8-2014

Comparing Models of Demographic Subpopulations

Jessica Jones Moehl
University of Tennessee - Knoxville, jjones70@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Jessica Jones Moehl entitled "Comparing Models of Demographic Subpopulations." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Geography.

Robert N. Stewart, Major Professor

We have read this thesis and recommend its acceptance:

Nicholas N. Nagle, Ronald Foresta

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Comparing Models of Demographic Subpopulations

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Jessica Jones Moehl
August 2014

ACKNOWLEDGEMENTS

I would like to thank Dr. Kao and Dr. Nagle for helping me with the implementations of their code and for providing helpful advice. I would also like to thank Amy Rose for letting me pick her brain for countless details and especially for letting me lean on her excellent editing skills. I thank also my advisor, Dr. Stewart, for his steering and helping me keep the course. Finally, thanks to my family for putting up with my increasingly grumpy self.

ABSTRACT

Understanding specific multi-dimensional demographics of populations in the United States at high resolutions is made difficult by the restriction of data released by the Census Bureau because of privacy concerns. Efforts to model these subpopulations have been increasing in recent years. These modeled populations have applications in decision making at all levels of government as well as in academia and the private sector. Two models have shown promising techniques for incorporating multiple levels of data to model sub populations in a meaningful way. These models, the Copula Model by Kao et al. (2012) and the Penalized Maximum Entropy Model by Nagle et al. (2014), have been applied in different study areas using different attributes. This paper provides a direct comparison which is needed to understand the strengths and weakness of each model as well as to assess the possibility of expanding their application nationally.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Section 1: Introduction

Publicly released Census data are often not both spatially and demographically detailed enough for researchers with questions at finer spatial scales (e.g. neighborhoods). This creates a data gap because researchers and planners across the public health, transportation, and public policy domains need data with both high spatial and demographic resolutions to understand the intricacies of the populations they serve. U.S. Census data are provided in different spatial hierarchies as indicated in Figure 1. The finest scale is blocks followed by block groups, tracts, and ultimately PUMAs. As the spatial detail becomes finer, demographic detail is lost. Their relationships are also shown in Figure 1. Figure 1B highlights the gap in publically released U.S. Census data.

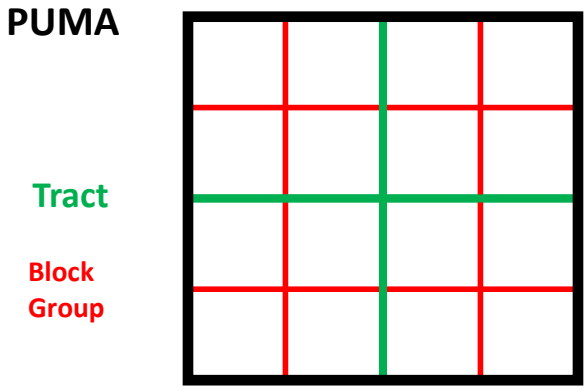

Figure 1 A: Block groups are within tracts which are within PUMAs.

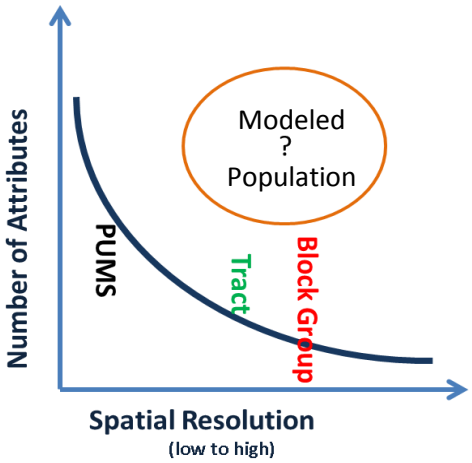

Figure 1 B: As spatial resolution increases, detailed attributes are reduced. Models fill the gap.

Table 1, adapted from Nagle et al. (2014), shows the data gap in clearer detail. For example, regional totals of black homeowners exist, but block group level counts do not. Researchers use various modeling techniques to fill this data gap and overcome the limitations of publicly released data (Wong 1992, Beckman et al. 1996, Williamson et al. 1998, Simpson and Tranmer 2005).

Table 1. Known joint distributions (PUMA) are distributed to smaller regions (Block Group 1 and Block Group 2) with the aid of summary tables (marginal data) from those sub-regions; adapted from Nagle et al. (2014).

	Block Group 1			Block Group 2			Region (PUMA)		
	Own	Rent	Total	Own	Rent	Total	Own	Rent	Total
Black	?	?	12	?	?	8	15	5	20
White	?	?	40	?	?	25	35	30	65
Total	19	33	52	31	2	33	50	35	85

The body of population modelling research is extensive. Many techniques are used, most of which principally rely on distributing a known population in a large zone to sub-regions within that zone with the help of ancillary information that describes how population should be distributed in these smaller zones. Tanton (2013) provides an overview of the history of many spatial microsimulation techniques. For example, Levy et al. (2014) use simulation to model the low income community which is vulnerable to economic stressors. Others have used simulation to understand future needs, such as aging adult care and child care, when demographic shifts are expected (Lymer et al. 2009, Harding et al. 2011). Iterative Proportional Fitting (IPF) is a very popular approach. Johnston and Pattie (1993) provide a useful examination of past efforts in geography which use IPF, alternatively called Entropy Maximizing procedures, and they also show how it is used to model voting patterns. Anderson (2013) provides a more recent elaboration on the history of IPF. Birkin and Clarke (1988) provide an early implementation of IPF and demonstrate its flexibility by creating summaries as well as individual units for further analysis. Wong (1992) shows how IPF can be impacted by the data distribution and categorization. Simpson and Tranmer (2005) improve the use of IPF by demonstrating its use in standard software over many dimensions. The IPF procedure has also been used in the transportation planning community by others such as Beckman et al. (1996), as part of the TRANSIMS model.

I compare two recent modelling advances in this paper: Copula Model (Kao et al. 2012) and P-MEDM (Nagle et al. 2014). Our objective is to assess the strengths and weaknesses of each model within the context of scaling to the nation level. To compare these models, I selected a study area comprised of varying geographic and demographic circumstances (e.g. rural, urban, affluent, poor, and mixed areas) that are expected when scaling nationally. By comparing these two models across this same heterogeneous study area, a direct comparison and assessment of model performance under a variety of conditions is possible. The model estimates are compared to American Community Survey (ACS) data

summaries. These summaries are referred to in this paper as ACS benchmarks. The ACS is a detailed demographic survey that is conducted continually. ACS data are aggregated across time and/or space to produce Census summaries for defined regions. For example, the Census single year summaries are only available for areas with 65,000 or more people (i.e. large areas), three year summaries are available for areas with 20,000 or more people (medium sized areas), and five year summaries for smaller areas/populations down to the block group level. Continual surveying allows the ACS five year data tables to be updated yearly while maintaining comparable small area spatial resolution provided by census long form summary files (U.S. Census 2006). These ACS summaries are used as benchmarks for model comparison because they cover the small geographies of interest to researchers and policy makers. However, because these data are based on a sample, they are uncertain and thus they have a margin of error which provides an opportunity to quantify the uncertainty, i.e. reliability of estimates.

Background information on the modeling techniques and evaluation methods is provided in the following section. Section 3 describes the data and study area. Section 4 contains the results and evaluation. Section 5 includes the summary and further considerations.

Section 2: Methods

Modeling Techniques

P-MEDM

The penalized maximum entropy model (P-MEDM) is a maximum entropy approach which incorporates uncertainty associated with estimates used in the maximum entropy fitting procedure (Nagle et al. 2014). Nagle et al. demonstrated this with an example that incorporated the ACS error estimates provided by the U.S. Census Bureau. The model is given by Nagle et al. (2014) as

$$max - \sum_{it} \frac{n}{N} \frac{w_{it}}{d_{it}} \log\left(\frac{w_{it}}{d_{it}}\right) - \sum_{k} \frac{e_k^2}{2\sigma_k^2}$$

subject to the relaxed pycnophylactic constraints

$$\sum_{it \in k} w_{it} = \widehat{Pop}_k + e_k$$

for each constraint k

where n is the sample size, N is the population size, and d_{it} is the prior estimate of the population w_{it}. In the second element of the equation, σ_k^2 is the variance of the uncertainty e_k, where in this paper k is the total target region population. I am solving for w_{it}, that is the number of individuals like sample record i in region t. The first element of the equation is the maximum entropy approach. This is equivalent to the iterative proportional fitting (IPF) procedure discussed in the previous section. Each sample record is distributed to the regions based on their likelihood of occurrence, given the constraints. This procedure makes no assumptions about which sample record occurs in which region because it relies only on the information provided by the constraints. The second element of the equation is the penalty term that allows for the inclusion of uncertainty associated with the constraints. The relaxed pycnophylactic constraints here say that when all of the estimates w_{it} are summed they will equal $\widehat{Pop}_k$, the expected total population, plus the error associated with the uncertain inputs. Pycnophylactic constraints (Tobler 1979) say that the pieces must add to the whole. Nagle et al. (2014) relax this by including the error term because without carrying the uncertainty present in the expected population through the modeling process, the estimates produced are given as certain, which is not the case. The model output is a collection of households equal to the expected number of households for each PUMA. Each sample household has been replicated in each block group according to its w_{it}. For further use, these samples can be used to summarize any variable or combination of variables provided in the sample microdata.

Copula Model

Another recent method by Kao et al. (2012) uses statistical copulas to model high resolution demographic data. The Copula Model was developed for use in transportation simulation models where IPF has been a popular method of synthesizing populations. Kao et al. (2012) note several problems of using IPF, notably that of empty cells when certain demographic combinations are not present, particularly when using a larger number of variables. Copulas, first described as such by Sklar (1959), are a popular statistical approach which uses known marginal distributions to create joint distributions (Kao et al. 2012). Copulas can be explained as functions that allow for the connection of multivariate distributions to their univariate margins. For example, let F represent an n-dimensional cumulative distribution function with univariate margins of $F_1, \dots F_n$. Sklar (1959) explains that there is an n-dimensional copula function C such that for an m-dimensional set of random variables y, $F(y_1, \dots, y_m) = C(F_1(y_1), \dots, , F_m(y_m)|W)$ where W is the copula parameter representing the correlation structure between the marginal variables (Trivedi and Zimmer 2006).

