An Info-Metrics Approach to Estimating the Supplemental Poverty Rates of Public Use Microdata Areas

Danielle Wilson

1 Introduction

The Supplemental Poverty Measure (SPM) is an extension of the Official Poverty Measure (OPM) that considers non-cash benefits, tax credits and necessary expenses when determining an individual's poverty status [6]. Annual supplemental poverty estimates rely on the Current Population Survey's Annual Social and Economic Supplement (CPS-ASEC), which collects detailed, individual-level data on income and expenses. A limitation of the Current Population Survey (CPS) is its sample size, which is insufficient for disaggregate geographic estimates below the state level. The American Community Survey (ACS) is a larger survey that can provide averages at both the state and Public Use Microdata Area (PUMA) level. However, the American Community Survey (ACS) does not ask respondents for detailed income data and is subsequently unable to alone produce supplemental poverty estimates.

In this paper I propose an information-theoretic model to estimate PUMA-level supplemental poverty rates using data from both surveys. The information-theoretic framework, also known as the info-metrics framework, is a constrained optimization one [9]. The framework's constraints, when appropriately specified, can reconcile the limitations of data from each survey. Data from the American Community Survey can be used to produce PUMA level supplemental poverty estimates, but they are noisy due the needed imputation of individual components of income. More accurate, but aggregate information from the CPS-ASEC is used to constrain noisy American Community Survey estimates. In other words, in the proposed information-theoretic model, out of sample aggregate CPS-ASEC estimates are used to constrain in-sample disaggregate American Community Survey estimates. The model uses the maximum entropy objective function to estimate the relative error of each PUMA-level supplemental poverty subject to the previously specified constraint. The maximum-entropy criteria provides the most conservative error estimates given information from both surveys. These error estimates are used to produce refined PUMA-level supplemental poverty rates.

The main benefit of using the American Community Survey (ACS) is its ability to produce estimates at the PUMA level. Official poverty rates are produced at this geographic

level without blanket imputation of income components and are readily available in the micro data [1]. However, the official poverty measure does not account for government aid or necessary expenses, both of which affect an individual's true poverty status. The official poverty measure also has a limited family definition which can overlook resource sharing among unmarried partners and their children [18]. The Supplemental Poverty Measure has a broader, nuanced definition of families, and considers government aid and expenses. However, this measure requires additional information to determine whether individuals are in supplemental poverty. Reliable supplemental poverty estimates can also be used to identify the most impactful additions and subtractions that either push or pull a family into or out of poverty. Refined or alternative estimates of supplemental poverty at disaggregate geographic levels can help researchers test the effectiveness of anti-poverty programs or answer poverty related questions.

The small area estimation literature provides similar solutions when sample data is limited or relatively unreliable [19]. These techniques traditionally use small area characteristics or predictive data to estimate the means, rates or proportions of interests. These methods are especially useful when data for particular small areas is missing or the outcome of interest is only available at an aggregate level. However, the information-theoretic model proposed in this paper requires no additional area characteristics to provide refinements of small area estimates. Subsequently, no assumptions regarding the homogeneity of areas or their characteristics need to be made [22].

The remainder of this paper is organized as follows. Section two introduces Supplemental Poverty Measure and discusses how the measure is applied using data from either survey. Section three presents the information-theoretic model, which uses data from both surveys at different geographic levels. Section four presents the results of a simulation experiment designed to mimic the strengths and weaknesses of each survey. Section five presents refined supplemental poverty rates using the model discussed in section three and an analysis comparing survey estimates to refined estimates.

2 The Supplemental Poverty Measure

This section will review how poverty is defined using the supplemental poverty measure (SPM) and compare its application in the CPS-ASEC and American Community Survey. An individual i is considered poor (i.e., $P_i = 1$) by this measure if the summation of their family's resources, $\boldsymbol{R}_i$, is less than the supplemental poverty threshold, $\boldsymbol{T_i}$. This threshold is defined by the Bureau of Labor Statistics and adjusted for family size, composition and geographic location [17]. In equation (1), P_i^* is the estimated poverty status of individual i, $\boldsymbol{R_i^*}$ is their family's observed sum of resources and T_i is their family's designated poverty threshold.

$$P_i^* = \begin{cases} 1, & \boldsymbol{R_i^*} < \boldsymbol{T_i} \\ 0, & \text{o.t.} \end{cases} \tag{1}$$

There are 13 components within an individual's family resource summation. Equation (2) defines each component considered by the supplemental poverty measure [5].

$$\begin{aligned} \boldsymbol{R}_i &= \sum_{j=1}^{13} c_{ij} \\ &= FamilyCashIncome_i + SNAP_i + WIC_i \\ &+ SchoolLunch_i + HousingSubsidy_i \\ &+ EnergySubsidy_i - FederalTaxes_i - StateTaxes_i \\ &- FICA_i - WorkExpenses_i - ChildCareExpenses_i \\ &- MedicalExpenses_i - ChildSupport_i \end{aligned} \tag{2}$$

Each component of an individual's family resource summation is noted as c_{ij}. The components $SNAP_i$ and WIC_i represent the value of aid provided by the Supplemental Nutrition Assistance Program (SNAP) and Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), respectively. $FICA_i$ represents the social security and Medicare contributions required by the Federal Insurance Contributions Act (FICA) and paid by all working members of individual i's family. The value of federal taxes paid, $FederalTaxes_i$, is net of any income tax or child tax credits.

2.1 Application of the Supplemental Poverty Measure

2.1.1 Using the CPS-ASEC

All terms in equation two are either observed or estimated using additional out-of sample information. Terms that are not directly observed use a combination of survey data from the CPS-ASEC, and public and administrative records to impute their value. The cash value of some components of an individual's resource summation, such as housing or energy subsidies, is explicit. However, the value of non-cash benefits, such as school lunch, are implicit and require estimation. The inherit value of non-cash benefits is typically unknown to recipients thus administrative program data is used to estimate the per-recipient value.

Expenses, such as taxes paid at the federal and state level, are also estimated as neither the Current Population survey, its Annual Social and Economic Supplement (CPS-ASEC), or the American Community Survey collect such information. The official supplemental poverty report uses the CPS-ASEC and simulation data to estimate each individual's federal and state tax contribution [6]. These simulation estimates are statistically matched to IRS public-use micro data [26]. If $FederalTaxes_i$ are the true and unknown value of taxes paid by individual i's family, then the estimated amount of federal taxes paid, $FederalTaxes_i^*$, can be expressed as

$$FederalTaxes_i^* = FederalTaxes_i + \epsilon_i \tag{3}$$

where ϵ_i is the unobserved error associated with the simulation's estimate.

Equation (2) is an individual's unobserved, true resource summation. However, their estimated resource summation using data from the CPS-ASEC is:

$$
\begin{aligned}
\boldsymbol{R}^*_{i,CPS} &= \sum_{j=1}^{5} c_{ij} + \sum_{j=7}^{13} \boldsymbol{c}^*_{ij} \qquad (4)\\
&= FamilyCashIncome_i + SNAP_i + EnergySubsidy_i \\
&- ChildCareExpenses_i - MedicalExpenses_i - ChildSupport_i \\
&+ \boldsymbol{WIC}^*_i + \boldsymbol{SchoolLunch}^*_i + \boldsymbol{HousingSubsidy}^*_i \\
&- \boldsymbol{FederalTaxes}^*_i - \boldsymbol{StateTaxes}^*_i \\
&- \boldsymbol{FICA}^*_i - \boldsymbol{WorkExpenses}^*_i \\
\\
&= FamilyCashIncome_i + SNAP_i + EnergySubsidy_i \\
&- ChildCareExpenses_i - MedicalExpenses_i - ChildSupport_i \\
&+ (\boldsymbol{WIC_i} + \boldsymbol{\epsilon_{i,7}}) + (\boldsymbol{SchoolLunch_i} + \boldsymbol{\epsilon_{i,8}}) + (\boldsymbol{HousingSubsidy_i} + \boldsymbol{\epsilon_{i,9}}) \\
&- (\boldsymbol{FederalTaxes_i} + \boldsymbol{\epsilon_{i,10}}) - (\boldsymbol{StateTaxes_i} + \boldsymbol{\epsilon_{i,11}}) \\
&- (\boldsymbol{FICA_i} + \boldsymbol{\epsilon_{i,12}}) - (\boldsymbol{WorkExpenses_i} + \boldsymbol{\epsilon_{i,13}})
\end{aligned}
$$

where components denoted with an asterisk, c^*_{ij} are estimated and thus contain an imputation error term, ϵ_{ij}. The components from the WIC and National School Lunch Program are estimated using characteristic data from the CPS-ASEC to determine eligibility. The cash value of both government aid programs is determined using administrative data on either average allocation or program cost [8]. The total value of housing assistance received, $HousingSubsidy^*_i$, is determined by simulation estimates that are statistically matched to administrative data from the U.S. Department of Housing and Urban Development [20] [12]. All components included in equation (4), whether directly observed via the CPS-ASEC or imputed, are readily available in micro data published by the U.S. Census Bureau [2].

