GeoDa Center
for Geospatial Analysis
and Computation

ARIZONA STATE UNIVERSITY

Uncertain Uncertainty: Spatial Variation in the Quality of American Community Survey Estimates

David C. Folch, Daniel Arribas-Bel,
Julia Koschinsky and Seth E. Spielman

2014

Working Paper Number 01

Uncertain Uncertainty: Spatial Variation in the Quality of American Community Survey Estimates

David C. Folch*
University of Colorado at Boulder

Daniel Arribas-Bel
VU University

Julia Koschinsky
Arizona State University

Seth E. Spielman
University of Colorado at Boulder

WORKING PAPER†
February 26, 2014

Abstract

The U.S. Census Bureau's American Community Survey (ACS) is the foundation of social science research, much federal resource allocation and the development of public policy and private sector decisions. However, the high uncertainty associated with some of the ACS's most frequently used estimates can jeopardize the accuracy of inferences based on these data. While there is high level understanding in the research community that problems exist in the data, the sources and implications of these problems have been largely overlooked. Using 2006–2010 ACS median household income at the census tract scale as the test case (where a third of small-area estimates have higher than recommend errors), we explore the patterns in the uncertainty of ACS data. We consider various potential sources of uncertainty in the data, ranging from response level to geographic location to characteristics of the place. We find that there exist systematic patterns in the uncertainty in both the spatial and attribute dimensions. Using a regression framework, we identify the factors that are most frequently correlated with the error at national, regional and metropolitan area scales, and find these correlates are not consistent across the various locations tested. The implication is that data quality varies in different places, making cross-sectional analysis both within and across regions less reliable. We also present general advice for data users and potential solutions to the challenges identified.

Key Words: American Community Survey, data uncertainty, income estimates, margin of error, spatial analysis

JEL Classification: C81, R12.

*Corresponding author: University of Colorado at Boulder, david.folch@colorado.edu

†Do not cite or quote without author's permission.

1 Introduction

The socio-economic, demographic and housing data produced by the U.S. Census Bureau (USCB) through the American Community Survey (ACS) are crucial inputs to social science research as well as many public and private sector decisions. However, this research and these decisions are complicated by the high margin of error (MOE) commonly found in the ACS estimates. High MOEs are especially common when the estimate reflects a small subset of the population or covers a small geographic area such as a census tract. For example, over 44 percent of the ACS census tract estimates of children in poverty have an MOE at least as large as the estimate itself.[1]

There is indeed growing recognition of the uncertainty associated with ACS estimates. This general recognition is the result of various academic articles (MacDonald, 2006; Salvo and Lobo, 2006; Citro and Kalton, 2007; Spielman et al., 2014; Bazuin and Fraser, 2013) and the USCB's own efforts to publicize the challenges of working with ACS data (U.S. Census Bureau, 2009a). However, the nature and causes of the uncertainty in the ACS are not widely understood. As we demonstrate in this research, a major cause for concern is that the uncertainty generally does not follow a random pattern across attributes and space. This implies that there is systematic variation in the quality of ACS data—some types of places have higher uncertainty than others.

This article studies the pattern of uncertainty in one attribute, 2006–2010 median household income at the census tract scale. This variable is of broad interest to academic, government and private sector researchers. Household income is a critical attribute to determine market demand for business analytics and public or nonprofit service eligibility, to make decisions about retail site location and to study income inequality, to name a few examples.

The following section empirically motivates the study with examples of patterns in ACS uncertainty. This is followed by a presentation of the spatial regression specifications and the data used. An analysis at the national, regional and metropolitan area levels identifies

[1] Children in poverty is defined as people under 18 living in a family whose income is below the poverty level and who are related to the householder by birth, marriage or adoption. In 32,332 census tracts (44.3 percent), excluding Puerto Rico, the MOE is greater than or equal to the estimate. 2006–2010 ACS data extracted from the National Historical Geographic Information System, table B17006.

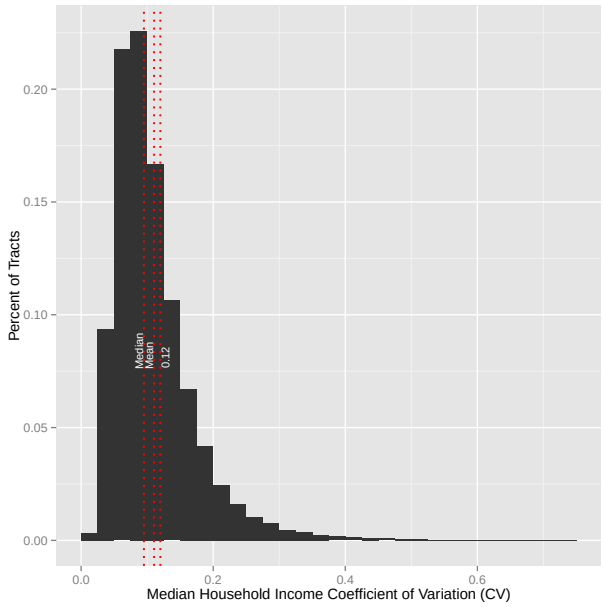
key spatial and demographic determinants of the variation in uncertainty. We follow this analysis with a discussion containing general advice for data users and potential solutions to the challenges identified. A conclusion summarizes the results and suggests directions for future research avenues.

2 Empirical Illustration

The distribution of uncertainty in the ACS is not random—uncertainty is not equally likely in all locations and for all demographic groups. Greater levels of median household income uncertainty exist in the South and Southwest of the United States, near city centers, and in places with lower median incomes (Figures 2, 4, 5 and 6). Interestingly, inner-cities and low income households, the groups and areas typically of most interest to social science researchers, are the specific subsets of the data that tend to be associated with the greatest uncertainty.

Median household income is a key indicator of socio-economic status and its uncertainty is the focus of our analysis. Uncertainty is measured by the coefficient of variation (CV), which is the standard error of an estimate divided by the estimate itself. It can be computed from published ACS data, and provides a relative measure of uncertainty—essentially the error measured as a percent of the estimate. There is no clear cutoff for what qualifies as an "acceptable" CV. A comprehensive report on the ACS (Citro and Kalton, 2007) produced for the National Research Council (NRC) states that the maximum acceptable CV should be in the 0.10–0.12 range, while noting that "what constitutes an acceptable level of precision for a survey estimate depends on the uses to be made of the estimate" (p.67). A white paper produced by the software company ESRI (ESRI, 2011) characterizes a CV below 0.12 as "high reliability", 0.12–0.40 as "medium reliability" and anything above 0.40 as "low reliability." Considering that these data are often used for allocating federal resources or the development of public policy, the more conservative values seem appropriate.