Kao et al. create synthetic households consistent with the dependence structure of a selection of variables from the PUMS microdata and fit these locally using known ACS summaries at the block group level. I am using the same six variables used by Kao et al. to generate the households used in our comparison. They are: household income, household size, number of workers, number of vehicles, highest educational attainment in the household, and total household travel time to work. Some copulas are not able to use non-continuous data, meaning that variables, such as number of vehicles, must be transformed (Panagiotelis et al. 2012). Also, continuous variables allow for a unique copula to represent the relationship between the joint distribution and the margins of the PUMS variables (Nelsen 2006). All of the variables of interest are used to define correlation structure and thus construct the synthetic households while only two, household income and household size, are used to fit the households in block groups. The copula as defined by the cumulative joint distribution of the six variables in the sample is used to generate synthetic households. The Copula Model scales the samples so they have uniform marginal distributions for the six variables, which allows for the correlation structure of the sample data to be preserved in the synthetic households at the PUMA level. This structure is not considered when the households are fit at the block group level. The attributes of all of the synthetic households in each block group are then summarized to create block group level joint distribution tables. The Copula Model can produce nonsensical results if left relatively unconstrained; for example, three workers can appear in a two person household.

Model Differences

These two models have some notable differences. The maximum entropy approach creates replicas of the existing microdata at finer spatial scales, while the Copula Model creates new households at finer spatial scales. These new households are often similar to the microdata households, but it is sometimes advantageous in transportation modelling research for households to have some variation. As implemented by Kao et al. (2012), households can have nonsensical results where, for example, three workers can appear in a two person household. An important difference between the two models is that the P-MEDM model incorporates uncertainty, present when data are created from estimates, in input data while the Copula Model does not. A potential shortcoming of the Copula Model for some applications is that it requires continuous variables, or variables that can be transformed into a continuous state. Thus, variables such as race and gender, which can be vital demographic elements in some research, are difficult to include. All of the differences identified here are important when considering deploying either model at a national scale.

Evaluation Methods

The relationship between variables in ACS summaries and the model estimates for a block group geography level is explained in three scenarios as shown in Table 2. In Scenario A, the models use the ACS summaries for certain variables to constrain their estimates which forces the model estimate to equal the known ACS summary value for that variable. In Scenario B the models estimates are produced for variables with known ACS summary variables but are not constrained by them, and thus do not have to match the known ACS summary value. This scenario is used in this paper to evaluate the how well models fit the ACS summaries for unconstrained variables. This serves as a proxy for Scenario C, as shown in Table 2, where models are used to create estimates for which no ACS summaries exist. Scenario C is the situation faced by researchers and planners for whom these models were developed.

Table 2. Three scenarios describe the relationship between ACS summaries used as Benchmarks and Model Estimates at the block group geography level.

ACS Summary (Benchmark)					Model Estimate			
Scenario A:				Known Constrained				
HH Size	1	2	3	Model Estimate forced to equal ACS Summary (Benchmark)	HH Size	1	2	3
BG 1	10	16	8		BG 1	10	16	8
BG 2	3	7	6		BG 2	3	7	6
Scenario B:				Known Unconstrained				
HH Vehicles	1	2	3	Model Estimate **NOT** forced to equal ACS Summary (Benchmark)	HH Vehicles	1	2	3
BG 1	10	16	8		BG 1	12	13	7
BG 2	3	7	6		BG 2	6	4	11
Scenario C:				Unknown				
HH owner's race	Black	White	Other	No ACS summary (Benchmark) Available	HH owner's race	Black	White	Other
BG 1	?	?	?		BG 1	15	1	4
BG 2	?	?	?		BG 2	5	6	9

The models are compared using methods introduced by Ruther et al. (2013) for model evaluation and validation and by using Moran's I. The methods used by Ruther et al. are: error in margin, residuals, and standard allocation error. I use these methods directly or in an altered form. The error in margin is useful for understanding the difference between the model allocation of variables over the entire study area and the summary table values. This measure allows for a general understanding of model performance, even for variables whose known higher resolution distribution is unavailable (Ruther et al. 2013). The residuals and standard allocation error (SAE) allow for more detailed comparisons of the model allocations with the actual population distribution at various scales (Ruther et al. 2013).

The residuals are calculated as $M_i - ACS_i$ where M_i is the model estimate for block group i and ACS_i is the ACS summary used as truth for block group i. The residuals are calculated for each block group for each model and compared at the block group level. The order in which I calculate the residuals makes the interpretation intuitive; Negative values indicate that the model estimate is lower than the ACS benchmark value for that block group, while positive numbers mean the model estimate was higher. I have adapted the Standard Allocation Error used by Ruther et al. because I am evaluating the measure for each block group instead of at the PUMA level. The SAE used by Ruther et al. is $\frac{\sum_i |ACS_i - M_i|}{\sum_i ACS_i}$ where the sum of absolute residuals is standardized by the sum of the ACS summaries. I adapted the formula in two ways: first, the order is the same as used in our residuals so that interpretation is intuitive and, second, I incorporated the margin of error associated with the ACS summaries because full enumerations are not available. The modified equation for Standardized Errors is $\frac{M_i - ACS_i}{moeACS_i}$ where the margin of error for the ACS benchmark in block group i is used for standardization. This allows for the uncertainty of the ACS benchmark to be incorporated and direction of the sign indicates whether the fit was high or low.

Normalization of the residuals is also used. The ACS summaries at the PUMA level used in the Error in Margin calculation are often available and when this higher level number is available, it provides an estimate to which modelled estimates at the block group level can be normalized. The formula for this process is: $M_i \left(\frac{\sum M_i}{\sum ACS_i}\right)$ for each block group model estimate M_i for the entire PUMA. This normalization keeps the population distribution from the models from being vastly different than the known higher level number. This normalization to a common number is helpful in this paper because it allows comparisons to be relative rather than absolute.

The Standardized Errors for each model are also mapped. Mapping the spatial structure of model performance helps identify how model strengths and weaknesses vary through space. This allows for quick identification of potentially anomalous areas. The Moran's I is also calculated for these mapped Standardized Errors in order to evaluate the residual patterns. Spatial autocorrelation in the residuals may indicate insufficiencies in the modeling techniques and/or potentially anomalous areas. These methods are newly applied to the Copula Model and applied to the maximum entropy model with different arrangements of variables on modern census data and thus are a novel application of this validation process. In the case of the Copula Model, this is the first time these methods will be applied. For the maximum entropy model, the validation procedures have been adapted to fit modern census

data where Ruther et al. (2013) used 1880 U.S. Census data. Additionally, I am applying the measures at the block group level. Visualization of the results is also an important part of this assessment, as patterns may appear which otherwise would be missed in a standard table comparison. Following Ruther et al. (2013), residuals are also mapped to allow easy tract to tract and block group to block group comparisons of model performance using the SAE. Application of these specific comparison techniques along with interpretation, recommendations, and suggestions for next steps represent the contribution of this work to the population modeling community.

Section 3: Data and Study Area

Data

For both the P-MEDM and Copula Model, Census ACS PUMS microdata samples and ACS five year summary files are used. The microdata for the 5 year interval represent a 5% sample which is much more desirable than the single year, 1% microdata samples because of the increased sample size and spatial resolution of input tables. The ACS five year summary tables are used to align with the temporal period represented by the microdata. Some counties in the U.S. have PUMA/tract/block group misalignment problems. These "boundary incongruity" (Voss et al. 1999) issues and methods for handling them are discussed by Zandbergen and Ignizio (2010) and will need to be considered for national implementation. Because this study focuses on comparison, additional data challenges such as boundary incongruity are avoided, thus allowing the results to be interpreted in a more straightforward manner.

The constraining variables used in the study are total household income and household size. Using these variables is beneficial because they are variables often used in the research community (Mohammadian et al. 2010, Beckerman et al. 1996). A non-constraining variable, the number of vehicles per household, is used for univariate evaluation and number of vehicles by household size is used for bivariate evaluation. In practice, the variable combinations to be modeled will necessarily be those for which estimates do not exist at the desired level of geography. However, for evaluation and comparison purposes, it is necessary to model variables for which summaries exist at the desired level of geography.

Study Area

The location of the study is Jefferson County, Kentucky, which encapsulates the city of Louisville. There are five Public Use Micro Areas (PUMAs) within the county. This county offers a mix of landscapes with primarily urban PUMAs, as well as PUMAs with a mix of urban and rural. This allows for the evaluation of the models with different types of geographic boundaries. I examined the structure of the variables used in the modeling to determine if the area was suitable for model implementation and comparison. As shown in Figure 3, the structure of household income among and within the PUMAs in Jefferson County, Kentucky seems to have some variability. PUMA 1701 has a much greater proportion of the population in the lower income classes than the rest of the PUMAs. The distribution within the PUMAs is also heterogeneous. This indicates that Jefferson County is capable of offering the spatial variability necessary to vigorously test robustness of the models.

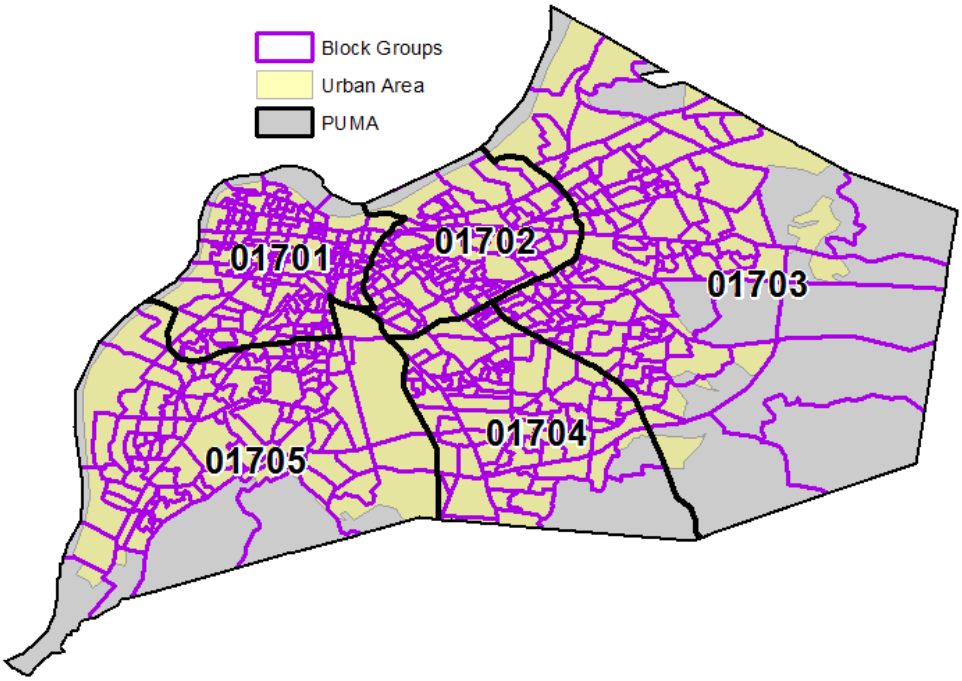

Figure 2. Jefferson County, Kentucky has several PUMAs with a mix of urban and rural areas which provides the variation in block group type necessary to compare model performance.

