Validating Small Area Population Estimates Using Historical Census Data

Matt Ruther, University of Colorado at Boulder
Galen Maclaurin, University of Colorado at Boulder
Stefan Leyk, University of Colorado at Boulder
Barbara Buttenfield, University of Colorado at Boulder
Nicholas Nagle, University of Tennessee

Corresponding Author:

Matt Ruther
Department of Geography
University of Colorado at Boulder
110 Guggenheim UCB 260
Boulder, CO 80309-0260
matthew.ruther@colorado.edu
Cell: 513-262-2948
Fax: 303-492-7501

March 1, 2013

Acknowledgment

This research is funded by the National Science Foundation: "Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata", project BCS-0961598 awarded to University of Colorado at Boulder and University of Tennessee at Knoxville. Funding by NSF is gratefully acknowledged.

Abstract

This paper is a validation of a methodology which spatially allocates Census microdata to census tracts, based on known, aggregate tract population distributions. To protect confidentiality, public-use microdata contain no spatial identifiers other than the Public Use Microdata Area (PUMA) in which the individual or household is located. Confirmatory information including the location of microdata households can only be obtained in a Census Research Data Center (CRDC). Due to restrictions in place at CRDCs, a procedure for validating the spatial allocation methodology needs to be implemented prior to accessing CRDC data. This study uses historical census data for which a 100% count of the full population is available at a fine spatial resolution to develop such a procedure and to gain knowledge of the behavior of the model under different specifications. The spatial allocation is performed using a microdata sample of records drawn from the full 1880 Census enumeration and synthetic summary files created from the same source. The results of the allocation are then validated against the actual values from the 1880 100% count. The results indicate that the validation procedure provides useful statistics, allowing an in-depth evaluation of the household allocation and identifying optimal configurations for model parameterization.

Introduction

Census public-use microdata possess an attribute richness which should make them tremendously useful to researchers interested in demographic small area estimation. However, they are underutilized, largely due to their coarse spatial resolution. The smallest identifiable geographic area contains a minimum of 100,000 individuals, which may require significant compromise as to the geographic nature of a demographic study. Research which focuses on smaller geographic areas mostly relies on a limited number of aggregate population characteristics provided by the Census Bureau in summary tables and cross-tabulations at the census tract or block group level. In order to better exploit the attribute richness of Census microdata at finer spatial scales, spatial allocation methods, which allocate microdata households to small areas and generate summary statistics for these smaller geographic units using the attributes of the allocated microdata households, may be used (Johnston and Pattie 1993; Ballas et al. 2005; Assunçao et al. 2005). Small area estimates, which contain extensive detail on the underlying population, are in great demand, and are important to research on demographic and social processes, such as migration, impoverishment, and human-environmental interactions.

A persistent shortcoming in the use of such allocation methods for deriving demographic small area estimates is the lack of confirmatory validation. There are often few, if any, sources against which to compare the estimated fine-scale population counts and the associated distributions of population characteristics. The main reason for this absence of fine-resolution comparison data is the confidentiality protection in census surveys which precludes the release of confirmatory data. Therefore an urgent need exists to derive measures for an objective validation of the accuracy and precision of fine-resolution population estimates following a spatial allocation. While demographic estimates based on U.S. Census data and geographies

may be validated at a Census Research Data Center (CRDC), the expense of accessing a CRDC and the necessary confidentiality restrictions in place at the CRDC mandate that the validation process, which is not trivial, be fully realized prior to its implementation at the CRDC.

This article describes a first validation procedure for demographic small area estimates derived from spatially allocated household microdata. The estimation methodology used here has been originally developed to spatially allocate Public Use Microdata Sample (PUMS) households to census tracts, using maximum entropy methods that are constrained by known, aggregate tract population distributions (summary statistics) (Leyk et al. 2013). Fine-scale validation data for contemporary Censuses are available only at a CRDC. In contrast, historical Census data from 1880 are publicly available, and these data contain the full demographic detail of a 100% count of the population. This historical data is used to (1) generate a nested data structure comparable to contemporary census data (i.e., a 5% microdata sample and small area population summary statistics), (2) run the allocation model, and (3) examine and validate model performance.

In the context of methodological validation, the 1880 Census presents a unique opportunity, as the publicly available data include the full count of the population at a fine spatial resolution. The spatial structure of the 1880 Census data is comparable, although not identical, to that of contemporary censuses, and the collected population characteristics are similar. Thus the performance of spatial microdata allocation procedures can be objectively evaluated and interpreted to better understand the quality of finer resolution demographic estimates and how they reflect underlying population characteristics when the model parameters are changed. In order to mimic the data available in contemporary censuses, a random 5% sample of population is drawn from the full 1880 Census enumeration (comparable to current PUMS data) and

"synthetic" summary tables are created from the same source (comparable to SF3 files). The spatial allocation procedure will be performed on these historical data using different combinations of constraining variables, and the results will be validated against the actual values from the 100% population count.

The primary purpose of this article is to evaluate the performance of a spatial allocation methodology which generates small area estimates, through comparisons of these estimates with actual population counts and investigation of model residuals and their geographic variation. Further, the paper will shed light on the evaluation process itself, highlighting important considerations in parameter selection and their influence on resulting estimates for different population attributes. These considerations are crucial in designing a robust validation process prior to undertaking the validation of the allocation results using contemporary census data at a CRDC. Data from the 1880 Census are utilized here as an easily accessible and appropriate surrogate for contemporary census data; as such, the priority in this analysis is neither in historical interpretations of these allocation results nor in drawing substantive conclusions regarding demographic processes in 1880. This article focuses on confirmatory testing that can be directly reproduced using contemporary public-domain Census data, as well as confidential data in a CRDC, and provides preliminary validation measures for spatial allocation methods.

Background

Small area estimation using Census microdata

Matching the distribution of spatially allocated survey data to known census population distributions has been widely employed in small area estimation in the geographical and other social sciences, using a variety of reweighting algorithms or other allocation techniques. To

date, much of this research has occurred in the United Kingdom (Johnston and Pattie 1993; Williamson, Birkin, and Rees 1998; Ballas et al. 2005; Smith, Clarke, and Harland 2009) and Australia (Melhuish, Blake, and Day 2002; Tanton et al. 2011). Of particular relevance to the current study is recent work that focuses on the definition of appropriate goodness-of-fit measures to assess the accuracy of synthetic or reweighted microdata (Williamson, Birkin, and Rees 1998; Voas and Williamson 2001). A general shortcoming in validating the performance of such models is the lack of a "true" population against which the allocation results can be compared. Beckman, Baggerly, and McKay (1996) apply an iterative proportional fitting (IPF) technique to 1990 U.S. Census data and demonstrate that estimated tract-level household distributions are concordant with tract-level summary statistics released by the Census Bureau. However, it is important to note that they validate their estimates against a different sample drawn from the same population, not against the 100% population count. Melhuish, Blake, and Day (2002) use a reweighting process that allocates Australian household survey data which lack locative information to small census districts based on the known socio-demographic profiles of these small geographies. Their evaluation of the results suggests that the allocated populations correctly match the 100% population counts for most districts, but data to evaluate the joint distributions for most population characteristics are not publicly available. Hermes and Poulsen (2012) provide a current and general overview of the use of microdata reweighting techniques in generating small area estimates.

Maximum entropy microdata allocation

A methodology to allocate demographic microdata to small enumeration areas such as census tracts using decennial U.S. Census data has been recently described (Nagle et al. 2012;

Leyk et al. 2013). In this approach, maximum entropy methods impute a set of tract-specific sampling weights for each microdata record. These weights are constrained to match the known (i.e., publicly available) tract-level distributions for a number of population characteristics. The weights imputation is thus guided and influenced by this chosen set of constraining variables. The sampling weights for each microdata household sum across all tracts to the design (or household) weight provided by the Census Bureau. As the design weight reflects the expected number of households in the Public Use Microdata Area (PUMA) that are similar to a given microdata record, each constructed sampling weight can be interpreted as the number of households of this "type" that can be expected in the respective census tract.

The maximum entropy imputation of sampling weights proceeds analogous to iterative proportional fitting, but uses nonlinear optimization to improve computational efficiency (Malouf 2002). Given a set of N microdata household attribute values X_i and a set of probabilities p_{ij} that a household randomly selected from the population has attributes similar to those of PUMS household i and is located in census tract j, it is possible to impute the k-th tract-level attribute value by the equation $\sum_{ij} N p_{ij} X_{jk}$. At the outset of modeling, the probabilities p_{ij} are unknown. The imputation is constrained such that the imputed probabilities reproduce tract-level populations given in Census Summary Files (SF3). This constraint is satisfied using the equation:

$$\max \sum_i \sum_j \left(N \cdot p_{ij}\right) \log\left(N \cdot \frac{p_{ij}}{d_{ij}}\right) \quad \textbf{(1)}$$

$$\text{subject to} \quad \sum_i N \cdot p_{ij} \cdot X_{jk} = y_{jk} \quad \text{and} \quad \sum_{ij} p_{ij} = 1$$

for all households *i*, tracts *j*, and attributes *k*. The y_{jk} are tract-level summaries of attribute *k* from the Summary Files (SF3), and d_{ij} are prior estimates of tract-level (i.e., for each tract *j*) weights for each PUMS record *i*, which are subject to reweighting (Leyk et al. 2013).

Once allocated, the microdata household characteristics can be summarized to (1) create revised estimates of tract-level demographic summary statistics, (2) generate summary statistics of attributes not available in Summary Files, and (3) compute new cross-tabulations. In Leyk et al. (2013) the revised summary statistics were compared to original tract population distributions from the Census-produced SF3 files (based on a 1-in-6 sample), and allocation ambiguity was evaluated for each household as a function of the distribution of imputed sampling weights over all census tracts. While correlations between the revised tract-level summaries and original tract summary statistics were found to be high and statistically significant for constraining and non-constraining variables, a full validation could not be conducted without access to the full population details stored and maintained at a CRDC.

In this paper, the same weights imputation technique will be applied to a sample of households from the 100% count of the 1880 Census. These households will be allocated to enumeration districts according to their exact imputed sampling weights. From these allocations, revised summary statistics are computed for each enumeration district. These revised tables are then compared against the true aggregated population attributes from the full (100%) population count.