2.1.2 Using the ACS

The American Community Survey collects less income information than the CPS-ASEC. Consequently, all resource components in the supplemental poverty measure other than family cash income need to be imputed. The estimated resource summation in the American Community Survey is:

$$
\begin{aligned}
\boldsymbol{R}^*_{i,ACS} &= c_{i1} + \sum_{j=2}^{12} \boldsymbol{c}^*_{ij} \qquad (5)\\
&= FamilyCashIncome_i + (\boldsymbol{SNAP_i} + \boldsymbol{\epsilon_{i,2}}) + (\boldsymbol{EnergySubsidy_i} + \boldsymbol{\epsilon_{i,3}}) \\
&- (\boldsymbol{ChildCareExpenses_i} + \boldsymbol{\epsilon_{i,4}}) - (\boldsymbol{MedicalExpenses_i} + \boldsymbol{\epsilon_{i,5}}) \\
&+ (\boldsymbol{WIC_i} + \boldsymbol{\epsilon_{i,6}}) + (\boldsymbol{SchoolLunch_i} + \boldsymbol{\epsilon_{i,7}}) + (\boldsymbol{HousingSubsidy_i} + \boldsymbol{\epsilon_{i,8}}) \\
&- (\boldsymbol{FederalTaxes_i} + \boldsymbol{\epsilon_{i,9}}) - (\boldsymbol{StateTaxes_i} + \boldsymbol{\epsilon_{i,10}}) - (\boldsymbol{FICA_i} + \boldsymbol{\epsilon_{i,11}}) \\
&- (\boldsymbol{WorkExpenses_i} + \boldsymbol{\epsilon_{i,12}})
\end{aligned}
$$

Note that child support is not included the above equation. The American Community Survey collects no information on whether (or not) an individual pays child support. Consequently, the value of child support paid cannot be imputed and included in this application of the supplemental poverty measure.

In the CPS-ASEC, there is imputation related uncertainty regarding seven of the 13 components needed to determine an individual's poverty status. However, in the American Community Survey, there is uncertainty regarding *all but one* of the 11 common components. Consequently, individual level supplemental poverty estimates are arguably more reliable in the CPS-ASEC as more information on each family is collected and fewer components of their resources need imputed in order to determine their supplemental poverty status.

2.2 Calculating and Comparing Supplemental Poverty Rates

Aggregate poverty rates can be calculated using the definition of resources specified by the supplemental poverty measure. Rates at different geographic levels can be related through weighted averages. However, the CPS-ASEC can only produce disaggregate estimates at the state level. Thus, comparison of rates with the American Community Survey can only be made at this level. A natural model constraint is specified by comparing rates at different levels and from different surveys at the state-level.

Using the empirical definition of resources in the CPS-ASEC, $R^*_{i,CPS}$, the estimated supplemental poverty rate for each state, s, is

$$r^*_{s,CPS} = \frac{\sum_{i=1}^{n_{s,CPS}} P^*_{is,CPS}}{n_{s,CPS}} \tag{6}$$

where $n_{s,CPS}$ is the survey's sample size in state s. $P^*_{is,CPS}$ is either one or zero and captures the estimated poverty status of each CPS-ASEC sampled individual i who lives in state s. The supplemental poverty rate for each PUMA, m, in state s using the American Community Survey can similarly be calculated as

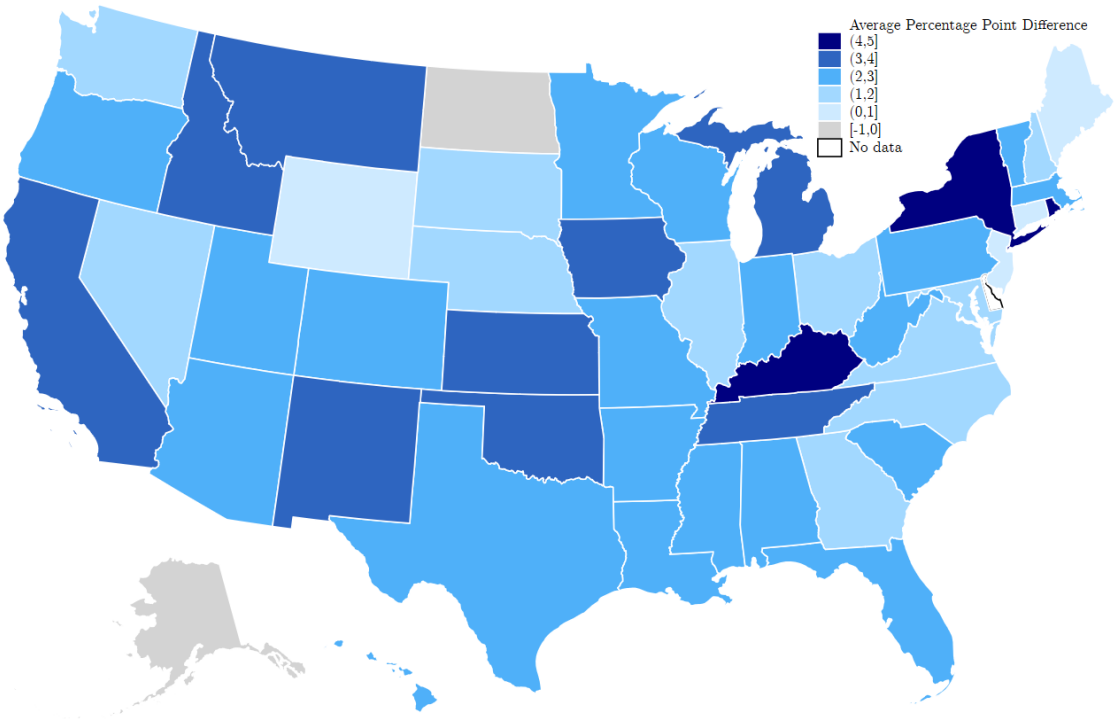
$$r^*_{ms,ACS} = \frac{\sum_{i=1}^{n_{ms,ACS}} P^*_{ims,ACS}}{n_{ms,ACS}} \tag{7}$$

where $n_{ms,ACS}$ is the sample size from PUMA m in state s, and $P^*_{ims,ACS}$ is the observed poverty status of each individual in PUMA m in this survey. Each individual's supplemental poverty status in the American Community Survey is determined by their family's resources, $R^*_{i,ACS}$, as defined in equation (5).

If both the CPS-ASEC and American Community Survey samples are representative of the same overall population, supplemental poverty rates at the state level should be approximately equal. However, as discussed in Fox, Glassman and Pacas [15], supplemental poverty rates differ annually at the national level and by demographic subgroups. Figure 1 shows significant variation in state level supplemental poverty estimates between surveys. From 2016 to 2018 the average supplemental poverty rate in 15 states differed by

more than three percentage points between surveys. For context, state-level supplemental poverty rates ranged from 6.7 to 18.2 percent from 2016 to 2018 using CPS-ASEC micro data. An underestimate of supplemental poverty by three percentage points can account for up to a quarter of the estimate.

Figure 1: Difference Between the CPS-ASEC and ACS SPM Rates by State: 2016 - 2018



Note: Due to limited sample sizes, state level estimates using the CPS-ASEC require three year averages. CPS-ASEC three year averages were compared to ACS three year averages. Differences were calculated by subtracting the average state CPS-ASEC estimate from the average state ACS estimate. Data on the state of Delaware are excluded due to data quality concerns in 2017. For more is available at `https://www.census.gov/programs-surveys/acs/technical-documentation/errata/120.html`.
Source: Current Population Survey, Annual Social and Economic Supplement and American Community Survey, 2016 to 2018.