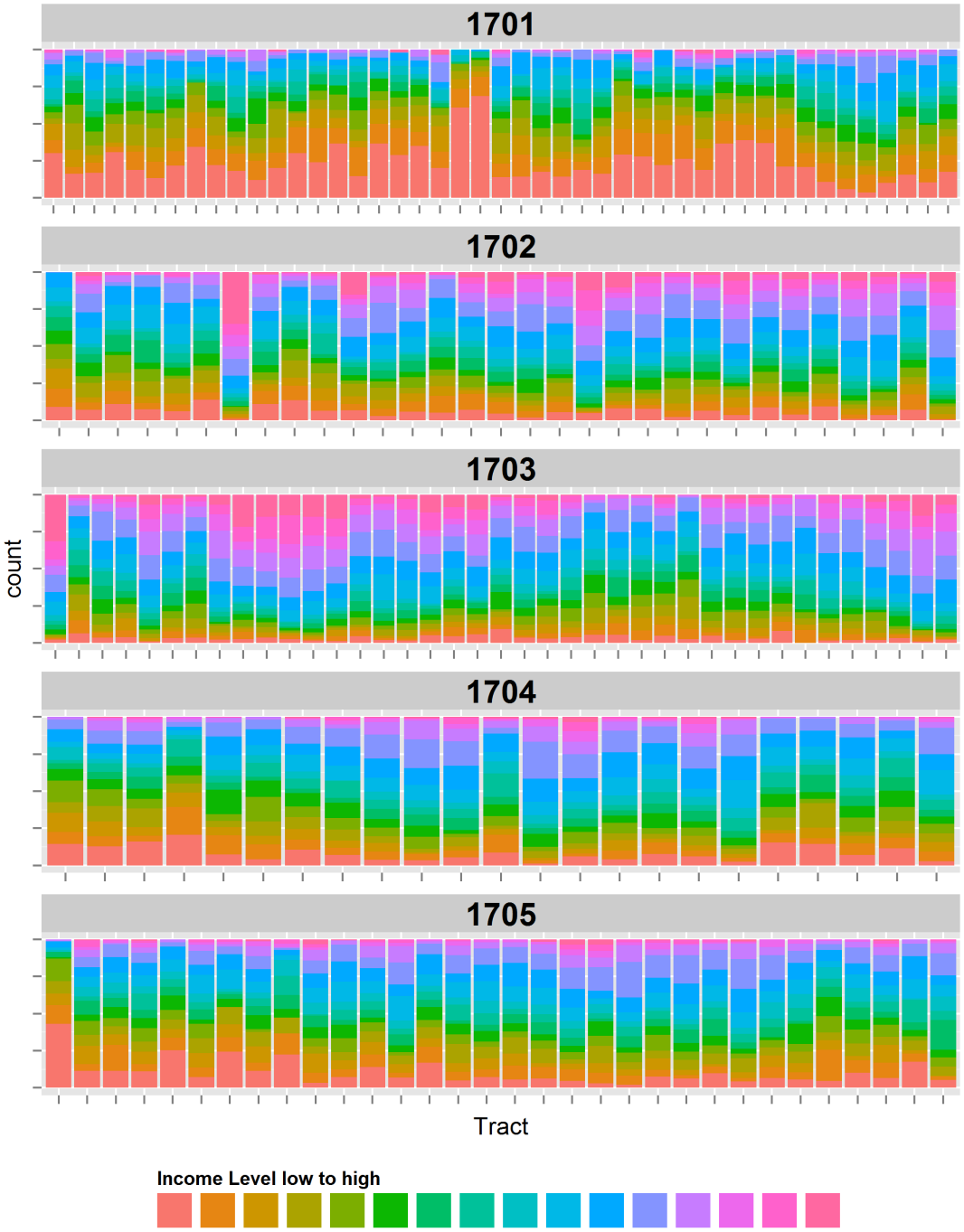

Figure 3. Tract Level income distributions for all five PUMAs show variation within and between PUMAs.

Section 4: Results and Evaluation

Univariate: Single Vehicle Households

Error in Margin

The first measure, error in margin, is a PUMA level measure that shows how well each model predicts the overall count of a variable. Table 3 shows the counts for each model in each PUMA and the ACS benchmarks of single vehicle households. Also shown in Table 3 are the PUMA level margins of error limits for the ACS benchmark for single vehicle households. The P-MEDM output is much closer to the ACS Benchmark value while the Copula Model results are outside of the margin of error for the PUMA level ACS benchmark. These PUMA level summaries, used as benchmarks for this study, are often available for multivariate combinations where block group and tract level summaries are not available. As such, this is the only level of validation available for these models.

Table 3. Error in Margin for single vehicle households for all PUMAs

PUMA	Model output		Benchmark	ACS Benchmark Margin of Error	
	Copula	P-MEDM	ACS benchmark	Low	High
1701	21,188	23,546	23,655	22,655	24,655
1702	17,734	19,629	19,617	18,896	20,338
1703	28,016	30,596	30,742	29,776	31,708
1704	14,487	16,203	16,691	15,908	17,474
1705	18,411	19,570	19,693	18,832	20,554

Residuals

The next measure, the residuals, shows how far above or below the ACS benchmark each of the model's outputs reach. This is a simple measure, but it is especially useful because it retains the units of the data which allows for a qualitative assessment of the fit. For example, a model estimate of 200 households for an area with a "true" value of 160 households may be within the acceptable bounds of the ACS benchmark margin of error (+/- 50) but the researcher may feel that 40 is unacceptable. I

calculated residuals for each block group, for all five PUMAs. The Error in Margin showed that the Copula Model estimates were lower that the P-MEDM and the ACS benchmark for all five PUMAs. They were outside of the margin of error. This indicates that the residuals at the block group level may be systematically lower, too. To make the residual comparison more realistic and useful for relative comparison, the modeled estimates are scaled, i.e. normalized, to the ACS benchmark for the PUMA level. This is a step that would likely be taken in practice and prevents the analysis at the block group level from being a direct reflection of the fact the PUMA overall total for the P-MEDM model is higher than the Copula Model, as seen in the Error in Margin. Figure 4 shows the residuals for all block groups in PUMA 1702. The gray line represents the margin of error for each block group. Plotting the raw residuals allows the researcher to evaluate the fit of the models in the units of the estimate. In Figure 4, most of the model estimates for single vehicle households are within 100 households of the ACS benchmark and almost all of the modeled results are also within the margin of error line. There is one case, block group 0106002, where the model results are 200 households below the ACS benchmark. This is close to the margin of error for that block group, but still may not be an acceptable result for the researcher.

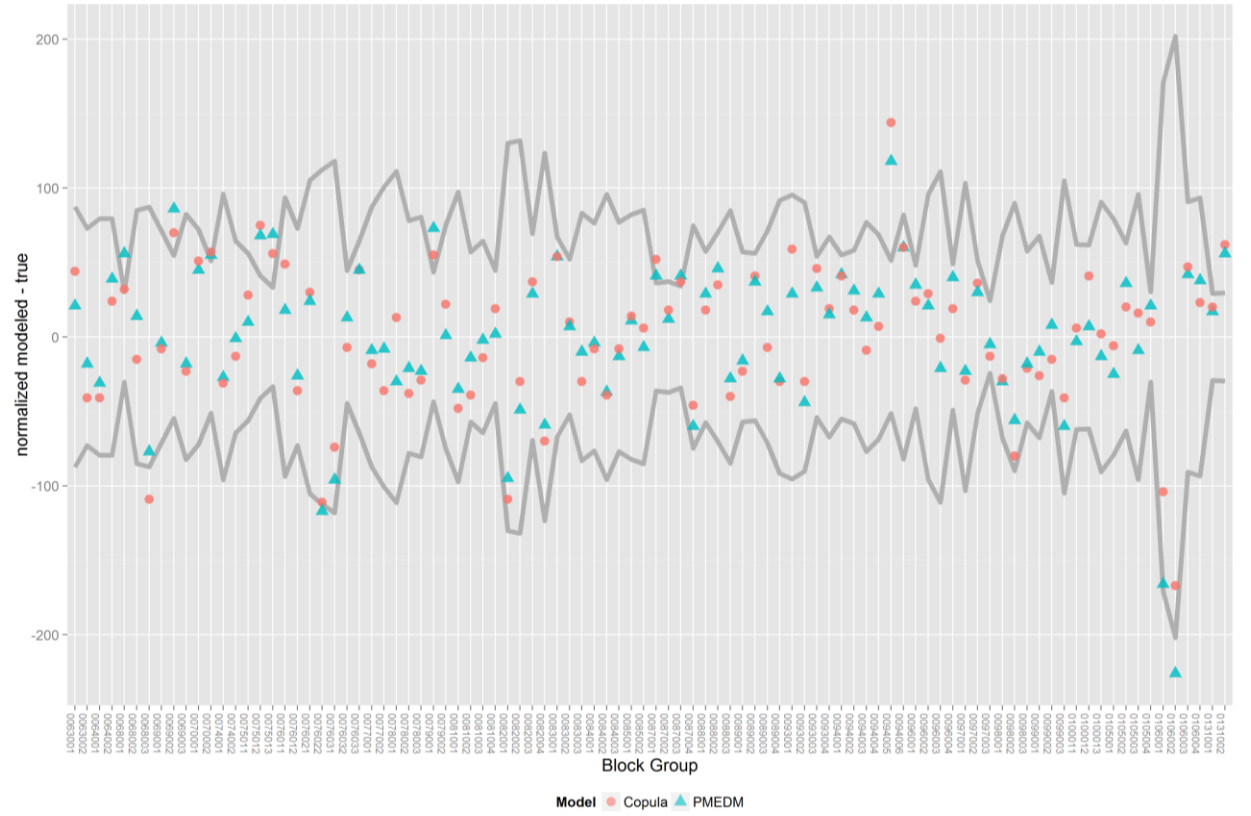

Figure 4. The normalized residuals for all block groups in PUMA 1702.