The context of the 1880 Census

The 1880 Census is considered the first high-quality enumeration of the U.S. population and full individual records from this historical census have been digitally transcribed and made available online (Goeken et al. 2003; Ruggles et al. 2010). Important for the research reported here, the 1880 Census records contain household microdata including spatial identifiers for the geographic units – enumeration districts – in which the households were located as well as the spatial boundaries of these districts. Although neither PUMAs nor census tracts were yet defined in 1880, State Economic Areas (SEAs) and enumeration districts (EDs) represent a similar spatial data structure as can be found in contemporary censuses. SEAs, which consist of single counties or groups of contiguous counties, were defined for the 1950 Census and retroactively applied to prior censuses by the Minnesota Population Center (Bogue 1951; Ruggles et al. 2010). SEAs were designed to have a minimum population of 100,000 people, much like contemporary PUMAs, although the retrospective definition of SEAs to the 1880 Census may result in substantially different population sizes. SEAs were divided into minor subdivisions known as EDs, similar to contemporary census tracts but slightly smaller; these districts corresponded to the area that a door-to-door enumerator could cover during the Census period. EDs are fully nested in and completely enclosed by SEAs. The similarity between SEAs and PUMAs, and enumeration districts and census tracts, allows the 1880 Census to serve as a reasonable substitute for more current censuses to run and validate the allocation methods.

Although the questions on the 1880 Census covered a wide array of social and demographic characteristics, there are differences in attribute coverage in the 1880 Census relative to recent censuses. Notably, the 1880 Census carried no questions regarding income or housing tenure, and the results from the tendered questions on educational attainment and literacy were not digitally transcribed. This lack of direct measures of socioeconomic status may

require the use of less distinct related data, such as occupational class or standing, in the construction of constraining variables to guide the imputation of sampling weights. The purpose of the constraining variables and the procedure used to select them are described in the Methods section below.

Because the number of observed attributes found in each individual record is quite large, the validation and discussion of the spatial allocation results will focus on selected benchmark variables that are commonly used by and likely of particular interest to demographic researchers. These benchmark variables include the gender, age, race, and marital status of the householder; the full list of benchmark variables and their categorizations are shown in Table 1. These benchmark variables and the categorizations used in this study are believed to be fairly representative of the full range of population characteristics available in this census. To clarify, while the benchmark variables include some variables that will be used as constraining variables in the allocation procedure, the function of the remaining benchmark variables is to serve as validation instruments.

Methods

The 1880 Census data were used to run the weights imputation and spatial allocation model for Hamilton County, Ohio. This county was chosen based on its stable boundaries over time and the fact that it was coextensive with a single SEA (SEA 336). Although the 1880 Census did not define households in the same way as is done in contemporary Censuses, variables describing household composition were added retrospectively by the Minnesota Population Center during data transcription (Ruggles et al. 2010). There were 68,160 households (comprising 313,702 individuals) in Hamilton County in the 100% count of the 1880 Census.

Household characteristics were identified using the records for all individuals listed as person number one (head of household), and all references to household or householder refer to the attributes of this individual. Individuals living in group quarters, who are not considered household members in current Censuses, are considered household members in this study. Hamilton County was divided into 135 EDs, which contained, on average, 505 households (or approximately 2,300 individuals). A 5% sample, similar to a contemporary PUMS, was randomly drawn from the full count of households in the SEA, and each household in this sample was assigned a design weight (household weight) of 20. This "pseudo-PUMS" (N=3,408) comprises the analytical sample used in the maximum entropy procedure, which is subsequently spatially allocated among the 135 EDs covering the county.

Prior to running the weights imputation, a crucial task is the selection of constraining variables. The procedure to select the constraining variables is described first. Next, three different measures are described that can be used to validate the imputation and allocation results for different combinations of constraining variables. As noted before, this study focuses on the validation process; technical details about the maximum entropy weights imputation and allocation process beyond the above summary are described in Nagle et al. (2012) and Leyk et al. (2013).

Finding meaningful constraining variables

The constraining variables in the maximum entropy weights imputation should ideally delineate different household-level residential patterns; this will increase the variability in the underlying data that can be explained and result in more accurate estimates. Population characteristics (such as gender) that are similarly distributed among EDs are unlikely to produce

satisfactory allocation results when used as constraints, since there is little variation to exploit. In addition, the inclusion of multiple highly correlated variables may be unnecessary, as highly correlated variables will likely be redundant in explaining variation in the underlying population distribution. The choice of constraining variables represents a difficult problem in survey sampling that has found limited attention to date and there is no standard method in place.

Bivariate correlations of ED-level population characteristics are calculated as one obvious way of assessing highly correlated variables that would be unsuitable constraining variables if applied in concert. Principal component analysis (PCA) is used to examine how much variation in the data is explained by the different population characteristics, and thus to identify the variables that may be most useful as constraints. While PCA is commonly used to reduce the dimensionality in a given set of data, it may also be helpful in describing the associations between the variables present in the data (Jolliffe 2002; Demšar et al. 2013). Finally, a segregation index, the index of dissimilarity (D), is computed at the ED-level to determine those variables that may represent appropriate constraints. The index of dissimilarity is a measure of the evenness of the distribution of two groups (Massey and Denton 1988), and may therefore be helpful in determining which variables best differentiate (or segregate) household residential patterns. The dissimilarity index is commonly interpreted as the proportion of individuals of one group who would have to move to reproduce within each ED the distribution of the two groups that exists within the entire area. Dissimilarity index values range from 0 to 1, with values tending towards 1 indicative of more highly segregated groups and values tending towards 0 suggesting low levels of segregation among the groups. The index is calculated for each of the benchmark variables across all of the EDs within Hamilton County.

Establishing a validation procedure

Weights imputation is performed using different combinations of constraining variables to examine the sensitivity of the allocation model to the number and types of constraints applied. As noted in the Methods section above, the weights imputation redistributes among the 135 EDs the original design weight for each household in the pseudo-PUMS sample, and then iteratively reweights the ED-level weights to match the aggregate summary statistics for each ED. Although these imputed weights are not required, and in reality are unlikely, to be whole numbers, the sum of the weights for a particular household record across all EDs will be equal to the expected number of households of that type (with 'type' characterized by the constraining variables used) in the SEA. Aggregating the imputed weights over those households exhibiting a particular attribute (e.g., foreign born household heads) within each ED will result in a revised summary statistic for that ED. This revised summary statistic will match exactly the actual count derived from the full enumeration if this attribute has been used as a constraining variable. An important component of the validation task then is to establish how well the revised summary statistics for household attributes not used as constraints replicate the actual number of households with those attributes in each ED. Following each model run, revised summary tables were generated by ED for the attributes of interest (benchmark variables as described above) based on the allocated microdata. The revised summary tables were compared to summary tables constructed from the 100% enumeration of the population. To examine the accuracy of allocation results from different perspectives, three goodness-of-fit statistics were calculated, as described below.

Error in Margin:

The actual number of households in the entire study area exhibiting a particular population characteristic will be compared to the total allocated number of households with the same characteristic to assess how well individual variables are being allocated overall. This difference is designated the error in margin. While the error in margin reveals little about the performance of the allocation procedure in reproducing the accurate population distribution within EDs, substantial differences between total household counts and allocated household counts will indicate variables for which the model critically fails. Importantly for the implementation of the allocation model with current Census data, the actual margin can be easily calculated in most cases based on publicly available data even if the other goodness-of-fit statistics described below cannot be derived. For example, when using 2011 ACS 1- or 3-year population estimates, census tract-level data is not released, but PUMA-level summary statistics are available. Thus, the performance of a tract-level allocation using this dataset cannot be evaluated at the tract scale; however, the actual PUMA population totals may be compared to the allocated PUMA totals to get a sense of the overall performance of the model. For such cases it is important to examine how well errors in margin correspond to the standardized absolute error (SAE) or z-statistics described below, which quantify the error in the distribution. These measures are sometimes irretrievable from contemporary censuses without access to confidential data.

Residuals and Standardized Allocation Error (SAE):

The residual is the difference within an ED between the actual population count and the allocated population count. Standardized Allocation Error (SAE) is the sum over all EDs of the absolute residuals standardized by the total expected population:

$$\frac{\sum_i |U_i - T_i|}{\sum_i U_i} \quad \textbf{(2)}$$

where U_i is the actual count of the population in ED_i and T_i is the allocated count of the population in ED_i. SAE will generally fall between 0 and 2, with values closer to 0 indicating a better fit between the actual and allocated distributions. Because the allocated margin is not required to match the actual margin for non-constraining variables, the SAE could, in theory, be greater than 2 for these variables. SAE was used to test the performance and accuracy of a variety of model specifications (e.g., different variables used as constraining variables) and to compare them. The SAE may also be computed for individual EDs, or for individual estimates within an ED. In this case, the SAE is similar to a coefficient of variation, which is calculated as the standard error of an (mean) estimate divided by the estimate itself.

Modified z-statistic (as described in Williamson, Birkin and Rees (1998)):

The modified z-statistic can be used to compare a table representing the actual joint distribution (or cross-tabulation) of two population characteristics with a table representing the allocated joint distribution of those population characteristics. The z-statistic is calculated for each corresponding pair of table cells, with significant values representing those elements in the distribution of the particular population characteristics for which the allocation procedure is performing inadequately. The modified z-statistic is calculated by

$$Z_{ij} = \frac{(r_{ij} - p_{ij})}{\sqrt{\frac{p_{ij}(1 - p_{ij})}{\sum_{ij} U_{ij}}}} \quad \text{where} \quad p_{ij} = \frac{U_{ij}}{\sum_{ij} U_{ij}} \quad \text{and} \quad r_{ij} = \frac{T_{ij}}{\sum_{ij} U_{ij}} \qquad \textbf{(3)}$$

where *i* and *j* are measurable characteristics of the population within an ED, U_{ij} is the actual count for cell *ij* in the ED and T_{ij} is the allocated count for cell *ij* in the ED. Population characteristics for which the actual and allocated distributions are poorly matched may require further consideration, such as additional constraining variables to be incorporated into the model.

The above three measures will highlight those variables which show unusual behavior within the allocation procedure and make it possible to carry out an in-depth validation based on the available full population count. Of particular interest is the level of accuracy with which non-constraining variables can be estimated. An important question is whether one can differentiate between those non-constraining variables which are strongly correlated with one or more constraining variables, and those which are seemingly unrelated to any of the constraining variables. This will provide important insight for the selection process of constraining variables and the configuration of the allocation model. The described validation procedure will also indicate whether the accuracies of the ED estimates for different population characteristics exhibit geographic heterogeneity through the compilation of residual maps, and whether the goodness-of-fit for an allocated distribution, as measured by the SAE, can be inferred from the error in margin.