As previously discussed, although the income information provided by American Community Survey is relatively limited, it's sample size is large enough to produce estimates at both the state and PUMA level. In this survey, state-level supplemental poverty rates can be written as a weighted sum of their PUMA-level supplemental poverty rates:

$$r^*_{s,ACS} = \frac{\sum_{m=1}^{M} r^*_{ms,ACS} \times n_{ms,ACS}}{n_{s,ACS}} \tag{8}$$

where M is the number PUMAs in each state s, $r^*_{ms,ACS}$ is the estimated PUMA level supplemental poverty rate in the ACS, $n_{ms,ACS}$ is the number of individuals sampled in PUMA m, and $n_{s,ACS}$ is the number of individuals sampled in state s. Additional imputation at the individual level in the American Community Survey has cascading effects on aggregate supplemental poverty rates. Focusing on the effect of individual-

level imputation on PUMA-level estimates, the difference between the estimated and true supplemental poverty rates can be expressed as follows:

$$r^*_{sm,ACS} = r_{sm,ACS} + \gamma_{sm,ACS} \tag{9}$$

where $r^*_{sm,ACS}$ is the estimated poverty rate, $r_{sm,ACS}$ is the true, but unknown, poverty rate, and $\gamma_{sm,ACS}$ is imputation-related noise. Incorporating this PUMA-level relationship into the state-level supplemental poverty rate gives

$$r^*_{s,ACS} = \frac{\sum_{m=1}^{M} \left(r^*_{sm,ACS} - \gamma_{sm,ACS} \right) \times n_{ms,ACS}}{n_{s,ACS}}. \tag{10}$$

Equation (10) is the American Community Survey's estimate state-level supplemental poverty in terms of its weighted PUMA-level counterparts and additional error from imputed resource components. This equation will serve as the basis for the fundamental constraint in the proposed information-theoretic model.

The CPS-ASEC provides an arguably better estimate of state-level supplemental poverty. More reliable information from this survey can consequentially be incorporated into the relationship of poverty at different geographic levels. Substituting the state-level supplemental poverty rate from the CPS-ASEC into the left-hand side of equation (10) gives:

$$\boldsymbol{r^*_{s,CPS}} = \frac{\sum_{m=1}^{M} \left(r^*_{sm,ACS} - \gamma_{sm,ACS} \right) \times n_{ms,ACS}}{n_{s,ACS}} \tag{11}$$

In the equation above, the PUMA-level ACS estimates on the right hand side are now constrained by the realtively more accurate state-level CPS-ASEC estimate on the left hand side. Equation (11) is an observed constraint that relates information at different geographic levels from both surveys. Without the state-level information from the CPS-ASEC, the imputation-related noise affecting American Community Survey estimates are otherwise indeterminable.

3 Information-Theoretic Model

The purpose of the information-theoretic model is to estimate the imputation related noise, $\gamma_{sm,ACS}$, within disaggregate American Community Survey estimates. The objective function, also known as the decision function, within the proposed information-theoretic model is a maximum entropy one. In this context entropy represents the information (or lack thereof) associated with the probability of an outcome [21]. Entropy is synonymous with uncertainty and is inversely related to information, which comes from the model's specified constraints [9]. The objective function maximizes entropy associated with the imputation related noise subject to information from both the CPS-ASEC and the American Community Survey. In other words, the maximum entropy criteria provides the most

conservative estimate of imputation-related noise relative to the poverty rates observed in both surveys and at various geographic levels.

Equation (11), which relates state-level supplemental poverty rates from the CPS-ASEC to PUMA-level rates from the American Community Survey, serves as the primary observed constraint in the proposed information-theoretic model. The objective of this model is to estimate the probability that imputation-related noise takes on a certain value. In other words, the model estimates a discrete probability distribution associated with values that the imputation-related error can be. The discrete distribution of possible imputation-related error values is assumed be one of three outcomes, and be symmetric and centered around zero. As specified in Table 1, the maximum value the imputation-related error can be is C, a positive real number, and the minimum value the error can be is $-C$. The expected value of the imputation-related error is:

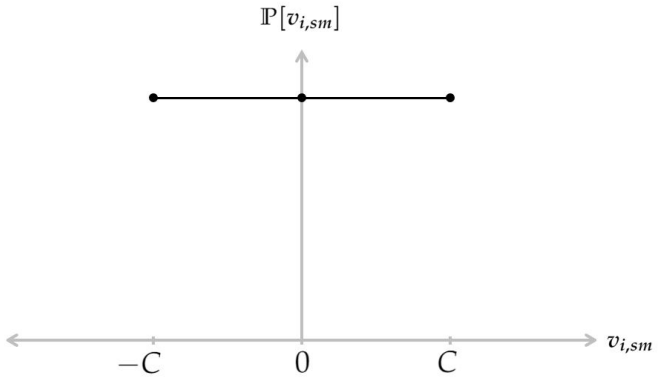
$$\begin{aligned} \mathbb{E}[\gamma_{sm,ACS}] &= v_{1,sm} \cdot \mathbb{P}[v_{1,sm}] + v_{2,sm} \cdot \mathbb{P}[v_{2,sm}] + v_{3,sm} \cdot \mathbb{P}[v_{3,sm}] \\ &= -C \cdot w_{1,sm} + 0 \cdot w_{2,sm} + C \cdot w_{3,sm} \\ &= -C \cdot w_{1,sm} + C \cdot w_{3,sm} \end{aligned} \tag{12}$$

where each probability, $\mathbb{P}[v_{i,sm}]$, is positive and all probabilities sum to one. Without any information from, for example, an observed constraint, the probability distribution of the imputation-related error is uniform. However, information from an observed constraint changes the weights ($w_{k,sm}$; $k = 1, 2, 3$) associated with each outcome, which in turn changes the expected value of the error.

Table 1: Discrete Distribution of Imputation-Related Error, $\gamma_{sm,ACS}$

	$\boldsymbol{v_{1,sm}}$	$\boldsymbol{v_{2,sm}}$	$\boldsymbol{v_{3,sm}}$
$\boldsymbol{v_{i,sm}}$	$-C$	0	C
$\mathbb{P}[\boldsymbol{v_{i,sm}}]$	$w_{1,sm}$	$w_{2,sm}$	$w_{3,sm}$

Figure 2: Distribution of Imputation-Related Error With No Observed Constraint



Using the observed constraint, the information theoretic model for each state s is:

$$\max_{w} H(W) = -\sum_{m=1}^{M}\sum_{k=1}^{3} w_{k,sm} log(w_{k,sm}) \tag{13}$$

subject to

$$r^*_{s,CPS} = \frac{\sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{k=1}^{3} w_{k,sm} v_{k,sm}\right] N^*_{sm,ACS}}{N^*_{s,ACS}} \tag{14}$$

$$\sum_{k=1}^{3} w_{ksm,ACS} = 1; m = 1, ..., M \tag{15}$$

$$w_{k,ms} \geq 0; m = 1, ..., M \text{ and } k = 1, 2, 3 \tag{16}$$

Equation (13) is the maximum entropy decision function [11] [13]. The first constraint, equation (14), is the observed constraint specified in section two and relates information from different geographic levels and surveys. However, unlike in equation (11), the observed constraint in this model uses the sample weighted populations of each PUMA ($N^*_{ms,ACS}$) and state ($N^*_{s,ACS}$) rather than their respective sample sizes. Sample weights ensure that estimates represent the true underlying population. In application, rates in the observed constraint, which come from survey micro-data, are also assumed to be weighted. Equations (14) and (15) are normalization constraints that ensure that the estimated values of $w_{k,sm}$ satisfy the requirements of a discrete probability distribution.

