRLN, and then evaluate them based on predictions of various features of the tract-level distributions. Then, in Section 3.2, we use our modeling framework to estimate U.S. census tract-level income distributions using 2015 ACS 5-year period estimates associated with features of tract-level income distributions, and compare these to held out estimates to evaluate PRLN and L-PRLN.

Both of these exercises are designed to make a "fair" comparison between PRLN and L-PRLN. That is, recall that compared to PRLN, L-PRLN is able to use more of the available estimates, account for the uncertainty in those estimates, and provide uncertainty quantification. These are necessary properties to solve our income segregation problem. However, it is important to emphasize that here we compare the performance of L-PRLN to PRLN based only on point estimates, and restrict L-PRLN to use a much more limited subset of the estimates than it is capable of incorporating.

3.1 Simulation study

We construct a synthetic population for our simulation study, and repeatedly sample from it using a stratified random sample based on the strata defined by the 2014 PUMS. We do not fully describe how the synthetic population is generated here; instead, see Appendix C of the Supplementary Material for a detailed description. Additionally, the `R` code (R Core Team, 2020) used to generate the population is included in the Supplementary Material. Figure A.2 in the Supplementary Materials contains maps of the true tract-level means, medians, and standard deviations of income for the synthetic population.

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. Then the synthetic ACS estimates are created using the 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, 2017b,c). 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 each fifth percentile estimate (5th, 10th, etc.) as well as the Gini index so that we can compare them to model-based estimates of the same quantities.

Then we fit L-PRLN using `Rstan` (Stan Development Team, 2016) to do MCMC via NUTS with four chains, and after a warm-up of 4,000 iterations per chain for tuning and burn-in, a further 4,000 iterations per chain were kept as draws from the posterior distribution. Both the mean and the median of the posterior predictive distribution for each percentile were taken as model-based estimates. Additionally, we fit PRLN on the synthetic bin estimates. This yields four estimates of each percentile: the mean and median of the L-PRLN posterior predictive distribution, constructed as in Section 2.3; the PRLN estimate; and the direct estimate. We computed the following four metrics for all four estimates: root mean square error (RMSE), mean absolute deviation (MAD), root mean square percentage error (RMSPE), and mean absolute percentage error (MAPE). All four metrics were com-

puted over all iterations of the simulation study and all tracts of the synthetic population simultaneously.

	Estimator	P5	P10	P15	P20	P25	P30	P35	P40	P45	P50
MAD	P. Mean	8.58	15.38	17.86	15.24	14.80	11.08	-2.00	-10.45	-11.25	-7.21
	P. Median	6.67	10.64	12.02	10.20	11.20	9.51	0.91	-6.27	-8.25	-5.63
	PRLN	1.55	-1.41	-2.31	-4.80	-2.94	-1.80	-3.03	-4.81	-4.00	-1.15
MAPE	P. Mean	3.90	12.99	15.37	12.14	13.17	10.51	-0.43	-7.98	-9.96	-6.44
	P. Median	3.50	8.40	9.60	6.86	9.98	8.44	2.25	-4.01	-6.84	-4.69
	PRLN	-0.83	-1.35	-2.22	-5.62	-2.42	-2.39	-2.73	-4.29	-3.92	-1.02
RMSE	P. Mean	7.61	17.94	23.53	20.07	13.43	7.09	-4.98	-11.71	-12.43	-9.78
	P. Median	6.43	12.86	17.14	15.59	11.05	7.20	-1.60	-7.32	-9.29	-7.77
	PRLN	-0.60	-2.98	-2.36	-3.94	-3.75	-2.95	-4.41	-4.76	-3.88	-2.17
RMSPE	P. Mean	2.67	14.64	19.55	16.63	12.91	8.11	-1.35	-7.34	-9.91	-8.22
	P. Median	2.73	9.82	13.43	11.81	10.24	7.64	1.43	-3.27	-6.74	-6.05
	PRLN	-3.34	-2.59	-2.10	-4.31	-3.28	-3.18	-3.83	-4.03	-3.68	-1.97

Table 1: Percentage difference in a variety of metrics between several estimates and the direct estimates for the first half of the income distribution. The estimates considered include the original Pareto-linear procedure (PRLN) the posterior predictive mean from L-PRLN (P. Mean), and the posterior predictive median from L-PRLN (P. Median). Negative numbers indicate that the method is doing better than the direct estimates.

Tables 1 and 2 display each of these metrics, expressed as a percentage of the same metric for the corresponding direct estimates. For example, PRLN had an RMSE for the 5th percentile 0.60% lower than that of the direct estimate, while it had a MAD for the 5th percentile 1.55% higher than that of the direct estimate. Note that the direct estimates are what our hypothetical data user would like the ACS to publish, but they are not available.