Results

The selection of constraining variables

The first step in the allocation process is the selection of those variables that will be used as constraints. Although the digitally transcribed 1880 Census includes fewer variables than more contemporary censuses, there is greater flexibility in choosing constraining variables using the 100% population count because univariate and joint distributions of any variables of choice can be constructed. Thus this step is not limited by the summary tables produced by the Census Bureau. As noted above, while the choice of constraining variables should be grounded in theory, there are analytical techniques that may guide the selection process. In this study segregation indices, bivariate correlations, and principal component analysis were used to determine the "optimal" constraining variables i.e., variables with higher potential explanatory power that are not strongly correlated with each other.

Table 2 displays the index of dissimilarity, measured at the level of the ED using the aggregate summary tables, for each of the benchmark variables. Some variables, including the urban/rural dichotomy, residence in group quarters, and residence on a farm, display very high levels of segregation, due to their natural geographical disparity. However, several benchmark variables are highly correlated, and the inclusion of multiple highly correlated variables as constraints would be redundant. Examples of highly correlated variables include urban residence and farm residence (Spearman ρ=-0.64) and group quarters status and single status (Spearman ρ=0.69). The full correlation matrix for all benchmark variables is displayed in Appendix 1.

Principal component analysis (PCA) provides another method of selecting relevant and non-superfluous constraining variables. The results from the PCA run on the ED-level aggregate summary tables for the 19 benchmark variables suggest that five underlying latent variables explain more than 85% of the variation in the benchmarks. These five principal components all have eigenvalues greater than 1; the sixth principal component has a substantially smaller

eigenvalue.[1] The original variables that can be identified as loading most heavily on these five principal components include those describing urban status, group quarters status, nativity status, age, and occupational status.

Based on the analysis of segregation and the PCA, it is possible to exclude variables that contain redundant information and to select five variables as potential constraining variables: The urban status of the householder (urban vs. rural), the group quarters status of the householder (group quarters vs. non-group quarters), the occupational status of the householder (non-worker, low-skill worker, medium skill worker, high-skill worker), the nativity status of the householder (foreign born vs. non-foreign born), and the race of the householder (non-white vs. white). Although householder race was not one of the five variables identified in the PCA, this variable was used in place of age, as it displayed a higher level of segregation than did most of the age categories. The exclusion of age as a constraint variable allows for its use in the confirmatory validation that follows. The occupational status variable was divided into four categories, one for householders outside of the labor force and three corresponding to low, medium, and high occupational status; the other four constraint variables describe dichotomous relationships.

[1] PCA is commonly used to reduce the dimensionality (number of variables) of a dataset by creating new variables (principal components) that are combinations of the original variables and that are uncorrelated with each other. The principal components should retain as much of the variation in the dataset that is explained by the original variables as possible. Eigenvalues are the sample variances of the principal component scores. The rubric of retaining only those principal components with eigenvalues greater than 1 (in cases where the PCA was run on a correlation matrix) is known as Kaiser's Rule (Kaiser 1960; Jolliffe 2002).

As noted before, the primary purpose of this paper is to describe a method of validating small area estimates using perfect and complete census information and to examine the case in which such complete information is not available, such as in contemporary censuses. A second purpose, however, is to assess how changing estimation parameters affect model performance and the estimates themselves. To this end, while the model with five constraints will form the base model, models with 2-4 constraining variables will also be estimated. This step-wise modeling approach will allow evaluation of the sensitivity of the estimation procedure to changes in the model parameterization. Because adding constraints likely increases the accuracy of the spatial allocation process in reproducing the actual 100% population distribution, a natural inclination would be to add constraining variables until the supply of available constraints was exhausted. However overfitting of the maximum entropy model through the inclusion of an excessive number of constraining variables may lead to inefficiency and non-convergence. This can be particularly true in cases where the univariate or joint population distributions (such as summary statistics from the Census SF3 or American Community Survey) for constraint variables used in the maximum entropy imputation include sampling error or imputed data. The univariate and joint population distributions used herein are created from the 100% population count and do not contain sampling error; thus non-convergence of the maximum entropy model is not a concern. This issue will be salient however in the context of contemporary Census-produced summary tables, since sampling error and imputed values are likely to be present in these data.

Following the maximum entropy imputation, the set of imputed weights is applied to allocate households to specific EDs. The imputed weight for a single household in a single ED represents the expected number of households of that type within that ED. The row sum of the

imputed weights will be equal to the original household design weight. Allocation can proceed by assigning fractional parts of households in strict adherence to the imputed weights, or by rounding the imputed weights to integers and relaxing the strict adherence (Leyk et al. 2013). The method applied here utilizes the exact imputed household weights.

Post-allocation results: Comparison of allocated distributions to actual distributions

Once allocation is complete, revised ED-level population counts can be generated. Figure 1 displays the total population counts in the SEA, compared to the allocated population counts following the maximum entropy allocation model with five constraining variables. The variables used as constraints are listed first (within grey area), followed by the additional benchmark variables. While the 100% population counts and the allocated population counts for constraining variables are by design required to be the same, this chart highlights how the allocated total counts for the other benchmark variables are very close to their actual counts. For example, the actual number of male householders in Hamilton County is 54,932 (80.59% of all householders), while the number of male householders predicted by the model is only slightly larger, at 54,999 (80.69% of all householders).

The largest absolute errors in margin occur in the number of households with five or more children, for which 931 households are over-allocated (˜9.1% error in margin), and in the number of married householders, over-predicted by 589 households (˜1.2% error in margin). Other than the variable denoting married householders, the only benchmark variable with an error in margin greater than 5% is the number of householders younger than age 18 (˜6.5% error in margin).

Figure 2 displays the SAE metrics for those benchmark variables *not* used as constraining variables in the five-constraint maximum entropy model. By design the SAE for variables used as constraints are 0. Although many of the benchmarks appear to be well allocated by this measure, two variables have noticeably poorer fits: Householders younger than age 18 and farm households. The SAE is equivalent to the mean residual divided by the mean actual number of households. On average, the number of allocated households in an ED is within approximately 20% of the actual number of households in that ED, for most benchmark variables.

As mentioned above, the maximum entropy procedure was also run with different numbers of constraining variables, to evaluate how additional constraints affect the distribution of allocation errors. Figure 3 displays the SAEs of various population attributes for the maximum entropy models with 2-4 constraining variables, as well as the SAEs for the baseline five-constraint model. As before, these SAEs fall to zero when the variable is used as a constraint in the model. In general, the addition of constraining variables to the model reduces the SAE for the benchmark variables, although the magnitude of the decrease appears to depend on the relationship between the benchmark variable and the newly added constraint. For example, the error in the allocation of farm households drops substantially when occupational status is added as a constraining variable (most farm householders have low occupational status), while the error in the allocation of native-born households is greatly reduced when foreign-born head of household is added as a constraint. Several benchmark variables, including those representing ages above 18, gender, and marital status, exhibit little change when additional constraints are added to the model. These benchmarks are largely uncorrelated with any of the constraining variables and generally have small errors under any of the model specifications.

Of central importance in the evaluation of model performance is the association between the error in margin and the SAE. Figures 4-7 highlight the relationships between the errors in margin of the benchmark variables (x-axis) and their average ED-level SAEs (y-axis), for the models with 2-5 constraining variables. The constraining variables themselves are excluded from these graphs, to allow accurate comparisons across different model specifications. A linear regression line is provided to summarize the point relationships for each model. The errors in margin and the SAE exhibit a positive association in each of the four models, with the strength of the correlation decreasing as additional constraints are added to the maximum entropy model. The large SAEs observable in the 2-constraint model are greatly reduced in subsequent models, while the error in margin shows nearly constant magnitude across different model specifications.

Post-allocation results: Geographic heterogeneity in benchmark variable allocation errors

The model with five constraints results in only two benchmark variables (householder age 0-17 and households with 5+ children) having an error in margin greater than 5% and only four benchmark variables (householder age 0-17, farm households, native-born householder, and single householder) having SAE values greater than 20%. Maps of the allocation errors in these poorly performing variables were created at the scale of the ED to assess whether spatial heterogeneity or local clustering was present in the errors. Figures 8-12 display the standardized residuals, by ED, for those benchmark variables that have high SAEs or high errors in margin. The focus of these maps is on the EDs in the denser, central portion of the county, which comprise most of the city of Cincinnati. The extant outset maps display the whole county as a reference. EDs are shaded according to their SAE in the five constraint model, with lighter EDs indicating lower allocation errors and darker EDs indicating greater allocation errors.

Residuals for householders age 0-17 (Figure 8) and households with 5+ children (Figure 9) appear to be largest in the south-central portion of the county, which encompasses the city of Cincinnati. While single householders (Figure 10) were also misallocated to the largest extent in this locale, large residuals for single householders seem to be clustered on the outskirts of the central city. Perhaps the most distinct clustering of allocation residuals occurs for the benchmark variables of native-born householder (Figure 11) and farm households (Figure 12). There are large errors in the allocation of native born households in the EDs just north of the historic central business district of Cincinnati, while farm households are highly misallocated in the majority of the downtown EDs.

Post-allocation results: Comparison of joint distribution of age and gender

To this point, only allocation errors in the univariate distributions of the group of benchmark variables have been explored. However, researchers are often interested in the joint distributions of variables; indeed, one anticipated goal from developing spatial allocation models is the ability to estimate joint distributions of variables for which none had previously existed. To assess the accuracy of the spatial allocation in duplicating the actual joint distributions of variables, the z-statistic described above may be used. The z-statistics for the joint distribution of two variables not used as constraints, age of the householder and gender of the householder, were calculated within the EDs and the results are displayed in Table 3 for the two EDs with the largest populations. Age and gender were chosen for this example because of their relevance to demographers and the fact that the joint distribution of these variables (at the household level) is absent from contemporary Census-produced aggregate summary tables.

