Danielle Wilson

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

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. The U.S. Census Bureau annually produces SPM estimates using the Current Population Survey's Annual Social and Economic Supplement (CPS-ASEC). Although the CPS-ASEC collects detailed income and relationship data, it can only produce estimates at the national and state level. Although the larger American Community Survey (ACS) can produce estimates at a more disaggregate level, it does not collect the detailed data necessary to produce SPM estimates. In 2020, the Census Bureau published a series ACS micro-data sets with imputed values for the necessary components considered by the SPM. However, the ACS estimates differ from their CPS-ASEC counterparts at both the national and state level. These differences are likely due to the additional imputation needed in the ACS but their magnitude is unknown at the Public Use Microdata Area (PUMA) level. Using an information-theoretic approach proposed by Papila & Fernandez-Vasquez (2020), the relative error of disaggregate SPM rates in the ACS is estimated by constraining their weighted average to more reliable state-level CPS-ASEC SPM rates. The estimated (relative) errors are subsequently used to refine all original average PUMA ACS rates for 2016 to 2018.

AU ID: 4818768
Date: 05/02/2021

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 [2]. Annual SPM 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. Neither the monthly Current Population Survey (CPS) nor the the American Community Survey (ACS) ask respondents for such detailed data, leaving both surveys unable to alone produce SPM estimates.

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), however, is a larger survey that can provide averages at the Public Use Microdata Area (PUMA) level. In 2020, the Census Bureau published a series of American Community Survey data sets for years 2009 to 2018 with individual supplemental poverty estimates. Fox, Glassman and Pacas [7] detail how theses data sets are created solely using publicly available information. The authors supplement each annual American Community Survey data set with additional out of sample information to estimate each respondent's supplemental poverty status. These micro data sets can subsequently be used to produce supplemental poverty rates at both the state and Public Use Microdata Area (PUMA) level. However, imputed components of an individual's resources, which are otherwise known in the Current Population Survey's Annual Social and Economic Supplement (CPS-ASEC) and are required for determining an individual's supplemental poverty status, potentially introduce additional error into the American Community Survey's estimates. Consequently, the supplemental poverty estimates at the state level, and by national demographic sub-groups, vary between the the two surveys.

This paper proposes an application of an information theoretic model to reconcile the limitations of two surveys that each produce the same poverty estimate. More accurate, but limited, information from the Current Population Survey (CPS) does not need to be ignored when estimating the supplemental poverty rates using American Community Survey (ACS) data. Without any distributional assumptions regarding within state variation, the information theoretic approach uses out of sample data (from the CPS-ASEC) to constrain in-sample supplemental poverty estimates from the American Community Survey. The model specified in this paper estimates the relative error of Public Use Microdata Area (PUMA) supplemental poverty rates using the American Community Survery (ACS). These error estimates are subsequently used to provide alternative American Community Survey supplemental poverty rates. Papalia & Fernandez-Vasquez [8] show that this method of constraining disaggregate estimates with more reliable aggregate data can produce refinements of the former, regardless of spatial heterogentiy *and* dependence.

The main benefit of using the American Community Survey (ACS) is its ability to produce estimates at the PUMA level. Consequently, refinements or alternative estimates of supplemental poverty estimates at this disaggregate geographic level can help researchers interested in using this measure for poverty related research. Additionally, PUMA level data can be cross-walked to the county level, thus refined supplemental poverty rates can help answer questions about the effectiveness of local poverty-reducing policies.

The small area estimation literature provides many similar solutions when sample data is limited or relatively unreliable [10]. 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 [11].

The remainder of this paper is organized as follows. Section two will introduce the Supplemental Poverty Measure (SPM) and highlight its use in the Current Population Survey (CPS-ASEC) and the American Community Survey (ACS). Section three will compare the supplemental poverty rates in each survey and discuss how they can be related. The information theoretic model is specified in section four and the model's results are presented in section five. Section six contrasts the estimated effect of housing subsidies when using the traditional survey and information theoretic estimates. Section seven concludes the paper and discusses future work.

2 Defining Poverty Status

This section will review how poverty is defined using the supplemental poverty measure (SPM) and compare its application in the CPS-ASEC and ACS. 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 below a threshold, $\boldsymbol{T_i}$, defined by the Bureau of Labor Statistics [9] and adjusted for family size, composition and geographic residence. In equation (1) below, P_i is the estimated poverty status of individual i who has been sampled in a survey.

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

Individual i contributes to their family's total resource summation, $\boldsymbol{R_i}$, through income earned or the receipt of, for example, a non-cash benefit. Each family member's individual resource summation is equal to the sum of their family's pooled resources net of expenses. Consequently, individual i's poverty status is determined by their family's pooled resources, $\boldsymbol{R}_i$, and not solely their own resources.

There are 13 components within an individual i's family resource summation. Equation (2) defines each component j considered by the SPM:

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

where each component of an individual's family resource summation is noted as r_{ij}. $SNAP_i$ 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. Fox [2] discusses in detail how each component is defined in the official Supplemental PoverMeasure (SPM) produced by the Census Bureau.

2.1 Imperfect Information and Resource Estimation

2.1.1 Using the CPS-ASEC

All terms in equation two are either observed in the survey data or estimated using additional out-of sample information. Terms that are not directly observed use a combination of survey data, and non-survey public and administrative records. 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, requires 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 (CPS), its supplement, or the American Community Survey (ACS) collect such information. The official SPM report uses statistical simulations to determine each individual's estimated dollar value of paid federal and state taxes. These simulation estimates, which are inherently imperfect, use demographic data from the CPS-ASEC and are statistically matched to IRS public-use micro data. Let $FederalTaxes_i$ be the true and unknown value of taxes paid by individual i's family. The estimated amount of federal taxes paid can be represented as:

$$FederalTaxes_i^* = FederalTaxes_i + \epsilon_i$$

where $FederalTaxes_i^*$ is the imputed value of federal taxes paid by individual i's family in the data and ϵ_i is the unobserved error associated with the simulation's estimate. Similarly, the total value of housing assistance received is determined by simulation estimates that are statistically matched to administrative data from the U.S. Department of Housing and Urban Development.