It is also useful to think about these normalized residuals as a whole. For PUMA 1702, shown in Figure 4, the P-MEDM fitted results are closer to the ACS benchmark twice as often as the Copula Model results. Table 4 shows the count of times each normalized model was closer to the ACS benchmark by whether the block group was urban or rural. However, Figure 4 shows that "closer" is not considerably different. The P-MEDM model had more estimates that were closer to ACS benchmark in four out of five PUMAs. The one PUMA where the Copula Model had more results that were closer to the ACS benchmark was PUMA 1703 which has a mix of urban and rural block groups while the others are much more urban. Interestingly, the P-MEDM fit rural blocks more closely almost every time in that PUMA. Table 4 also shows the percentage and total number of times each model was outside the margin of error for each PUMA. These numbers are all fairly close with the P-MEDM having slightly more extreme values. The total times and percent of times each model was outside the margin of error by urban and rural is also shown in Table 4. Although the sample size of 20 rural blocks is a small number, both models were outside the bounds of the margin of error twice as often, which is certainly an important

observation. This indicates that because there is a mixture of urban and rural, the minority rural blocks are being fitted more poorly. It is possible that these models would fit the urban areas more poorly in mixed PUMAs that are predominately rural with some urban. Additional testing needs to be done to test this scenario. It is important to note that the models tend to trend together; when one is high, the other is high and neither accounts for all of the extremes.

Table 4. Count of times each model was closer to the ACS benchmark for single vehicle households in each block group by PUMA and urban/rural designation and number and percentage of times each model was outside margin of error.

	Closer to ACS Benchmark					Outside Margin of Error			
	Urban (N=536)		Rural (N=20)			Count		Percentage	
PUMA	Copula	P-MEDM	Copula	P-MEDM		Copula	P-MEDM	Copula	P-MEDM
1701	70	74	3	0		19	18	12.9	12.2
1702	34	60	0	0		11	12	11.7	12.8
1703	67	49	1	10		18	22	14.2	17.3
1704	35	44	0	2		14	15	17.3	18.5
1705	43	60	3	1		19	17	17.8	15.9
				Urban		74	77	13.8	14.4
				Rural		7	7	35.0	35.0

Spearman's Rank correlations were also calculated between model outputs and between the each model output and the ACS benchmark. Table 5 shows these correlations. The P-MEDM model is more highly correlated with the ACS benchmark in all PUMAs, although not by a great margin. The high correlation between the model outputs underlines the relationship between models at the block group level shown in Figure 4 of the residual plots.

Table 5. The Spearman's Rank Correlations between models and between models and ACS Benchmark for single vehicle households by PUMA.

PUMA	Copula/P-MEDM	Copula/ACS	P-MEDM/ACS
1701	0.9798	0.8366	0.8400
1702	0.9788	0.8709	0.8918
1703	0.9870	0.9267	0.9337
1704	0.9881	0.9216	0.9339
1705	0.9812	0.8417	0.8680

Standardized Error

The Standardized Error allows us to understand when model fit is extremely outside of normal. The Standardized Error is the residual divided by the margin of error. In this paper the margin of error is used to account for the uncertainty present in the ACS benchmark estimates. Figure 5 shows the block group distribution of Standardized Error for PUMA 1704. The Standardized Error is similar to a z-score which means the expected value is usually between -2 and 2. In PUMA 1704, as in all of the other PUMAs, almost all values are within that range. Figure 5 shows that the models tend to trend together with the high values being no exception. Table 6 shows the summary statistics for the Standardized Error values for both models in all PUMAs. PUMA 1701 had one block group with an extremely high value, pulling the max to double all of the other PUMA max values. In this case, the ACS truth was close to zero with the modeled values over 100.

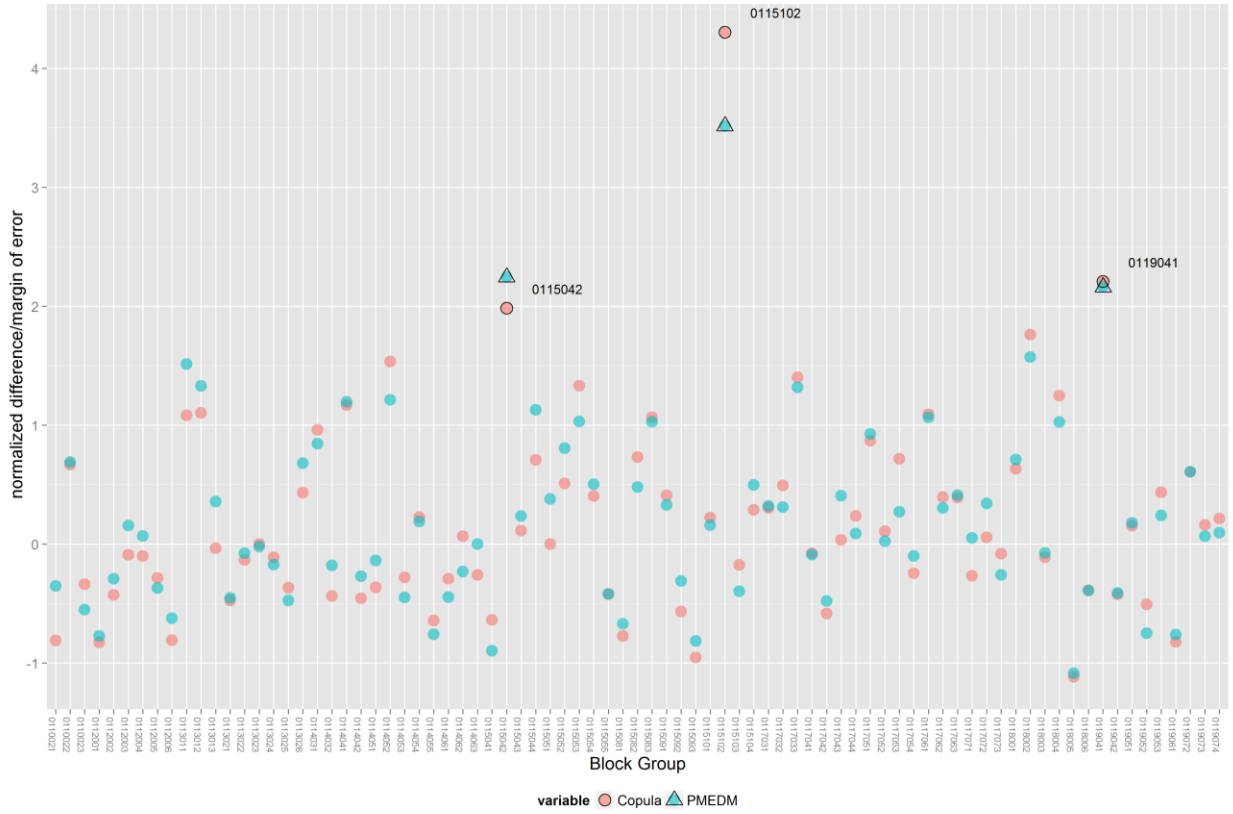

Figure 5. Block group distribution of Standardized Error for PUMA 1704.

Table 6. Summary statistics for the Standardized Error values for both models in all PUMAs.

	Copula					P-MEDM				
PUMA	Min	Max	Mean	Median		PUMA	Min	Max	Mean	Median
1701	-1.11	8.23	0.19	-0.06		1701	-1.19	8.91	0.17	-0.01
1702	-1.12	2.29	0.18	0.08		1702	-1.25	2.80	0.15	0.08
1703	-1.59	4.07	0.23	0.14		1703	-1.72	3.35	0.19	0.05
1704	-1.09	3.52	0.23	0.10		1704	-1.12	4.31	0.21	0.06
1705	-1.22	3.91	0.21	0.01		1705	-1.55	4.04	0.19	-0.04

Moran's I

Moran's I is used to assess whether the spatial distribution of a set of values is not random (Moran 1950). The null hypothesis is that the distribution is random, so a significant p-value means that the distribution shows spatial autocorrelation, i.e. it is different from random. For this study, Moran's I tests were conducted for the Standardized Errors, described in the Evaluation Methods section, instead of the residuals because differences in residuals between areas are affected by population differences which bias the test (Waldhor 1996). Table 7 shows the results of the Moran's I for each model in each PUMA. Overall, both models show similar results for each PUMA. Only PUMA 1702 has a significant p-value for either model, indicating the distribution is not random. PUMA 1705 was close to significant while the rest were not. One interesting note is the sign change on the Moran's I for PUMA 1704. The others are leaning in the positive direction, including PUMA 1702, where there is a non-random clustering relationship between block groups Standardized Errors. Figure 6 is a map of the P-MEDM Standardized Errors for PUMA 1704. There are several instances of very high differences near very low differences, which may cause the negative lean in the Moran's I statistic. Figure 7 is a map of the Copula Model Standardized Errors for PUMA 1702. In this map, the high values are often paired with other high values and the lowest values are often surrounded by other low values.

Table 7. Moran's I for each model in each PUMA.