The probabilities needed to determine the expected imputation-related error can be recovered using the following Lagrangian

$$\begin{aligned}\mathcal{L} \quad &= -\sum_{m=1}^{M}\sum_{k=1}^{3} w_{k,sm} log(w_{k,sm}) \\ &+ \lambda\left[r^*_{s,CPS} - \frac{\sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{k=1}^{3} w_{k,sm} v_{k,sm}\right] N^*_{sm,ACS}}{N^*_{s,ACS}}\right] \\ &+ \lambda_m\left[1 - \sum_{k=1}^{3} w_{k,sm}\right]\end{aligned} \tag{17}$$

The respective first order conditions are

$$\frac{\partial \mathcal{L}}{\partial w_{k,sm}} = -log(w_{k,sm}) - 1 - \lambda\left[\frac{-v_{k,sm} N^*_{sm,ACS}}{N^*_{s,ACS}}\right] - \lambda_m = 0 \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = r^*_{s,CPS} - \frac{\sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{k=1}^{3} w_{k,sm} v_{k,sm}\right] N^*_{sm,ACS}}{N^*_{s,ACS}} = 0 \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_m} = 1 - \sum_{k=1}^{3} w_{k,sm} = 0 \tag{20}$$

The normalized solution[1] for each probability and for every PUMA m is

$$w^*_{k,sm} = \frac{e^{\left[\frac{-v_{k,sm}N^*_{sm,ACS}}{N^*_{s,ACS}}\right]}}{e^{\sum_{k=1}^{3}\left[\frac{-v_{k,sm}N^*_{sm,ACS}}{N^*_{s,ACS}}\right]}}. \tag{21}$$

Each estimated probability from the model can be used to determine the expected value of the imputation-related noise. Substituting $w^*_{1,ms}$, $w^*_{2,ms}$, and $w^*_{3,ms}$ into equation (12) gives

$$\gamma^*_{sm,ACS} = \sum_{k=1}^{3} v_{k,sm}w^*_{k,sm} = -C \cdot w^*_{1,sm} + C \cdot w^*_{2,sm} \tag{22}$$

As foreshadowed in equation (9), the estimated imputation-related error can be subtracted from the American Community Survey's PUMA-level supplemental poverty rate to get a new, refined rate:

$$\tilde{r}_{sm,ACS} = r^*_{sm,ACS} - \gamma^*_{sm,ACS} \tag{23}$$

This new rate is not the true, unknown PUMA-level supplemental poverty. However, it is a refined estimate of supplemental poverty at this disaggregate level *relative to* more reliable, aggregate information from the CPS-ASEC.

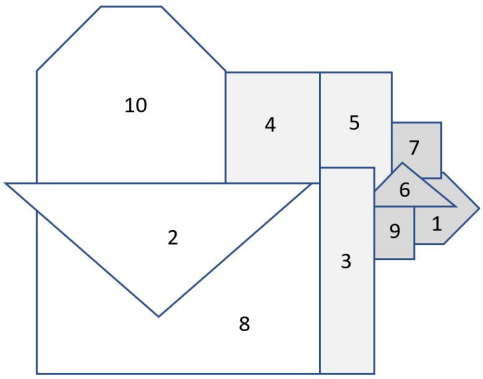
4 Simulation Experiment

A simulation was designed to test the efficacy of the proposed information-theoretic method. An artificial state with ten PUMAs was created. Following the U.S. Census Bureau's definition of a PUMA, each artificial area was given a random population between 100 to 200 thousand people [23]. Each individual in the simulation was randomly assigned a value of labor income conditional on their PUMA of residence. As shown in Figure 3, PUMAs were identified as being in one of three regions dependent on their size and concentration. For example, PUMAs 1, 7, 6, and 9 are the smallest and thus are the most densely concentrated. The center and skewness of each PUMA's labor income distribution increased with concentration. Labor income was also inflated by neighboring averages to simulate spatial dependence.

To mimic aid received from various government programs, a random fraction of qualifying individuals in the artificial population were assigned a value of benefit income. An unconditional fraction of individuals were also assigned a value representative of expenses. The cash value of government aid and expenses was randomly determined and not spatially dependent. All individuals in the sample were randomly assigned as being one of five types of people (i.e., type 1, 2,.., or 5). The distribution of type was conditional on relative labor income. Consequently, individual type was correlated with labor income. Total income was the sum of labor and aid income, net of expenses, and

[1] See Appendix for details.

Figure 3: Artificial State With Ten PUMAs



Note: All artificial PUMAs contain between 100 to 200 thousand simulated individuals. Region 1 in white contains PUMAs 10, 2 and 8, which are the least concentrated. Region 2 in light grey contains PUMAs 4, 5 and 3, which are moderately concentrated. Region 3 in dark grey contains the remaining PUMAs that are densely concentrated.

was used to determine each individual's true poverty status. An individual was in poverty if their total income was below the 33th percentile.

To simulate the CPS-ASEC, a sample of individuals was randomly drawn from the population by type. The probability of collecting respondent information was consequentially correlated with type in order to mimic nonrandom nonresponse bias. Each component of total income, labor income, government aid, and expenses, was observed with a randomly determined amount of noise. To simulate the American Community Survey, a larger sample of individuals was similarly drawn. However, each component of total income was observed with on average relatively more noise. Representative weights for each sampled individual in both simulated surveys were created. State-level poverty estimates were subsequently calculated for both simulated surveys. PUMA-level poverty estimates were solely calculated for the simulated American Community Survey. Table 2 contains descriptive statistics of estimates from both the simulated surveys at each regional level.

Table 2: Descriptive Statistics of Simulated Survey Estimates

			Sample Poverty Estimates/Rates					
Simulated Survey	Region	Number	Minimum	Median	Maximum	Range	Mean	S.D.
CPS-ASEC	States	100	6.65	12.50	18.94	18.94	12.36	2.99
ACS	States	100	7.50	13.56	20.24	20.24	13.37	3.07
ACS	PUMA	1000	1.71	11.12	31.82	31.82	13.38	7.66
True	PUMA	1000	0.92	9.85	30.07	30.07	11.98	7.39

This simulation was run 100 times. In each run the true poverty rates at both the state and PUMA-level were collected to test the effectiveness of the proposed information-theoretic model. Estimates from both simulated surveys were implemented into the information-theoretic model and refined poverty rates were produced. Characteristic data from each artificial PUMA were also collected. These data are not used in the latter model. However, they are needed by the Fay–Herriot model, an alternative small area estimation method, to improve disaggreate poverty estimates [4] [14]. The Fay-Herriot method improves a

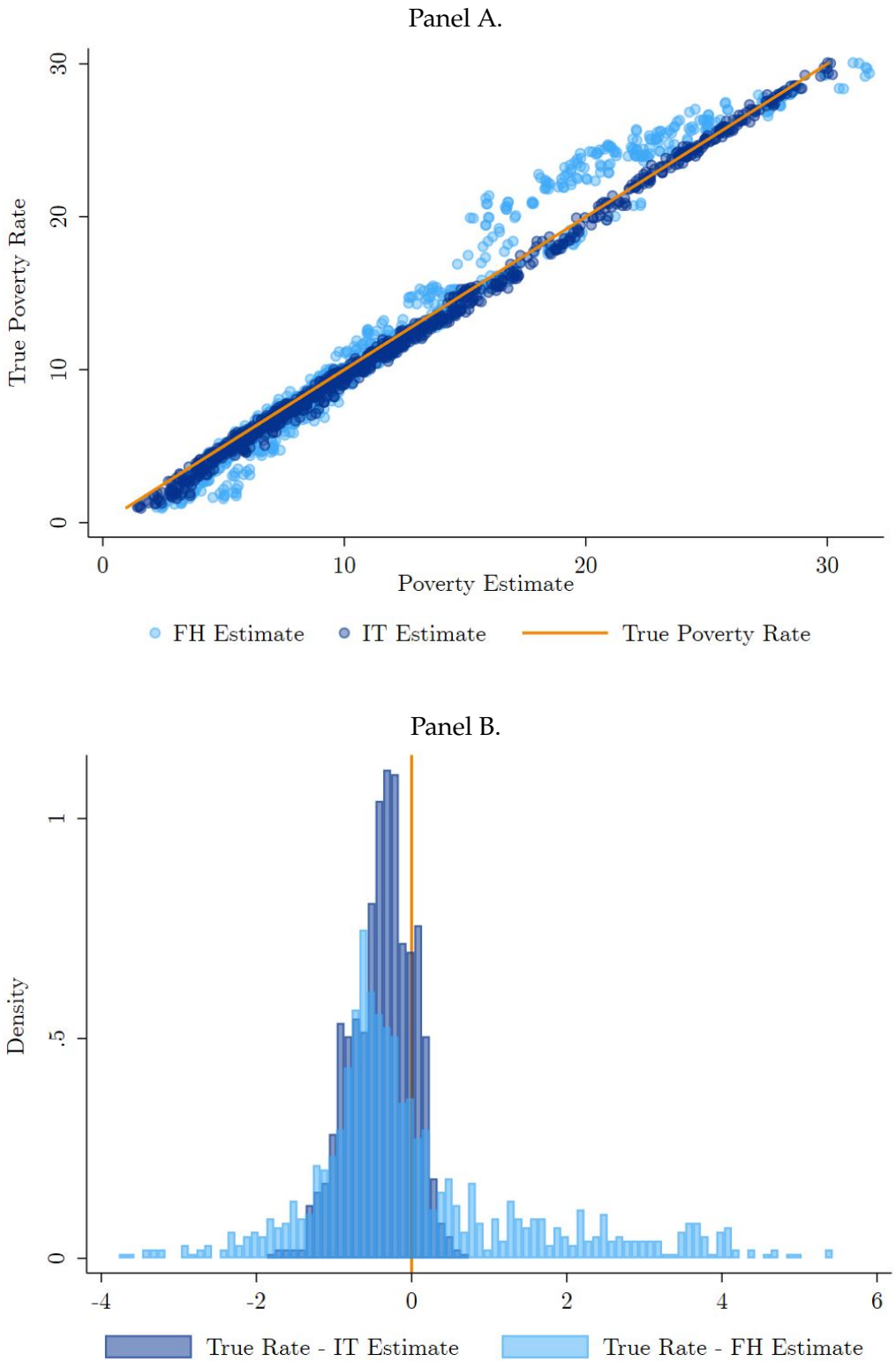
small area estimate by weighting it with respect to the variance of its sampling error, and averaging it with a weighted regression estimate produced using small area characteristics [10]. The variance of the sampling error is unknown but estimated using the sample variance of poverty within each PUMA [27]. The sample variance is also used to define the maximum and minimum error bounds needed in the information-theoretic model. Covariates used to improve the precision of simulated PUMA-level poverty estimates in the Fay-Herriot model include: sample shares of each type of individual at the PUMA level, sample poverty estimates at the state level, and sample shares of benefit income recipients at the PUMA-level. Following [24], poverty rates were transformed using the arcsine function prior to being regressed on these covariates.