If equation (2) is an individual's unobserved, true resource summation, their *observed* resource summation is:

$$\begin{aligned} \boldsymbol{R}^*_{i,CPS} &= \sum_{j=1}^{5} r_{ij} + \sum_{j=7}^{13} r^*_{ij} \qquad (3) \\ &= FamilyCashIncome_i - ChildCareExpenses_i - MedicalExpenses_i \\ &\quad - ChildSupport_i + EnergySubsidy_i + SNAP_i - (FederalTaxes_i + \epsilon_{i,7}) \\ &\quad - (StateTaxes_i + \epsilon_{i,8}) - (FICA_i + \epsilon_{i,9}) - (WorkExpenses_i + \epsilon_{i,10}) \\ &\quad + (HousingSubsidy_i + \epsilon_{i,11}) + (WIC_i + \epsilon_{i,12}) + (SchoolLunch_i + \epsilon_{i,13}) \end{aligned}$$

where components denoted with an asterisk, r^*_{ij} are imputed and thus contain an imputation error term, ϵ_{ij}. Consequentially, equation (1), the indicator function determining an individual's poverty status, for the CPS-ASEC would be

$$P^*_{i,CPS} = \begin{cases} 1, & \boldsymbol{R}^*_{i,CPS} = \sum_{j=1}^{5} r_{i,j} + \sum_{j=7}^{13} r^*_{i,j} < \boldsymbol{T_i} \\ 0, & \text{o.t.} \end{cases} \tag{4}$$

2.1.2 Using the ACS

The ACS collects less income information than the CPS-ASEC. Consequently, in addition to the seven total resource components that need estimated in the CPS-ASEC (as noted in equation (3)), the dollar values of $SNAP_i$, $EnergySubsidy_i$, $MedicalExpenses_i$ and $ChildCareExpenses_i$ also need to be estimated in the ACS. Equation (1) rewritten for the ACS would subsequently be

$$P^*_{i,ACS} = \begin{cases} 1, & \boldsymbol{R}^*_{i,ACS} = FamilyCashIncome_i + \sum_{j=2,\ j\neq 4}^{13} r^*_{i,j} < \boldsymbol{T_i} \\ 0, & \text{o.t.} \end{cases} \tag{5}$$

where the only term directly provided from the ACS is individual i's family cash income. Note that child support (i.e. $r^*_{i,4} = ChildSupport^*_i$) is not included in an individual's resource summation when using the ACS. This survey collects no information on whether or not an individual receives child support. Consequently, the value of child support paid is neither included not imputed in the ACS variation of Supplemental Poverty Measure (SPM). Section three discusses how this difference is accounted for in the information theoretic model.

In the CPS-ASEC, there is uncertainty regarding seven of the 13 terms needed to identify an individual's poverty status. However, in the ACS, there is uncertainty regarding *all but one* of the same 12 terms. Consequently, the individual level SPM estimates are arguably more reliable in the CPS-ASEC as fewer components need imputed. Additionally, imputed values of, for example, the value of a family's housing assistance are are not statistically matched to administrative data in the ACS. Other than family's cash income, all components of an family's total resources are estimated using publicly available data.

3 Calculating and Comparing Poverty Rates

Given that implementation of the Supplemental Poverty Measure (SPM) in the ACS does not include any information on child support, this component is also excluded from the CPS-ASEC variation of the measure. The poverty rates from the CPS-ASEC noted from this point forward are modified to exclude child support to ensure comparability between the implementation of the same estimate in different surveys. The exclusion of this component from the measure is not expected to significantly alter the supplemental poverty rates of geographic regions. For example, in 2019 consideration of this necessary expense changed the supplemental poverty status of approximately 300 thousand (0.09 percent of) individuals in the United States [2].

Using the empirical definition of poverty for the CPS-ASEC from equation (4), the rate of poverty for any 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 and $P^*_{is,CPS}$ is the observed poverty status for a CPS-ASEC sampled individual i who lives in state s. Similarly, the rate of poverty for PUMA

m in state s using the ACS is

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

where $r^*_{ms,ACS}$ is the SPM rate of PUMA m in state s, and $n_{ms,ACS}$ is the sample size from PUMA m and state s in the ACS.

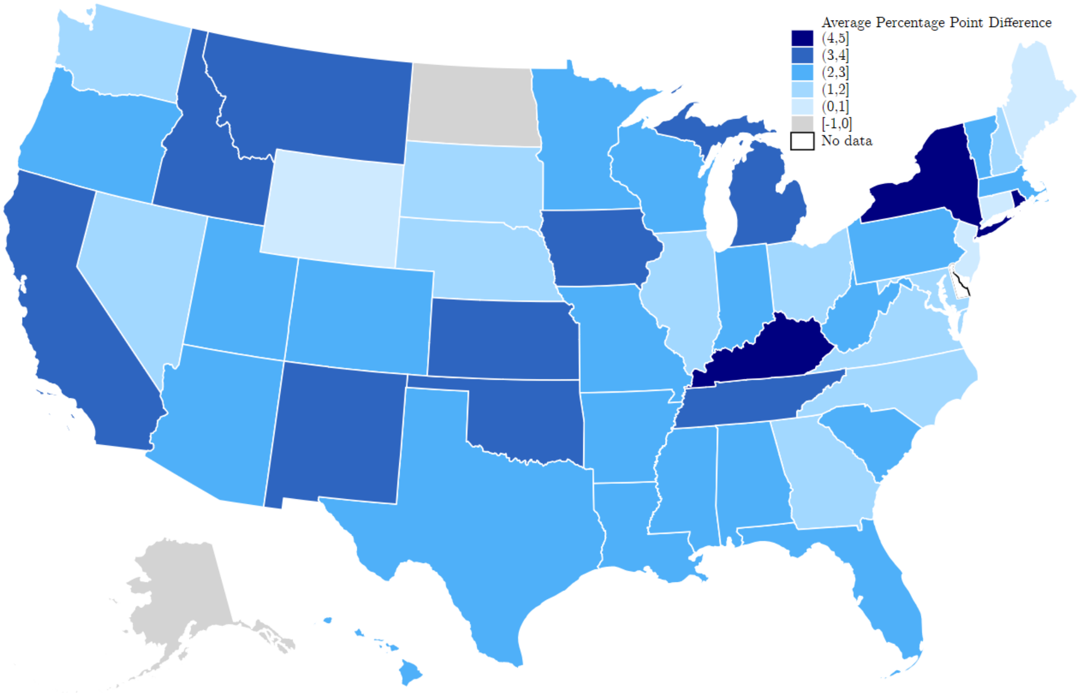
If both the CPS-ASEC and ACS samples are representative of the true population, SPM estimates by geography should be approximately equal. However, as discussed in Fox, Glassman and Pacas [7], SPM rates differ annually at the country level and by demographic characteristics. As shown in Figure 1, state level poverty estimates also vary substantially between surveys. From 2016 to 2018 the SPM rate in 15 states differs by at least three percentage points. As discussed in the previous section, although the income information provided by ACS is relatively limited, the survey's sample size is large enough to produce estimates at a more disaggregate level.