Figure 1 presents the distribution of the CV of median household income for over 70,000 census tracts in the continental U.S. for the 2006–2010 ACS data. The median CV is 0.095, with a slightly higher average of 0.110 due to the long tail in the distribution. Using the



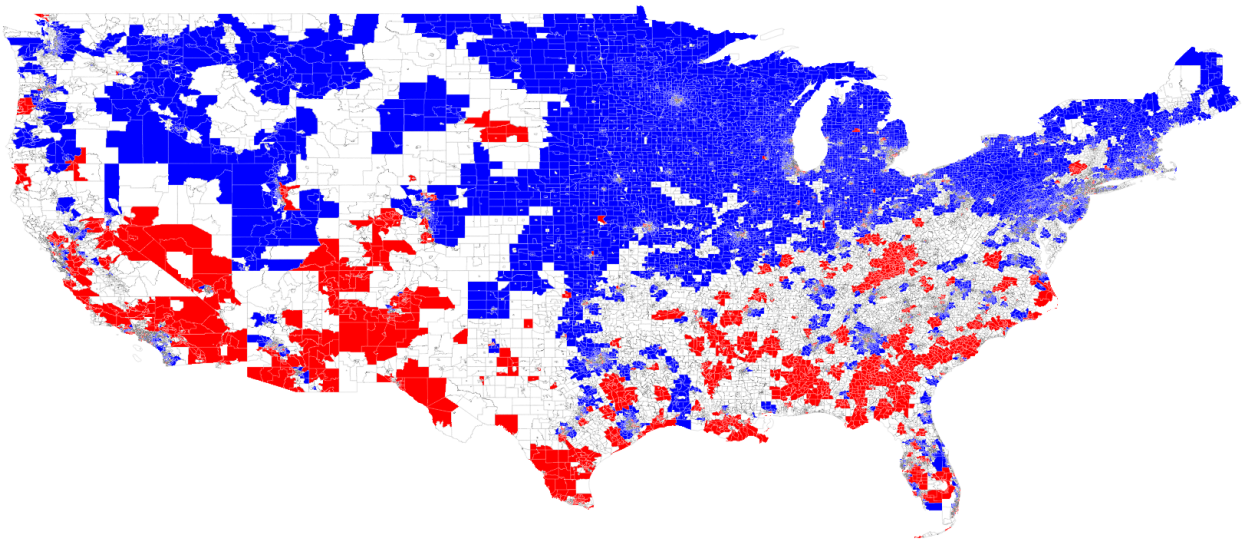
Note: Figure truncated at CV of 0.75.

Figure 1: Coefficient of Variation on Median Household Income, Continental U.S. Census Tracts ACS 2006–2010

high end of the NRC range (0.12), approximately one third (32.1%) of U.S. census tracts have too much uncertainty while two-thirds (67.9 percent) would be considered acceptable.[2]

While the quality of the ACS income estimates is concerning, the spatial patterns in data quality are even more troubling—some areas of the country are more prone to high uncertainty than others. Figure 2 shows local spatial autocorrelation in the quality of income estimates using the Local Moran's I statistic (Anselin, 1995). Concentrations of high quality (low uncertainty) income estimates are found in the north of the country (blue areas on the map), particularly the Midwest, while red clusters of low quality estimates (high uncertainty) are concentrated more in the South and Southwest. Red areas (hotspots) represent statistically significant concentrations of high uncertainty, and blue (coldspots) identifies low uncertainty concentrations. Since census tracts average approximately 4,000 people, at this scale the map shows the uncertainty pattern in low density parts of the country. Figure 3 zooms in further to present the CV distribution for nine metropolitan areas across the country. In Madison, Wisconsin, for example, approximately 75 percent of census tracts are

[2]These estimates are based on census tracts in the continental U.S. with outlier tracts removed (see Section 3.3).



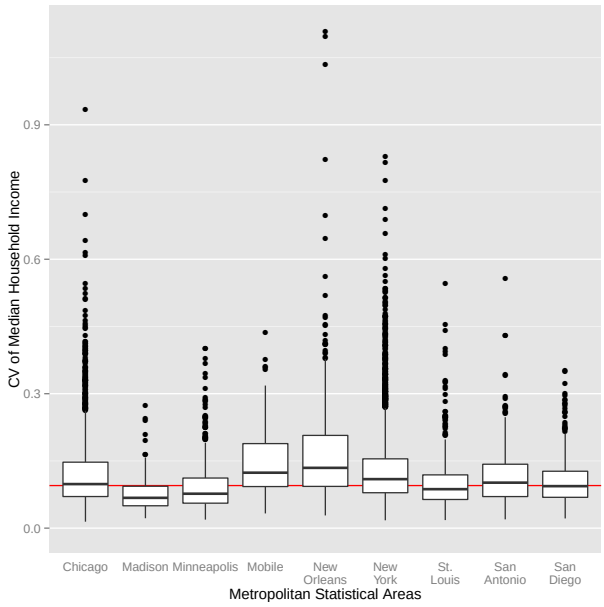
Note: Results based on the local Moran's I statistic (Anselin, 1995). Colored tracts represent statistically significant clusters at the 0.05 level, based on 999 random permutations of the data. Red tracts are clusters of high values, and blue tracts are clusters of low values.

Figure 2: Concentrations of Uncertainty on Median Household Income, Census Tracts ACS 2006–2010

below the national median uncertainty level; in contrast, nearly 75 percent of the New Orleans, Louisiana census tracts are above this level. Other metropolitan areas such as Chicago and San Diego have median levels nearly identical to the national median. This diversity in the magnitude and range of uncertainty across metropolitan areas can affect the reliability of cross sectional analyses.

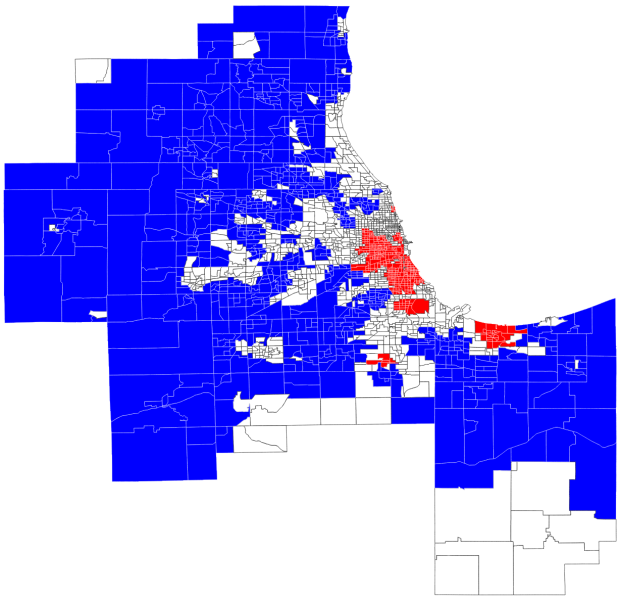
The spatial variation in attribute uncertainty is not simply a macro scale phenomenon that varies from region to region, but also manifests itself within regions. Figure 4 presents uncertainty hotspots and coldspots within the Chicago, Illinois metropolitan area. The map highlights multiple high uncertainty concentrations around central Chicago and one in central Gary, Indiana. The low uncertainty areas are generally located in the exurban periphery of the region, with notable exceptions such as the village of Oak Lawn, Illinois to the southwest of downtown Chicago.