In both EDs, concordance between the estimates obtained from the spatial allocation and the actual population values is fairly high. In ED 192, the most egregious errors are in the allocations of male and female heads of household age 19 and under and female heads of household age 65 and older. In this example, the young male age groups are under-allocated, while the young female age groups are over-allocated. These very young household heads are mostly group quarters residents, who are predominantly native born, white, and outside of the work force. Thus, based on the constraint variables used in the allocation, there is little variation with which to differentiate between the two genders. It is notable that the aggregate 100% count of males and females age 19 or under (282) is quite close to the aggregate allocated count for this same group (280). There are almost no young heads of household in ED 145 (likely explaining the improved allocation of this group in this ED), but significant distribution errors exist for a few of the older age groups. This error is more difficult to explain, as there is little observable pattern in under- and over-allocation between the genders or age groupings.

Although it is likely infeasible to examine the joint distributions of all variables over each and every ED in the sample, the information gleaned from the comparisons of a few EDs may be useful in restructuring the original optimization problem. In this case, adding a constraint variable for the gender of the householder may ameliorate some of the errors in ED 192, but may have no effect on the errors in ED 145. Ultimately, the decision on the acceptable level of error, including where it occurs in the distribution of the population characteristic(s), may depend on the research question under consideration.

Discussion and Concluding Remarks

The maximum entropy procedure detailed above aims to increase the utility of Census microdata in small area estimation by adding geographic detail to household microdata records. This spatially enhanced microdata can be used in the construction of revised summary tables which cover a wider range of population characteristics than those that are currently available as well as new joint distributions. However there has been little authentication of the results obtained from this spatial allocation model, a model which may comprise many different specifications, variables, and geographical contexts. Because the CRDC is the only available source with which to authenticate these revised tables, and due to the massive scope of this validation challenge, it is necessary to lay out a systematic procedure to accomplish the methodological steps prior to its undertaking at a CRDC. The purpose of this paper, then, is to establish the validation procedure, highlighting the performance of the model under different configurations using publicly available Census data from 1880 and drawing conclusions about how to transfer this framework to the more contemporary context. The results shown above suggest that the validation procedure provides useful statistics, allowing an in-depth evaluation of the accuracy of the household allocation model and highlighting some directions for future work.

One of the important conclusions from this assessment is that the addition of constraint variables improves model fit not just for the constraining variables themselves, but also for variables that are correlated with the constraining variables (Figure 3). For example, the addition of the foreign born variable as a constraint results in a decrease of nearly 50% in the SAE for the native born benchmark variable, with which it is highly correlated. This behavior can be leveraged, and the total number of constraints minimized, through a careful selection of constraining variables that share multiple high correlations with other variables. A second

significant conclusion is that smaller errors in margin are associated with overall better fitting distributions, although this relationship appears to deteriorate as additional constraints are added. Errors in margin are an easily calculable fit statistic, and since their computation requires no knowledge of the actual distribution of the population within the EDs (or tracts) in the SEA (or PUMA), they can be computed using publicly available data. This fact is highly beneficial, as it would allow for a preliminary validation of an allocation model with a specific set of parameters without any need to access confidential data. However, the number of variables over which this relationship was tested was small, some of the benchmark variables still displayed poor distributional fit, and the overall association between the errors in margin and the SAE was not remarkably strong.

While the intent here is not to develop the optimally fitting model for the 1880 Census data, it is instructive to consider the overall pattern of data fit that is being produced by the maximum entropy imputation and subsequent spatial allocation, as this model has been originally developed for use with contemporary Census data. In general, the estimates being produced by the model are quite promising. In its report on the use of American Community Survey data, the National Research Council (2007, p. 64) advises that coefficients of variation (CVs) in the range of 10-12% are acceptable for population estimates. While the SAEs shown above are not CVs in a strict sense, they are mathematically comparable. The results from Figure 2 show that nearly half of the non-constraint benchmark variables used in this study achieve this goal, with several others performing only marginally worse. In the context of contemporary ACS tract population estimates which have large variances, the estimates from this spatial allocation do not seem excessive for most of the benchmark variables surveyed.

Although the allocated counts for most benchmark variables display high concordance with the 100% counts, two variables, the number of householders age 0-17 and the number of farm households, are poorly allocated. The large allocation errors for these variables are somewhat surprising since both variables are highly correlated with constraint variables included in the model. The minor householder population is positively correlated with the group quarters population (ρ=0.53), while the farm household population is negatively correlated with the urban population (ρ=-0.64). The problem in the allocation of these two variables, then, appears to be that both describe relatively rare populations. Of the non-constraining benchmark variables examined in this paper, these two variables have by far the smallest sample counts ($N_{AGE\ 0\text{-}17}$ = 87, N_{FARM} = 219). Although both the group quarters variable and the non-white variable also have sample counts in this range, these variables are used as constraints; thus, the allocation errors for these variables are by design 0. The inability of the maximum entropy procedure to accurately allocate variables describing rare populations is troubling, as estimates for these variables may be the most desired (in the contemporary context, variables with small populations may be least likely to have Census-produced summary tables). Additional research is therefore warranted on whether these variables may be better estimated through a different post-imputation allocation and how they can be reliably identified based on model diagnostics.

In addition to overall measures of goodness of fit, the spatial distributions of model residuals for individual variables are very useful to determine where a model over- or under-predicts and to identify local clusters of small or large residuals. While in this study substantive discussion is not a priority the results demonstrate the usefulness of such maps for researchers who are interested in more detailed interpretations of residual distributions with regard to specific variables of interest.

As noted before, this article does not discuss substantive questions regarding demographic processes in 1880 due to a desire to focus on the validation procedure itself, rather than any inferences that may be drawn from the outputs of the spatial allocation. An important question is how the validation methods described above will translate from the 1880 Census to more current Censuses such as the ACS. Nothing in the validation procedure is specific to the data from 1880 (or to the chosen geography of Hamilton County, Ohio), although there are certainly differences between the 1880 Census and the current context.

The maximum entropy procedure requires constraining variables that occur within (and are comparable between) both the public-use microdata file and the Census-produced summary files. This caused no restriction in using the 1880 data, for which the "public-use" microdata file and summary files could simply be constructed from the 100% data. This will not be the case when using current data, although an examination of the 2006-2010 ACS public-use microdata and summary files reveals that it contains many of the constraining variables used in this analysis. A more persistent methodological problem may be the presence of sampling variance and imputed data in contemporary Census data. Because the full 1880 census was available for use in this analysis, there is no inherent uncertainty in the summary tables created. Sampling variance and data imputation in current Census-produced summary files could lead to convergence problems in the maximum entropy procedure, and may require model reparameterization. In the worst case scenario some potential constraining variables may have to be discarded, if their inclusion in the model repeatedly leads to non-convergence. This indicates an obvious need for uncertainty-sensitive modeling techniques that can handle inherent sampling variance in constraining variables.

Based on the insights from this study there are some general rules and actions that can be done prior to undertaking a model validation at a CRDC. The first is to develop a set of benchmark variables against which to evaluate the results. A limited number of benchmark variables were included in this validation analysis. It may be desirable to include additional variables in the full evaluation, particularly those variables which are uncorrelated with model constraints or which have small overall margins, as these benchmarks exhibited high residuals and SAEs. Next, those variables available for use as constraining variables can be determined using a combination of publicly available summary tables and PUMS documentation. Recall that constraining variables must be procurable in both the microdata and the summary tables. Variables which are likely to produce the most satisfactory results when used as constraints may be identified using bivariate correlations, measures of segregation, and PCA. The data necessary to run these identification tests are publicly available in Census-produced summary files. Following the selection of the constraining variables, the imputation may be run using the publicly available PUMS data. The imputed weights can then be used in the tract allocation, and the total margins for the allocated data can be compared to the actual margins to identify prominent errors and adjust the model accordingly. For those benchmark variables for which Census-produced summary tables or cross-tabulations are publicly available, SAEs and z-statistics can be computed to further adjust the model. Measures of error for benchmark variables and join distributions not publicly available will require evaluation at a CRDC.

Limitations and future steps

Some potential limitations with regard to the relationship between the historical and contemporary data may require further consideration. Relative to current censuses, the 1880

Census appears to include a less diverse population with more homogeneous residential patterns (less segregation), and thus the choice of constraining variables may need to be revisited for application to current data. While the results in this paper indicate that additional constraining variables have a beneficial impact on the reproduction of the correct population distribution for other non-constraining variables, it is still unclear what the optimal number of constraints might be. Additional work with current ACS data will allow determination of the point at which additional constraints may result in model non-convergence or increasing misallocation. The impact of population size of SEAs and EDs must also be reexamined to better understand the effect of population size on the maximum entropy method applied. This will also provide some indication how the method might be applied to different survey data, such as the National Health Interview Survey or the National Health and Nutrition Examination Survey, which are reported only for large geographies (i.e. national regions). Future research will also investigate differences in the validation results within rural and urban settings in more detail.

References

Assunçao, R.M., C.P. Schmertmann, J.E. Potter, and S.M. Cavenaghi. 2005. "Empirical Bayes estimation of demographic schedules for small areas." *Demography* 42(3):537-558.

Ballas, D., G. Clarke, D. Dorling, H. Eyre, B. Thomas, and D. Rossiter. 2005. "SimBritain: A spatial microsimulation approach to population dynamics." *Population, Space and Place* 11:13-34

Beckman, R.J., K.A. Baggerly, and M.D. McKay. 1996. "Creating synthetic baseline populations." *Transportation Research A* 30(6):415-429

Demšar, U., P. Harris, C. Brunsdon, A.S. Fotheringham, S. McLoone. 2013. "Principal component analysis on spatial data: An overview." *Annals of the Association of American Geographers* 103(1):106-128

Goeken, R., C. Nguyen, S. Ruggles, and W. Sargent. 2003. "The 1880 U.S. population database." *Historical Methods* 36(1):27-34

Hermes, K. and M. Poulsen. 2012. "A review of current methods to generate synthetic spatial microdata using reweighting and future directions." *Computer, Environment and Urban Systems* 36:281-290

Johnston, R.J. and C.J. Pattie. 1993. "Entropy-maximizing and the iterative proportional fitting procedure." *Professional Geographer* 45(3):317-322

Jolliffe, I.T. 2002. *Principal component analysis (2nd edition).* Berlin: Springer Verlag.