The CPS-ASEC can only provide SPM estimates at the state level while the ACS can provide estimates at both the state *and* PUMA level. Note that the ACS SPM rate at the state level, s, can be written as a weighted sum of SPM rates at the PUMA level, m:

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

where M is the number PUMAs in state s. The additional imputation at the individual level in the ACS adds uncertainty to the SPM rates at both geographic levels. Focusing on the former, the relationship between what is observed and what is true at the PUMA level is summarized by

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



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. Note that data on the state of Delaware are excluded due to data quality concerns in 2017. For more information see `https://www.census.gov/programs-surveys/acs/technical-documentation/errata/120.html`

the equation below:

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

where $r^*_{sm,ACS}$ is the observed poverty rate at the PUMA level in the ACS, $r_{sm,ACS}$ is the true poverty rate and $\gamma_{sm,ACS}$ is the imputation related noise. Substituting equation (9) into equation (8) gives

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

the empirical ACS supplemental poverty rate at the state level in terms of its weighted counterparts at the PUMA level.

4 Constructing the Model

Equation (10) summarizes the relationship between the percentage of individuals who are poor at the state level to those who are poor in each respective PUMA in the ACS. The additional imputation needed in the ACS introduces error at the individual level thus impacting both the PUMA and state estimates. However, the degree to which these estimates are impacted can only be compared at the state level, where there are more reliable estimates from the CPS-ASEC. Substituting the later into the left hand side of equation (10) gives

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

In the equation above, the ACS estimates on the right hand side are now constrained by the respective CPS-ASEC state estimate on the left hand side.

Equation (11) will serve as the primary observed constraint in the information theoretic model. The objective of the model will be to estimate the relative noise in the ACS PUMA level estimates from additional, but needed, imputation. Note that in such a model, the average error term in equation (11), $\gamma_{sm,ACS}$, must be redefined over a discrete support space [4] [3]. Assuming that the error distribution is symmetric and centered around zero, a support space of dimension three can be defined as $[v_1 \;\; v_2 \;\; v_3] = [-C \;\; 0 \;\; C]$, where $C \in \mathbb{R}^{>0}$. The corresponding probabilities of realizing each element of the support space are w_1, w_2, and w_3, respectively. The error term in equation (11) can thus be written as

$$\gamma_{sm,ACS} = \sum_{k=1}^{3} w_{ksm,ACS} v_{ksm,ACS} \tag{12}$$

where k corresponds to each element in the support space, $v_{kms,ACS}$ is the PUMA specific support space element, and $w_{kms,ACS}$ is the corresponding probability of $v_{kms,ACS}$. Without any information from, for example, an observed constraint, the probability distribution of the support space is otherwise uniform.

Samples from the CPS-ASEC and the ACS are provided with survey weights to produce representative population estimates. Poverty rate estimates using sample weights subsequently have standard errors. The standard errors associated with each weighed PUMA level supplemental poverty rate are used to define the support space for the error term in equation (12): $\boldsymbol{v_{sm,ACS}} = [v_{1sm} \;\; v_{2sm} \;\; v_{3sm}] = [-x \cdot SE_{sm} \;\; 0 \;\; x \cdot SE_{sm}]$, where SE_{sm} is the standard error specific to the weighted estimate of poverty from PUMA m in state s.

The results presented in section five use a support space defined by $x = 10$, i.e. $\boldsymbol{v_{sm,ACS}} = [-10 \cdot SE_{sm} \;\; 0 \;\; 10 \cdot SE_{sm}]$. This support space is wide enough to satisfy the observed constraint in equation (11) for all PUMA's in the United States. As shown in Figure 1, the state poverty rates between the two surveys vary by up to five percentage points. This large variation prevents the information theoretic model (described below) from finding a solution if the support space were narrowed. For example, if the support space were $[-4 \cdot SE_{sm} \;\; 0 \;\; 4 \cdot SE_{sm}]$, the information theoretic model would not be able to estimate the relative error of the PUMA level supplemental poverty rates in 11 states. Rhode Island, New York and Kentucky are among the states that have the highest percentage point differences in poverty rates between the two surveys *and* whose PUMA level poverty rates cannot be refined using the information theoretic model. Section five presents a sensitivity analysis among the states whose supplemental poverty rates were able to be refined using a narrower support space. When using a 99 percent confidence interval, only three (out of 1,919) PUMA's had had statistically different refined supplemental poverty estimates when using either the $[-10 \cdot SE_{sm} \;\; 0 \;\; 10 \cdot SE_{sm}]$ or the $[-4 \cdot SE_{sm} \;\; 0 \;\; 4 \cdot SE_{sm}]$ support spaces.

Using the $[-10 \cdot SE_{sm} \ \ 0 \ \ 10 \cdot SE_{sm}]$ and the observed constraint in equation (11), the information theoretic model for each state s is:

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

subject to

$$r^*_{s,CPS} = \frac{\sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{k=1}^{3} w_{ksm}v_{ksm}\right] N^*_{sm,ACS}}{N^*_{s,ACS}} \tag{11}$$

$$\sum_{k=1}^{3} w_{ksm,ACS} = 1 \ , \ w_{kms,ACS} \geq 0 \tag{14}$$

$$1 \geq \sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{\mathrm{k}=1}^{3} w_{\mathrm{ksm}}v_{\mathrm{ksm}}\right] \geq 0 \tag{15}$$

where $N^*_{s,ACS}$ and $N^*_{ms,ACS}$ are the weighted, observed populations of state s and PUMA m, respectively, and M is the total number of PUMAs in state s. The problem specified above maximizes the entropy objective function (i.e., equation (13)) subject to observed and normalizing constraints (i.e., equations (14) and (15)) [5] [6]. The entropy objective function produces the most uniform and least biased error distribution given the observed constraint, equation (11) [3]. In other words, the maximum entropy objective function produces the most conservative estimate for the probability distribution, $\boldsymbol{\hat{w}_{sm,ACS}} = [\hat{w}_{1sm} \ \ \hat{w}_{2sm} \ \ \hat{w}_{3sm}]'$, associated with the error support space $\boldsymbol{v_{sm,ACS}}$ for each PUMA m in state s. Using $\boldsymbol{\hat{w}_{sm,ACS}}$, the average error can be estimated as follows:

$$\hat{\gamma}_{sm,ACS} = \sum_{k=1}^{3} \hat{w}_{ksm,ACS} v_{ksm_ACS} \tag{16}$$