PUMA	Model	p-value	Moran's I statistic	Expected value
1701	Copula	0.1446	0.0388	-0.0068
	P-MEDM	0.2052	0.0295	
1702	Copula	0.0132	0.1260	-0.0108
	P-MEDM	0.0541	0.0881	
1703	Copula	0.2814	0.0230	-0.0079
	P-MEDM	0.3387	0.0143	
1704	Copula	0.8026	-0.0682	-0.0125
	P-MEDM	0.6974	-0.0468	
1705	Copula	0.0836	0.0700	-0.0094
	P-MEDM	0.0906	0.0675	

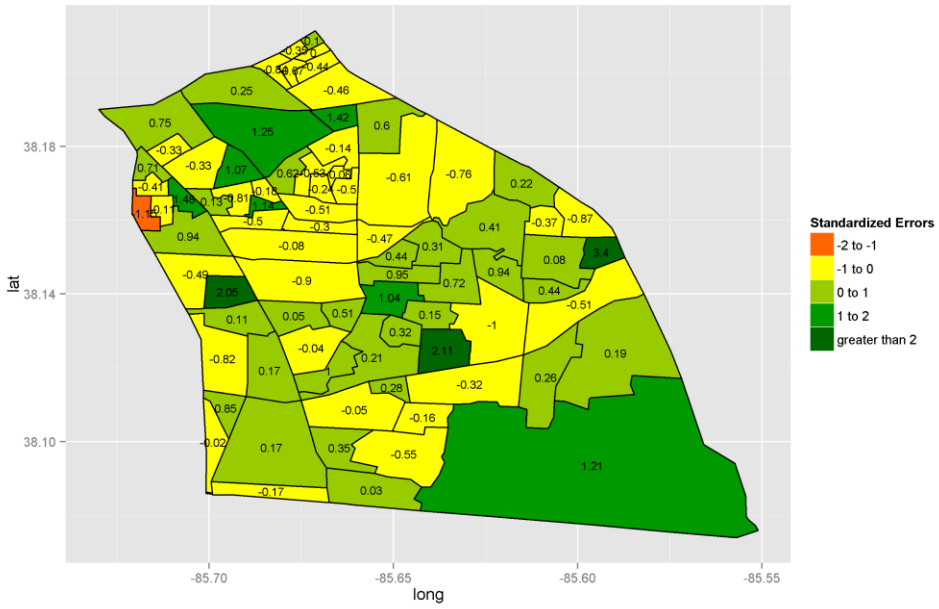

Figure 6. P-MEDM Standardized Errors for PUMA 1704 show dispersion indicated by Moran's I.

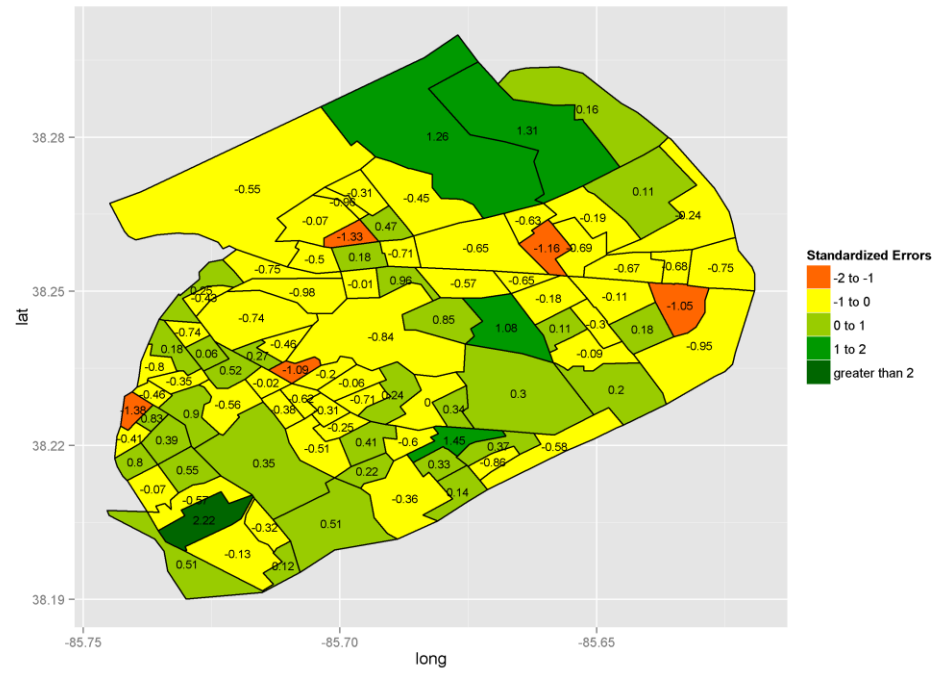

Figure 7. Copula Model Standardized Errors for PUMA 1702 show some clustering as indicated by Moran's I.

Bivariate: Two person household with one vehicle available

For the first set of tests I used a single variable. In reality, multiple variables will be more desirable, because the joint distributions of these are the ones that are not available at the small geographies such as the block group or tract. There are fewer summaries available for multivariate combinations so modeling is more important for these cases. For the second set of tests, the combination of household size and vehicles available is used. Examples and comparisons are made using two person households with one vehicle available.

Table 8 shows an example contingency table distribution from which this variable combination is a subset. The estimate for each tract is collected from a contingency table and used in this analysis. Because block group summaries are unavailable for use as benchmarks, tract level summaries, which are spatially coarser, are compared. The limited bivariate combination selection for use in this comparison illustrates the need for alternative methods for estimating populations at these small geographies.

Table 8. A contingency table shows the bivariate distribution for PUMA 1701. The subset of variables used in the bivariate analysis is highlighted.

PUMA 1701		Vehicles Available						
		0	1	2	3	4	5	6
Household Size	1	9922	8038	2959	536	80	7	2
	2	3932	6676	4205	982	143	7	3
	3	1323	3252	2675	772	142	11	2
	4	748	2051	1947	708	193	10	2
	5	179	677	798	315	99	10	0
	6	89	337	416	193	62	8	7
	7	39	127	172	88	26	4	1
	8	4	28	49	35	31	2	1
	9	1	0	2	3	1	0	0
	10	0	0	1	1	0	0	0
	11	0	2	6	6	11	0	0
	12	0	0	0	0	1	0	0

Error in Margin

The error in margin is a PUMA level measure that shows how well the overall count is estimated by the models. As seen in Table 9, the Copula Model is estimating much higher numbers of two person one vehicle households than either the P-MEDM or the ACS Benchmarks for the PUMA. Also for all of the PUMAs other than 1701, the Copula Model estimate exceeds the high end of the margin of error for

the ACS Benchmark while the P-MEDM estimate remains within the margin of error. The estimate produced by the copula model is especially high for PUMA 1703. As with the single variable, one of the models is consistently higher than the other. It should be noted that when a higher level summary is available, it is suitable to scale, i.e. normalize, to that summary number so that relative differences can be compared with the other metrics.

Table 9. Counts of two person single vehicle households in each PUMA for both models and the ACS Benchmark show Error in Margin.

PUMA	Models		Benchmark	Margin of Error Of Benchmark	
	Copula	P-MEDM	ACS	Low	High
1701	6,676	5,715	5,912	5,092	6,732
1702	5,776	3,448	3,604	2,979	4,229
1703	10,260	6,377	6,257	5,456	7,058
1704	5,353	3,676	4,114	3,400	4,828
1705	6,604	4,503	4,675	3,946	5,404

Residuals

The residuals show tract level deviation from the ACS Benchmark and provide an assessment of individual level fit. These were calculated at the tract level for all five pumas. There are no block group level summaries for this bivariate combination, so tract level summaries are used instead. Again, the model estimates are normalized to the ACS Benchmark for a relative comparison. This is done to overcome the absolute differences shown in the Error in Margin. Figure 8 shows the normalized residuals for all tracts in PUMA 1705. The gray line represents the margin of error for each tract. There are pairs of points for each tract. Most points fall within +/-50 of the estimate and a few of the points fall outside of the margin of error. This plot clearly shows that the models trend together. When one model is high, the other is similarly high. This same trend is seen in the other PUMA plots as well. Most instances of points falling above or below the margin of error line are similar to the points for tract 009000; the points are over 100 households from the ACS Benchmark, but the margin of error is very high as well. The remaining instances are similar to tract 012203, where the model estimates are over

100 households from the ACS Benchmark, but the margin of error is much lower. These are the cases where the results indicate that there may be underlying processes that are not being considered, which may be a problem for the researcher. For example, this tract is an urban tract that borders a rural tract, which may indicate change in the structure of the population at the tract level.

Table 10 shows the count of normalized model estimates that are closer to the ACS Benchmark for each PUMA by urban/rural. Also shown are the count and percent of times each normalized model estimate was outside the margin of error of the ACS Benchmark. The Copula Model often has more tract estimates that are closer to the ACS Benchmark. However, as Figure 8 shows, the closer model is often not especially different or better than the other. Both models have similar counts of times outside the MOE. Table 10 also shows that, when considering all PUMAs, both models are outside the margin of error over twice as often in rural tracts versus urban tracts. This indicates some difference between each model's ability to fit in the urban and rural areas in these PUMAs. In this particular case, having a small number of rural tracts in a mostly urban PUMA may make it harder for any model to fit the benchmark. This ratio of urban to rural geographies is not uncommon and these situations will need to be handled carefully when modeling at the nation scale, no matter which model is used. More broadly, there may be issues when any type of area is a minority type in a given PUMA. Examples may include PUMAs with mostly rural areas and a central town, colleges in small communities, as well as military bases and prisons where the populations would be very different than the surrounding populations.