In the lower portion of the income distribution, the PRLN estimate does the best according to most metrics, while the L-PRLN posterior median outperforms the posterior mean. In the middle of the distribution this completely reverses: PRLN does the worst, and the posterior mean outperforms the posterior median. In the upper portion of the distribution but still under the 90th percentile, PRLN does the best again, but the posterior mean still

	Estimator	P55	P60	P65	P70	P75	P80	P85	P90	P95	Gini
MAD	P. Mean	-3.95	-0.08	-0.60	0.91	2.56	7.38	3.52	-1.61	-1.71	14.58
	P. Median	-2.49	1.85	1.75	-0.08	5.03	10.49	6.25	0.87	5.07	12.14
	PRLN	-1.79	-3.17	-6.32	-4.27	-2.77	-0.00	-2.60	0.45	38.89	9.46
MAPE	P. Mean	-4.26	-0.75	-0.84	-0.12	1.17	5.93	2.84	-1.68	-1.85	15.19
	P. Median	-2.80	0.92	1.96	-1.20	3.91	8.59	5.46	0.85	4.93	12.68
	PRLN	-1.71	-2.85	-5.62	-3.59	-3.43	-0.95	-2.62	0.10	37.59	9.71
RMSE	P. Mean	-5.62	-2.61	-2.26	-2.66	-3.00	1.48	-2.48	-6.52	-8.79	10.55
	P. Median	-3.52	-0.20	-0.40	-2.50	-1.25	4.08	0.26	-3.66	-3.91	9.17
	PRLN	-2.12	-3.63	-5.70	-5.04	-4.68	-2.34	-4.08	5.25	57.91	8.96
RMSPE	P. Mean	-5.95	-3.69	-2.89	-4.17	-4.57	-0.47	-3.56	-6.66	-9.26	11.19
	P. Median	-4.05	-1.53	-0.66	-3.90	-2.53	1.70	-1.11	-3.82	-4.36	9.74
	PRLN	-2.05	-3.12	-4.51	-4.03	-5.43	-3.53	-4.34	3.76	53.92	9.35

Table 2: Percentage difference in a variety of metrics between several estimates and the direct estimates for the last half of the income distribution and the Gini coefficient. The estimates considered include the original Pareto-linear procedure (PRLN) the posterior predictive mean from L-PRLN (P. Mean), and the posterior predictive median from L-PRLN (P. Median). Negative numbers indicate that the method is doing better than the direct estimates.

outperforms the posterior median. In the 90th percentile, the posterior mean performs the best, while PRLN performs the worst. In the 95th percentile the same pattern holds, but PRLN performs disastrously bad. This is because if PRLN cannot guarantee an estimate for an α that is greater than one in the top bin, it assumes the bin is a point mass on the bin minimum. See Section 2.1 for details. This can drastically hurt PRLN's predictions in the upper tail, which we see here. L-PRLN does not have this problem since each α is constrained to be greater than one and is regularized away from one by the prior.

So in general, the best performing point-estimate depends on which region of the income distribution the data-user cares about. For the middle of the distribution or the far right tail, L-PRLN is superior, but everywhere else PRLN is superior. PRLN performs the best for the Gini coefficient, with the posterior median outperforming the posterior mean. For other measures of inequality and other functionals of the income distributions, which esti-

mate performs best will depend on how much they load on different regions of the income distribution. Note that this comparison deliberately limited L-PRLN by preventing it from using all of the available estimates – estimates that PRLN cannot use.

It is also important to emphasize that L-PRLN provides uncertainty estimates, which are unavailable in PRLN. As an illustration, Table 3 presents the coverage rates of 95% credible intervals for every fifth percentile, as well as the Gini coefficient. Two coverage rates were computed, one with the true population as reference values and one with the PRLN estimates as reference values. The first comparison shows that L-PRLN's intervals slightly undercover the truth; i.e., the 95% credible intervals cover about 80-90% of the time, but with better coverage in the lower portion of the income distribution. Note that L-PRLN has better coverage precisely where its point estimates do the worst. The second comparison shows that the PRLN measures in the lower part of the distribution are largely contained in the L-PRLN's 95% credible intervals. More precisely, L-PRLN's estimate and PRLN's estimate for a given percentile were statistically indistinguishable at least 60% of the time. This is an underestimate since it does not account for uncertainty in the PRLN estimates, but the statistical properties of PRLN are unknown.