This estimated error can then be converted to percentage points and subsequently subtracted from each observed PUMA level supplemental poverty rate. The result is a refined supplemental poverty rate at the PUMA level, $r^{\boldsymbol{R}}_{sm,ACS}$:

$$r^{\boldsymbol{R}}_{sm,ACS} = r^{*}_{sm,ACS} - \underbrace{(\hat{\gamma}_{sm,ACS} \cdot 100)}_{\text{Percentage Point Refinement}} \tag{17}$$

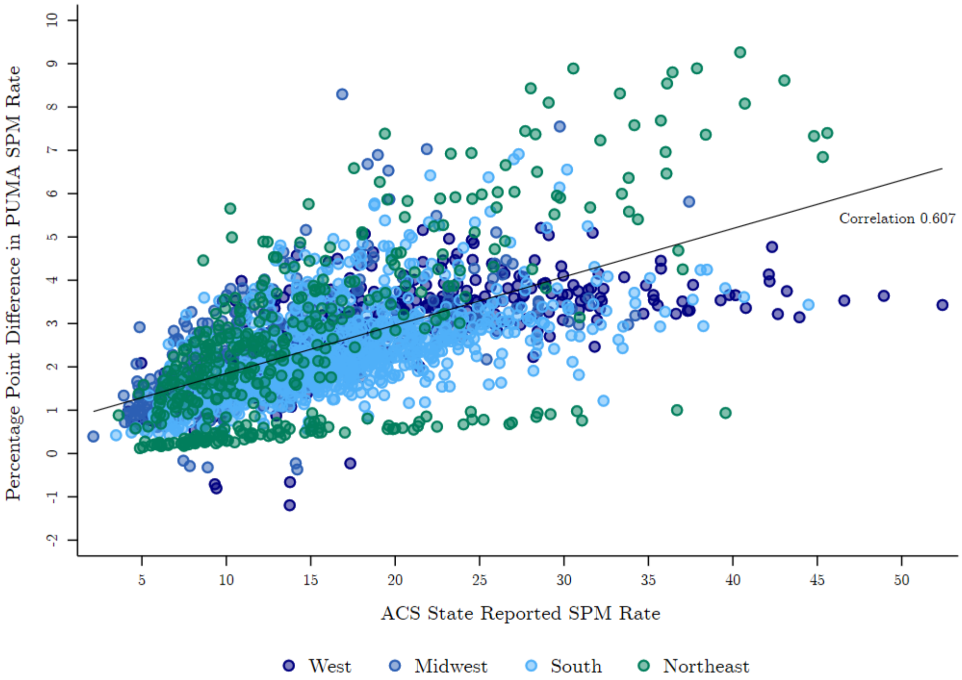
5 Results

5.1 Entropy Estimated Refinements

This section will review findings from the re-estimated PUMA SPM rates using the maximum entropy model specified in the previous section. 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 maximum entropy model, including those from the ACS, 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 subsequently used in this analysis.

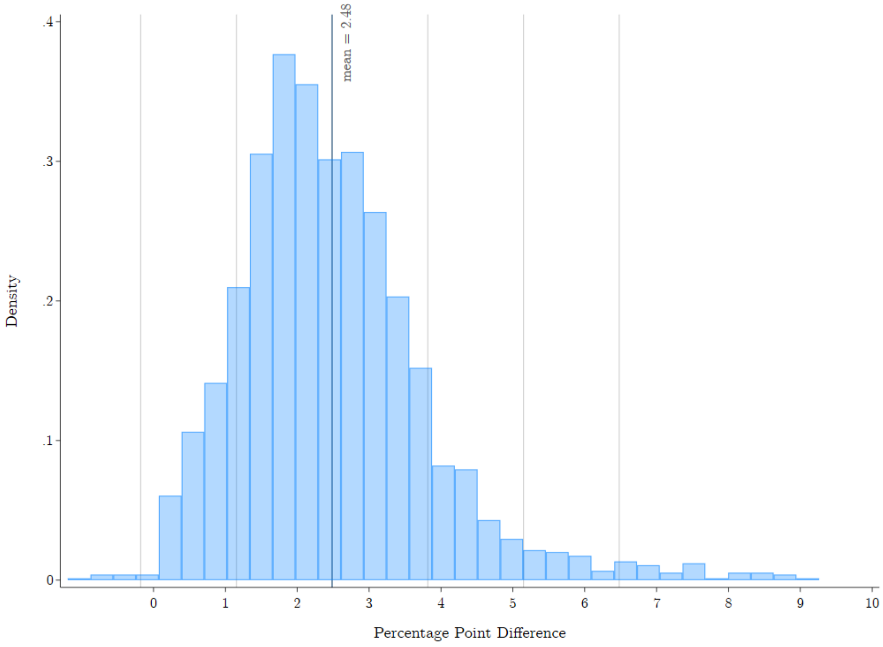
Figure 2 presents the relationship between every reported ACS SPM rate its refined entropy-estimated error, $\hat{\gamma}_{sm,ACS}$, in percentage points. Note that most refinements are positive as the ACS supplemental poverty rates are in most states greater than their CPS-ASEC SPM counterparts. As noted in equation (17), a positive refinement, $\hat{\gamma}_{sm_{ACS}} > 0$, once subtracted from the original ACS SPM rate, provides a lower refined supplemental poverty (i.e. re-estimated rate), $r^{\boldsymbol{R}}_{sm_{ACS}}$. The moderately strong correlation between a PUMA's original SPM rate and the magnitude of its refinement is unsurprising given the large differences observed in Figure 1 at the aggregate, state level. Variation in the magnitude of entropy-estimated refinements is greatest for PUMAs in the Northeast and seemingly increases with the ACS SPM rate.

Figure 2: Relationship Between ACS SPM Rate and Entropy Estimated Refinements: 2016 - 2018



Note that data on the state of Delaware are excluded due to data quality concerns in 2017. For more information see `https://www.census.gov/programs-surveys/acs/technical-documentation/errata/120.html`.

Figure 3: Distribution of Maximum Entropy Refinements: 2016 - 2018



Gray vertical lines indicate standard deviation from the mean.Note that data on the state of Delaware are excluded due to data quality concerns in 2017. For more information see `https://www.census.gov/programs-surveys/acs/technical-documentation/errata/120.html`.