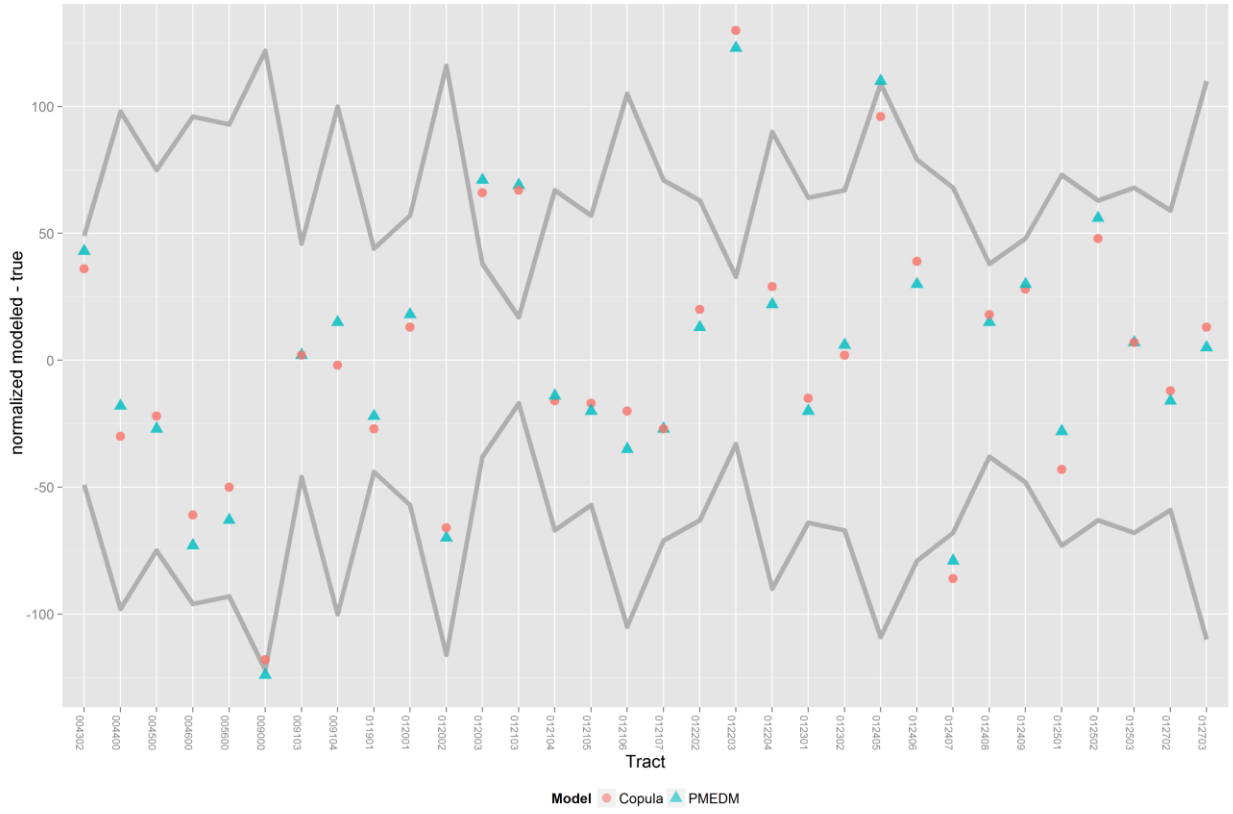

Figure 8. Normalized residuals for tracts in PUMA 1705.

Table 10. The number of tracts where each model was closer to the ACS Benchmark for all PUMAs, broken down by urban/rural designation.

	Closer to ACS Benchmark					Outside Margin of Error			
	Rural (N= 11)		Urban (N=159)			Count		Percentage	
PUMA5	Copula	P-MEDM	Copula	P-MEDM		Copula	P-MEDM	Copula	P-MEDM
1701	0	1	22	22		5	3	11.1	6.7
1702	0	0	19	12		3	3	9.7	9.7
1703	2	4	14	19		7	7	17.9	17.9
1704	1	0	13	9		2	2	8.7	8.7
1705	1	2	17	12		4	6	12.5	18.8
				Urban		18	18	11.3	11.3
				Rural		3	3	27.3	27.3

The Spearman's Rank correlation was assessed between the model outputs and between each model output and the ACS Benchmark as an additional comparison. Table 11 shows these correlations for all five PUMAs. The models themselves are highly correlated, while the P-MEDM model was always slightly more correlated with the ACS Benchmark. This reinforces the closeness of the modeled results shown by the residual plots in Figure 8.

Table 11. Spearman's Rank correlation for all 5 PUMAs for two person single vehicle household counts.

PUMA	Copula/P-MEDM	Copula/ACS	P-MEDM/ACS
1701	0.9267	0.7227	0.7366
1702	0.9428	0.7123	0.7433
1703	0.9697	0.6688	0.6930
1704	0.9686	0.6113	0.6513
1705	0.9767	0.7729	0.7790

Standardized Error

The Standardized Error allows us to understand when model fit is extremely outside of normal. Figure 9 shows the tract distribution of the Standardized Error for PUMA 1705. Much like a z-score, the expected value is somewhere between -2 and 2. Most tracts in all of the PUMAs have values within this range. PUMA 1705 has some exceptions. One notable exception is tract 012203. This is a tract where the residual plot in Figure 8 showed great disparity between the model estimate points and the margin of error. One other tract, 012103, has a similarly high Standardized Error. Using both the Standardized Error and the residual plots allow for the identification of both tracts that are potentially issues with the residuals maintaining the context of the original units. Table 12 shows the summary statistics for the Standardized Error values for both models across all PUMAs. PUMA 1705, as shown in Figure 9 has the highest values and the rest of the PUMAs have maximum values fairly close to the upper expected value for the Standardized Error.

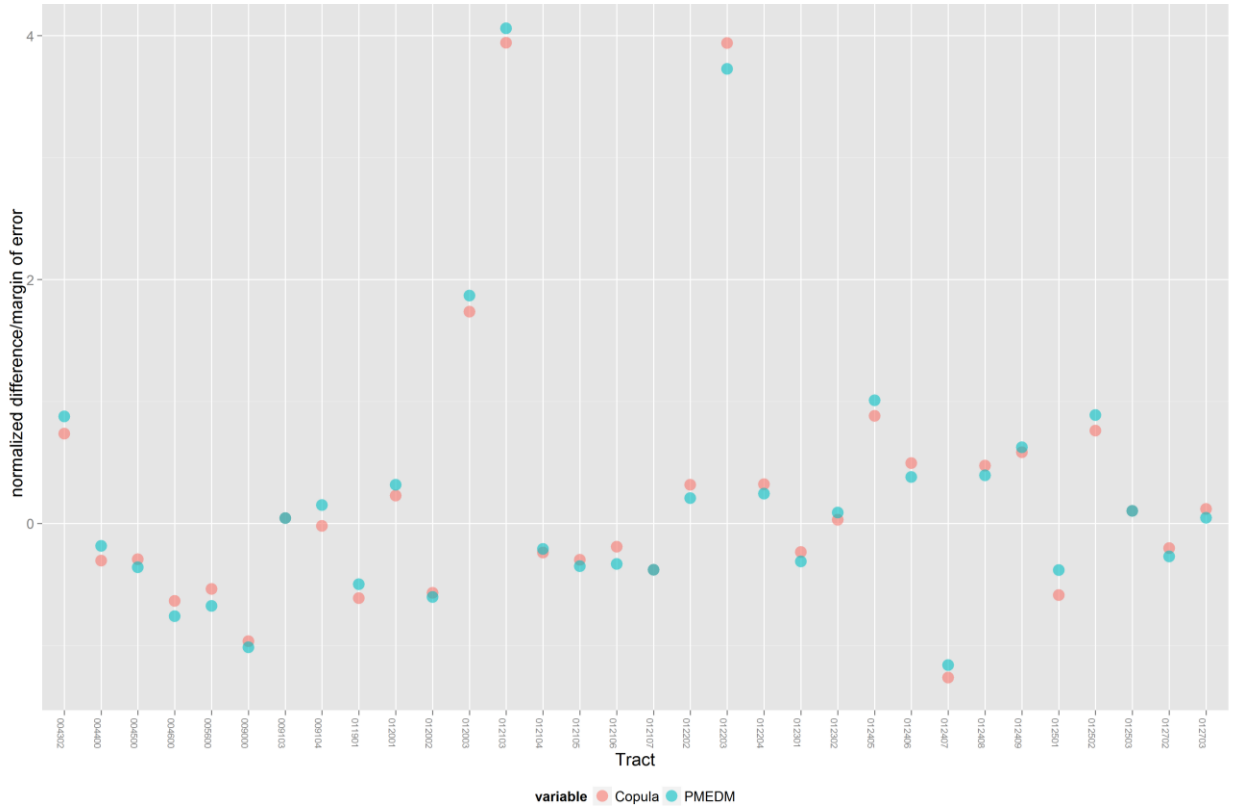

Figure 9. Normalized Standardized Errors for each tract in PUMA 1705 for two person single vehicle households.

Table 12. Summary Statistics for Standardized Errors for all PUMAs.

	Copula					P-MEDM			
PUMA	Min	Max	Mean	Median	PUMA	Min	Max	Mean	Median
1701	-0.99	1.29	0.07	0.05	1701	-1.14	1.83	0.08	0.00
1702	-0.96	2.78	0.13	-0.08	1702	-0.94	2.40	0.14	0.00
1703	-1.69	2.25	0.12	0.11	1703	-1.60	1.81	0.13	0.12
1704	-0.81	1.83	0.06	-0.20	1704	-0.89	2.25	0.07	-0.15
1705	-1.16	4.06	0.24	0.04	1705	-1.26	3.94	0.23	0.00

Moran's I

Moran's I tests were conducted for the Standardized Errors, which account for population, for each model in each PUMA to evaluate spatial autocorrelation. Results from the non-normalized Standardized Errors are reported here. Tests were done using the normalized data as well as row standardized weights matrix and the patterns were the same. Table 13 shows Standardized Errors for both models in PUMA 1703 as being significantly positively spatially autocorrelated. Weak amounts of spatial autocorrelation are seen in the Standardized Errors for both models in most of the PUMAs. PUMA 1702, however, is the exception. These Standardized Errors are more dispersed than expected according to the Moran's I statistic, but not at a significant level. Standardized Errors for PUMA 1702 for the Copula Model are shown in Figure 10. The highest Standardized Errors are not clustered and high and low values are often neighbors. Figure 11 shows a map of the Copula Model Standardized Errors for PUMA 1703, where the most significant p-value occurred. The lowest values are somewhat clustered in the west with most of the highest values occurring near the rural edge of the PUMA.

Table 13. Moran's I of Standardized Errors for all PUMAs for two person single vehicle households.

