Uncertainty in Demographic Small Area Estimates

Stefan Leyk[1], Barbara P. Buttenfield[1] and Nicholas N. Nagle[2]
[1]Department of Geography, University of Colorado, Boulder, CO, USA, {stefan.leyk},{babs}@colorado.edu
[2]Department of Geography, University of Tennessee, Knoxville, TN, USA, nnagle@utk.edu

***Abstract*—This paper describes a methodology to model uncertainties underlying the spatial allocation of demographic microdata to small areas such as census tracts or subunits of census tracts. The procedure models probable allocations for each single household to one of the small area units as derived from bootstrap analysis. The method allows evaluation of the uncertainty of the allocation dependent on the constraint variables used for the imputation of population weights which is based on empirical likelihood methods. Dasymetric mapping is demonstrated as an effective tool to spatially refine the allocation of households and shows great potential for more advanced spatial allocation processes.**

Keywords Small area estimates, uncertainty, dasymetric modeling, census data

I. INTRODUCTION

Small area statistics and spatially explicit demographic data are in increasing demand for establishing population distribution models at finer spatial scales. Increasing the accuracy of small area estimation procedures will improve understanding of spatial processes in a variety of problem domains.

Some initial experiments to model uncertainties underlying census demography are described in this paper. The aim of these experiments is a more reliable allocation of individual observations for which little or no locational data is available to one of a set of aggregation units, such as census tracts or landcover patches. Such a systematic exploration of locational uncertainty in the context of small area estimates has relevance to many types of complex spatial problems where social and natural systems interact.

The presented uncertainty analysis is based on empirical likelihood methods currently under development that allocate PUMS census microdata records to finer spatial resolutions than are publicly available (Nagle et al., 2010). Microdata records contain a multitude of demographic attributes at household resolution, but the spatial location is unspecified except to indicate which Public Use Microdata Area (PUMA) unit (roughly 25 census tracts; Fig. 1A). We allocate microdata records to a specific census tract within the PUMA via empirical likelihood methods, and to sub-tract resolution via dasymetric modeling with ancillary data. Our methods respect census tract summary statistics. We concentrate on the uncertainty analysis of the allocation results by quantifying measures of robustness, consistency and sensitivity as derived from bootstrapping. We demonstrate the allocation of microdata to sub-tract level based on dasymetric mapping that uses landuse and landcover as ancillary, limiting variables (Eicher and Brewer, 2001; Mennis and Hultgren, 2006). The implementation of associations between demographic attributes of allocated microdata households and different ancillary variables to develop more advanced dasymetric models is briefly discussed. The presented problem is a common instance of the spatial interpolation problem; and can be addressed for census demography only by inference, at present (Nagle et al., 2010).

II. DATA AND METHODS

We use data of the 2000 U.S. Census from Davidson County, TN, to demonstrate the allocation process, present some experimental results of the uncertainty assessment and some mapping outcomes from the dasymetric refinement.

A. Imputation of microdata weights

We present a reweighting approach (Williamson et al, 1998; Hynes et al, 2008), which overcomes the limitation of providing only a single allocation of micro data households to census tracts, but allows us to calculate allocation probabilities that determine how likely it is that a household is in each census tract given the tract-level summary statistics. These probabilities represent the quantifiable effect of uncertainty in the allocation problem. Uncertainty in this problem arises because the census summary statistics provide only aggregate locational information, and many different allocations of the microdata are possible with respect to the summary statistics. A method that accounts for this uncertainty, and allocates microdata stochastically, rather than deterministically, is appropriate.

A household record in the PUMS is written as $h_i = \{x_{i1}, \ldots, x_{iK}\}$, where h_i, $i = 1, \ldots, N$ represents the i-th household, and $\{x_{i1}, \ldots, x_{iK}\}$ represents the set of household- and individual-level characteristics, such as race, income, sex, etc. We determine the probabilities p_{ij} that household *i* is located in census tract *j* in the considered PUMA using *maximum empirical likelihood* (Owen, 2001) estimates. We constrain this solution by ensuring that the expected allocation of the microdata to census tracts preserves the tract-level summary tables.

Let the summary data for the *j*-th census tract be $t_j = \{x_{j1}, \ldots, x_{jK}\}$, where $\{x_{j1}, \ldots, x_{jK}\}$ represents the set of tract-level summary statistics. The constrained maximum likelihood can then be written as:

$$\max \prod_{i,j} p_{ij}; \quad subject\ to \sum_{i} p_{ij} x_{ik} = x_{jk} \qquad (1)$$

This research is funded by NSF Award Number 0961598 "Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata".

This set of constraints ensures that the expected allocation preserves the census tract summaries for the characteristics incorporated. The maximum empirical likelihood is found via the method of Lagrangian maximization to the constrained log likelihood function (Nagle et al., 2010).

