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

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 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 CPS-ASEC is its sample size, which is insufficient for disaggregate geographic estimates below the state level. The 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 ACS data sets for years 2009 to 2018 with individual supplemental poverty estimates. Fox, Glassman and Pacas [4] detail how theses ACS data sets were created solely using publicly available information. These data sets can subsequently be used to produce SPM averages at both the state and PUMA level. Fox, Glassman and Pacas [4] supplement each annual ACS data set with additional out of sample data to estimate each respondent's supplemental poverty status. However, imputed components an individual's total resource summation, some of which are otherwise known in the CPS-ASEC and required for determining an individual's poverty status, potentially introduce additional error into the ACS SPM estimates. Consequentially, SPM estimates at the state level, and by national demographic sub-groups, vary between the CPS-ASEC and ACS.

Using an information theoretic (IT) approach, ACS SPM PUMA level estimates will be refined subject to more accurate CPS-ASEC state level estimates. Without any distributional assumptions regarding within state variation, the IT approach can use both in-sample data (from the ACS) and out of sample data (from the CPS-ASEC) to (i) estimate the error of ACS SPM rates and (ii) provide re-estimated PUM level SPM rates. Papalia & Fernandez-Vasquez [5] show that 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 ACS is its ability to produce estimates at the PUMA level. Consequently, refinements of SPM 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 SPM rates can help answer questions about the effectiveness of local poverty-reducing policies.

2 Defining Poverty Status

This section will review how poverty is defined using the SPM and compare its application in the CPS-ASEC and ACS. An individual i is considered poor by the SPM if the summation of their family's resources is below a threshold, $\boldsymbol{T_i}$, defined by the Bureau of Labor Statistics [1] 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 either the CPS-ASEC or the ACS.

$$p_i^* = \begin{cases} 1, & \boldsymbol{R}_i = \sum_{j=1}^{13} r_{ij} < \boldsymbol{T_i} \\ 0, & \text{o.t.} \end{cases} \tag{1}$$

$\boldsymbol{R_i}$ is the sum of income, benefits and expenses of all family members. 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, $\boldsymbol{R}_i$. Consequently, individual i's poverty status is then determined by their family's pooled resources, $\boldsymbol{R}_i$ and not than 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:

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

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.

2.1 Imperfect Information and Resource Estimation

2.1.1 Using the CPS-ASEC

All terms in equation two are either observed or estimated using additional survey data. Terms that are estimated use a combination of survey data, non-survey public data and administrative records. Non-cash benefits, and select cash benefits and expenses require estimation. Non-cash benefits, such as school lunch, have an inherit value that is typically unknown to recipients. The CPS-ASEC collects data on program participation and researchers at the Census Bureau subsequently estimate the dollar value received by each non-cash program participant.

Expenses, such as taxes paid at the federal and state level, also estimated as neither survey collects such information. The official SPM report uses simulation data 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 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 my 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:

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

where components denoted with an asterisk, r^*_{ij} are imputed. Consequentially, equation (1) 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 $SNAP_i$, $EnergySubsidy_i$, $ChildSupport_i$, $MedicalExpenses_i$ and $ChildCareExpenses_i$ also need to be estimated in the ACS. Equation (1) rewritten for the ACS

would subsequently be

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

where the only term directly provided from the ACS is individual i's family cash income. 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 13 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

Using the empirical definition of poverty for the CPS-ASEC, 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 poverty status, defined by equation (4), for an 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} 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 [4], 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 the equation below:

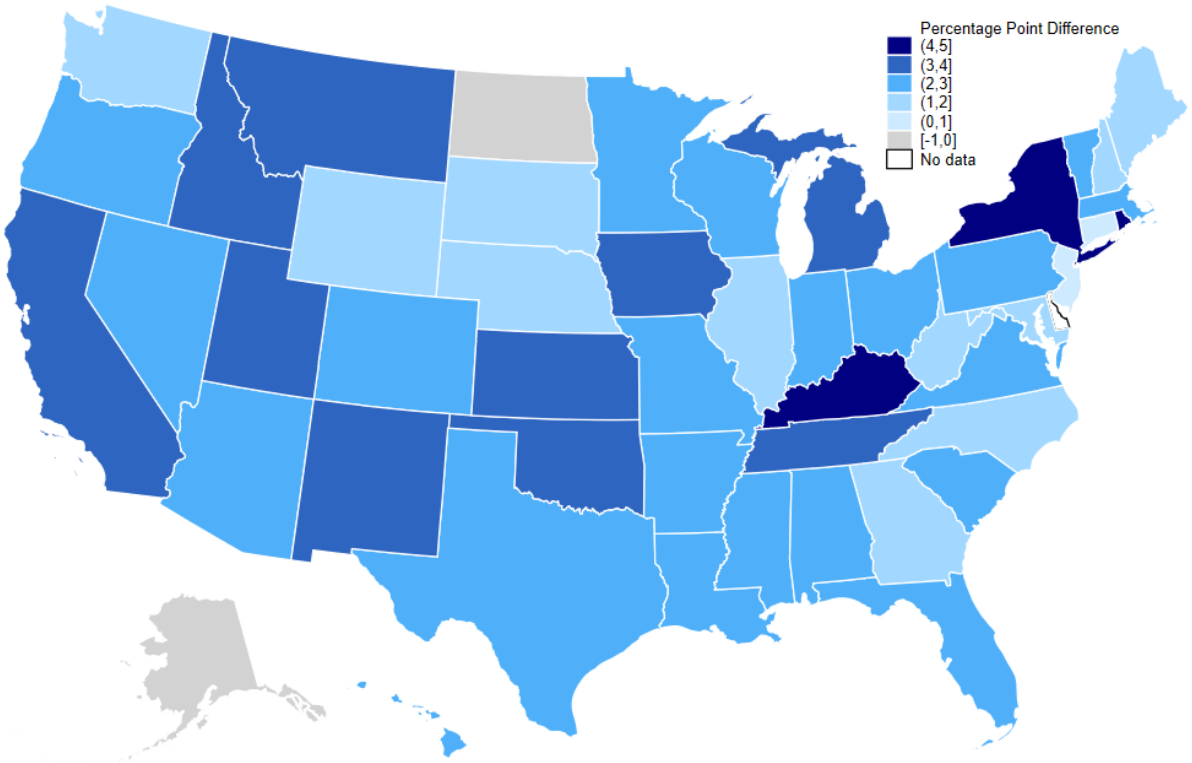
$$r^*_{sm,ACS} = r_{sm,ACS} + \mu_{sm,ACS} \tag{9}$$

Note that substituting equation (9) into equation (8), the weighted sum of SPM rates at the PUMA

level, gives

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

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



Due to insufficient 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`

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 equation (10) gives

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

where the ACS estimates on the right hand side are now constrained by the respective CPS-ASEC state estimate on the left hand side.

Note that in an information theoretic model, the average error term in equation (11), $\mu_{sm,ACS}$, must be redefined over a discrete support space. For example, a support space of dimension three can be defined as $(v_1, v_2, v_3) = (-C, 0, C)$, where $C \in \mathbb{R}$, and the corresponding probabilities of realizing each of these values is w_1, w_2, and w_3. The error term in equation (11) can thus be written as

$$\mu_{sm,ACS} = \sum_{p=1}^{3} w_{psm,ACS} v_{psm,ACS} \tag{12}$$

where p corresponds to each element in the support space.

Samples from each survey are provided with weights to produce representative estimates. Sample sizes and SPM rates can be re-weighted to be consistent with Census population data. SPM rates subsequently have corresponding standard errors within each survey. These standard errors at the PUMA level will be used to define the support space for the error term in equation (9): $\boldsymbol{v_{sm,ACS}} = (v_{1sm}, v_{2sm}, v_{3sm}) = (-10SE, 0, 10SE)$. Using this support space and the observed

constraint in equation (11), the maximum entropy problem for each state s is:

$$\max_{\boldsymbol{w}} H(P,W) = -\sum_{m=1}^{M}\sum_{p=1}^{3} w_{psm} log(w_{psm})$$

subject to

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

$$\sum_{p=1}^{3} w_{psm,ACS} = 1 \ , \ w_{pms_ACS} \geq 0$$

$$1 \geq \sum_{m=1}^{M}\left[r^*_{sm,ACS} - \sum_{p=1}^{3} w_{psm}v_{psm}\right] \geq 0$$

where $N^*_{s,ACS}$ and $N^*_{ms,ACS}$ are the weighted, observed populations of state s and PUMA m, respectively.