PUMA	Model	p-value	Moran's I statistic	Expected Value
1701	Copula	0.0900	0.0916	-0.0227
	P-MEDM	0.4461	-0.0110	
1702	Copula	0.7175	-0.0881	-0.0333
	P-MEDM	0.7711	-0.1041	
1703	Copula	0.0411	0.1370	-0.0263
	P-MEDM	0.0154	0.1728	
1704	Copula	0.0787	0.1037	-0.0455
	P-MEDM	0.1979	0.0482	
1705	Copula	0.1546	0.0604	-0.0323
	P-MEDM	0.3242	0.0096	

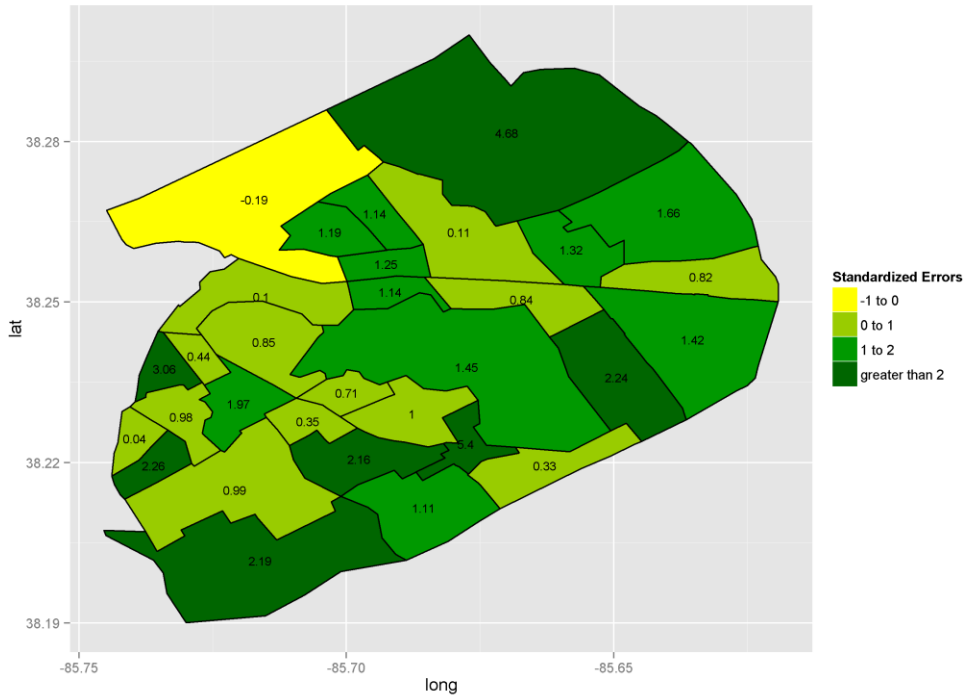

Figure 10. Standardized errors for Copula Model for PUMA 1702 show dispersion indicated by Moran's I.

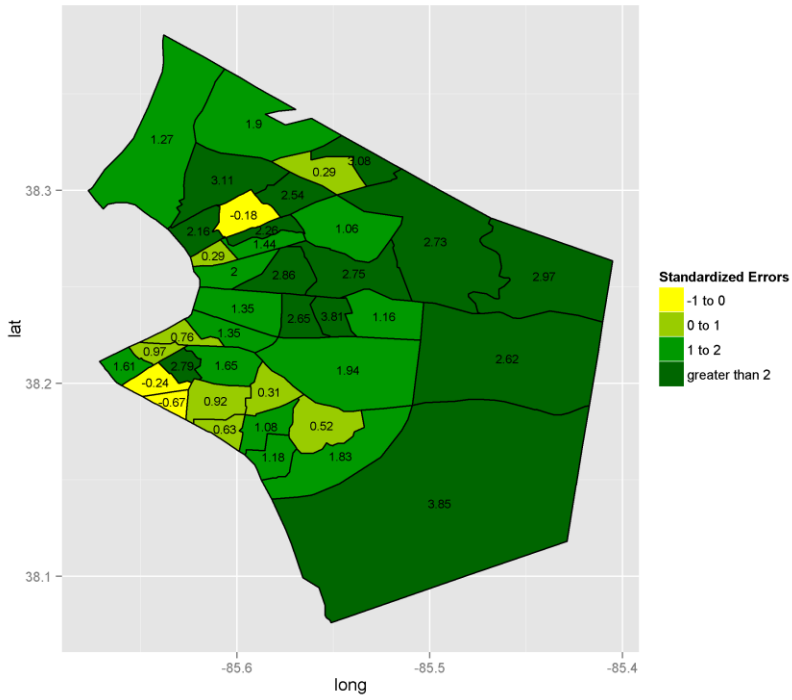

Figure 11. Standardized Errors for Copula Model in PUMA 1703 show clustering indicated by Moran's I.

Section 5: Summary and Further Considerations

Summary

First and foremost, when choosing the better of two options, one must decide what better means. For this study, fitting the estimates well is certainly important. The measures used here most certainly allow for that evaluation. In all PUMAs for both the univariate and bivariate tests, the P-MEDM model had a closer overall fit, as shown by the error in margin. Based on this, the P-MEDM model would be the better choice because it is more stable. There are some situations however, where an overall total may be known and used, making the model outputs effectively weights. Thus, the absolute fit is not as important as the relative fit. This relative fit is evaluated by the normalized residual plots in this paper and the correlations of the residuals. The better model according to these measures is less clear; both models were very similar in their relative distributions. These situations, while relevant, are specific, meaning the P-MEDM is still most often the better choice.

The spatial autocorrelation found using Moran's I helped indicate underlying differences between areas in the PUMAs, particularly in the case of PUMA 1703 where urban/rural differences may be contributing to the over or under prediction by the models. Further exploration at the block group and tract level may also indicate the limitations of the ACS estimates as a benchmark. Figure 12 shows an up close picture of the block group with the maximum Standardized Error in PUMA 1701. The housing structure of nearby block groups is similar but the ACS Benchmarks are very different. This indicates that the ACS estimates used as benchmarks may have limitations.

The overall performance of the P-MEDM indicates that this model is likely to be more stable when considering implementation at a national scale. This model takes far less time to compute and also provides more flexibility in the variable selection. However, as indicated by the Standardized Errors and the Moran's I, areas with variations in the make up of blocks groups and tracts may prove problematic, as was the case in PUMAs with uneven mixtures of urban and rural areas.

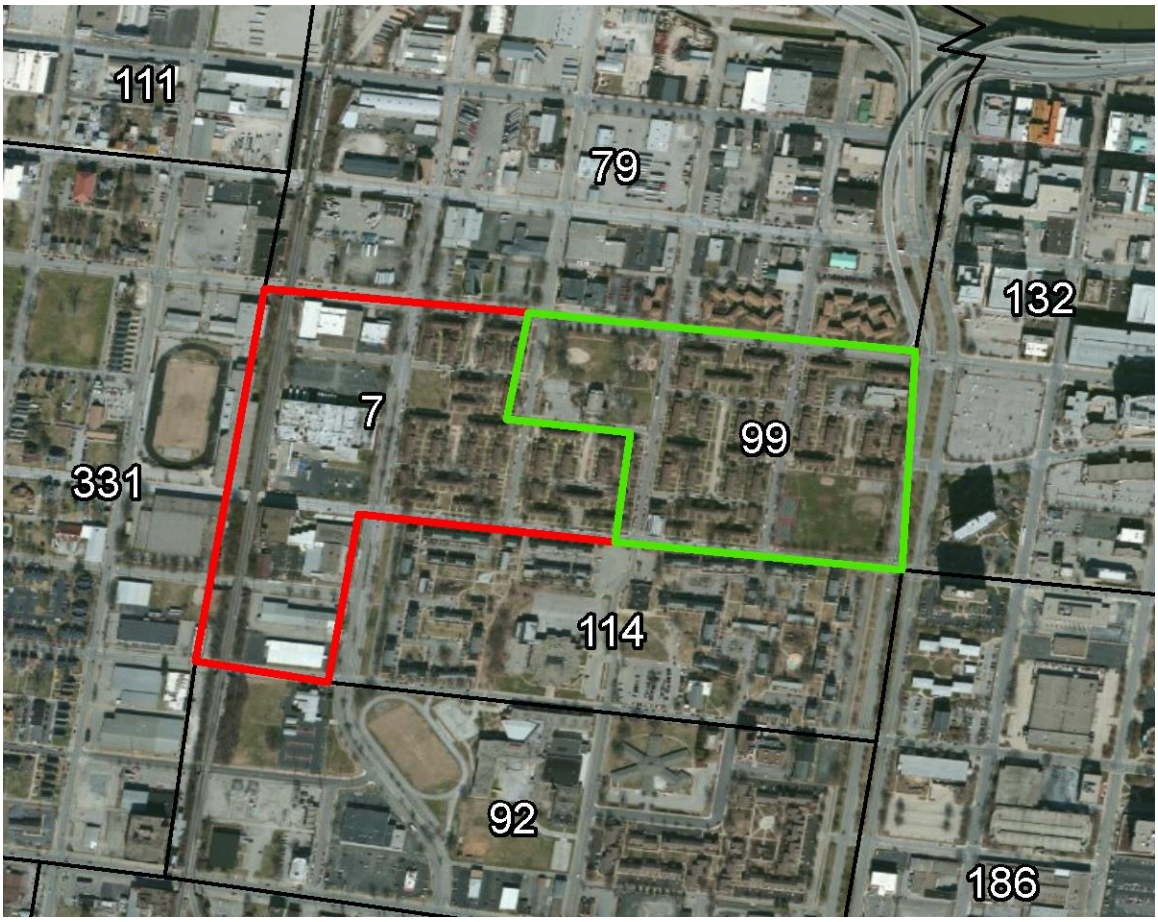

Figure 12. The block group in red has a very high Standardized Error. Nearby block groups with similar housing structures show very different ACS Benchmarks (labeled in white) indicating possible limitations of the ACS Benchmarks.

Further Considerations

Although this research produced a useful comparison of two promising models, there are some considerations that must be addressed to scale either model to a national implementation. An important problem is that geographic boundary alignment issues occur frequently. Most happen when areal units are not required to nest, which results in a single small area occurring in multiple large areas, i.e. one block group overlapping two PUMAs. For the data types used in this paper, alignment issues exist between PUMA boundaries from the 2000 vintage and the block group and tract boundaries, especially those from the 2010 census, because pre-2010 PUMAs were not required to nest with other Census geographies. These "boundary incongruity" issues have been discussed and remedies have been explored (Voss et al. 1999, Zandbergen and Ignizio, 2010). This issue affects all releases of ACS data. The 2010 PUMA boundaries were designed to fix this issue, but data releases, such as the 3-year and 5-year

summaries, spanning years where two sets of boundaries were used, will have misalignment issues as well. This means that five year microdata are likely to have two sets of underlying geographies. A possible solution would be splitting these data, fitting a model, and combining the results, which would double the effort for a national implementation.

Another consideration of the data produced by either model is that when a model over estimates in one area, it will underestimate somewhere else. This is true because the population balance is maintained at the higher level. Figure 13 shows this relationship for the bivariate model outputs in PUMA 1701. I found and discussed in Section 4 an over estimation of two person single vehicle households by the Copula Model. Figure 13 shows that other single vehicle counts for the Copula Model were under estimated as were the two person two vehicle households. Scaling (normalizing) at the PUMA level when possible is a way this issue could be mitigated. By addressing these further issues and exploring the areas where the comparisons indicated deficiencies in the modeling, national level implementation will be highly achievable.