3.2 Application to the American Community Survey

We fit PRLN and L-PRLN to 2015 ACS 5-year period estimates of features of tract-level income distributions for all tracts in five separate PUMAs: 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), PUMA 600 in Montana (a sparsely populated rural PUMA), and 3706 in New York (a poor urban PUMA in New York City). Figure A.3 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 models using each of the bin estimates described in Section 2, as well as the mean and median estimates. We held out estimates of the 20th, 40th, 60th, 80th, and 95th percentile, as well as the Gini coefficient to validate the models. To fit each model we used

Estimand	Population	PRLN	Estimand	Population	PRLN
P5	0.92	0.86	P55	0.79	0.60
P10	0.90	0.81	P60	0.79	0.62
P15	0.89	0.78	P65	0.82	0.69
P20	0.90	0.75	P70	0.80	0.71
P25	0.90	0.74	P75	0.80	0.72
P30	0.89	0.73	P80	0.81	0.73
P35	0.88	0.71	P85	0.85	0.76
P40	0.87	0.69	P90	0.85	0.79
P45	0.85	0.65	P95	0.88	0.75
P50	0.82	0.63	Gini	0.95	0.84

Table 3: Coverage rates of 95% credible intervals from the tract level model for each quantity of interest, averaged over tracts. Coverage rates are computed taking the true population value as the reference value (Population), and taking the PRLN estimate as the reference value (PRLN).

`Rstan` (Stan Development Team, 2016) to do MCMC via NUTS 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.

For L-PRLN, we construct the posterior predictive mean and median for each estimand, as in Section 2.3. We compare each of these estimates as well as estimates from PRLN to each of the held out estimates using the same four metrics as in Section 3.1: RMSE, RMSPE, MAD, and MAPE, all computed across tracts. Tables D.1–D.5 of the Supplementary Materials contain these metrics for each of the five PUMAs we considered. Note that for some tracts, some of the held out estimates were missing – particularly the 95th percentile, and mainly in the IL PUMA.

For most estimands in most tracts, and according to most metrics, L-PRLN does about the same or slightly worse than PRLN. The main exceptions are in either tail of the distribution, where for some tracts the difference between PRLN and L-PRLN is more magnified.

L-PRLN especially has trouble relative to PRLN in the lower tail. On the other hand, L-PRLN often performs better than PRLN for the Gini coefficient, and in particular in the IL PUMA it performs much better for the 95th percentile and consequently for the Gini coefficient. This is due to the phenomenon discussed in Section 3.1, where PRLN sometimes significantly incorrectly estimates the distribution in the upper bin. Additionally, in the CO PUMA, the L-PRLN outperforms PRLN in the middle of the distribution.

4 INCOME SEGREGATION INDICES

Now we turn to our motivating problem: estimating income segregation indices using ACS data. Households are segregated by income to the extent that households with similar incomes choose to live near each other. To measure this, Reardon (2011); Reardon and Bischoff (2011) construct the rank-order information theory index. The basic idea of the index is to construct an entropy measure of the income distribution for an entire metro area, and then construct the same measure for the income distributions for each census tract in the metro area. Then the index is a weighted sum of the relative differences in this entropy measure between each tract and the metro area.

Formally, let $F_i(y)$ denote the CDF of income for census tract i, where $i = 1, 2, \ldots, I$ indexes all census tracts in a given metro area. Then we assume that the income distribution for the metro area, denoted by $F(y) = \sum_{i=1}^{I} w_i F_i(y)$, is a population weighted mixture of the tract-level income distributions, where w_i is the proportion of the metro area's population in tract i. Next, define $E(G||F)$ as the integrated binary entropy from the CDF F to the CDF G, i.e.

$$E(G||F) = \int_{-\infty}^{\infty} e[F(y)]dG(y) \tag{6}$$

where $e(p) = -p\log(p) - (1-p)\log(1-p)$ is binary entropy. Then the rank-order information theory index, denoted by H_R, can be defined as

$$H_R = \sum_{i=1}^{I} w_i \frac{E(F||F) - E(F||F_i)}{E(F||F)}. \tag{7}$$

Since H_R is based on entropy, it is better understood as a measure of the differences in *diversity* of the income distributions between the tract-level and metro-level (Roberto, 2015). Indeed, the following example illustrates the point. Suppose that households in the metro area only have one of two incomes: $y = 30,000$ and $y = 100,000$. In the entire metro area $P(y = 30,000) = 2/3$, while in tract i, $P_i(y = 30,000) = 1/3$. Then for tract i we have

$$\begin{aligned} E(F_i||F) &= -e[F_i(30,000)]P(y = 30,000) - e[F_i(100,000)]P(y = 100,000) \\ &= -e[1/3]\frac{2}{3} - e[1]\frac{1}{3} = -e[2/3]\frac{2}{3} = E(F||F) \end{aligned}$$

since $e(p) = e(1-p)$. So tract i contributes nothing to the metro area's segregation index even though it has a much higher concentration of rich households than the entire metro area.

To remedy this, Roberto (2015) proposes the KL divergence index. Let f denote the PDF associated with F above, and similarly for f_i and F_i. Then the KL divergence index can be defined as

$$D = \sum_{i=1}^{I} w_i D(f_i||f), \qquad D(g||f) = \int_{-\infty}^{\infty} \log \frac{g(y)}{f(y)} g(y) dy \tag{8}$$

where $D(g||f)$ is the KL divergence from the PDF f to the PDF g. In other words, the divergence index is the population weighted sum of the divergences from the metro-level income distribution to each of the tract-level distributions.