Table 3: Descriptive Statistics of Small Area Estimates

	PUMA-Level Poverty Estimates					
	Minimum	Median	Maximum	Range	Mean	S.D.
True	0.92	9.85	30.07	29.15	11.98	7.39
Survey Estimates	1.71	11.12	31.82	30.11	13.38	7.66
Fay-Herriot Estimates	2.22	10.06	31.74	29.52	11.95	6.52
Information-Theoretic Estimates	1.43	10.18	30.22	28.78	12.37	7.27

Table 3 compares statistics from the distribution of true poverty rates to those from the distribution of (i) survey estimates, (ii) Fay-Herriot estimates and (iii) information-theoretic estimates. Both the Fay-Herriot and information-theoretic models produce estimates that are on average closer to the truth than survey estimates. The average PUMA-level poverty estimate from either model is not statistically different from the true average of 11.98 percent. The distribution of Fay-Herriot poverty estimates is narrower than the true distribution and more closely centered around the true mean. However, Figure 4 shows that individual PUMA-level information-theoretic estimates are on average closer to true rates when compared to the competing Fay-Herriot estimates. Figure 4, Panel A, plots each model's estimate relative to its true rate for every PUMA. The orange 45-degree line corresponds to the true distribution of PUMA-level poverty rates. The information-theoretic estimates (in dark blue) follow more closely the distribution of true rates than the Fay-Herriot estimates (in light blue). Each model estimate was subsequently subtracted from its corresponding true poverty rate to determine the magnitude of deviations across PUMAs. Figure 4, Panel B, compares the distribution of these deviations for both the Fay-Herriot and information-theoretic estimates. Fay-Herriot estimates diverged on average 1.1 percentage points from true estimates. Meanwhile, information-theoretic estimates diverged on average 0.4 percentage points from true estimates. The variation in the magnitude of theses deviations was also approximately three times larger among Fay-Herriot estimates. Overall, the information-theoretic model produces more accurate PUMA-level estimates without area-level characteristics. The information-theoretic model only requires relatively more accurate, aggregate estimates in order to provide refined poverty estimates.

Figure 4: Simulation Results - Comparing Small Area Estimates



Note: Fay-Herriot (FT) estimates are in light blue and information-thereotic estimates (IT) are in dark blue.

5 Application

5.1 Data

In 2020, the Census Bureau published a series of augmented American Community Survey micro-data sets with individual-level supplemental poverty estimates. Fox, Glassman and Pacas [15] supplement each original annual American Community Survey data set with needed additional, but imputed, components of resources. The authors detail how additional components (i.e., variables) are created using publicly available out of sample information and how they can be used to determine the supplemental poverty rates at both the state and Public Use Microdata Area (PUMA) level.

Having shown that the information-theoretic approach can effectively refine noisy small area estimates, the proposed method is applied to all PUMA-level supplemental poverty rates produced using the augmented American Community Survey data sets. Aggregate and relatively more informed state level estimates from the the CPS-ASEC are used to constrain these PUMA-level supplemental poverty rates. The Census Bureau recommends using three year averages when producing state estimates with the CPS-ASEC. Consequently, all estimates needed for the implementation of the information-theoretic model, including those from the American Community Survey, are pooled from 2016 to 2018. In 2019 the Census Bureau began releasing CPS-ASEC micro data using a redesigned survey and an updated processing system. Micro data from prior year's surveys were published as a "bridge" and "research" files, and are used in this analysis [3]. To further ensure comparability of supplemental poverty estimates between surveys, CPS-ASEC estimates will not consider child support paid. As discussed in Section 2, the American Community Survey does not collect any information on child support paid so cannot impute its value. The exclusion of this component from the measure is not expected to significantly alter supplemental poverty rates. For example, in 2020 consideration of this expense changed the supplemental poverty status of approximately 0.06 percent of the United States population [7].

5.2 Results

The proposed information-theoretic model is used to refine the supplemental poverty rate of all PUMAs in the United States. Each PUMA has it's own discrete distribution of imputation-related error values. These values are defined using the sample standard error of each PUMA's estimated supplemental poverty rate. Each estimated poverty rate is computed using replicate survey weights. These weights are produced using random subsamples, also known as replicates, of the primary sample, which in effect simulate repeated full samples [25]. Variability between full sample estimates and replicate estimates are used to compute more informed sample standard errors. Table 4 details the discrete distribution used to estimate the imputation-related error of each PUMA supplemental poverty rate. The maximum value the imputation-related error can take is $10 \cdot s_m$, where

s_m is the sample standard error of n_{sm} observations in PUMA m and state s. The minimum value the imputation-related error can take is $-10 \cdot s_m$.

Table 4: Discrete Distribution of Imputation-Related Error, $\gamma_{sm,ACS}$, in Application

	$v_{1,sm}$	$v_{2,sm}$	$v_{3,sm}$
$v_{i,sm}$	$-10 \cdot s_m$	0	$10 \cdot s_m$
$\mathbb{P}[v_{i,sm}]$	$w_{1,sm}$	$w_{2,sm}$	$w_{3,sm}$

Note: s_m is the sample standard error of n_{sm} observations in PUMA m and state s.

As shown in Figure 1, the state-level supplemental poverty rates produced using the CPS-ASEC and the American Community Survey vary up to 4.7 percentage points. Consequently, narrower bounds prevent the model from producing estimates of imputation-related noise that satisfy the primary observed constraint in equation (14). A sensitivity analysis was done to ensure that refined PUMA-level poverty estimates were not subject to varying bounds. When the range of the discrete distribution used to estimate the imputation-related error was reduced to $[-4 \cdot s_m, 4 \cdot s_m]$, refinements were attainable for 1,919 PUMAs in 40 of the 56 states. Refinements were not attainable in states such as New York, Rhode Island and Kentucky, where state-level supplemental poverty rates vary over four percentage points between surveys. Of the 1,919 PUMAs whose supplemental poverty rates could be refined using the narrower range of imputation-related error, only nine were statistically different at the 95 percent confidence level from the refined rates produced using the initial range of $[-10 \cdot s_m, 10 \cdot s_m]$. The majority of these nine PUMAs were also in Philadelphia City, where there was a data collection error in the 2017 American Community Survey [2]. PUMA's in Philadelphia City and the state of Delaware are excluded from any analysis due to data collection errors.

Figure 5 presents the distribution of estimated refinements from the information-theoretic model. The average (unrefined) PUMA-level supplemental poverty rate in the U.S. from 2016 to 2018 was 15.7 percent. The average maximum entropy refinement was approximately 2.5 percentage points. In other words, constraining PUMA-level poverty estimates from the American Community Survey to more reliable, aggregate estimates from the CPS-ASEC decreases the average PUMA-level supplemental poverty rate by approximately 20 percent. Figure 6 compares the magnitude of estimated refinements to state-level percentage point differences in supplemental poverty by survey. Intuitively, larger PUMA-level refinements estimated by the information-theoretic model are correlated with larger differences at the aggregate state level. Figure 7 illustrates the heterogeneity of estimated PUMA-level supplemental poverty refinements across the country and within states. The American Community Survey data does not reveal whether respondents live in metropolitan or urban areas. However, the area of PUMAs can be used as a proxy for concentration. Area is weakly correlated with the magnitude of the estimated refinements

[2] More details on American Community Survey data collection errors are available at `https://www.census.gov/programs-surveys/acs/technical-documentation/errata.html`.