Kaiser, H.F. 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement* 20:141-151

Leyk, S., B.P. Buttenfield, and N. Nagle. 2013. "Modeling ambiguity in Census microdata allocations to improve demographic small area estimates." *Transactions in Geographic Information Science*: DOI: 10.1111/j.1467-9671.2012.01366.x

Logan, J.R., J. Jindrich, H. Shin, and W. Zhang. 2011. "Mapping America in 1880: The urban transition historical GIS project." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44(1):49-60

Massey, D.S. and N.A. Denton. 1988. "The dimensions of residential segregation." *Social Forces* 67(2):281-315

Melhuish, T., M. Blake, and S. Day. 2002. "An evaluation of synthetic household populations for Census Collection Districts created using optimization techniques." *Australasian Journal of Regional Studies* 8(3):369-387

Nagle, N.N., B.P. Buttenfield, S. Leyk, and S.E. Spielman. 2012. "An uncertainty-informed penalized maximum entropy dasymetric model." Proceedings of the 7th International Conference on Geographic Information Science (GIScience 2012), Columbus, OH, September 18-21, 2012

National Research Council. 2007. *Using the American Community Survey: Benefits and challenges.* Panel on the Functionality and Usability of Data from the American Community Survey, Citro, C.F. and G. Kalton (eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press

Ruggles, S., J.T. Alexander, K. Genadek, R. Goeken, M.B. Schroeder, and M. Sobek. 2010. *Integrated public use microdata series: Version 5.0 [machine-readable database]*. Minneapolis: University of Minnesota.

Smith, D.M., G.P. Clarke, and K. Harland. 2009. "Improving the synthetic data generation process in spatial microsimulation models." *Environment and Planning A* 41:1251-1268

Tanton, R., Y. Vidyattama, B. Nepal, and J. McNamara. 2011. "Small area estimation using a reweighting algorithm." *Journal of the Royal Statistical Society, Series A* 174(4):931-951

Williamson, P., M. Birkin, and P.H. Rees. 1998. "The estimation of population microdata by using data from small area statistics and samples of anonymised records." *Environment and Planning A* 30:785-816

Voas, D. and P. Williamson. 2001. "Evaluating goodness-of-fit measures for synthetic microdata." *Geographical & Environmental Modelling* 5(2):177-200

Table 1: Benchmark Variables for Validation of Spatial Allocation Validation

Benchmark	# of Categories	Measurement	Record Count (PUMS N = 3,408)
Age of Householder	4	Age 0-17	87
		Age 18-34	976
		Age 35-49	1,278
		Age 50+	1,067
Gender of Householder	1	Male	2,747
Race of Householder	1	Non-White	151
Marital Status of Householder	2	Single	305
		Married	2,542
Presence of Children in Household	2	Any Children	2,528
		5+ Children Present	555
Nativity of Householder[1]	2	Native Born	872
		Foreign Born	1,918
Occupational Status of Householder[2]	4	Non-Worker	637
		Low-Skill	997
		Medium-Skill	909
		High-Skill	865
Group Quarters Status of Household[3]	1	Group Quarters	293
Urban Status of Household[4]	1	Urban Household	2,788
Farm Status of Household	1	Farm Household	219

[1] Native born refers to individuals born in the U.S. with parents who were born in the U.S. Foreign born refers to individuals not born in the U.S. A third grouping, U.S.-born household heads whose parents were foreign born, is not considered here.

[2] Occupational standing is measured using the occupational earnings score variable, with the observed variable broken into three tertiles (Low-Skill, Medium-Skill, and High-Skill). Non-workers were individuals outside of the labor force.

[3] Because households were not defined in the 1880 Census, the contemporary distinction between group quarters and households is not relevant here.

[4] The converse of "Urban" is "Rural", distinct from but correlated with the "Farm" designation.

Table 2: Segregation Indices for Hamilton County, Ohio
(Diversity measured by Enumeration District)

Variable	D
Urban vs. Rural	1.00
Farm vs. Non-farm	0.81
Group vs. Non-group	0.81
Male vs. Female	0.14
White vs. Non-white	0.53
Single vs. Non-single	0.25
Married vs. Non-married	0.13
Children present vs. No children present	0.13
5+ Children present vs. Less than 5 children present	0.15
Foreign born vs. Non-foreign born	0.28
Native vs. Non-native	0.39
Occupation: Non-worker vs. All other	0.13
Occupation: Low-skill vs. All other	0.27
Occupation: Medium-skill vs. All other	0.19
Occupation: High-skill vs. All other	0.14
Age: Age 0-17 vs. All other	0.76
Age: Age 18-34 vs. All other	0.07
Age: Age 35-49 vs. All other	0.06
Age: Age 50+ vs. All other	0.09

Note: The urban/rural dichotomy has an index of dissimilarity of 1 because EDs are wholly classified as either urban or rural, with the classification extending to all households within the district. While no such "perfect" constraining variables will exist in contemporary Census data, this variable was nevertheless retained as a constraint.

Table 3: Comparison of Allocated Age and Sex Distribution to 100% Count Distribution

Enumeration District 192

	Male			Female		
Age	100% Count	Allocated	z-score	100% Count	Allocated	z-score
0-14	132	99	*-3.09	29	82	*10.49
15-19	89	23	*-7.35	32	76	*8.35
20-24	84	73	-1.26	13	30	*4.85
25-29	125	103	*-2.11	32	31	-0.19
30-34	115	112	-0.30	22	22	0.00
35-39	107	96	-1.13	22	31	*2.01
40-44	100	85	-1.59	26	24	-0.42
45-49	70	64	-0.74	22	21	-0.22
50-54	51	59	1.15	18	25	1.72
55-59	39	33	-0.98	11	15	1.23
60-64	22	33	*2.37	10	10	0.00
65+	24	33	1.86	4	20	*8.07

Enumeration District 145

	Male			Female		
Age	100% Count	Allocated	z-score	100% Count	Allocated	z-score
0-14	0	0	0.00	0	0	0.00
15-19	0	2	*2.00	2	0	-1.42
20-24	41	31	-1.60	2	3	0.71
25-29	88	79	-1.02	1	5	*4.01
30-34	117	101	-1.60	7	7	0.00
35-39	118	110	-0.80	10	16	*1.97
40-44	90	114	*2.69	16	16	0.00
45-49	76	98	*2.66	10	23	*4.26
50-54	95	86	-0.99	33	29	-0.79
55-59	60	59	-0.13	15	21	1.64
60-64	50	48	-0.29	21	18	-0.71
65+	45	38	-1.07	27	20	-1.49

Note: * $p < .05$

Figure 1: Comparison of Actual Population Count to Allocated Count, Model with Five Constraints