The spatial concentration of uncertainty in Chicago is not confined to that region, but is indicative of a general pattern across U.S. metropolitan areas. To see this we pool all census tracts from the 150 largest metropolitan areas into 100 bins (percentiles) based on



Note: Two outlier census tracts not shown: New York with CV of 2.73 and Chicago with CV of 1.41

Figure 3: Distribution of Uncertainty on Median Household Income, Selected Metropolitan Area Census Tracts ACS 2006–2010



Note: Results based on the local Moran's I statistic (Anselin, 1995). Colored tracts represent statistically significant clusters at the 0.05 level, based on 999 random permutations of the data. Red tracts are clusters of high values, and blue tracts are clusters of low values.

Figure 4: Concentrations of Uncertainty on Median Household Income, Chicago Metropolitan Area Census Tracts ACS 2006–2010

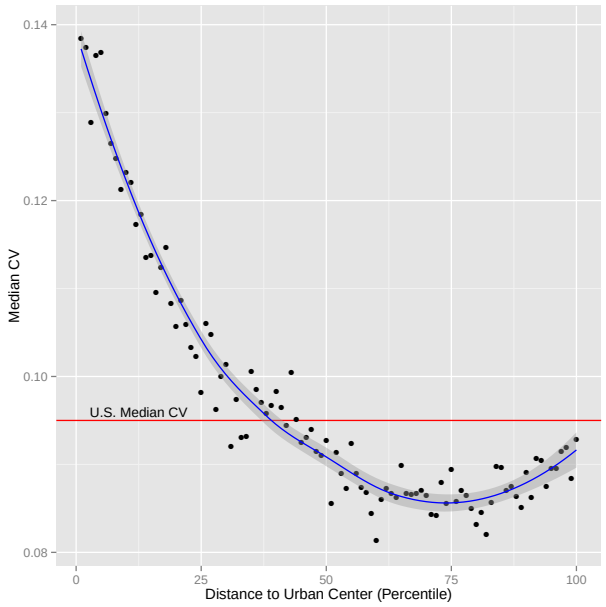


Figure 5: Variation in Uncertainty on Median Household Income Coefficient of Variation Based on Distance from Urban Center, Largest 150 Metropolitan Areas Census Tracts ACS 2006–2010

distance from their respective city centers.[3] Each census tract is assigned to a bin based on its relative distance from the city center. Relative distance is used on the X axis because MSAs vary greatly in size. Figure 5 shows the median CV value from each of these 100 bins. There is a steep decline in uncertainty as distance initially increases from urban cores, which eventually moderates and then begins increasing again when reaching the peripheries of the regions.

In addition to a clear spatial structure, uncertainty of median household income in the U.S. major metropolitan areas also displays a pattern across income levels. Figure 6 is similar in design to Figure 5, except in this case we group census tracts based on 100 income bins for their respective MSAs. This allows us to control for inter-metropolitan variations in income. The results show that uncertainty in median household income declines as median household income increases. The similar patterns in Figures 5 and 6 are likely related, as some MSAs' lower income residents live closer to the urban core. The increasing level of uncertainty at the urban periphery is likely caused by the diversity of exurban locations, which can range from wealthy suburban enclaves to lower income agricultural communities.

[3]See Table 1 for notes on the data source.

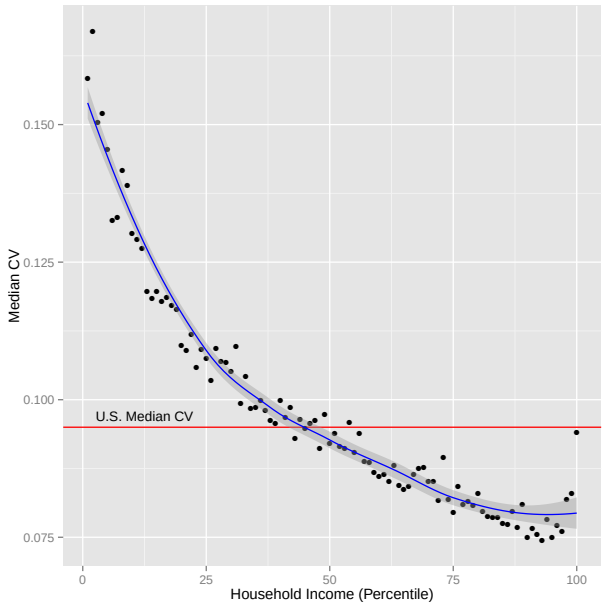


Figure 6: Variation in Uncertainty on Median Household Income Coefficient of Variation Based on Increasing Income, Largest 150 Metropolitan Areas Census Tracts ACS 2006–2010

3 Methodology and Data

The previous section showed that there are clear social and geographic patterns in the quality of ACS median household income estimates. We next adopt a regression framework in an effort to identify the determinants of these patterns. In order to assess this process at different degrees of spatial resolution, the analysis is conducted at three levels: the nation, large-scale regions, and metropolitan areas.

3.1 National Model Specification

We begin with a national equation that allows us to obtain a global picture of the process:

$$\log y_i = \alpha + \delta \log inc_i + \phi \log hu\_respond_i + \beta \log X_i + \gamma \log E_i + \vartheta \log T_i + u_i \qquad (1)$$

where y_i represents the margin of error (provided by the USCB) of the estimate of income, inc_i, in tract i. Uncertainty is explained by ACS response level ($hu\_respond_i$), a set of socio-demographic variables (X_i), a group of characteristics of the housing environment (E_i) and features relating to the structure and definition of the census tract as a statistical entity

(T_i). $(\alpha, \delta, \phi, \beta, \gamma, \vartheta)$ is a vector of parameters and u_i is the error term. Table 1 summarizes the description of the variables and their sources. The equation takes a log-log specification that allows coefficients to be interpreted as elasticities, expressing the percentage change expected on y_i given a one percentage increase in the explanatory variable.