The average difference between the PUMA level ACS SPM rate and the re-estimated, refined rate is approximately 2.5 percentage points. In other words, the average refinement subtracted from each observed PUMA's ACS SPM rate, $r^*_{sm,ACS}$, is 2.5 percentage points. Figure 3 shows the distribution of differences across all PUMAs except for those in Delaware due to state specific data quality concerns. Interestingly, among the 42 PUMAs with refinements greater than three standard deviations above the average adjustment, approximately 80 percent were in one of three Rust Belt states: New York, Pennsylvania and Michigan.

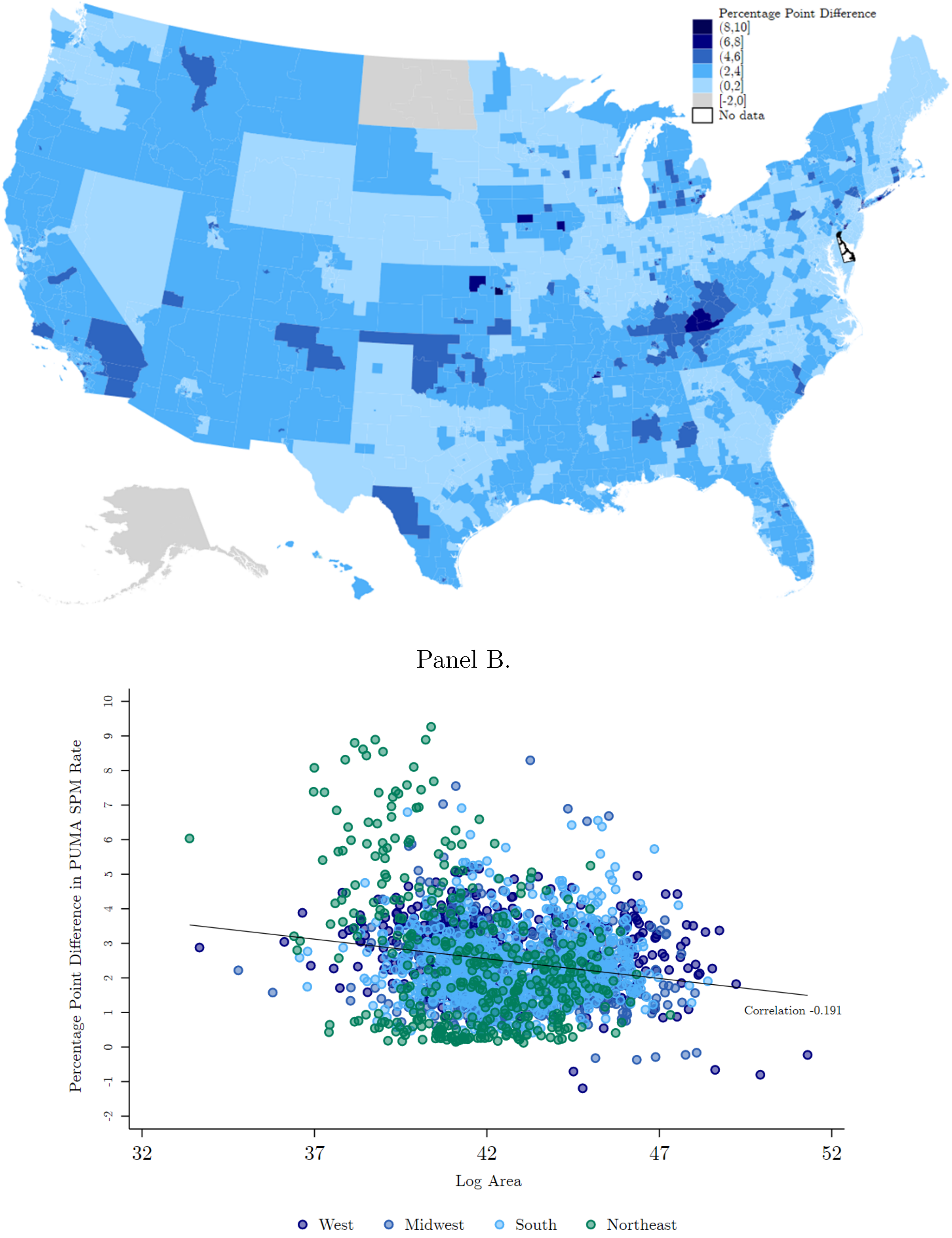
Figure 4, Panel A maps the percentage point entropy-estimated refinement, $(\hat{\gamma}_{sm,ACS} \cdot 100)$, for all PUMAs in the United States. The map presents another interesting observation: the largest refinements are not concentrated in highly populated areas. PUMAs must contain at least 100 thousand people and no more than 200 thousand. Consequently, cities are typically encompassed in smaller geographically sized PUMAS. Figure 3, Panel B presents the relationship between the area of a each PUMA and the magnitude of its entropy-based refinement. The correlation between both variables is negative but small, suggesting that neither cities or rural areas are singularly driving higher entropy-based refinements.

The following model was run to determined the relationship between the magnitude of an entropy-based refinements to each PUMA's population, geographical area and aggregate region:

$$Refinement_i = \beta_0 + \beta_1 ln(Area_i) + \beta_2 ln(Population_i) + \beta_3 SPM\_Rate_i + \boldsymbol{\beta R_i} + \epsilon_i \qquad (18)$$

where i denotes each PUMA in the U.S. except for those in the state of DE. Regional controls are included to capture any unobserved characteristics, such as variation in tax or government aid policy by state, both of which are considered by the the SPM and imputed in the ACS.

Figure 4: Magnitude of Entropy Estimated Refinements Across All PUMAS in the United States: 2016 - 2018



Panel A.

Panel B.

Note that data on the state of Delaware are excluded due to data quality concerns in 2017. For more information see `https://www.census.gov/programs-surveys/acs/technical-documentation/errata/120.html`.

Table 1 presents the estimated coefficients from the above model using aggregate Census-defined regional controls and state controls. The latter model suggests that a ten percent increase in geographic area is expected to decrease the magnitude of the entropy-based refinement by approximately 0.005 percentage points. This result confirms that PUMA size, a proxy for regional concentration, is not meaningfully correlated with the size of the entropy estimated refinement. Increases in either population or original ACS supplemental poverty rates are also not expected to meaningfully increase the magnitude of the entropy estimated refinement.