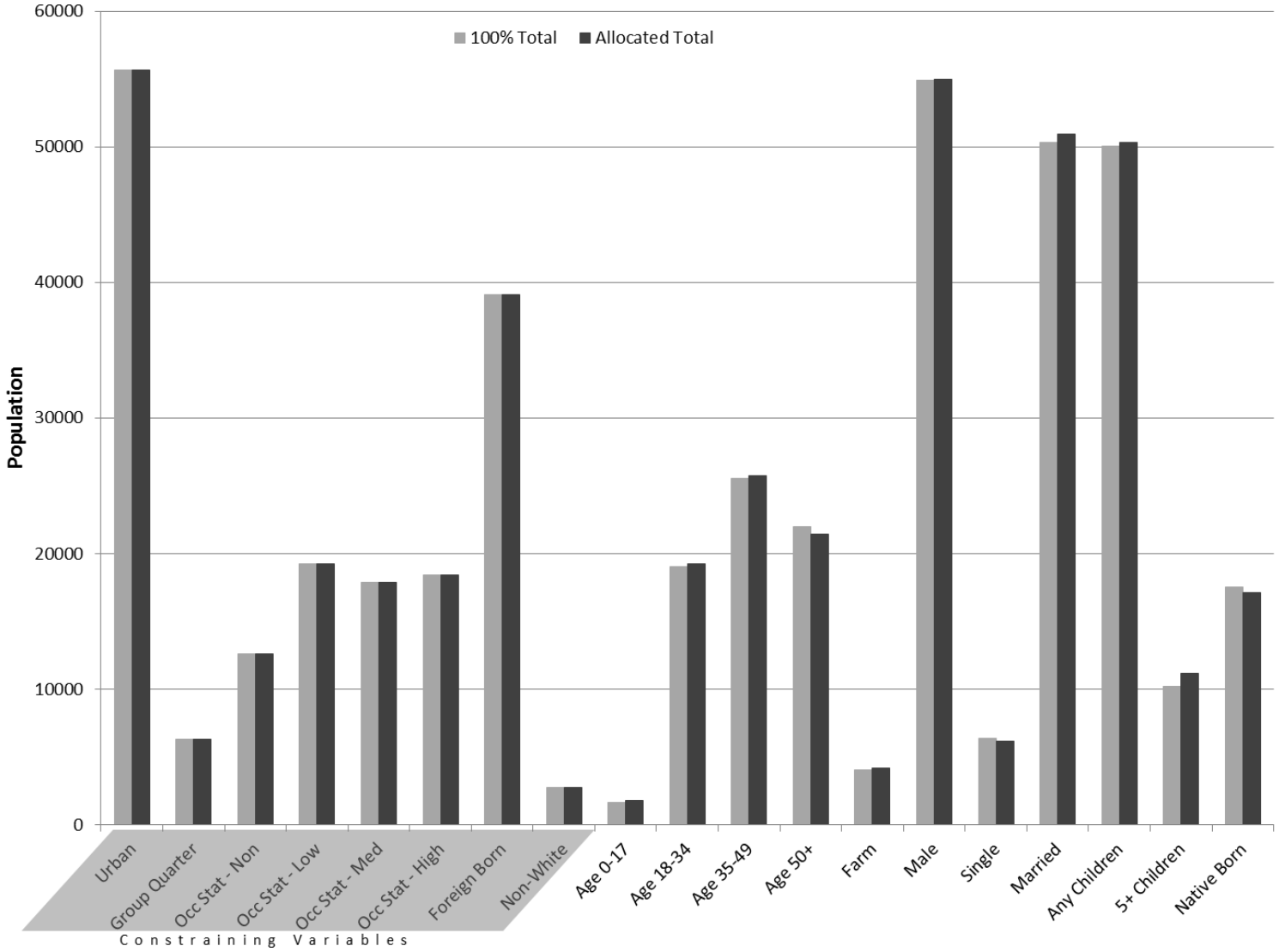


Figure 2: Standardized Allocation Error, Model with Five Constraints

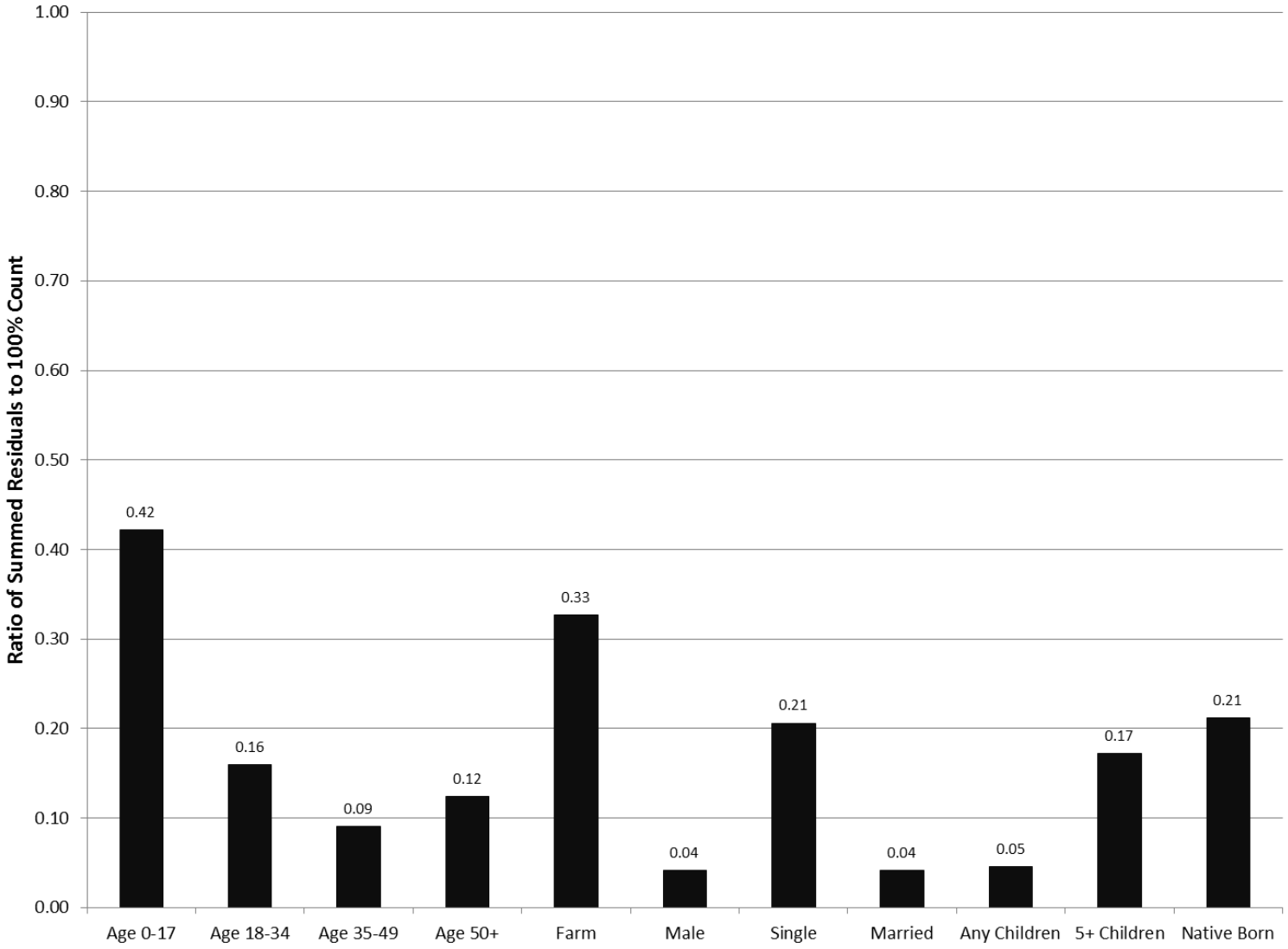


Figure 3: Comparison of Standardized Allocation Error for Different Constraint Variable Specifications

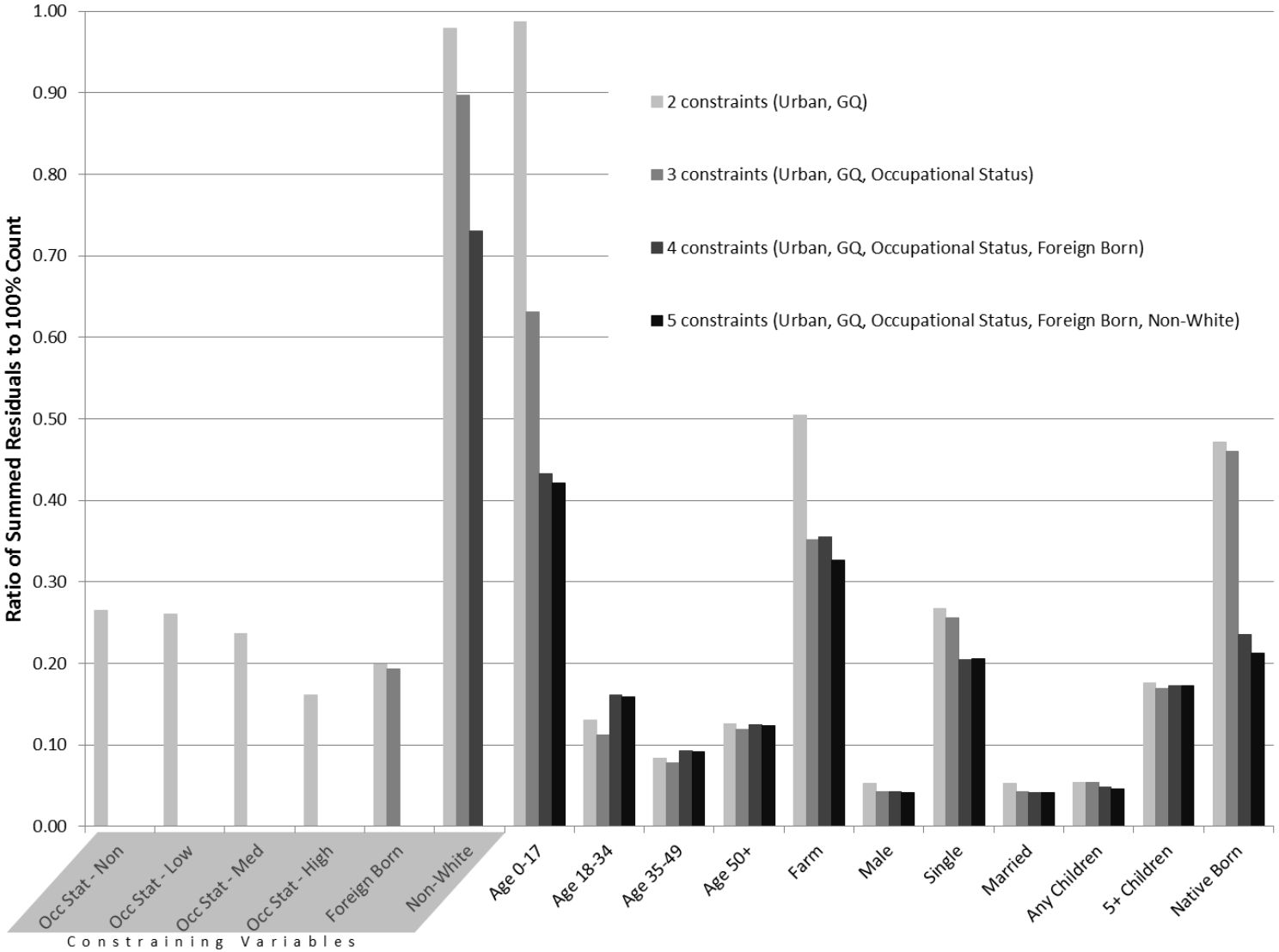


Figure 4: Model with 2 Constraints: Error in Margin (Ratio of Residual to Actual Count) by Error in Distribution (Ratio of Summed Absolute Residuals to Actual Count)

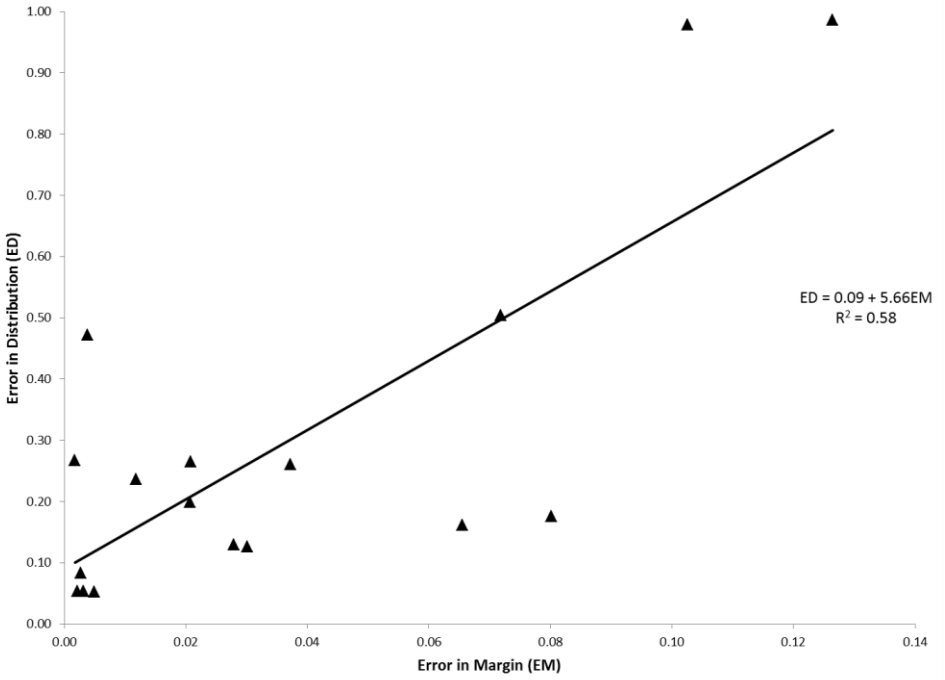


Figure 5: Model with 3 Constraints: Error in Margin (Ratio of Residual to Actual Count) by Error in Distribution (Ratio of Summed Absolute Residuals to Actual Count)