Given the fine-grained scale of census tracts, it is likely that some of the unobserved characteristics captured by the error term are spatially correlated, in which case the estimates loose precision (Anselin, 1988). To account for this form of spatial dependence, we assume a spatial auto-regressive error term of the following structure:

$$u_i = \lambda \sum_j w_{ij} u_j + \epsilon_i \tag{2}$$

where w_{ij} is the ijth element of a spatial weights matrix W that formally represents the spatial connectivity of tracts, and ϵ_i is an i.i.d. and well-behaved disturbance. All the results shown relate to a matrix built using the common queen contiguity criterion, under which two observations are neighbors, and thus assigned a weight of one, if they share a border of any length, including a single point. This matrix is then standardized so every row sums to one, effectively converting $\sum_j w_{ij} u_j$ into the average value of u in the surroundings of i. These models are estimated with a generalized method of moments, following the recent approach proposed by Arraiz et al. (2010), which suggests an estimator robust to spatial autocorrelation and heteroskedasticity.

3.2 Regional and Metropolitan Model Specifications

To explore geographical variation in the determinants of uncertainty we maintain the census tract as the unit of analysis, and apply a spatial regimes approach that determines whether factors associated with uncertainty play a different role in different places. The baseline model of Equation 1 is expanded with spatial regimes yielding the following:

$$\log y_{ir} = \alpha_r + \delta_r \log inc_{ir} + \phi_r \log hu\_respond_{ir} + \beta_r \log X_{ir} + \gamma_r \log E_{ir} + \vartheta_r \log T_{ir} + u_{ir}, \tag{3}$$

Figure 7: U.S. Census Bureau Divisions

where the subscript r indicates membership to a given region; otherwise the specification remains the same. This means we obtain a set of $(\alpha_r, \delta_r, \phi_r, \beta_r, \gamma_r, \vartheta_r)$ parameters by region. Also, we subset the spatial weights matrix W so only neighbors within the same region remain connected. The combination of both is equivalent to running separate regressions for each group of observations under the same r subscript. The advantage of this framework is that it allows for a significance test of the stability of estimates across regions by means of the spatial Chow test (Anselin, 1990).

In order to allow for sufficient variation across space while keeping the number of parameters to interpret tractable, we use census divisions (Figure 7), which partition the U.S. into nine separate regions. This regimes approach helps assess the spatial variation in correlates of uncertainty and, as the discussion below shows, helps shed light on the results of the national model.

In the last step of the analysis, we explore spatial differences in more detail. The census divisions are now replaced by the 150 largest metropolitan statistical areas (MSAs) in the country. MSAs are defined by the U.S. Office of Management and Budget, and approximate a functional urban region. Although MSAs do not cover the entire extent of the country, these 150 MSAs were home to approximately 84 percent of the U.S. population in 2010. We model these using the regime approach presented in Equation 3 where the subscript r now represents an MSA.

3.3 Data

We know that the characteristics of places, the population mix, the built environment, etc., do not vary randomly in space (Jargowsky, 1997; Briggs, 2005). Our concern in this model relates to potential covariation in these attributes and ACS uncertainty. Ideally there would be no correlation, a result indicating that uncertainty at the census tract scale is not a systematic function of the place.

The dependent variable is the MOE on median household income. Due to the complexity of the sampling and weighting processes used to compute ACS estimates, the USCB estimates the MOE using an empirical approach. The successive differences replication approach recomputes all ACS estimates 80 times using different base weights on the completed surveys each time, and then uses all this information to compute the variability in the actual estimate (see U.S. Census Bureau, 2009b, Chapter 12 for more details). To control for the magnitude of the MOE, the income estimate itself is included as an explanatory variable in Equations 1 and 3. These are the only two variables in the model taken from ACS estimates. We specifically do not include other ACS variables in the strategy design because they are expected to contain similar uncertainty problems as what we are trying to model—it is important that the explanatory factors of uncertainty are, as much as possible, unaffected by measurement error. As a result, the pool of potential explanatory variables is heavily constrained, so we turn to the 2010 decennial census and a 2010 restricted-use database from the U.S. Department of Housing and Urban Development (HUD). Since the ACS data are collected over five years, there is some degree of temporal mismatch; however, the extent to which this affects our results is limited since the variables used are rather persistent over time and thus offer a good approximation to establish a long-term relationship, as specified in Equations 1 and 3. The variables used and their sources are summarized in Table 1.

When considering potential determinants of uncertainty in ACS estimates we first consider the total number of surveys collected in the particular tract (`hu_respond`). More responses are expected to reduce margin of error. Beyond this purely sampling aspect of tracts, we consider characteristics of the place along three dimensions: socio-demographics, housing and residential environment and tract structure and diversity. The first dimension

considers the residents and household structure of the place, along with a proxy for low income in the form of the number of federally subsidized housing units (`hud_total`).[4] The second dimension captures variables on the types of housing units in the place (rental units, vacant units and group quarters population) and also a measure of how urban the place is (`urban_hsu`). The final dimension investigates the census tract as a unit of analysis by looking at variation in land area, whether the tract population was stable over the previous decade (proxied by tracts that did not change shape) and three variables measuring diversity. Diversity is measured using the Simpson index (also known as the Herfindahl index), which takes higher values if the population is spread more evenly across groups and lower ones if the population is more concentrated in a single group. We measure diversity in resident age, race of residents and household size, with the expectation that more diverse places will have higher uncertainty since the population is not of uniform type.

The unit of analysis is always the census tract. Although there are high level parameters that constrain tract delineation (U.S. Census Bureau, 1994), exceptions do occur. For this reason we only include census tracts with a population greater than 500 and with more than 200 housing units. In general this excludes areas such as national parks, lakes, prisons, large college dormitories, etc. We also exclude tracts that have a household income estimate, but no associated MOE; these generally occur when the median household income estimate is at the reporting bounds of $2,499 or $250,001. Combined, these steps remove 1,186 tracts. The geography is further constrained to the lower continental U.S., leaving 71,353 census tracts for the analysis.

4 Findings

The results from the national and regional regressions are presented in Table 2 along with the Chow test on each variable. To help interpret the large number of coefficients for the MSA regressions, we summarize the results in Figures 8 and 9. The former presents the count of MSAs in which a particular variable is significant, and whether it has a positive or negative

[4]Federally subsidized housing programs include public housing (traditional and HOPE VI), multi-family housing (including housing for the elderly and disabled, Section 202 and 811), and vouchers (predominantly Housing Choice Vouchers for tenants). Address-level records were aggregated to the tract level.