The problem specified above maximizes the entropy objective function subject to observed and normalizing constraints. Without any additional information, such as administrative data, the entropy objective function produces the most uncertain and least biased error distribution given the observed constraint, equation (11) [3]. In other words, the maximum entropy objective function produces the most conservative estimates for the probability distribution, $\hat{\boldsymbol{w}}_{\boldsymbol{sm,ACS}} = (\hat{w}_{1sm}, \hat{w}_{2sm}, \hat{w}_{3sm})$, associated with the error support space $\boldsymbol{v}_{\boldsymbol{sm,ACS}}$ for each PUMA m in state s. Using $\hat{\boldsymbol{w}}_{\boldsymbol{sm,ACS}}$, the average error can be estimated as follows:

$$\hat{\mu}_{sm,ACS} = \sum_{p=1}^{3} \hat{w}_{psm,ACS} v_{psm_ACS} \tag{13}$$

This estimated error can then be subtracted from each observed PUMA level SPM rate to get a

refined rate, $r^{\boldsymbol{R}}_{sm,ACS}$:

$$r^{\boldsymbol{R}}_{sm,ACS} = r^{*}_{sm,ACS} - \hat{\mu}_{sm,ACS} \tag{14}$$

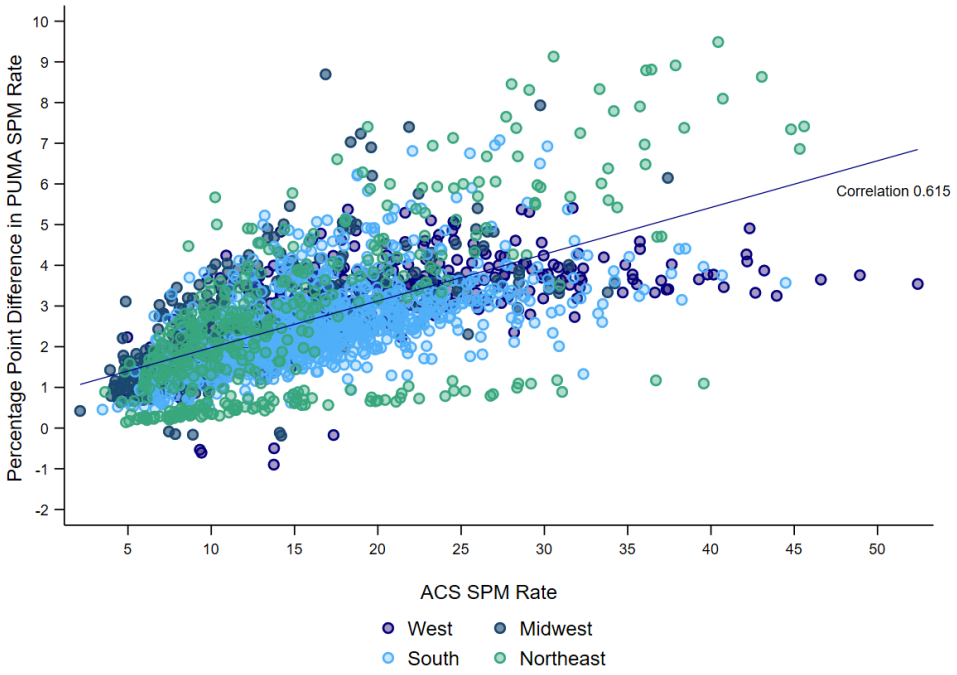
5 Results

This section will report aggregate 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 years 2016, 2017 and 2018. In 2019 the Census Bureau began releasing CPS-ASEC micro data using a redesigned survey and an updated processing. Micro data from prior year's surveys were published as a "bridge" and "research" file and used for this analysis.