4.1 Correlates of income segregation

Reardon and Bischoff (2011) investigate the correlates of income segregation as measured by H_R. In particular, they are interested in whether income inequality, as measured by the Gini index, is correlated with income segregation. They consider the largest 100 metro areas in the U.S. by population, and fit a variety of regression models controlling for various covariates. We focus on a portion of their Table 4, which reports the results of severale regression models, of which we focus on three: one for all families, one for black families only, and one for white families only. In these models they control for the year of the census, various metro-year and race-metro-year covariates, and include metro fixed effects. They find

a stable positive relationship between the Gini coefficient and H_R. Further, the strength of this relationship is about the same for white families alone as it is for black families alone. Our aim is to attempt to rerun these regressions using recent ACS data, then run them again replacing H_R with the divergence index.

Since we use ACS data, our controls and data differ in several ways in general. First, we use the top 100 metro areas by population according to the 2018 ACS 5-year period estimates of population. This list may not be identical to the list used by Reardon and Bischoff (2011). Second, we use ACS estimates for households instead of families because more of the required variables are available, though Reardon and Bischoff (2011) note that they would have preferred to do a household level analysis, but it was not possible due to data limitations. Finally, we only use a single year of ACS 5-year period estimates. The ACS is not old enough to have more than two years of non-overlapping 5-year period estimates. We use the 2018 5-year period estimates. 2013 5-year period estimates are also available, though the definitions of several covariates differ across vintages. To avoid this complication, we only use a single year of estimates. This leaves us with no within-metro variation in any of the three regression models, so we omit the metro area fixed effects. Otherwise, we attempt to faithfully include every covariate in the regression of Reardon and Bischoff (2011) in our own regressions. Appendix E of the Supplementary Materials describes how each covariate was sourced from the 2018 ACS 5-year period estimates, including how standard errors were constructed if necessary. For black households and white households, two covariates are not available: percentage of households with a female householder, and the Gini index. We omit the female householder covariate in the black households and white households regressions for this reason. However, we take advantage of L-PRLN to construct the metro-level Gini index for black households and white households only along with its standard error using the available metro-race-level income estimates. See Appendix F for a description of how this was performed.

To construct both H_R and D for a given metro area, we first fit L-PRLN to the household ACS 5-year period estimates of variations features of the household income distribution for each tract in that metro area. For the household income distributions, we use bin

estimates with the same boundaries as in Section 3.1, mean and median estimates, estimates of the 20th, 40th, 60th, 80th, and 95th percentile, estimates of the income shares of the quintiles of the income distribution as well as the top 5% of the income distribution, and an estimate of the Gini coefficient (U.S. Census Bureau, 2020a,b,c,d,e). For many tracts, some of these estimates or their standard errors are not available. For those tracts we proceed with whichever estimates with standard errors are available. In all cases if the mean estimate or its standard error was not available for the tract, or if the ACS estimate of the population of households was less than 100, the tract was omitted from the analysis. The same procedure was applied to estimating the tract-level income distributions of black households alone, and of white households alone, again as long as there were at least 100 households of the given race in the tract according to the ACS. The only available tract-race-level household income distribution ACS estimates are the bin, mean, and median estimates. The same priors as in Section 2.5 were used, except in the black households models, country-level bin estimates for only black households were used to center the prior on the bin probabilities, and similarly for the white households models.

We cannot use the approach in Section 2.3 to estimate H_R and D using the L-PRLN income distribution estimates because both H_R and D will yield nonsensical results if each tract's income distribution does not have the same support. So instead we treat both indices as a function of the underlying tract-level parameters. Then we approximate the integrals in (6) and (8) for each draw from the posterior distribution using importance sampling techniques – see Appendix G for details. The result of this process is that for a given metro area, we obtain a joint posterior sample of the index and the standard error associated with approximating the integrals. Our approach is the same for computing H_R and D by race in a given metro area.

4.2 Error-in-variables regression

To fit the regressions, we must contend with two complications that were not present in Reardon and Bischoff (2011). First, the response and each covariate of each regression is measured with error, though in each case the standard error is known. Second, instead of

observing the response and its standard error, we observe a sample from the joint posterior distribution of the response and its standard error.