Variables	Description	Source
`inc`	Median household income estimate	2006–10 ACS
`hu_respond`	Housing units responding to ACS	2006–10 ACS
Socio-demographics (X)		
`hud_total`	Federally subsidized housing units	HUD
`child_fam`	Families with children	2010 Census
`black`	African-American population	2010 Census
`hisp`	Hispanic population	2010 Census
Housing and Residential Structure (E)		
`group_pop`	Population in group quarters	2010 Census
`urban_hsu`	Urban housing units	2010 Census
`vacant`	Vacant housing units	2010 Census
`renter`	Rental housing units	2010 Census
`hsu`	Total housing units	2010 Census
Tract Structure and Diversity (T)		
`area`	Land area	2010 Census
`tr_nochange`	2000–10 tract boundary stability (dummy)	2000–10 Census†
`hhSize_simp`	Household size diversity (Simpson index)	2010 Census
`age_simp`	Resident age diversity (Simpson index)	2010 Census
`race_simp`	Racial/ethnic diversity (Simpson index)	2010 Census
`dist2center`*	Distance to urban core	2010 Census‡
`population`	Total residents	2010 Census
Dependent Variable		
Income Error	Median household income margin of error	2006–10 ACS

All variables measured in levels unless noted in the table. Natural logarithms are taken of each variable before use in the econometric model.

* Only included in metropolitan area regressions

† Computed by authors based on physical census tract boundary changes between 2000 and 2010 reported in the 2010 Census Tract Relationship Files.

‡ Urban centers identified using the U.S. Geological Survey's Geographic Names Information System. The latitude-longitude marker for the first city listed in the MSA name is extracted from the database, and then distances are computed from each census tract centroid to the urban center.

Table 1: Description of Dependent and Independent Variables

effect on MOE; while the latter shows the spatial distribution of the significant variables along with the direction and magnitude of the effect on MOE. The models perform reasonably well: the pseudo-R^2 index[5] for the national regression is 0.45, with the corresponding values for the regional regressions ranging from 0.38 to 0.51, and for the MSA models they range from 0.21 to 0.73. Some broad trends emerge from the results, which we present below.

The key variables associated with MOE are income (`inc`) and response level (`hu_respond`). These are the only two variables that are significant in the national model and all nine census divisions (called "regions" going forward). These two variables are also significant in 68 percent and 91 percent of the MSA regressions respectively. Furthermore, when these two variables are significant, their directional impact on the margin of error is consistent: negative impact for response level and positive for income. At the same time, there is wide variation in the significance level and direction for the other variables when comparing across regions and MSAs. For the national model, a 1 percent increase in `inc` is associated with a 0.8551 (Table 2) increase in MOE. This indicates that the magnitude of the error on income increases at a slower rate than income itself, which leads to lower relative uncertainty (measured by the coefficient of variation) in higher income places. For the country as a whole, a one percent increase in the number of responding housing units results in a half percent reduction (-0.5091) in MOE, which reinforces the premise that more raw data from which to build the estimates results in lower uncertainty in those estimates. When estimating the national model with these two variables only, we find the coefficients on `inc` and `hu_respond` to be 0.7096 and -0.4715 respectively. The removal of the covariates in X, E and T from Equation 1 has little influence on response level, but without these variables the influence of `inc` on MOE is much stronger, i.e. further away from 1.

Higher levels of renters (`renter`), subsidized housing (`hud_total`), vacant units (`vacant`) and group quarters population (`group_pop`), which includes prisons, college dormitories and military barracks, tend to be associated with higher MOE. None of these are significant in all regions and MSAs but, when they are, the direction of influence tends to be positive. A notable exception is that `renter` has a negative association in four California MSAs. The significant Chow tests on most of these variables indicate that while the impacts are similar,

[5]This is the squared correlation between the actual and predicted dependent variable.

	U.S.	Pacific	Mountain	West North Central	West South Central	East North Central	East South Central	New England	Middle Atlantic	South Atlantic	Chow Test
constant	2.0645***	1.9615***	2.8648***	2.0508***	1.8556***	2.7187***	2.0106***	1.6093***	2.1303***	2.1593***	
inc	0.8551***	0.8809***	0.8424***	0.8185***	0.8499***	0.8534***	0.8131***	0.8798***	0.8520***	0.8518***	
hu_respond	-0.5091***	-0.4794***	-0.4778***	-0.5021***	-0.4342***	-0.4620***	-0.4272***	-0.4689***	-0.4958***	-0.3963***	***
hud_total	0.0132***	0.0100**	0.0041	0.0048	0.0123***	0.0176***	0.0230***	0.0293***	0.0207***	0.0156***	**
child_fam	-0.1082***	-0.1278***	-0.0148	-0.1330***	-0.2112***	-0.0307	-0.2500***	-0.1339**	-0.0424	-0.1149***	***
black	0.0001	-0.0093*	-0.0256***	0.0088	-0.0155***	0.0077**	-0.0119**	-0.0275**	-0.0104**	-0.0153***	***
hisp	-0.0036	0.0128	-0.0032	-0.0344***	-0.0202***	-0.0102*	-0.0333***	0.0154	-0.0008	-0.0072	***
group_pop	0.0083***	0.0092***	0.0097***	0.0029	0.0048*	0.0068***	0.0070**	0.0147***	0.0170***	0.0073***	***
urban_hsu	-0.0069***	-0.0015	0.0023	0.0014	0.0033	-0.0002	0.0003	0.0020	-0.0090***	-0.0074***	**
vacant	0.0585***	0.0378***	0.0663***	0.0757***	0.0921***	0.0826***	-0.0126	0.0551***	0.0736***	0.0727***	***
renter	0.0367***	0.0136	0.0582***	0.0513***	0.0192	0.0342***	0.0286	0.0099	0.0104	0.0429***	*
hsu	0.1466***	0.2378***	0.0890	0.0412	0.0353	-0.0099	0.3215***	0.2772***	0.1390***	-0.0046	***
area	0.0043***	0.0072*	0.0174***	0.0119**	0.0122**	-0.0059	0.0361***	-0.0138	-0.0320***	0.0138***	***
tr_nochange	0.0160***	0.0147	0.0282*	-0.0065	0.0124	-0.0078	-0.0161	0.0195	-0.0324**	-0.0039	*
hhSize_simp	0.2202***	0.2192***	0.0917	0.1437*	0.2520***	0.1159**	0.3864***	0.3905***	0.1906***	0.2234***	*
age_simp	-0.0018	-0.0527	-0.0271	0.0549	0.0527	-0.2220***	0.0540	-0.2331**	-0.1413**	0.0907*	***
race_simp	0.0150	0.0042	0.0697*	0.0720	-0.0203	-0.0023	0.1210***	0.1013*	0.0145	-0.0015	**
population	-0.1421***	-0.2187***	-0.2572***	0.0038	0.0231	-0.1338**	-0.1569**	-0.2307***	-0.1477***	-0.0847**	**
λ	0.1473***	0.1240***	0.0918***	0.0317	0.1207***	0.0994***	0.0389	0.0652**	0.1009***	0.1119***	***
Pseudo R^2	0.4546	0.4502	0.4404	0.4101	0.4185	0.4290	0.3843	0.4445	0.5078	0.4351	
N	71353	10193	5182	5244	8012	11644	4399	3326	9922	13431	