Figure 2 presents the relationship between the reported ACS SPM rate and the the percentage point difference between the former and the refined entropy-based estimate for every PUMAs. Note that most refinements are positive as the ACS SPM rates are greater than their CPS-ASEC SPM counterparts for most states. A positive refinement, $\hat{\mu}_{sm_ACS} > 0$, once subtracted from the original ACS SPM rate, provides a lower refined SPM *rate* (i.e. re-estimated SPM 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 percentage point difference is greatest for PUMAs in the Northeast and seemingly increases with the ACS SPM rate.

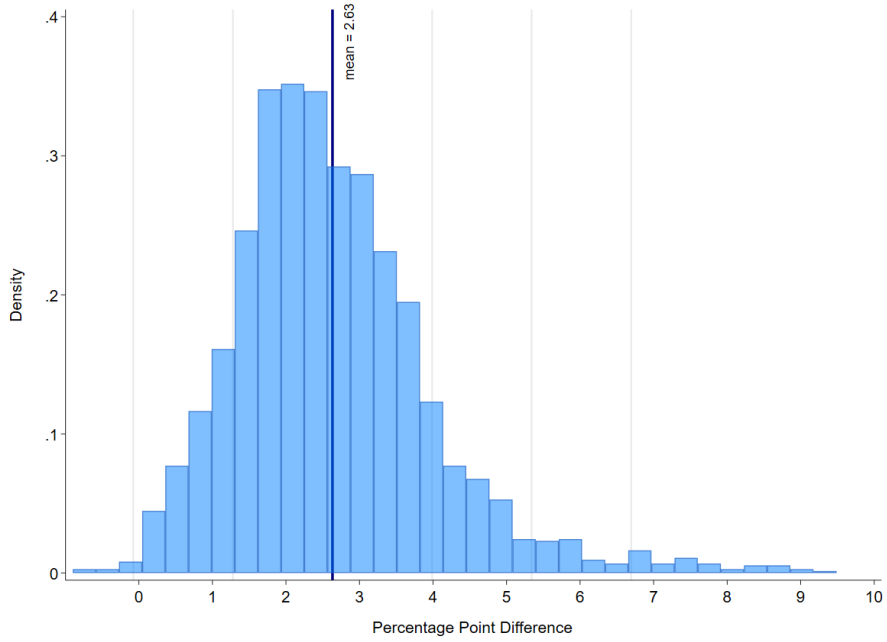
The average difference between the PUMA level ACS SPM rate and the re-estimated, refined rate is approximately 2.6 percentage points. In other words, the average refinement subtracted from a PUMA's ACS SPM rate is 2.6 percentage points. Figure 3 shows the distribution of differ-

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`. The average (standard deviation) percentage point entropy-based refinement, as presented on the horizontal axis, is 2.83 (1.00), 2.38 (1.26), 2.53 (1.04) and 2.62 (2.04) for the West, Midwest, South and Northeast, respectively. Excluding the Northeast PUMAs reduces the correlation coefficient to 0.592.

Figure 3: Distribution of Differences Between Reported PUMA (ACS) SPM Rates and 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`.

-ences across all PUMAs except for those in Delaware due to state specific data quality concerns. Interestingly, among the 41 PUMAs with refinements greater than three standard deviations above the average adjustment, 71 percent were in one of three Rust Belt states: New York, Pennsylvania and Michigan.

Figure 4, Panel A maps the refinements shown in Figure 3, i.e. the result of the maximum entropy model specified in section four 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:

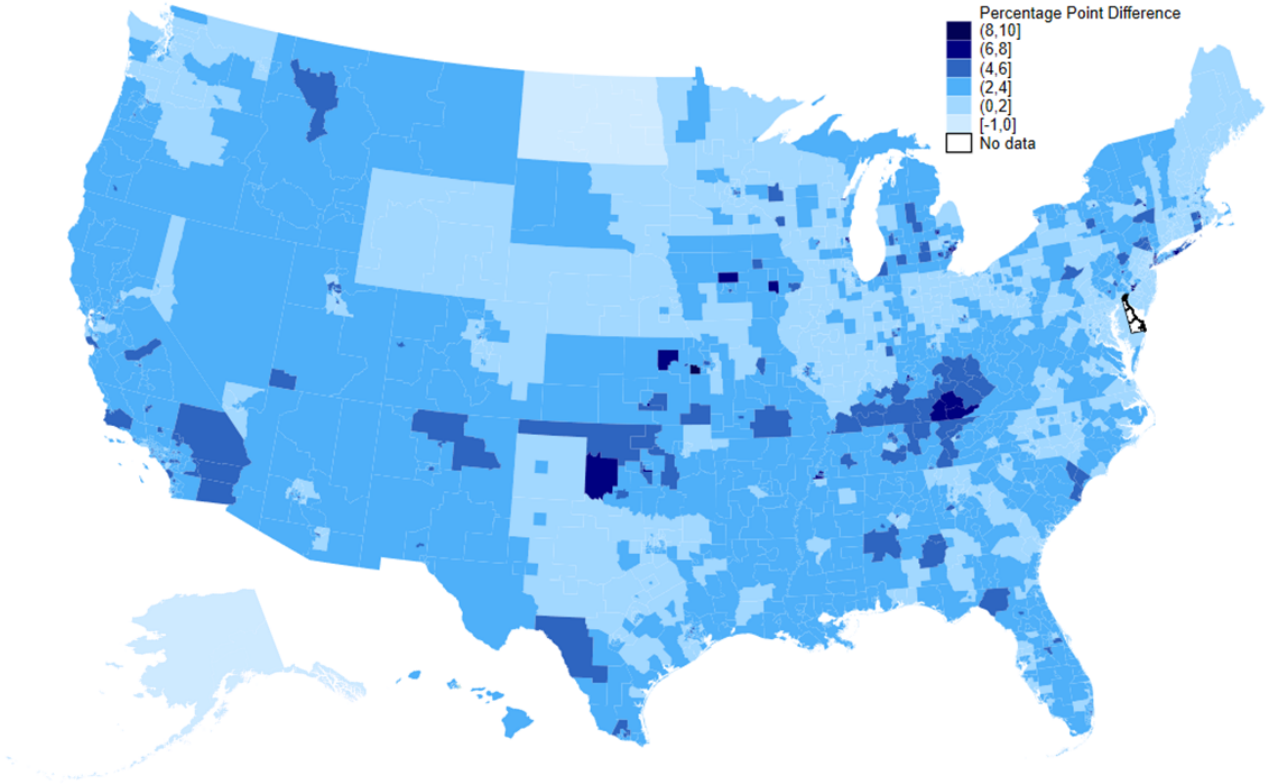
$$Refinement_i = \beta_0 + \beta_1 ln(Area_i) + \beta_2 ln(Population_i) + \beta_3 SPM\_Rate_i + \boldsymbol{\beta R_i} + \epsilon_i \qquad (15)$$

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.

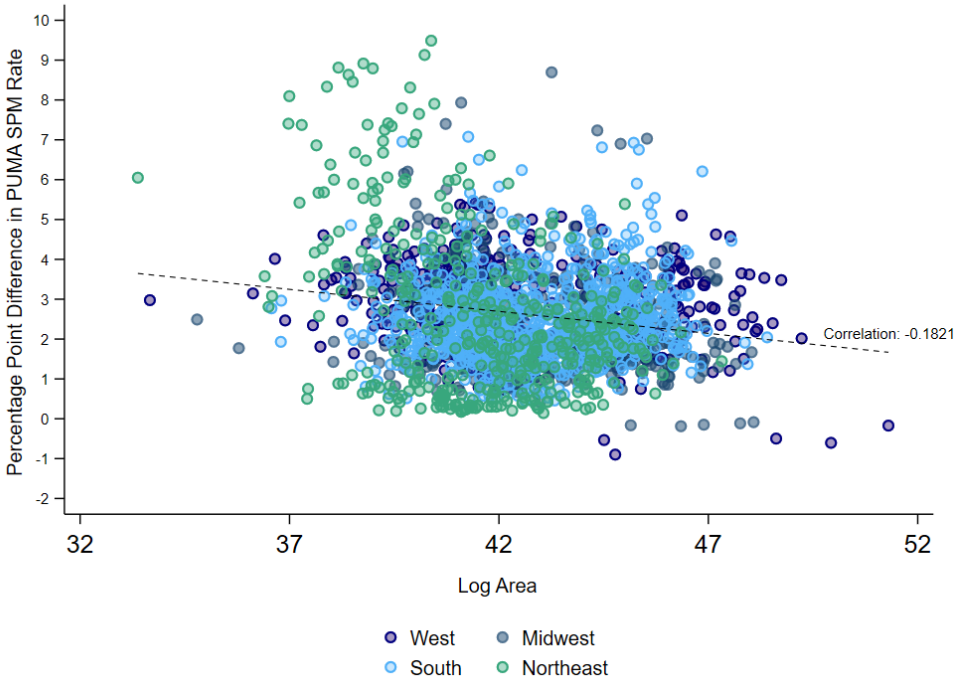
Table 1 presents the results of the above model using either aggregate Census-defined regional controls or state controls. The latter model suggests that if geographic area increases by ten percent, the magnitude of the entropy-based refinement is expected to increase by 0.005 percentage