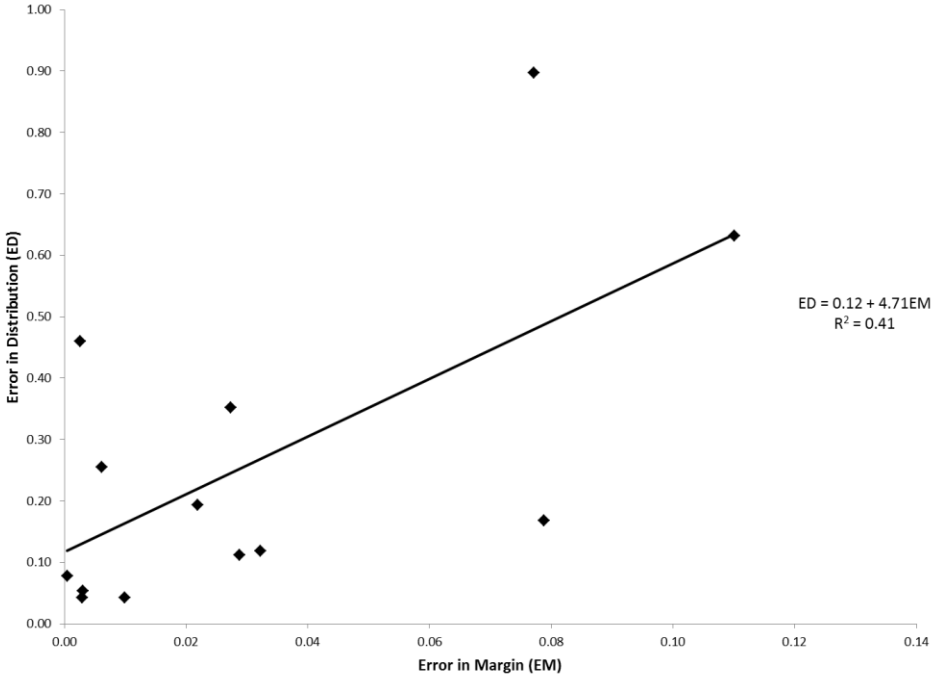


Figure 6: Model with 4 Constraints: Error in Margin (Ratio of Residual to Actual Count) by Error in Distribution (Ratio of Summed Absolute Residuals to Actual Count)

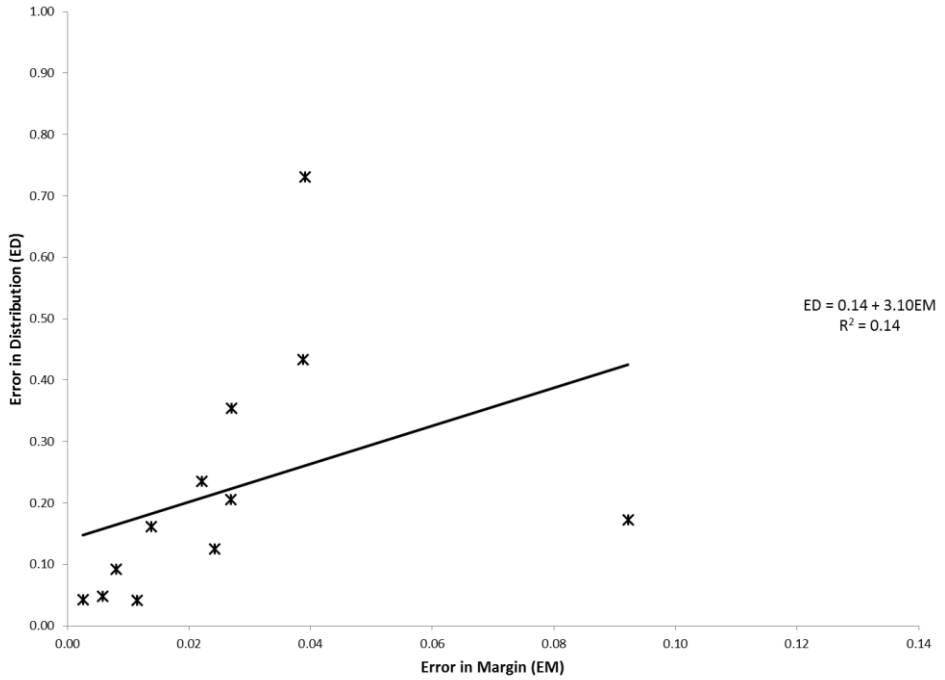


Figure 7: Model with 5 Constraints: Error in Margin (Ratio of Residual to Actual Count) by Error in Distribution (Ratio of Summed Absolute Residuals to Actual Count)