Notes: Dependent variable: median household income margin of error
All variables in logarithms
Significance: *0.10, **0.05, ***0.01
Pseudo R^2 is the squared correlation between the actual and predicted dependent variable

Table 2: National and Regional Regression Results

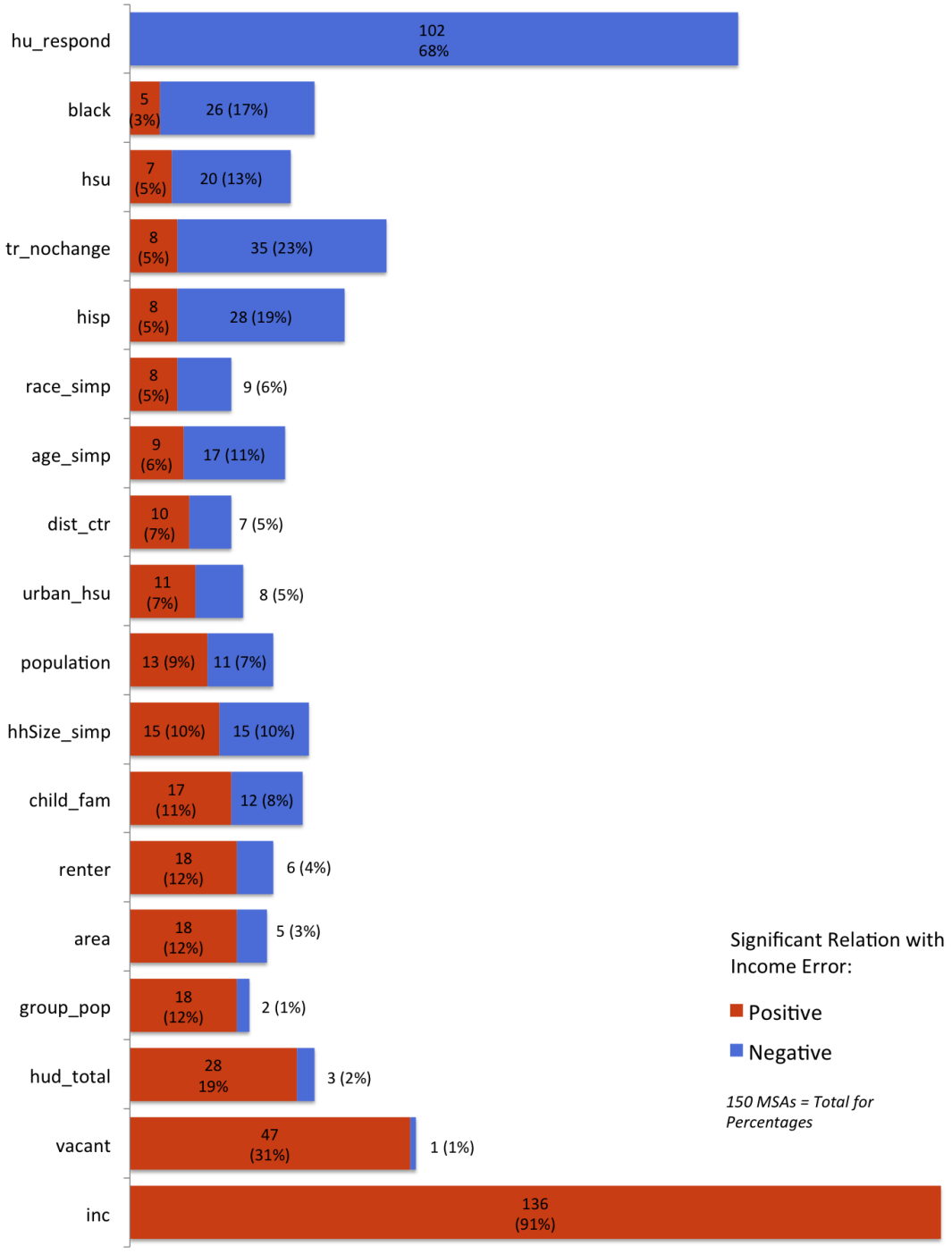


Figure 8: Counts of Significant Regression Coefficients, Metropolitan Areas

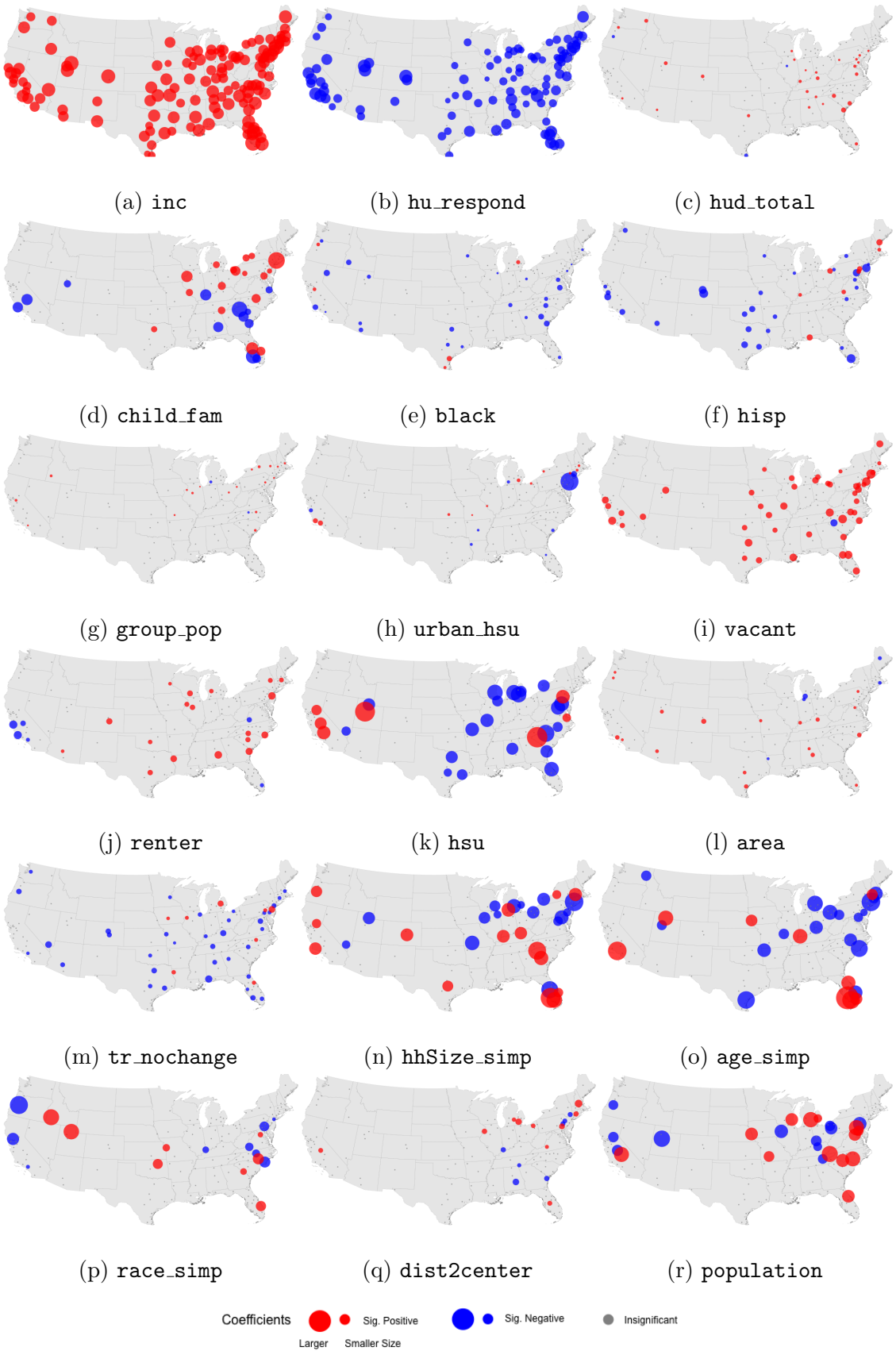


Figure 9: Magnitude and Spatial Distribution of Significant Regression Coefficients, Metropolitian Areas

the coefficients are statistically different across regions. These four variables are indicators of population churn, and not necessarily growth or decline of areas. Population stability from 2000 to 2010, as proxied by the dummy variable `tr_nochange`, captures aggregate change, and presents a less clear picture. In the national model population stability leads to more error, it is largely insignificant at the regional level, but tends to point to less error at the MSA level.

The race and ethnicity variables are insignificant in the national model, but are significant in many of the regions and MSAs. The magnitude of the African American population (`black`) is significant (0.05 level) in seven regions and 31 MSAs, while the Hispanic population (`hisp`) is significant in only three regions but 36 MSAs. Generally, increases in these populations correspond to decreases in MOE, but exceptions exist. For example, all MSAs where `hisp` is associated with increasing MOE are in the eastern half of the country. A third race variable, `race_simp`, captures diversity within the tract, but is only significant in the East South Central region and 17 MSAs. Of note is that when it is significant the magnitude of the coefficient tends to be relatively large. Overall, these variables tend to point to less error in places with increased representation of one ethnic group.

Two variables look at urban character of tracts: `area` and `urban_hsu`. Greater land area is typically associated with more error. That being said, `area` is either not significant or negative in the densest regions, e.g. East North Central, New England and Middle Atlantic, and all but one MSA with a significant negative coefficient are located in these parts of the country. The impact of more urban housing units, as defined by the USCB, has a less clear impact on MOE. It is negatively correlated in the national model and for the Middle Atlantic and South Atlantic, and split nearly evenly for the 19 MSAs where it is significant. These results might be related to a greater ease of identifying and collecting surveys in denser locations.

Beyond these nationally consistent trends, there is considerable variation in the magnitude and significance of the remaining coefficients across regions and MSAs. Even when controlling for factors expected to vary from region to region, we still find that one model does not provide the same explanations for the MOE pattern in all regions. A global Chow test shows that we can reject the null hypothesis that all coefficients are the same across

regions. No two regions have the same set of significant explanatory variables, and only income (`inc`) and response rate (`hu_respond`) are significant in all nine regions and the U.S.

5 Implications

The results from the previous sections show that while the uncertainty in the ACS is not random, the form of its structure varies from region to region and MSA to MSA. The key implication therefore is that any broad "solution" a user of ACS data might consider must be flexible enough to accommodate the idiosyncratic nature of the error itself.