The solution to the first issue is an EIV regression; e.g., see Carroll et al. (2006), Arima et al. (2015), and the references therein. But to use that, we first need to solve the second problem using the variance decomposition formula. Let $\boldsymbol{\theta}$ denote all unknown parameters of the tract-level L-PRLN models for a given metro area, let $d^*(\boldsymbol{\theta})$ denote a segregation index as a function of those parameters, let d denote our estimate of that index, and let $h(\boldsymbol{\theta})$ denote the corresponding standard error. Then conditional on the model parameters we have $d|\boldsymbol{\theta} \sim \mathrm{N}(d^*(\boldsymbol{\theta}), h^2(\boldsymbol{\theta}))$. Then we can write $\mathrm{E}[d] = \mathrm{E}[d^*(\boldsymbol{\theta})]$ and $\mathrm{var}[d] = \mathrm{E}[h^2(\boldsymbol{\theta})] + \mathrm{var}[d^*(\boldsymbol{\theta})]$. Given a sample $\{(d_m, h_m^2) : m = 1, 2, \ldots, M\}$ from the joint posterior of $(d, h^2(\boldsymbol{\theta}))$, we can approximate these quantities by

$$\begin{aligned}
\mathrm{E}[d] &\approx \overline{d} = \frac{1}{M}\sum_{m=1}^{M} d_m \\
\mathrm{var}[d] &\approx \overline{h}^2 = \frac{1}{M}\sum_{m=1}^{M} h_m^2 + \frac{1}{M-1}\sum_{m=1}^{M}(d_m - \overline{d})^2.
\end{aligned}$$

Then for simplicity we assume a simple measurement error model using these quantities:

$$\overline{d} \sim \mathrm{N}(d^*, \overline{h}^2),$$

where d^* is the true underlying index. This approach works for both H_R and D, and allows us to completely reduce the regression problem to EIV regression.

The EIV regression model we employ can be written as follows. Let $i = 1, 2, \ldots, I$ index metro areas, let $\overline{d}_i$ denote either segregation index for that metro area, and let $\overline{h}_i$ denote its associated standard error. Let $\boldsymbol{x}_i$ denote a vector of covariates for the metro area, with $\boldsymbol{S}_i$ the associated (diagonal) error covariance matrix. Further, let d_i^* and $\boldsymbol{x}_i^*$ denote the latent true values of the index and covariates, respectively. Then the model is given by (9)

$$\begin{aligned}
\overline{d}_i|\boldsymbol{x}_i, d_i^*, \boldsymbol{x}_i^* &\sim \mathrm{N}(d_i^*, \overline{h}_i^2) \\
\boldsymbol{x}_i|d_i^*, \boldsymbol{x}_i^* &\sim \mathrm{N}(\boldsymbol{x}_i^*, \boldsymbol{S}_i) \\
d_i^*|\boldsymbol{x}_i^* &\sim \mathrm{N}(\alpha + (\boldsymbol{x}_i^*)'\boldsymbol{\beta}, \tau^2) \\
x_{ij}^* &\overset{ind}{\sim} \mathrm{N}(\mu_j, \sigma_j^2),
\end{aligned} \tag{9}$$

for $i = 1, 2, \ldots, I$, where $j = 1, 2 \ldots, J$ indexes covariates. To complete the model we need priors on α, $\boldsymbol{\beta}$, τ^2, the μ_js, and the σ_j^2s.

We specify priors on the standardized regression coefficients for ease of interpretation and elicitation, and on the corresponding standardized versions of all other parameters, i.e. on $\widetilde{\beta}_j = \beta_j s_{x_j}/s_{\overline{d}}$, $\widetilde{\mu}_j = (\mu_j - \overline{x}_j)/s_{x_j}$, and $\widetilde{\sigma}_j = \sigma_j/s_{x_j}$ for $j = 1, 2 \ldots, J$, and on $\widetilde{\alpha} = (\alpha - \overline{\overline{d}} + \overline{\boldsymbol{x}}'\boldsymbol{\beta})/s_{\overline{d}}$ and $\widetilde{\tau} = \tau/s_{\overline{d}}$. Using this parameterization, we employ the independent priors listed in (10)

$$
\begin{aligned}
&\widetilde{\alpha} \sim \mathrm{N}(0, 100^2) && \\
&\widetilde{\beta}_j \overset{iid}{\sim} \mathrm{N}(0, 3^2) && \text{for } j = 1, 2, \ldots, J \\
&\widetilde{\tau} \sim \mathrm{N}^+(0, 0.8^2) && \\
&\widetilde{\mu}_j \overset{iid}{\sim} \mathrm{N}(0, 3^2) && \text{for } j = 1, 2, \ldots, J \\
&\widetilde{\sigma}_j \overset{iid}{\sim} \mathrm{N}^+(0, 2^2) && \text{for } j = 1, 2, \ldots, J.
\end{aligned} \tag{10}
$$

The priors on the $\widetilde{\beta}_j$s imply that for any covariate, we are 68% sure that a one standard deviation change in the covariate will result in no more than a three standard deviation change in the response. The prior on $\widetilde{\alpha}$ implies that we are 68% certain that the intercept will be within 100 sample standard deviations of the response from the sample mean of the response. The half-normal prior on $\widetilde{\tau}$ implies that we are 68% certain that the error variance will be no more than 0.8 times the the total sample variance of the response. All of these priors are loose relative to typical expectation of regressions in the social sciences, but still provide a small amount of regularization.