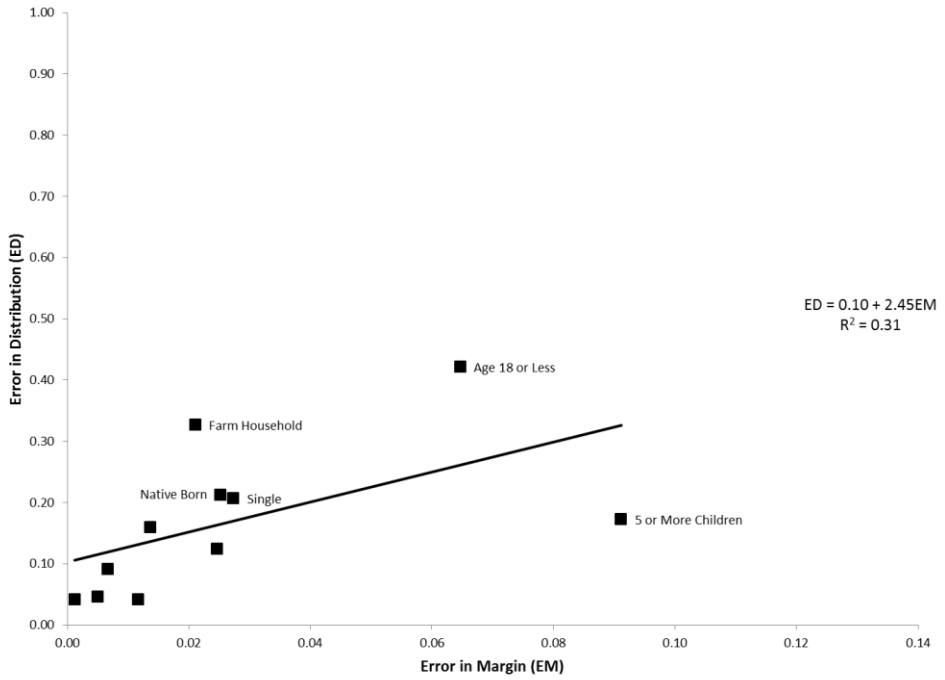


Figure 8: Standardized Allocation Error for Householders Age 0-17

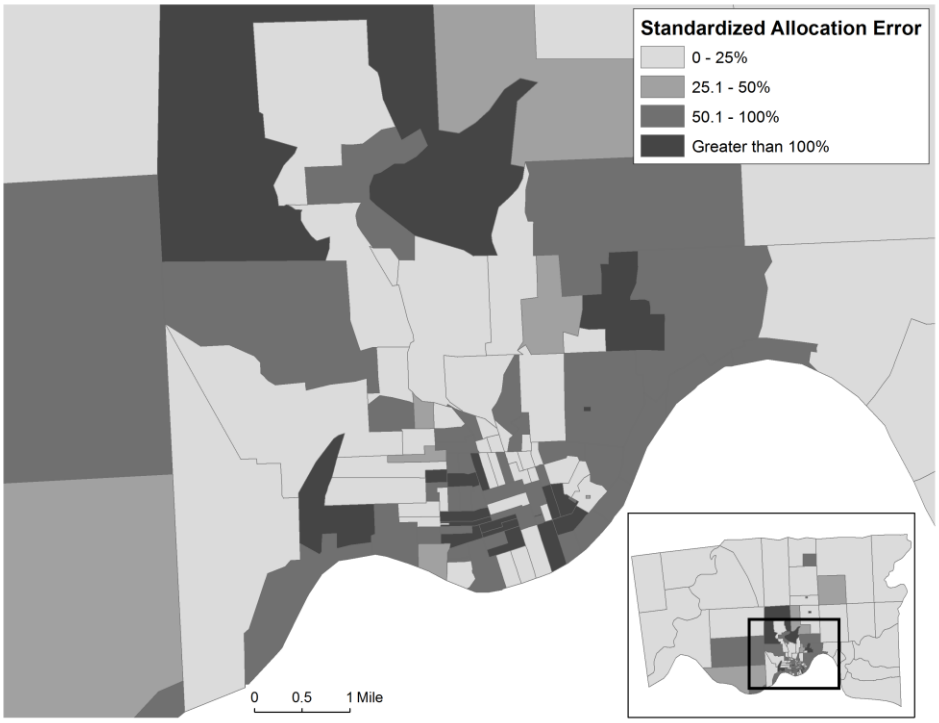


Figure 9: Standardized Allocation Error for Households with 5+ Children

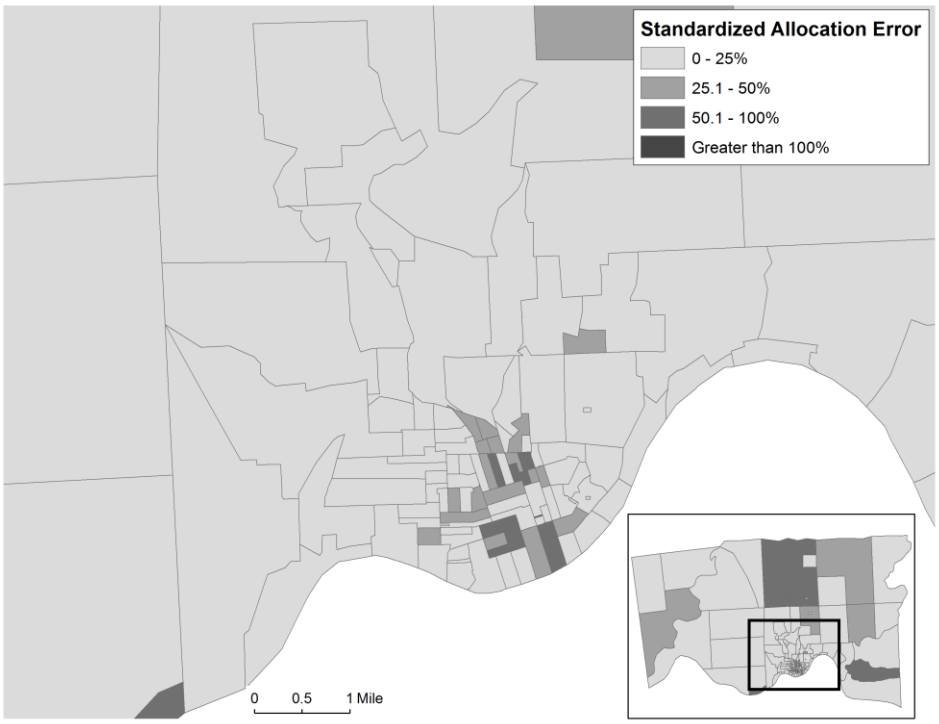


Figure 10: Standardized Allocation Error for Single Households

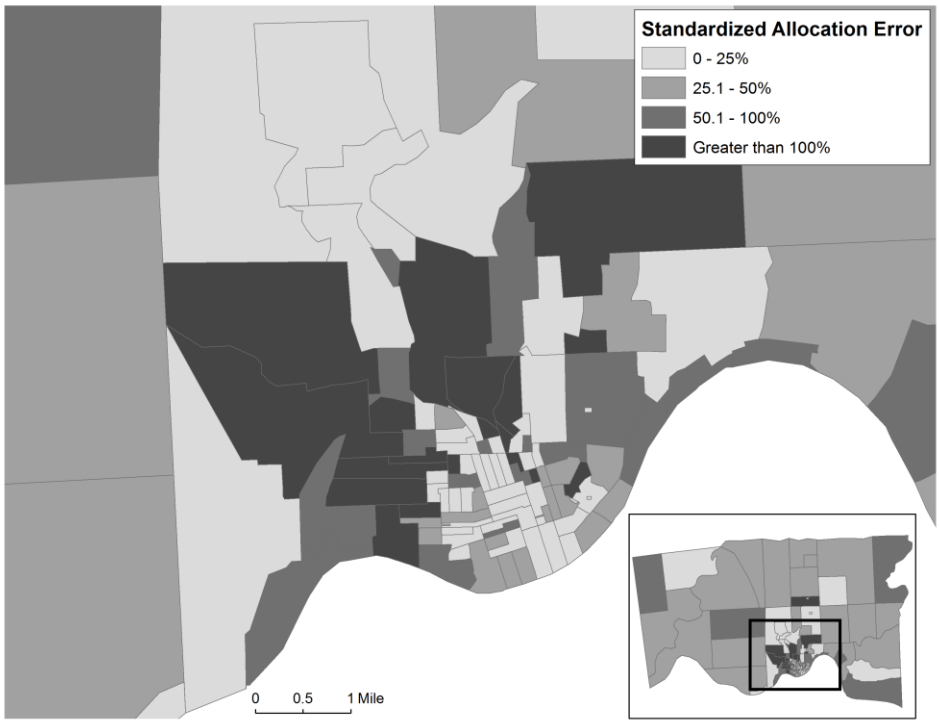


Figure 11: Standardized Allocation Error for Native Born Households

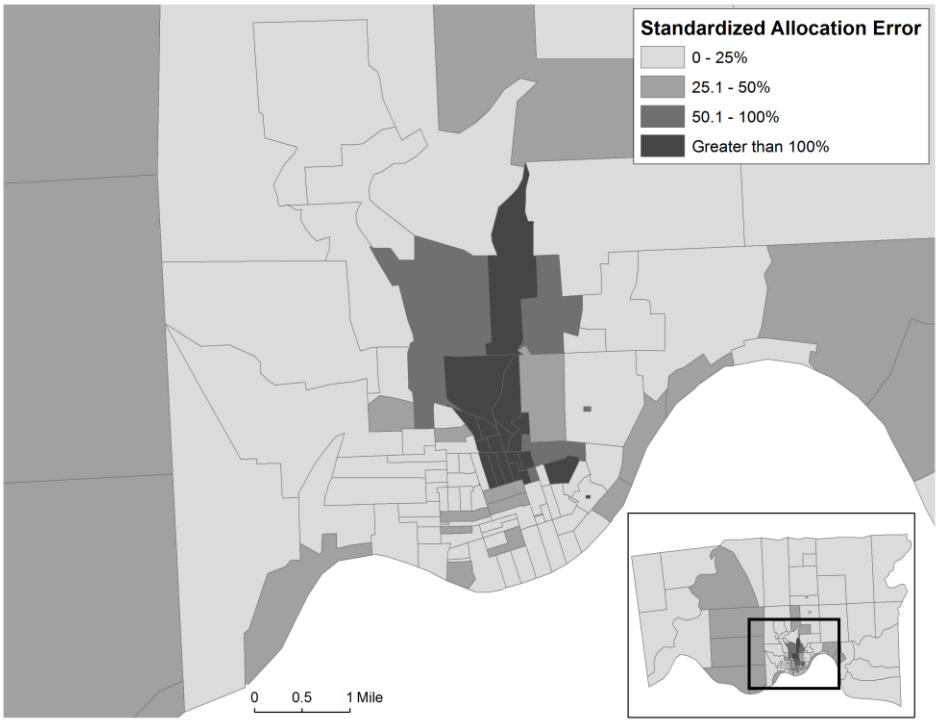
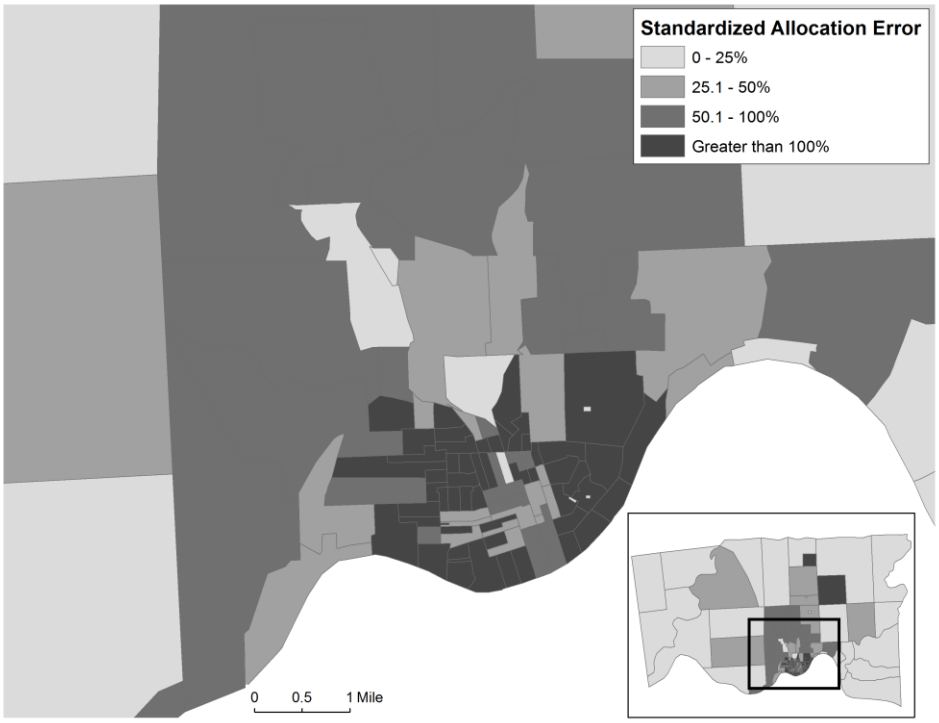


Figure 12: Standardized Allocation Error for Farm Households



Appendix 1: Spearman Correlation Coefficients for Benchmark Variables (Variables Measured as Proportion of Households in Enumeration District Exhibiting the Characteristic)

	Age 0-17	Age 18-34	Age 35-49	Age 50+	Male	Non-white	Single	Married	Any Children	5+ Children
Age 0-17	1.00									
Age 18-34	0.15	1.00								
Age 35-49	-0.30	0.00	1.00							
Age 50+	-0.40	-0.60	-0.30	1.00						
Male	-0.37	-0.11	0.24	0.19	1.00					
Non-white	0.12	0.06	0.10	-0.08	-0.25	1.00				
Single	0.55	0.05	-0.37	-0.22	-0.59	0.41	1.00			
Married	-0.51	-0.07	0.35	0.20	0.92	-0.29	-0.75	1.00		
Any Children	-0.49	-0.11	0.43	0.19	0.65	-0.48	-0.84	0.75	1.00	
5+ Children	-0.29	-0.26	0.31	0.27	0.68	-0.31	-0.51	0.66	0.72	1.00
Native	0.00	-0.20	-0.09	0.22	0.06	0.56	0.28	0.00	-0.29	-0.12
Foreign	-0.13	0.21	0.23	-0.11	-0.01	-0.44	-0.34	0.08	0.34	0.20
Non-worker	0.36	-0.19	-0.32	0.06	-0.53	0.09	0.39	-0.58	-0.44	-0.42
Low-skill	0.07	-0.12	-0.14	0.14	0.15	0.36	0.21	0.02	-0.13	0.12
Med-skill	-0.17	0.37	0.38	-0.26	0.11	-0.33	-0.40	0.23	0.37	0.14
High-skill	-0.18	0.16	0.43	-0.15	0.02	-0.03	-0.26	0.17	0.22	0.00
Group Quarters	0.53	0.11	-0.21	-0.35	-0.57	0.28	0.69	-0.62	-0.64	-0.47
Urban	0.00	0.35	0.31	-0.42	-0.36	-0.10	-0.08	-0.20	0.03	-0.23
Farm	-0.12	-0.19	-0.15	0.24	0.54	0.09	-0.04	0.40	0.08	0.31

	Native	Foreign	Non-worker	Low-skill	Med-skill	High-skill	Group Quarters	Urban	Farm
Native	1.00								
Foreign	-0.92	1.00							
Non-worker	0.03	-0.10	1.00						
Low-skill	0.49	-0.46	-0.25	1.00					
Med-skill	-0.62	0.66	-0.18	-0.65	1.00				
High-skill	-0.11	0.16	-0.12	-0.60	0.34	1.00			
Group Quarters	0.05	-0.10	0.34	0.07	-0.27	-0.05	1.00		
Urban	-0.52	0.50	0.03	-0.64	0.57	0.60	0.12	1.00	
Farm	0.49	-0.47	-0.21	0.42	-0.38	-0.23	-0.21	-0.64	1.00