We can split the use of ACS data into two categories: 1) direct reporting of the data and 2) statistical reporting of the data. Direct reporting is the presentation of raw ACS estimates in tables, charts and maps. For example, a table comparing the number unemployed residents for selected census tracts or a map of unemployment levels by census tract for a city. The challenge of integrating the uncertainty then becomes a communication issue—the goal is to add value to the estimates through clear communication of the uncertainty. The most straightforward examples of this are adding a column of MOEs to a table of estimates or including a second map that shows the MOEs. However, these approaches may not be the most effective means of communicating uncertainty information to all audiences, especially those unfamiliar with the interpretation of survey data. It is possible that a more coarse representation of the uncertainty could be used, such as a red-yellow-green light icon attached to each estimate in a table indicating how much caution should be used when interpreting the value. In a mapping context, interactions of color intensity or hatching overlays could be used to create a single map that integrates the estimates and MOEs (see for example Sun and Wong, 2010).

A subcategory of direct reporting is the computation of ratios, proportions and sums of ACS data. The USCB provides equations for the computation of MOEs on these user generated estimates in the appendix of U.S. Census Bureau (2009a). Once these are computed the challenge then reverts back to modes of communication of the uncertainty. In all cases, if the uncertainties for some or all of the estimates being reported are high, some form of communication of this information should be included. Striking a balance between too much

and not enough information is a challenge that continues to be studied (see for example Wong and Sun, 2013).

Correlations, regression coefficients and other forms of model output fall into the second category: statistical reporting of ACS data. These types of measures are fraught with hidden statistical issues when the input data are measured with error, as is the case in the context of ACS data. The primary issue is attenuation bias, which causes the magnitude of a statistic (e.g. correlation or regression coefficient) to be reduced when one or more variables are measured with error. Corrections for the correlation coefficient have existed for over a century (Spearman, 1904), but not without controversy (Muchinsky, 1996). In the regression context, the issue is more complex. Standard econometric theory assumes that explanatory variables are deterministic and measured without error. The presence of a variable with error in the design matrix has the potential to taint all the regression coefficients in unpredictable ways (Greene, 2003, chapter 5). The main problem then lies in the interpretation of regression coefficients, which are likely to be biased if the ACS error is large. Various errors-in-variables and instrumental variables types approaches have been considered as a potential fix for this problem; the spatial structure embedded in the ACS is another option to consider in order to ameliorate the data challenges (see Anselin and Lozano, 2008 for a discussion on the topic). One strength of ACS data, as compared to other cases of error in measurement, is that we know the magnitude of the error on each estimate, and can leverage this information in the model specification. Future research along these lines could provide interesting insights for more robust inferences using ACS data. When it comes to the dependent variable, consequences are less severe since the measurement mismatch is transferred to the error component of the regression specification, and this can then be appropriately modelled to account for its structure.