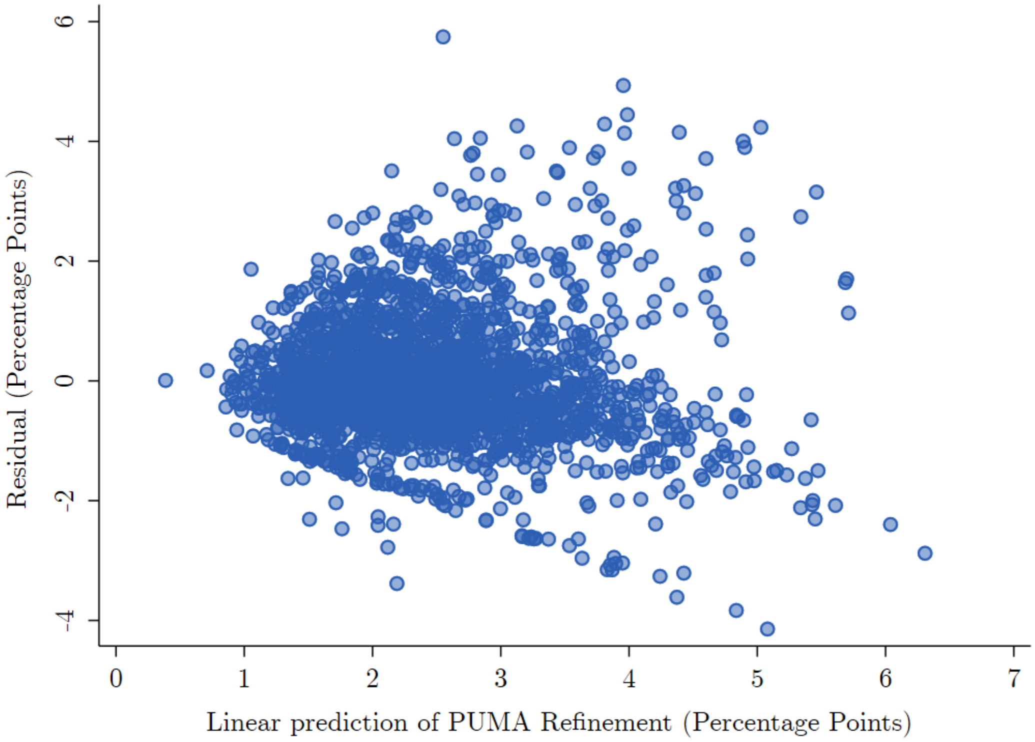
Table 1: Magnitude of PUMA-level Maximum Entropy Refinement

	(1) Census Regions	(2) State Controls
Log Geographic Area	0.00296	-0.0495***
	(0.0103)	(0.00716)
Log Population	0.525***	0.509***
	(0.0902)	(0.0592)
SPM Rate (Perct. Points)	0.0860***	0.0799***
	(0.00626)	(0.00405)
West	-0.428***	
	(0.0658)	
Midwest	-0.252***	
	(0.0681)	
South	-0.634***	
	(0.0588)	
Cons.	-4.360***	-2.227***
	(0.944)	(0.632)
N	2345	2345
R-sq	0.4038	0.7686
Adj. R-sq	0.4022	0.7634
Root MSE	1.0297	0.64789

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$.
Dependent variable is in percentage points. PUMAs in the state of DE not inlcuded due to data quality concerns.

Note that between both models presented in Table 1, the use of state controls rather than Census regional controls eliminates more prediction error. The relatively lower root mean squared error suggests that the model with state controls captures more variation in entropy-estimated refinements. However, as illustrated in Figure 5, heteroskedasticiy does exist suggesting that regression model (18) is likely omitting key covariates that could otherwise account for the magnitude of entropy-estimated refinements. State controls may be capturing other regional specific characteristics that potentially reflect demographic composition or government aid policy, and subsequently influence imputation error in the ACS supplemental poverty rates.

Figure 5



Overall, the results above suggest that PUMA characteristics implicitly considered by the information theoretic model have little impact on the relative estimated error of PUMA specific SPM rates (i.e. ($\hat{\gamma}_{sm,ACS} \cdot 100$)). Furthermore, without additional data, the magnitude of refinements, showing little meaningful relationship to population and geographic size, are least-biased estimators of the supplemental poverty error in the ACS *relative to* the more accurate rates produced by the CPS-ASEC.

5.2 Sensitivity Analysis

As discussed in section four, various support spaces were tested to ensure that the results from the maximum entropy model using the $[-10 \cdot SE_{sm}\ \ 0\ \ 10 \cdot SE_{sm}]$ support space are robust. As defined in equation (17), $r^{\boldsymbol{R}}_{sm,ACS}$ is the refined supplemental poverty rate of PUMA m in state s using the entropy-estimated refinement, $\hat{\gamma}_{sm,ACS}$. Let $r^{\boldsymbol{R},10}_{sm,ACS} = r^{\boldsymbol{R}}_{sm,ACS}$, where the superscript "10" specifies the size of the support space, $[-10 \cdot SE_{sm}\ \ 0\ \ 10 \cdot SE_{sm}]$, used to calculate the refined rate. The same maximum entropy refinements were again estimated using the $\boldsymbol{v_{sm,ACS}} = [-x \cdot SE_{sm}\ \ 0\ \ x \cdot SE_{sm}]$ support space for values $x = 4, 5, 7$ and 8. Consequentially, the following refined supplemental poverty rates were produced: $r^{\boldsymbol{R},4}_{sm,ACS}$, $r^{\boldsymbol{R},5}_{sm,ACS}$, $r^{\boldsymbol{R},7}_{sm,ACS}$, and $r^{\boldsymbol{R},8}_{sm,ACS}$.

Large differences in the observed state-level supplemental poverty rates between the CPS-ASEC and the ACS prevent the information theoretic model from estimating corresponding PUMA-level refinements. Table 2 summarizes the number of states (including the District of Columbia) whose PUMA supplemental poverty rates could be refined. Fewer states can be considered by the information theoretic model as the range of the support space narrows.