$(r = -0.19)$, suggesting that imputation-related noise is not confined to any type of rural or urban community.

Figure 5: Distribution of Estimated Refinements, $\gamma^*_{sm,ACS}$: 2016 - 2018

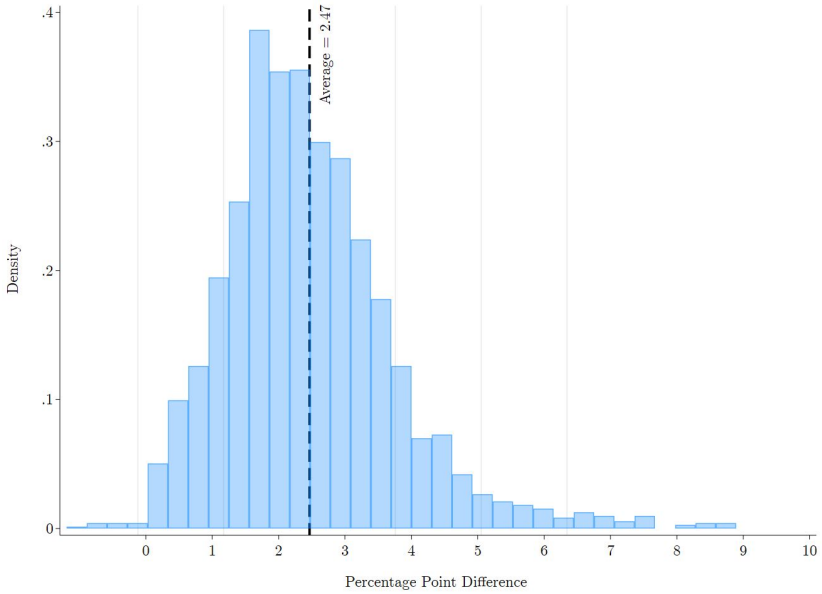
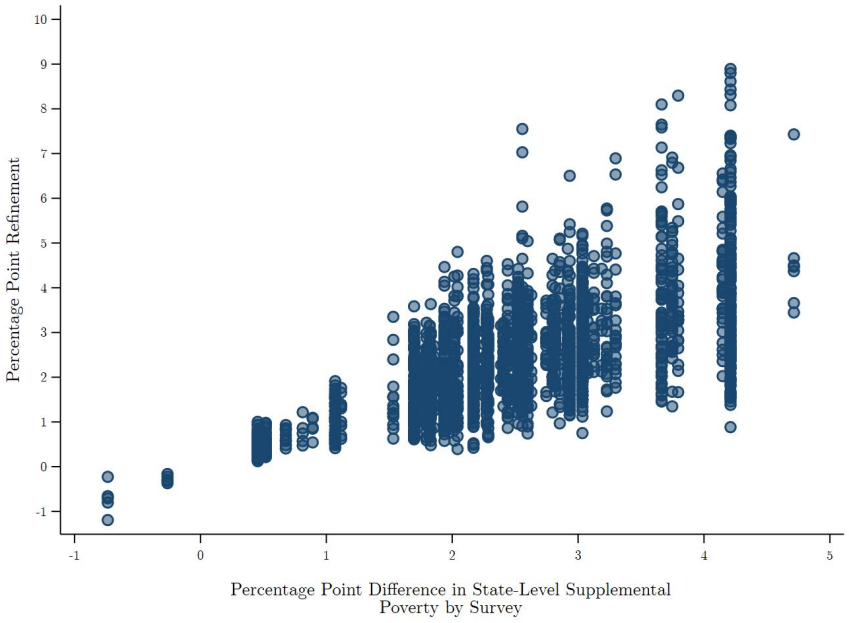
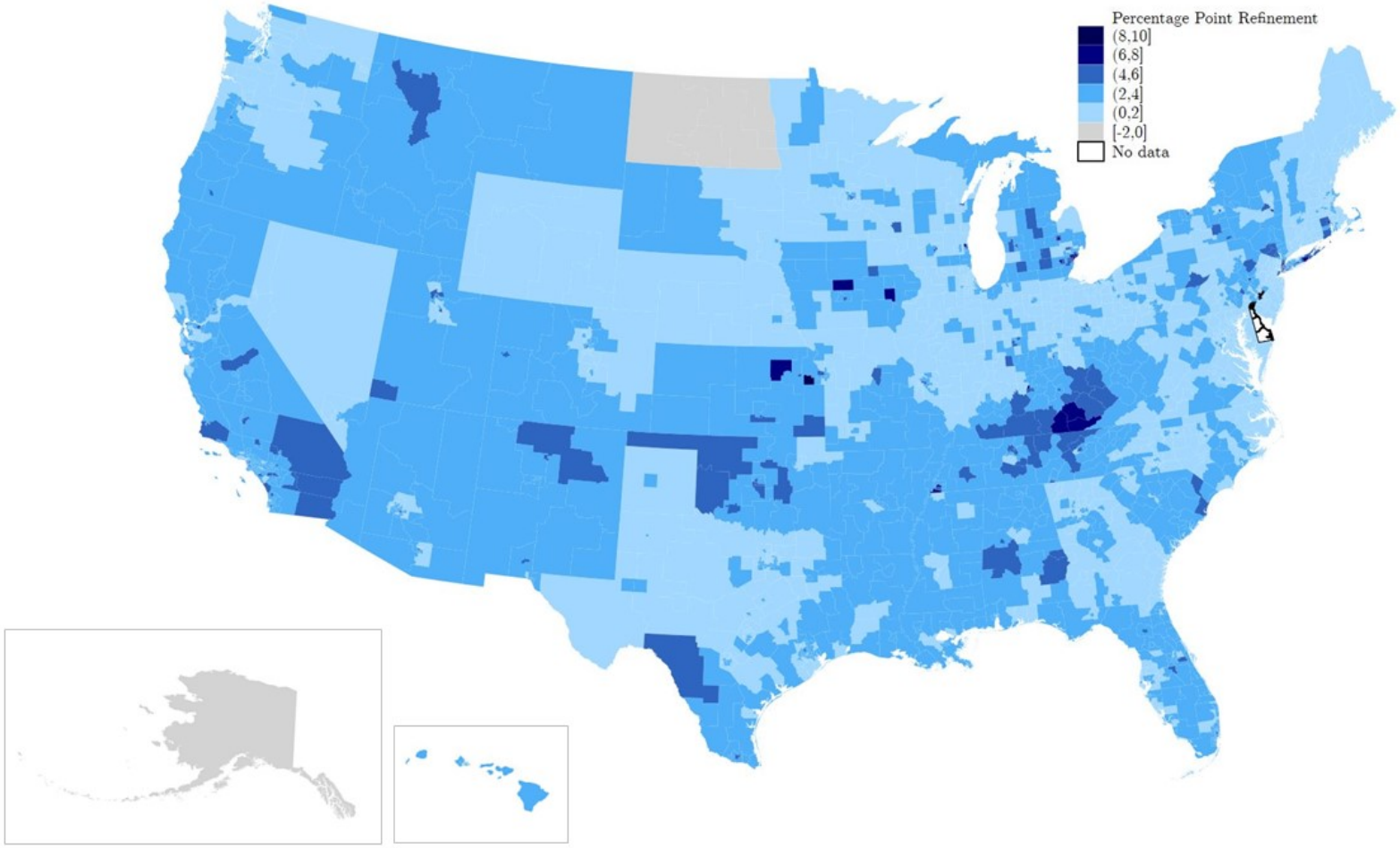


Figure 6: Estimated Refinements Versus State Level Differences by Survey: 2016 - 2018



Source: American Community Survey and Current Population Survey, Social and Economic Supplements, 2016 - 2018.
Note: PUMAs in the state of Delaware and in Philadelphia city are excluded due to data collection errors in the American Community Survey. Grey vertical lines in Figure 5 denote standard deviations from the mean.

Figure 7



Source: American Community Survey and Current Population Survey, Social and Economic Supplements, 2016 - 2018.

Note: PUMAs in the state of Delaware and in Philadelphia city are excluded due to data collection errors in the American Community Survey.