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`. In Panel B, the correlation coefficient drops from -0.18 to -0.12 when the Northeastern PUMAs are excluded.

points. This result confirms that PUMA size, a signal of regional concentration and cities, is not expected to meaningfully increase the size of the entropy estimated refinement (i.e. estimated error as defined in equation 14). Similarly, a ten percent increase in population or a percentage point increase in a PUMA's SPM rate has little estimated effect on the magnitude of the entropy based refinement.

Table 1: Magnitude of PUMA-level Refinement: 2016 - 2018

Regional Controls:	Census Regions	State
	(1)	(2)
Log Geographic area	0.00757	-0.0513***
	(0.0105)	(0.00742)
Log Poulation	0.552***	0.543***
	(0.0981)	(0.0077)
SPM Rate (Perct. Points)	0.0882***	0.0831***
	(0.00859)	(0.00690)
West	-0.467***	
	(0.0830)	
Midwest	-0.225**	
	(0.0630)	
South	-0.630***	
	(0.0786)	
Cons.	-4.734***	-2.227**
	(0.956)	(0.783)
N	2,345	2,345
R-sq	0.411	0.761
adj. R-sq	0.409	0.755
Root MSE	1.0396	.66921
F	201.2	128.3

Robust 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, linear heteroskedasticiy does exist suggesting that the model is possibly

omitting key regressors. Although model (2) can better estimate the magnitude of a PUMA specific refinement, regional controls may be capturing other state specific characteristics that potentially reflect government aid policy, and influence imputation in the ACS and subsequently measurement error at the PUMA level.

6 Conclusion

Overall, these results suggest that PUMA characteristics such as size and population density, have little impact on the estimated error of PUMA specific SPM rates. Within PUMA-level characteristics, such as demographic characteristics, which at the individual level are considered by the imputation of needed variables such as government aid receipt, are more likely contributors to ACS SPM error. Furthermore, without additional data, the magnitude of refinements, showing little meaningful relationship to population and geographic size, are least-biased estimators of SPM error in the ACS *relative to* the more accurate rates produced by the CPS-ASEC.

References

[1] U.S. Census Bureau. *Poverty Thresholds*. 2020. URL: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html. 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).

[2] Liana Fox. *The Supplemental Poverty Measure: 2019*. Tech. rep. U.S. Census Bureau, 2020.

[3] Amos Golan. "Rational Inference: A Constrained Optimization Framework". In: *Foundations of Info-Metrics*. Oxford: Oxford University Press, 2018. Chap. 2, pp. 10–24.

[4] Brian Glassman Liana Fox and José Pacas. *The Supplemental Poverty Measure using the American Community Survey*. Working Paper 2020-09. U.S. Census Bureau, 2020.

[5] Rosa Bernardini Papalia and Esteban Fernandez-Vazquez. "Forecasting Socioeconomic Distributions on Small-Area Spatial Domains for Count Data". In: *Advances in Info-Metrics: Information and Information Processing across Disciplines*. Ed. by Min Chen et al. Oxford: Oxford University Press, 2020. Chap. 9, pp. 240–263.