For each original refined supplemental poverty rate, $r^{\boldsymbol{R},10}_{sm,ACS}$, a corresponding simple random sample standard error was calculated and used to create a confidence interval at the 99, 95 and 90 percent level [1]. The refined rates of $r^{\boldsymbol{R},4}_{sm,ACS}$, $r^{\boldsymbol{R},5}_{sm,ACS}$, $r^{\boldsymbol{R},7}_{sm,ACS}$, and $r^{\boldsymbol{R},8}_{sm,ACS}$ where subsequently tested subject to each confidence interval centered around $r^{\boldsymbol{R},10}_{sm,ACS}$. Figure 6 presents the results

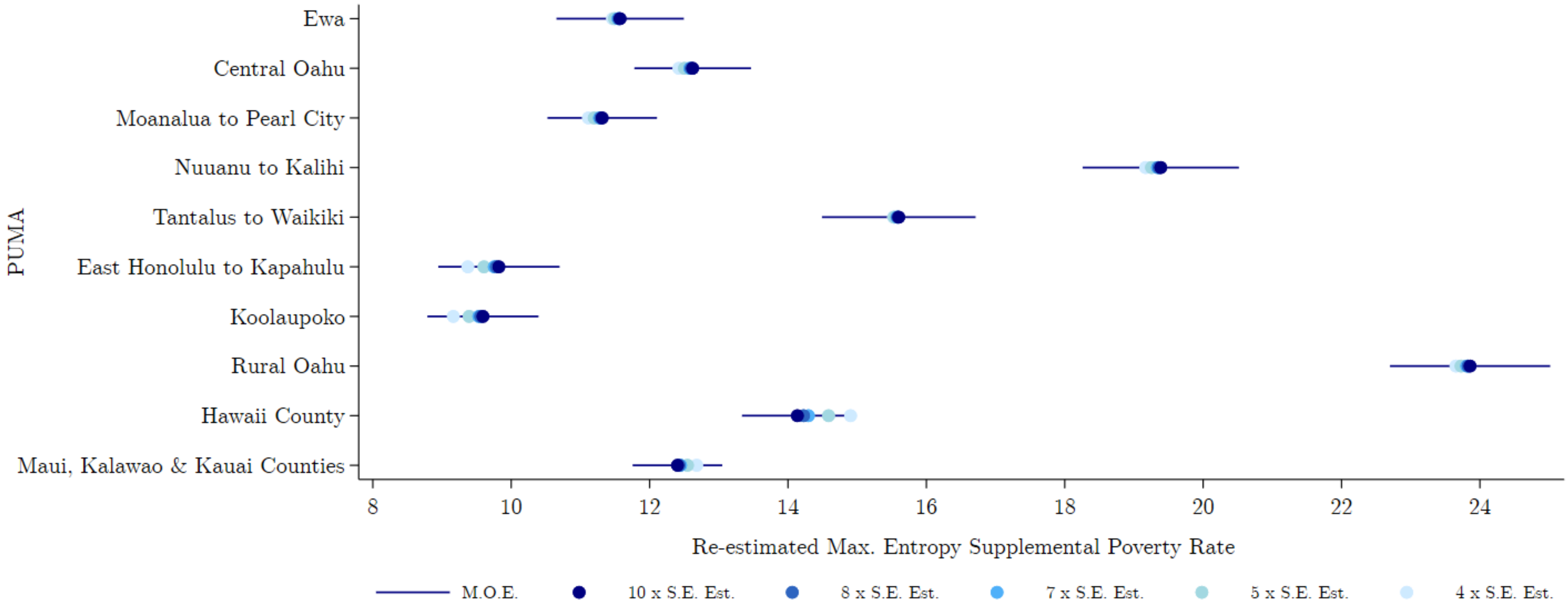
Table 2: Number of Regions Used in the Information Theoretic Model, by Support Space Size

Support Space Size, x	Number of Regions Estimated
4	40
5	45
7	50
8	51
10	51

of this test using a 90 percent confidence interval for all PUMAs in Hawaii. Hawaii is one of the 40 states in which the narrowest support space can be used to calculate its PUMA level poverty rate refinements. As illustrated below, the margin of error captures all estimates implying that the original refined poverty rate, $r_{sm,ACS}^{\boldsymbol{R},10}$, is robust for all PUMAs. Table 3 summarizes the the number of PUMAs whose original refined supplemental poverty rate was not robust. When the original refined rates are compared to their counterparts from the narrowest support space, $r_{sm,ACS}^{\boldsymbol{R},4}$, using the most conservative confidence level (i.e., the 90 percent confidence interval), only 33 out of the corresponding 1,919 PUMAs are significantly different.

[Figure 6 and Table 3 are on the next two pages.]

Figure 6



The margin of error (M.O.E.) forms the 90 percent confidence interval centered at the $r^{\boldsymbol{R},10}_{sm,ACS}$ estimate (in the same shade of dark blue).

Table 3: States with Statistically Different PUMA Refined Supplemental Poverty Rates

Support Space $[-4 \cdot SE_{sm}\ \ 0\ \ 4 \cdot SE_{sm}]$

90% C.I.		95% C.I.		99% C.I.	
State	Num. PUMAs	State	Num. PUMAs	State	Num. PUMAs
Arizona	1	California	2	California	1
California	7	Pennsylvania	6	.	.
Montana	1	Wisconsin	1	.	.
Pennsylvania	22	.	.	.	.
Wisconsin	2	.	.	.	.
Total	33	Total	9	Total	1

Support Space $[-5 \cdot SE_{sm}\ \ 0\ \ 5 \cdot SE_{sm}]$

90% C.I.		95% C.I.		99% C.I.	
State	Num. PUMAs	State	Num. PUMAs	State	Num. PUMAs
California	1	California	1	Iowa	1
Iowa	10	Iowa	5	Oklahoma	1
Montana	1	Oklahoma	7	.	.
Oklahoma	1	.	.	.	.
Tennessee	4	.	.	.	.
Total	17	Total	7	Total	2

Support Space $[-7 \cdot SE_{sm}\ \ 0\ \ 7 \cdot SE_{sm}]$

90% C.I.		95% C.I.		99% C.I.	
State	Num. PUMAs	State	Num. PUMAs	State	Num. PUMAs
New York	1	.	.	.	.
Total	1	Total	0	Total	0

None of the $r^{\boldsymbol{R},8}_{sm,ACS}$ estimates were statistically different from their $r^{\boldsymbol{R},10}_{sm,ACS}$ counterparts.