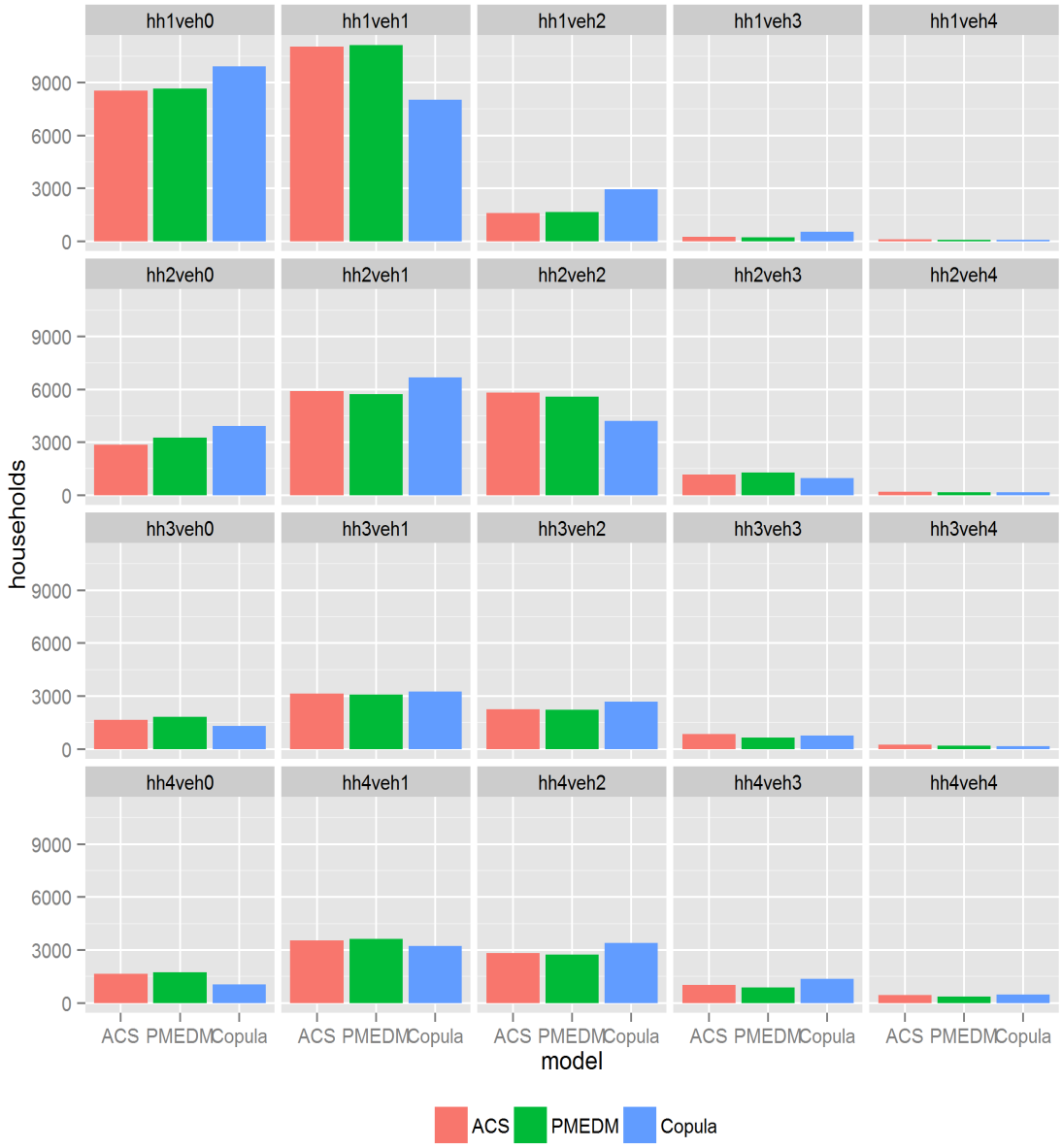

Figure 13. The full joint distribution of household size and vehicles available for PUMA 1701 has 20 cases. When a model estimate is high in one combination of variables, it will be low elsewhere.

Finally, the Copula Model as implemented in Kao et al. (2012) does not take advantage of the weights that are provided with the PUMS data by the U.S. Census. These weights play an important role in the P-MEDM model and they are meant to give a complete representation of structure of all of the households in the PUMA. Some microdata households in the PUMA are expected to occur more often than others, as indicated by the associated weight. Ignoring these weights effectively gives each

microdata household equal weight, which distorts the relationships between the PUMS microdata households. Table 14 shows the unweighted (upper right) and weighted (lower left) correlations between variables used in the Copula Model in PUMA 1701. The weighted correlation values are lower 11 out of 15 times.

Table 14. Correlations between variables used in the Copula Model for 1701. Unweighted correlations, as used in the in Kao et al. 2012, are shown on the upper right. On the lower left, the weighted alternatives are lower 11 out of 15 times.

	Household Income	Persons In HH	Workers in Family	Vehicles Available	Max Education	Total Travel Time
Household Income		0.4206	0.5186	0.6047	0.3924	0.5476
Persons In HH	0.3231		0.5810	0.4149	0.1781	0.3966
Workers in Family	0.5324	0.5409		0.4619	0.3462	0.6551
Vehicles Available	0.5587	0.3105	0.4661		0.3047	0.3884
Max Education	0.4124	0.1470	0.3280	0.2975		0.3130
Total Travel Time	0.5449	0.3402	0.6558	0.3649	0.3036	

References

Anderson, Ben. “Estimating Small-Area Income Deprivation: An Iterative Proportional Fitting Approach.” In Tanton, R. e. (2013). Spatial Microsimulation: A Reference Guide for Users. In K. e. Edwards & SpringerLink (Eds.): Dordrecht : Springer Netherlands : Imprint: Springer.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating Synthetic Baseline Populations. *Transportation Research Part A: Policy and Practice*, 30(6), 415-429. doi: http://dx.doi.org/10.1016/0965-8564(96)00004-3

Birkin, M., & Clarke, M. (1988). SYNTHESIS -- a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A*, 20(12), 1645-1671.

Harding, A., Vidyattama, Y., & Tanton, R. (2011). Demographic change and the needs-based planning of government services: projecting small area populations using spatial microsimulation. *Journal of Population Research*, 28(2-3), 203-224. doi: 10.1007/s12546-011-9061-6

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

Kao, S.-C., Kim, H. K., Liu, C., Cui, X., & Bhaduri, B. L. (2012). Dependence-Preserving Approach to Synthesizing Household Characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, 2302(-1), 192-200. doi: 10.3141/2302-2

Levy, J. I., Fabian, M. P., & Peters, J. L. (2014). Community-Wide Health Risk Assessment Using Geographically Resolved Demographic Data: A Synthetic Population Approach. *PLoS ONE, 9*(1), e87144. doi: 10.1371/journal.pone.0087144

Lymer, S., Brown, L., Harding, A., & Yap, M. (2009). Predicting the need for aged care services at the small area level: The CAREMOD spatial microsimulation model. *International Journal of Microsimulation*, 2 (2), 27–42.

Mohammadian, A., Javanmardi, M., & Zhang, Y. (2010). Synthetic household travel survey data simulation. *Transportation Research Part C: Emerging Technologies*, 18(6), 869-878. doi: 10.1016/j.trc.2010.02.007

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric Modeling and Uncertainty. *Annals of the Association of American Geographers*, 104(1), 80-95. doi: 10.1080/00045608.2013.843439

Nelsen, R. B. (2006). An Introduction to Copulas (Second Edition.. ed.). New York, NY: New York, NY : Springer Science+Business Media, Inc.

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair Copula Constructions for Multivariate Discrete Data. *Journal of the American Statistical Association, 107*(499), 1063-1072. doi:10.1080/01621459.2012.682850

Ruther, M., Maclaurin, G., Leyk, S., Buttenfield, B., & Nagle, N. (2013). Validation of spatially allocated small area estimates for 1880 Census demography. *Demographic Research,*29(22), 579-616. http://www.demographic-research.org/Volumes/Vol29/22/ doi: 10.4054/DemRes.2013.29.22

Simpson, L., & Tranmer, M. (2005). Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software*. *The Professional Geographer*, 57(2), 222-234. doi: 10.1111/j.0033-0124.2005.00474.

Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris,* Vol. 8, pp. 229–231.

Tanton, R. e. (2013). Spatial Microsimulation: A Reference Guide for Users. In K. e. Edwards & SpringerLink (Eds.): Dordrecht: Springer Netherlands: Imprint: Springer.

Tobler, W. R. (1979). Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association, 74*(367), 519-530. doi: 10.2307/2286968

U.S. Census Bureau. 2006. Design and Methodology American Community Survey U.S. Government Printing Office, Washington, DC. Accessed 14 November 2013. https://www.census.gov/history/pdf/ACSHistory.pdf

Voss, P.R., D.D. Long, and R.B. Hammer. (1999). When census geography doesn't work: Using ancillary information to improve the spatial interpolation of demographic data. Center for Demography and Ecology, University of Wisconsin – Madison, Working Paper No. 99-26.

Waldhor, T. (1996). The Spatialautocorrelation Coefficient Moran's I Under Heteroscedasticity. *Statistics in Medicine*, 15(7-9), 887-892.

Wong, D. (1992). The Reliability of Using Iterative Proportional Fitting Procedures. *The Professional Geographer*, 44: 340-348.

Zandbergen, P. A., and Ignizio, D. A. (2010). Comparison of Dasymetric Mapping Techniques for Small-Area Population Estimates. *Cartography and Geographic Information Science*, 37(3), 199-214. doi: 10.1559/152304010792194985

Trivedi P.K and Zimmer, D. M. (2005). "Copula Modeling: An Introduction for Practitioners." Foundations and Trends in Econometrics 1(1): 1-111.

Vita

Jessica grew up in East Tennessee and completed her undergraduate degree in Geography at the University of Tennessee in 2008. Since then, she has worked in various capacities with the Geographic Information Science and Technology (GIST) Group at Oak Ridge National Laboratory (ORNL) where she will continue to work upon completion of her degree. Her research centers around population distribution and dynamics at various scales globally. She is married and lives with her husband in Roane County, Tennessee.