The USCB identifies two user level approaches for reducing the uncertainty in ACS estimates: combine attributes or combine geographic areas (U.S. Census Bureau, 2009a). Combining attributes can be accomplished by collapsing finer grain measures (say racial/ethnic unemployment levels) into aggregate categories (such as "white" and "non-white"). If the substantive research project does not allow for attribute collapsing, then geographic areas can be combined into "regions." Both of these approaches will almost always result in a

reduction in the CV over the original data, but they also reduce the amount of information available for analysis. This trade-off is readily apparent when combining attributes and straightforward to implement in spreadsheet software; in contrast, combining geographic areas does not offer this same ease. As the number of areas grows, the possible ways to combine contiguous observations into regions grows at an increasing rate meaning that, for all but trivially small problems, it is not possible to examine all possible combinations of areas to find the "best" solution. Folch and Spielman (2014) propose a multivariate and multiple criteria regionalization algorithm based on the max-p approach (Duque et al., 2012) that can be used to group areas together such that the estimates on each region meet some predetermined CV threshold and also minimize information loss due to the grouping of geographic areas.

6 Conclusion

By now the higher levels of uncertainty associated with ACS estimates compared to the decennial census are widely recognized. However, when we think of measurement error we generally hope that it is of the well behaved type that follows a random pattern, equally likely in all locations. As this work has shown, such a perspective should not be taken when using ACS data. Uncertainty is not only clustered over space, but the characteristics of places with high margins of error vary from region to region.

The results in Section 2 suggest that analyses of urban core areas involving more diverse households with lower incomes are especially vulnerable to data quality problems—this impacts a large proportion of sociological, urban and policy research. In contrast, more affluent and homogeneous suburban areas have better income estimates but less need for improvements in public services. By means of regression, we find that the typical sampling rules hold nationally: higher response rates are associated with less error. What was less expected is that the socio-demographic and built environment characteristics of places are also associated with the precision with which median household income is measured. Furthermore, the particular correlates of MOE vary regionally and by metropolitan area, which precludes meaningful national summaries and recommendations because the kind of error analysts will

encounter, and its drivers, vary by study area.

On the other hand, a position that ACS income estimates generally have too much error at the tract level to be useful in research and decision making at all is also not warranted: if one's area of interest is one with lower-than-average margins of error (such as some areas in the North), it might have enough accuracy to be included in an analysis. But to determine this requires a review of the error associated with the estimates in an area of interest rather than assuming *a priori* that errors will be too high or low at the tract level.

Users of ACS data are thus left to identify workarounds to these data quality issues. For higher-income areas, commercial vendors (such as InfoUSA) offer good quality data on areas with strong consumer expenditures, but these are the same areas with relatively good ACS data. Regionalization solutions that can reduce uncertainty, as discussed in the previous section, will likely result in larger regions in the urban core than in more affluent suburbs, precluding small-scale urban analysis. These gaps are key because they disproportionately affect the most vulnerable neighborhoods, and reduce our ability to study rising inequalities between low and high income areas.

Acknowledgements

David C. Folch and Seth E. Spielman acknowledge financial support from the National Science Foundation (award number 1132008). Julia Koschinsky acknowledges financial support from the U.S. Department of Housing and Urban Development. The authors are solely responsible for the accuracy of the statements and interpretations contained in this publication. Such interpretations do not necessarily reflect the views of the Government. This work used the Python Spatial Analysis Library (Rey and Anselin, 2007, www.pysal.org).

References

Anselin, L. (1988). *Spatial econometrics*. Kluwer Academic Publishers Dordrecht.

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, 30(2):185–207.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2):93–115.

Anselin, L. and Lozano, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, 34(5):5–34.

Arraiz, I., Drukker, D., Kelejian, H., and Prucha, I. (2010). A spatial Cliff-Ord-type model with heteroskedastic innovations: small and large sample results. *Journal of Regional Science*, 50(2):592–614.

Bazuin, J. T. and Fraser, J. C. (2013). How the ACS gets it wrong: The story of the American Community Survey and a small, inner city neighborhood. *Applied Geography*, 45(December):292–302.

Briggs, X. d. S., editor (2005). *The Geography of Opportunity: Race and Housing Choice in Metropolitan America*. Brookings Institution Press, Washington, D.C.

Citro, C. F. and Kalton, G. (2007). *Using the American Community Survey: benefits and challenges*. National Academies Press, Washington, D.C.

Duque, J. C., Anselin, L., and Rey, S. J. (2012). The max-p-regions problem. *Journal of Regional Science*, 52(3):397–419.

ESRI (2011). The American Community Survey. Technical report, ESRI.

Folch, D. C. and Spielman, S. E. (2014). Identifying regions based on flexible user-defined constraints. *International Journal of Geographical Information Science*, 28(1):164–184.

Greene, W. (2003). *Econometric analysis*. Prentice Hall Upper Saddle River, NJ.

Jargowsky, P. (1997). *Poverty and Place: Ghettos, Barrios, and the American City*. Russell Sage Foundation, New York, NY.

MacDonald, H. (2006). The American Community Survey: Warmer (more current), but fuzzier (less precise) than the decennial census. *Journal of the American Planning Association*, 72(4):491–503.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56(1):63–75.

Rey, S. J. and Anselin, L. (2007). PySAL: A python library of spatial analytical methods. *The Review of Regional Studies*, 37(1):5–27.

Salvo, J. J. and Lobo, A. P. (2006). Moving from a decennial census to a continuous measurement survey: factors affecting nonresponse at the neighborhood level. *Population Research and Policy Review*, 25(3):225–241.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Spielman, S. E., Folch, D. C., and Nagle, N. N. (2014). Patterns and causes of uncertainty in the American Community Survey. *Applied Geography*, 46:147–157.

Sun, M. and Wong, D. W. S. (2010). Incorporating data quality information in mapping American Community Survey data. *Cartography and Geographic Information Science*, 37(4):285–299.

U.S. Census Bureau (1994). Geographic areas reference manual. Technical report, U.S. Census Bureau.

U.S. Census Bureau (2009a). *A Compass for Understanding and Using American Community Survey Data: What Researchers Need to Know*. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2009b). *Design and Methodology: American Community Survey*. U.S. Government Printing Office, Washington, D.C.

Wong, D. W. and Sun, M. (2013). Handling data quality information of survey data in GIS: A case of using the American Community Survey data. *Spatial Demography*, 1(1):3–16.