5.3 Application of Refined Estimates

Refined supplemental poverty rates produced using the information-theoretic model provide an alternative measure of the economic well-being of small areas. The supplemental poverty measure considers government aid thus can be used to measure the poverty-reducing impact of specific programs. This section provides an example of how refined supplemental poverty rates can provide an alternative picture of the effectiveness of housing subsidies in the state of Maryland. Although this policy has a relatively small impact on the national-level poverty rate [7], its relationship to housing prices, which are spatially dependent and higher in urban ares, is informative about the potential improvements these refinements may provide. Counterfactual estimates are produced by first subtracting the cash value of a specific aid program from an individual's resource summation, then comparing this reduced resource summation to their original poverty threshold. For example, let $\boldsymbol{R}_{i,HS} = \boldsymbol{R}_i - HousingSubsidy_i$, the resource summation of individual i without their housing subsidy. The counterfactual estimated poverty status of individual i is now determined by:

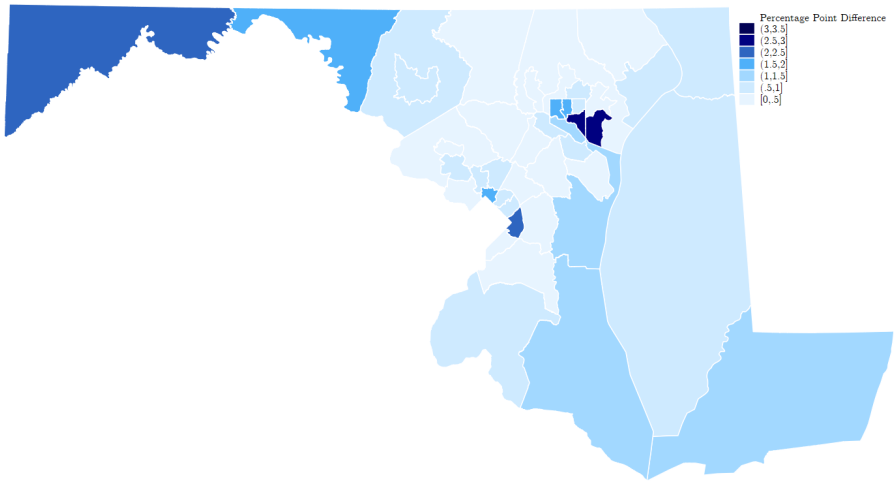
$$P^*_{i,HS} = \begin{cases} 1, & \boldsymbol{R}^*_{i,HS} < \boldsymbol{T_i} \\ 0, & \text{o.t.} \end{cases} \tag{19}$$

The counterfactual supplemental poverty status of all sampled individuals is then used to produce counterfactual supplemental poverty estimates. Subtracting the counterfactual poverty rate, $r^{*,NH}_{sm,ACS}$, of PUMA m in state s from its original counterpart, $r^*_{sm,ACS}$, provides the estimated poverty reducing effect of housing subsidies in percentage point form. Figure 8, Panel A shows the effect of housing subsidies in the state of Maryland using only ACS data. Panel B shows the effect of the same government aid program using refined estimates. Panel C compares the poverty reducing effect of housing subsides as estimated by the ACS to the same effect using refined rates.

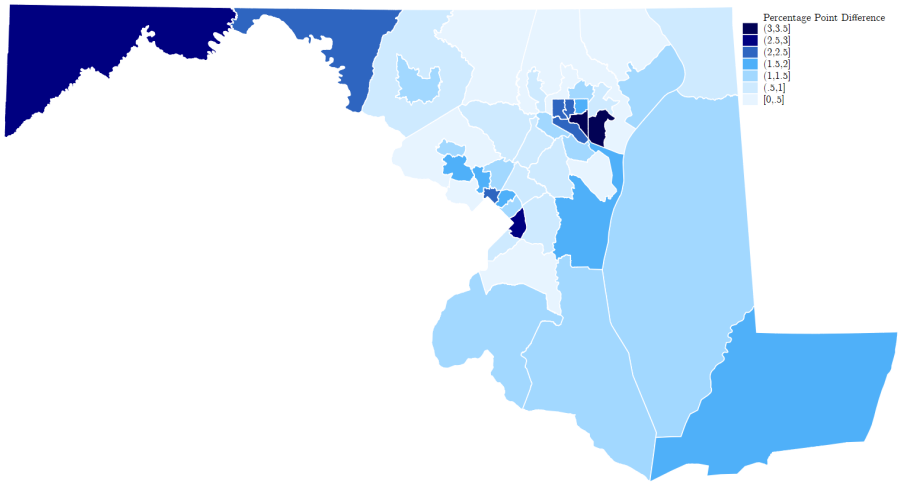
Refined supplemental poverty rates predict that American Community Survey rates are on average underestimating the effect of housing subsidies by approximately 0.5 percentage points. This magnitude is small but, as previously discussed, to be expected considering the relatively modest impact this government aid program has nationwide. More interestingly, refined supplemental poverty rates suggest that the survey is specifically underestimating the impact of housing subsidies by 0.5 to 1 percentage points in PUMAs surrounding Washington D.C. and the city of Baltimore (see Figure 7, Panel C). The information theoretic model considers no additional area characteristics, or information on average housing prices or difference in cost of living. Regardless, refinements reasonably recognize regions where the reducing impact of housing subsidies may be underestimated. This result suggests that the information-theoretic model is inherently handling spatial heterogeneity and dependence as shown in the simulations by Papalia & Fernandez-Vasquez [16].

Figure 8: Estimated Effect of Housing Subsidies in Maryland: 2016 - 2018

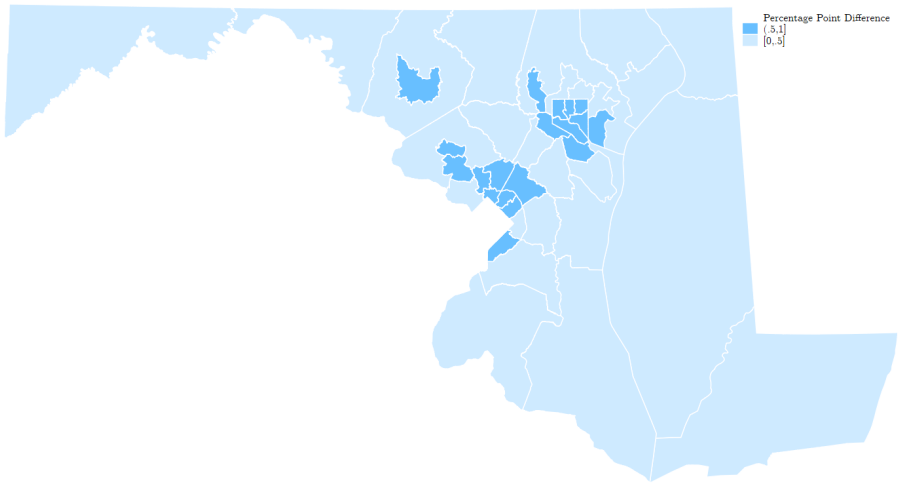
Panel A.
Using Supplemental Poverty Rates from American Community Survey (ACS)



Panel B.
Using Refined Supplemental Poverty Rates s



Panel C.
Comparison of Estimated Effects Between ACS Rates and Refined Rates



6 Conclusion

This paper proposes an information theoretic model to refine small area estimates using aggregate out of sample data. Supplemental poverty estimates from two surveys at different geographic levels are used in one model. Although small area estimates are noisy, they can be used in conjunction with more reliable estimates at the aggregate geographic level to produce refined rates. Unlike traditional small area estimation methods, no additional small area-level characteristics are needed to produced these refined estimates. Simulation results in Section four show that information-theoretic refinements are closer to true small area poverty rates than Fay-Herriot refinements.

Refined supplemental poverty rates produced using the information theoretic model are relative to detailed data in the CPS-ASEC. They should consequently be interpreted as conservative and alternative measures of poverty at the PUMA level. They complement estimates produced solely using the American Community Survey, and can be used in a variety of policy and research applications. Unlike Official poverty estimates, refined supplemental poverty rates can help answer questions regarding the effectiveness of local poverty-reducing programs considered by the measure. PUMA-level refined supplemental poverty estimates may also be cross-walked to the county level and used to explicitly test changes in anti-poverty policy.

References

[1] U. C. Bureau, *2019 ACS PUMS data dictionary*, 2020. [Online]. Available: https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2019.pdf.