The priors on the covariate means and standard deviations are similarly loose, and can be thought of as empirical Bayes priors. The priors on the μ_js are loosely centered on the sample means of the x_js, and the priors on the σ_js allow for a wide range of variation around the sample standard deviations of the x_js.

4.3 Results

Figures H.4, H.5, and H.6 in Appendix H demonstrate that the divergence and information theory indices substantially disagree about the relative ranking of metro areas in terms of

income segregation. This demonstrates that Roberto (2015)'s criticism of the information theory index is not merely a theoretical curiosity, but instead that there is significant mismeasurement of income segregation. As a result, we should expect the EIV regression results to differ as well.

Table 4 contains posterior summaries of the Gini index EIV regression coefficient for each model fit — see Appendix H for detailed tables including every covariate. In similar information index regressions, Reardon and Bischoff (2011, Table 4) found that the regression coefficient on the Gini index to be 0.56 for all families, 0.47 for black families, and 0.47 for white families. Our regressions use more recent household-level data, do not have exactly the same covariates, and do not have metro fixed effects. Despite this, our results for all households are broadly consistent with the results of Reardon and Bischoff (2011) for all families. Our results for black and white households are somewhat different. In both cases the 95% credible intervals are much wider, and contain zero. This is largely due to higher standard errors for the black households and white households ACS estimates compared to the standard errors of the all households ACS estimates. The covariates simply contain less useful information for the black households and white households regressions.

However, the Gini index regression coefficients also appear to be much closer to zero for black households and white households. The upper end of the 95% credible intervals do not contain Reardon and Bischoff (2011)'s estimates, and for black households the posterior mean and median are both negative. This may be due to differences in model specification since our black households and all households regressions are missing the female head of household covariate since it is not available in the ACS. Additionally, our regressions do not include metro area fixed effects since we have only one year of data, though this difference is present in the all households regressions as well. That said, we take these regressions as a baseline to compare with divergence index regressions using the same model specification.

The results for the divergence index are significantly different. The coefficient on the Gini index is larger in all cases, though for making these comparisons the standardized coefficients displayed in Table 5 is more meaningful. In that case, the Gini index regression coefficient is similarly sized in the all households regressions, though there is more uncertainty in the

Households	Mean	SD	2.5%	25%	50%	75%	97.5%
Information theory index							
All	0.427	0.058	0.313	0.389	0.427	0.467	0.540
Black	-0.079	0.125	-0.328	-0.161	-0.078	0.004	0.164
White	0.145	0.111	-0.072	0.071	0.146	0.220	0.366
Divergence index							
All	0.763	0.157	0.458	0.657	0.762	0.868	1.074
Black	1.403	0.742	-0.036	0.899	1.402	1.898	2.872
White	0.683	0.240	0.207	0.523	0.683	0.843	1.155

Table 4: Posterior summaries of raw EIV regression coefficients for the Gini index.

divergence index regression. The Gini index coefficients in the black households and white households regressions are once again smaller than in the all households regressions, but this difference is much less extreme for the divergence index than for the information theory index. In fact, the 95% credible interval for white households is strictly above zero, and for black households only just contains zero inside the lower bound. The upshot is that the Gini index appears to be positively associated with the divergence index for black households only and white households only, while the same is not true for the information theory index. That said, there is enough uncertainty that we cannot rule out that there is no meaningful difference between the divergence index coefficients and the information theory index coefficients.

5 DISCUSSION

L-PRLN serves its purposes well. It interpolates the income distribution nearly as well as the original PRLN when forced to use a restricted subset of the available estimates. However, it has several added benefits. First, our L-PRLN is able to take advantage of a wider variety of tract-level estimates than PRLN, including quantile and moment estimates. PRLN is fundamentally limited to using only bin estimates. Second, unlike PRLN, L-PRLN takes into account the standard errors of the tract-level estimates. Finally, while PRLN can only

Households	Mean	SD	2.5%	25%	50%	75%	97.5%
Information theory index							
All	0.566	0.076	0.414	0.515	0.565	0.618	0.714
Black	-0.100	0.157	-0.412	-0.203	-0.098	0.005	0.207
White	0.138	0.106	-0.069	0.067	0.139	0.209	0.348
Divergence index							
All	0.580	0.120	0.348	0.500	0.580	0.661	0.817
Black	0.379	0.200	-0.010	0.243	0.378	0.512	0.775
White	0.399	0.140	0.121	0.306	0.399	0.493	0.675

Table 5: Posterior summaries of standardized EIV regression coefficients for the Gini index.

provide point estimates, L-PRLN provides uncertainty quantification through the posterior distribution.