6 Application of Alternative Estimates

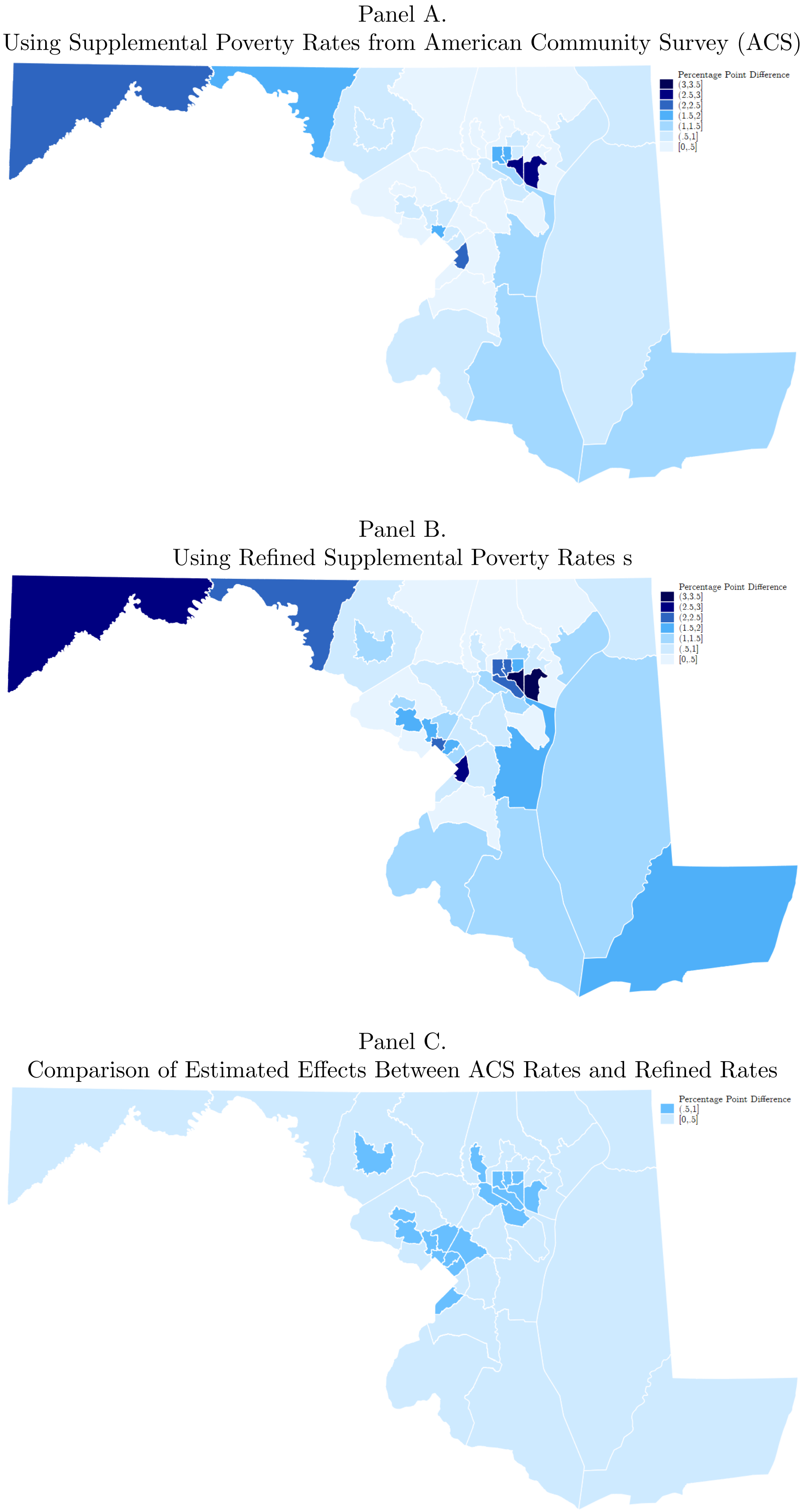
This section will review how refined PUMA-level supplemental poverty rates can be used for policy analysis. This example will focus on the poverty reducing impact of housing subsidies in the state of Maryland. Although this policy has a relatively small impact on the national poverty rate [2], its relationship to housing prices, which are spatially dependent and higher near urban ares, is informative about the potential improvements entropy-based refinements may provide.

Each year the official supplemental poverty report presents estimates on the poverty reducing impact of all government aid programs considered by the measure [2]. These 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 designated poverty threshold. For example, let $\boldsymbol{R}_{i,HS} = \boldsymbol{R}_i - HousingSubsidy_i$, the resource summation of individual i without their housing subsidy. This reduced resource summation is then substituted into equation (2). The counterfactual estimated poverty status of individual i is now determined by:

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

Following the same method described in section 2, the counterfactual supplemental poverty statuses of all sampled individuals are then used to produce alternative 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 7, 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 entropy-based refined estimates. Panel C compares the poverty reducing effect of housing subsides as estimated by the ACS to the same effect using entropy-based refined rates. The entropy-based refined supplemental poverty rates predict that housing subsidies have a larger poverty reducing effect across all PUMAs in Maryland. Note that the refined rates suggest that the ACS estimates are on average

Figure 7: 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

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. However, note that the entropy-based refined poverty rates suggest that the ACS is 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 information on average housing prices, cost of living or spatial relationships, yet it is able to recognize regions in which the poverty reducing impact of housing subsidies may be underestimated. This result suggests that the maximum entropy refinements are inherently handling spatial heterogeneity and dependence as shown in the simulations by Papalia & Fernandez-Vasquez [8].

7 Conclusion

This paper proposes an information theoretic model to refine small area estimates using aggregate out of sample data. The supplemental poverty estimates from two surveys at different geographic levels are considered in one model. Unlike traditional small area estimation methods, no additional data on the area is needed. The same information at an aggregate level, but from a more reliable survey, is instead used to improve the small area estimates from another survey. The entropy-based 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 alone by either survey. This research can be extended by incorporating estimates of the same measure of poverty from the Survey of Income and Program Participation (SIPP) into the information theoretic model. Further extensions should also include simulations to mimic the sampling procedures and subsequent limitations of each survey. Simulation results will allow for better understanding of the accuracy of the entropy based refinements.

References

[1] M. Davern and J. Strief, *Ipums user note: Issues concerning the calculation of standard errors (i.e., variance estimation) using ipums data products.* [Online]. Available: `https://international.ipums.org/international/resources/misc_docs/user_note_variance.pdf`.

[2] L. Fox, "The supplemental poverty measure: 2019," U.S. Census Bureau, Tech. Rep., 2020.

[3] A. Golan, "Rational inference: A constrained optimization framework," in *Foundations of Info-Metrics*, Oxford: Oxford University Press, 2018, ch. 2, pp. 10–24.

[4] A. Golan, G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, ser. Second Edition. West Sussex, England: Wiley, 1996.

[5] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.

[6] 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.

[7] 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.

[8] 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.

[9] *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`.

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

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