[2] ——, *Annual social and economic supplements*. [Online]. Available: https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html.

[3] *CPS ASEC redesign and processing changes*, https://www.census.gov/data/datasets/time-series/demo/income-poverty/cps-asec-design.html.

[4] R. E. Fay and R. A. Herriot, "Estimates of income for small places: An application of James-Stein procedures to census data," *Journal of the American Statistical Association*, vol. 85, pp. 398–409, 1979. DOI: https://doi.org/10.1080/01621459.1979.10482505.

[5] L. Fox, *Data dictionary, Supplemental Poverty Measure: 2018*, 2019. [Online]. Available: https://www.census.gov/content/dam/Census/topics/income/supplemental-poverty-measure/spm-data-dictionary.pdf.

[6] ——, "The Supplemental Poverty Measure: 2019," U.S. Census Bureau, Tech. Rep., 2020.

[7] L. Fox and K. Burns, "The Supplemental Poverty Measure: 2020," U.S. Census Bureau, Tech. Rep., 2021.

[8] L. E. Fox and D. Wilson, "Impact of using state average WIC values in the Supplemental Poverty Measure," U.S. Census Bureau, Working Paper 2020-16, 2020.

[9] A. Golan, *Foundations of Info-Metrics*. Oxford: Oxford University Press, 2018.

[10] C. Halbmeier, A.-K. Kreutzmann, T. Schmid, and C. Schröder, "The fayherriot command for estimating small-area indicators," *The Stata Journal*, vol. 19, no. 3, pp. 626–644, 2019. DOI: 10.1177/1536867X19874238.

[11] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957. DOI: https://doi-org.proxyau.wrlc.org/10.1103/PhysRev.106.620.

[12] P. D. Johnson, T. Renwick, and K. Short, "Estimating the value of federal housing assistance for the Supplemental Poverty Measure," U.S. Census Bureau, Working Paper 2010-13, 2011.

[13] R. D. Levine, "An information theoretical approach to inversion problems," *Journal of Physics A: Mathematical and General*, vol. 13, no. 1, pp. 91–108, 1980.

[14] H. Li and P. Lahiri, "An adjusted maximum likelihood method for solving small area estimation problems," *Journal of Multivariate Analysis*, vol. 101, pp. 882–892, 2010.

[15] B. G. Liana Fox and J. Pacas, "The Supplemental Poverty Measure using the American Community Survey," U.S. Census Bureau, Working Paper 2020-09, 2020.

[16] R. B. Papalia and E. Fernandez-Vazquez, "Forecasting socioeconomic distributions on small-area spatial domains for count data," in *Advances in Info-Metrics: Information and Information Processing across Disciplines*, M. Chen, J. M. Dunn, A. Golan, and A. Ullah, Eds., Oxford: Oxford University Press, 2020, ch. 9, pp. 240–263.

[17] *Poverty thresholds*, The 2016 and 2017 thresholds, shares, and means were produced by Juan Munoz; earlier years' results were producted by Marisa Gudrais. This work is conducted under the guidance of Thesia I. Garner. Munoz (and Gudrais prior to 2017) and Garner work in the Division of Price and Index Number Research (DPINR), Bureau of Labor Statistics (BLS)., 2020. [Online]. Available: `https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html`.

[18] A. J. Provencher, "Unit of analysis for poverty measurement: A comparison of the Supplemental Poverty Measure and the Official Poverty Measure," U.S. Census Bureau, Working Paper 2011-22, 2011.

[19] J. Rao and I. Molina, *Small Area Estimation*, ser. Second Edition. Hoboken, New Jersey, US: Wiley, 2015.

[20] T. J. Renwick, "Estimating the value of federal housing assistance for the Supplemental Poverty Measure: Eliminating the public housing adjustment," U.S. Census Bureau, Working Paper 2017-38, 2017.

[21] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[22] A. Tarozzi and A. Deaton, "Using census and survey data to estimate poverty and inequality for small areas," *The Review of Economics and Statistics*, vol. 91, no. 4, pp. 773–792, 2009.

[23] "Understanding and using the American Community Survey Public use microdata sample files: What data users need to know," U.S. Census Bureau, Research Report, 2020.

[24] C. C.-C. Valencia, J. Encina, and P. Lahiri, "Poverty mapping for the Chilean comunas," in *Analysis of Poverty Data by Small Area Estimation*, M. Pratesi, Ed., Hoboken, NJ: Wiley, 2016, ch. 20, pp. 379–403.

[25] "Variance estimation," in *Design and Methodology Report*, U.S. Census Bureau, 2014, ch. 12, pp. 1–8. [Online]. Available: `https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html`.

[26] L. Wheaton and K. Stevens, "The effect of different tax calculators on the Supplemental Poverty Measure," The Urban Institute, Research Report, 2016.

[27] Y. You and B. Chapman, "Small area estimation using area level models and estimated sampling variances," *Survey Methodology*, vol. 32, no. 1, pp. 4–10, 2006.

Appendix

The associated Lagrangian to solve the information-theoretic model proposed in section three is:

$$\begin{aligned}\mathcal{L} \quad &= -\sum_{m=1}^{M}\sum_{k=1}^{3} w_{k,sm} log(w_{k,sm}) \\ &+ \lambda \left[r^*_{s,CPS} - \frac{\sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{k=1}^{3} w_{k,sm} v_{k,sm}\right] N^*_{sm,ACS}}{N^*_{s,ACS}} \right] \\ &+ \lambda_m \left[1 - \sum_{k=1}^{3} w_{k,sm} \right]\end{aligned} \tag{17}$$

The respective first order conditions are

$$\frac{\partial \mathcal{L}}{\partial w_{k,sm}} = -log(w_{k,sm}) - 1 - \lambda \left[\frac{-v_{k,sm} N^*_{sm,ACS}}{N^*_{s,ACS}} \right] - \lambda_m = 0 \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = r^*_{s,CPS} - \frac{\sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{k=1}^{3} w_{k,sm} v_{k,sm}\right] N^*_{sm,ACS}}{N^*_{s,ACS}} = 0 \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_m} = 1 - \sum_{k=1}^{3} w_{k,sm} = 0 \tag{20}$$

Rearranging the first order condition with respect to each probability, $w_{k,sm}$, gives

$$log(w_{k,sm}) = -1 - \lambda \left[\frac{-v_{k,sm} N^*_{sm,ACS}}{N^*_{s,ACS}} \right] - \lambda_m$$

$$w_{k,sm} = e^{-(1+\lambda A_{k,sm} + \lambda_m)} \tag{A.1}$$

where $A_{k,sm} = \left[\frac{-v_{k,sm} N^*_{sm,ACS}}{N^*_{s,ACS}} \right]$. Substituting (A.1) into the normalization constraint, equation (20), gives

$$1 = \sum_{k=1}^{3} e^{-(1+\lambda A_{k,sm} + \lambda_m)}$$

$$log(1) = \sum_{k=1}^{3} log\left(e^{-(1+\lambda_m)} \cdot e^{-\lambda A_{k,sm}} \right)$$

$$0 = \sum_{k=1}^{3} -(1+\lambda_m) + log\left(e^{-\lambda A_{k,sm}} \right)$$

$$1 + \lambda_m = -\lambda \sum_{k=1}^{3} A_{k,sm}$$

$$\frac{-(1+\lambda_m)}{\sum_{k=1}^{3} A_{k,sm}} = \lambda \tag{A.2}$$

Substituting the Lagrangian, λ, into (A.1) then gives

$$w^*_{k,sm} = e^{-\left(1+\left(\frac{-(1+\lambda_m)}{\Sigma_{k=1}^{3} A_{k,sm}}\right)A_{k,sm}+\lambda_m\right)}$$

$$w^*_{k,sm} = e^{-(1+\lambda_m)} \cdot e^{(1+\lambda_m)} \cdot e^{\left(\frac{A_{k,sm}}{\Sigma_{k=1}^{3} A_{k,sm}}\right)}$$

$$w^*_{k,sm} = \frac{e^{A_{k,sm}}}{e^{\sum_{k=1}^{3} A_{k,sm}}} \tag{A.3}$$

after substituting $A_{k,sm}$ into the simplified equation above,

$$w^*_{k,sm} = \frac{e^{\left[\frac{-v_{k,sm} N^*_{sm,ACS}}{N^*_{s,ACS}}\right]}}{e^{\sum_{k=1}^{3}\left[\frac{-v_{k,sm} N^*_{sm,ACS}}{N^*_{s,ACS}}\right]}} \tag{21}$$