While we employ L-PRLN to construct income segregation indices, it can be used to construct any other feature of income distributions of interest. For example, sociologists and economists are interested in a variety of measures of income ineqality and income segregation, and use a variety of methods to estimate them not limited to PRLN (Kennedy et al., 1996; Jargowsky, 1996; Mayer et al., 2001; Hardman and Ioannides, 2004; Watson, 2009). These approaches tend to suffer from the same limitations as PRLN, and L-PRLN can be applied to estimating them as well.

L-PRLN can also be generalized and applied to other types of variables. For example, it could be used to interpolate the age distribution, for which there are often a selection of bin estimates available. To do this only requires appropriate choices for the f_ks in (1). Each f_k could be a truncated normal density, though in practice the age distribution should be investigated to determine an appropriate choice. Many choices will require estimation of more parameters per bin than in the PRLN density. In order to handle this, it may be necessary to reduce the number of knots so that there are more bin estimates than knots. The framework can also be applied to data from sources other than the Census Bureau as well. The key is that there are a wide variety of available estimates of different distributional

features at the area-level. These will typically be bin estimates, but many other estimate types could be used.

Based on the simulation study in Section 3.1 and out-of-sample performance on held out estimates in Section 3.2, neither PRLN nor L-PRLN performed uniformly superior than the other when L-PRLN was restricted to a subset of available estimates. L-PRLN performed the best in the middle and far right tail of the distribution, with PRLN typically performing better elsewhere. This is likely due to how informative the Dirichlet prior is on the knot probabilities. As noted in Section 2.5, a more informative prior was necessary in this case to help facilitate NUTS. In particular, note that for some census tracts, the bin estimate for one or more income categories is zero. Without an informative prior, these probabilities will be estimated to be close to zero and NUTS will go into the extreme tails of the transformed space, causing numerical and sampling problems. The informative prior regularizes those estimates away from zero and prevents the computational problem. This leads to a loss of predictive accuracy, although this is reflected in the uncertainty estimates that are provided by L-PRLN. Further, note the knots in L-PRLN are set equal to the boundaries defining the bins for the bin estimates. This is done for computational convenience but is not necessary. Indeed, knot selection is a potential avenue for improving L-PRLN. Naively, it seems as though spacing the knots roughly equally in the quantile domain would alleviate the problem with probabilities being estimated close to zero, and improve the quality of the model. In model fits not reported here, we found that this degrades model performance despite the looser priors, suggesting that there are other factors important for knot selection. The number and spread of available tract-level estimates should fundamentally constrain the optimal number and placement of the knots in some way, but precisely how is an area of future research.

We turn now to the empirical application constructing income segregation indices and estimating the association between them and the Gini index. The analysis in Reardon and Bischoff (2011) cannot be performed for more recent years due to the elimination of the decenial census long form. Instead, ACS estimates are available, but their standard error must be taken into account. Because of its deficiencies, PRLN is not suitable for

constructing income segregation indices using ACS data. Unlike PRLN, L-PRLN allows us to use all available estimates, account for uncertainty in those estimates, and propagate that and other sources of uncertainty into the estimated indices. Additionally, L-PRLN played a secondary role by allowing us to construct metro-level Gini indices for black households only and white households only using ACS estimates, while again propagating uncertainty into the indices so that it could be accounted for in the EIV segregation index regressions. The segregation indices themselves disagreed substantially about the relative ranking of metro areas in terms of income segregation, illustrating that Roberto (2015)'s criticisms of the information theory index are well-founded.

The results of the regressions were also instructive. Our regression results were meaningfully different when using the divergence index, though only for black households only and white households only. Using more recent ACS data and the information theory index in a somewhat different model specification, we were not able to reproduce Reardon and Bischoff (2011). Namely, that the Gini index and the information theory index were positively associated for both black households only and white households only, though there is a lot of uncertainty in our estimates. However, using the divergence index we do see a positive association with the Gini index for both black households only and white households only. The upshot is that we confirm one of the central conclusions of Reardon and Bischoff (2011) – that increased income inequality predicts increased income segregation even within racial groups – but only with a proper measure of income segregation and not with their original index.

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: Latent PRLN density functionals. Includes formulas for all of the relevant functionals of the latent PRLN density referenced in the paper, including

mean, variance, CDF, quantile function, income shares, and Gini index.

Appendix C: Generating the synthetic population. Includes details about how the synthetic population was generated in the simulation study in Section 3.1.

Appendix D: Evaluating Point Estimates. Includes tables evaluating L-PRLN and PRLN point estimates on a variety of metrics from comparison using ACS data in Section 3.2.

Appendix E: Segregation index EIV data. Includes details on how the data for the segregation index EIV regressions were sourced from the ACS for Section 4.

Appendix F: Household level Gini index estimation by race. Includes details on metro-level and metro-race-level household income Gini indices were estimated for use in Section 4.

Appendix G: Computing segregation indices. Includes details on how both the information theory index and the divergence index were computed as a function of model parameters in the posterior distribution for use in Section 4.

Appendix H: Segregation index results. Includes detailed tables of regression coefficients for each of the income segregation index EIV regressions we performed in Section 4.

Github Repository: `https://www.github.com/simpsonm/latentprln` Includes all code for the all models discussed and used in the paper, and for reproducing our results, including `R` code for the Pareto-linear procedure, and code for downloading and cleaning ACS tables.

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.

Arima, S., Datta, G. S., and Liseo, B. (2015). "Bayesian estimators for small area models when auxiliary information is measured with error." *Scandinavian Journal of Statistics*, 42, 2, 518–529.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). "An error-components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association*, 83, 401, 28–36.

Betancourt, M. and Girolami, M. (2015). "Hamiltonian Monte Carlo for hierarchical models." *Current Trends in Bayesian Methodology With Applications*, 79, 30.

Braithwaite, J. (2015). "Sexual violence in the backlands: Toward a macro-level understanding of rural sex crimes." *Sexual Abuse*, 27, 5, 496–523.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Dagum, C. (1977). "A New Model of Personal Income Distribution: Specification and Estimation." *Economie Appliquee*, 413–437.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). "Micro-level estimation of poverty and inequality." *Econometrica*, 71, 1, 355–364.

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.

Fay III, R. E. and Herriot, R. A. (1979). "Estimates of income for small places: an application of James-Stein procedures to Census data." *Journal of the American Statistical Association*, 74, 366a, 269–277.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.

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.

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.

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15, 1, 1593–1623.

Jargowsky, P. A. (1996). "Take the money and run: Economic segregation in US metropolitan areas." *American Sociological Review*, 61, 6, 984–998.

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.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC.

Kooperberg, C. and Stone, C. J. (1992). "Logspline density estimation for censored data." *Journal of Computational and Graphical Statistics*, 1, 4, 301–328.

Marchetti, S., Tzavidis, N., and Pratesi, M. (2012). "Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators." *Computational Statistics & Data Analysis*, 56, 10, 2889–2902.

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.

Molina, I. and Rao, J. (2010). "Small area estimation of poverty indicators." *Canadian Journal of Statistics*, 38, 3, 369–385.

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.

Neal, R. (2011). "MCMC using Hamiltonian dynamics." In *Handbook of Markov chain Monte Carlo*, eds. S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, 113–162. CRC press.

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.

R Core Team (2020). *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.

Roberto, E. (2015). "The Divergence Index: A Decomposable Measure of Segregation and Inequality." *arXiv Preprint arXiv:1508.01167*.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Singh, S. and Maddala, G. (1976). "A Function for Size Distribution of Incomes." *Econometrica: Journal of the Econometric Society*, 963–970.

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.

Stone, C. J. et al. (1994). "The use of polynomial splines and their tensor products in multivariate function estimation." *The Annals of Statistics*, 22, 1, 118–171.

Tarozzi, A. and Deaton, A. (2009). "Using census and survey data to estimate poverty and inequality for small areas." *The review of economics and statistics*, 91, 4, 773–792.

Tzavidis, N., Salvati, N., Pratesi, M., and Chambers, R. (2008). "M-quantile models with application to poverty mapping." *Statistical Methods and Applications*, 17, 3, 393–411.

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.

— (2017a). "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.

— (2017b). "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.

— (2017c). "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.

— (2017d). "Table B19080: Household Income Quintile Upper Limits, 2011 – 2015 American Community Survey." https://data.census.gov/.

— (2017e). "Table B19082: Shares of Aggregate Household Income by Quintile, 2011 – 2015 American Community Survey." https://data.census.gov/.

— (2017f). "Table B19083: Gini Index of Income Inequality, 2011 – 2015 American Community Survey." https://data.census.gov/.

— (2017g). "Table S1901: Income in the Past 12 Months, 2011 – 2015 American Community Survey." https://data.census.gov/.

— (2017h). "Table S2503: Financial Characteristics, 2011 – 2015 American Community Survey." https://data.census.gov/.

— (2020a). "Table B19080: Household Income Quintile Upper Limits, 2014 – 2018 American Community Survey." https://data.census.gov/.

— (2020b). "Table B19082: Shares of Aggregate Household Income by Quintile, 2014 – 2018 American Community Survey." https://data.census.gov/.

— (2020c). "Table B19083: Gini Index of Income Inequality, 2014 – 2018 American Community Survey." https://data.census.gov/.

— (2020d). "Table S1901: Income in the Past 12 Months, 2014 – 2018 American Community Survey." https://data.census.gov/.

— (2020e). "Table S2503: Financial Characteristics, 2014 – 2018 American Community Survey." https://data.census.gov/.

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.