B. Sensitivity and consistency of simulated microdata allocations

Using these estimated allocation weights p_{ij} we conduct random spatial allocations of the microdata households to census tracts and use bootstrapping to generate a frequency distribution for the allocation of each household to individual census tracts. The imputed weights used for allocation are based on an exemplary set of constraint variables as explained above.

To allocate a household we generate a random number between zero and the cumulative sum of imputed weights for this household over the total number of census tracts. The microdata household is allocated to the tract for which the random number is less than the cumulative imputed tract weight but greater than the cumulative imputed weight of the former tract. For example, if three census tracts carry weights of 0.24, 0.30 and 1.20, the random thresholds are 0.24, 0.54, and 1.74. The allocation process is repeated 1,000 times and frequency distributions are tabulated. To assess uncertainty inherent in this allocation process we analyze the sensitivity and consistency at two different levels.

First, we observe the number of microdata households that have been allocated in average to each census tract within the PUMA during simulation, and calculate summary statistics, which allow to estimate the robustness of the allocation process by comparison to tract summaries as well as the creation of boxplots. We create maps using average frequencies and their variations to visualize the spatial distribution of the uncertainty in our allocation results on the tract level (Fig. 1B).

Second, to evaluate the consistency and sensitivity of the spatial allocation on the household level, we create a random sample of 10 households and record how frequently they are allocated to each tract during the simulation. These frequencies are plotted to develop a better understanding of the household level situation. For each single household a probability surface is created to indicate the probability of this household to be allocated to the different census tracts in the PUMA. Based on the observed frequencies of individual microdata households, we derive measures to quantify the consistency and sensitivity of the allocation decision. We calculate a confusion index ($CI \in [0,1]$) to measure the robustness of the allocation decisions for each single household. The confusion index *CI* is modified from its form commonly used in fuzzy set theory (Burrough et al. 2001) and has the form:

$$CI_i = (F_{Max2_i} / F_{Max1_i}) \qquad (2)$$

Where F_{Max} and F_{Max2} are the two highest normalized allocation frequencies in the range [0,1] that could be found

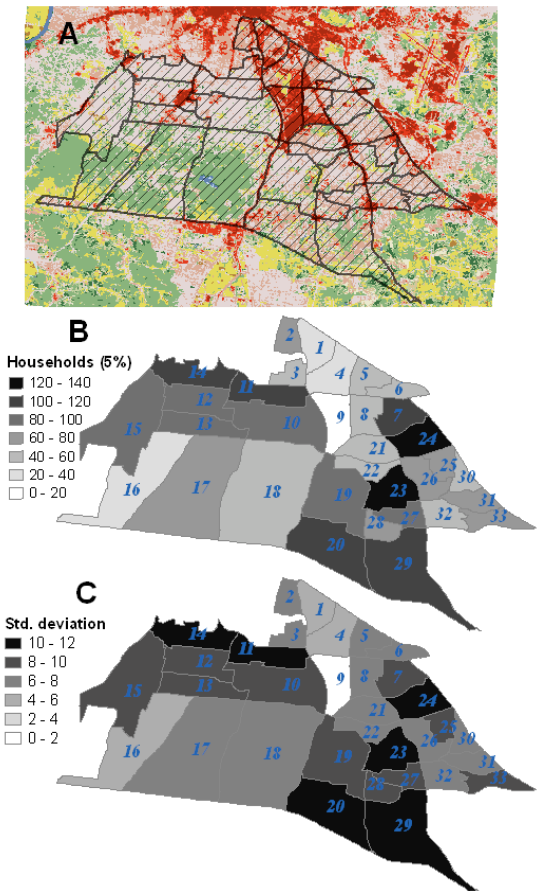

Figure 1. (A) Map of the 33 census tracts of the PUMA 2203 on top of the NLCD 2001 data; (B) Mean frequency of allocated microdata households to census tracts from 1000 allocation runs; (C) Standard deviation of allocation frequencies.

for the allocation of microdata household *i* to all census tracts. *CI* indicates how much confusion there is in making the final decision to place a household in one census tract by comparing the two highest frequencies with which the household was allocated to census tracts during simulation. If *CI* is close to one, confusion is very high because the household has been allocated to at least two different census tracts with similar frequency. A *CI* close to zero indicates that confusion is very low; the household has been allocated to one tract most of the times, and the allocation decision is less ambiguous.

A variety measure is calculated to indicate the number of different tracts the household was allocated to during simulation. The measure includes a tract only if the allocation frequency was greater than 75% of the maximium allocation frequency observed. The variety measure Var_i has the form:

$$Var_i = \sum_{j=1}^{n} I(F_{i,j} > F_{Max\,i} * 0.75) \qquad (3)$$

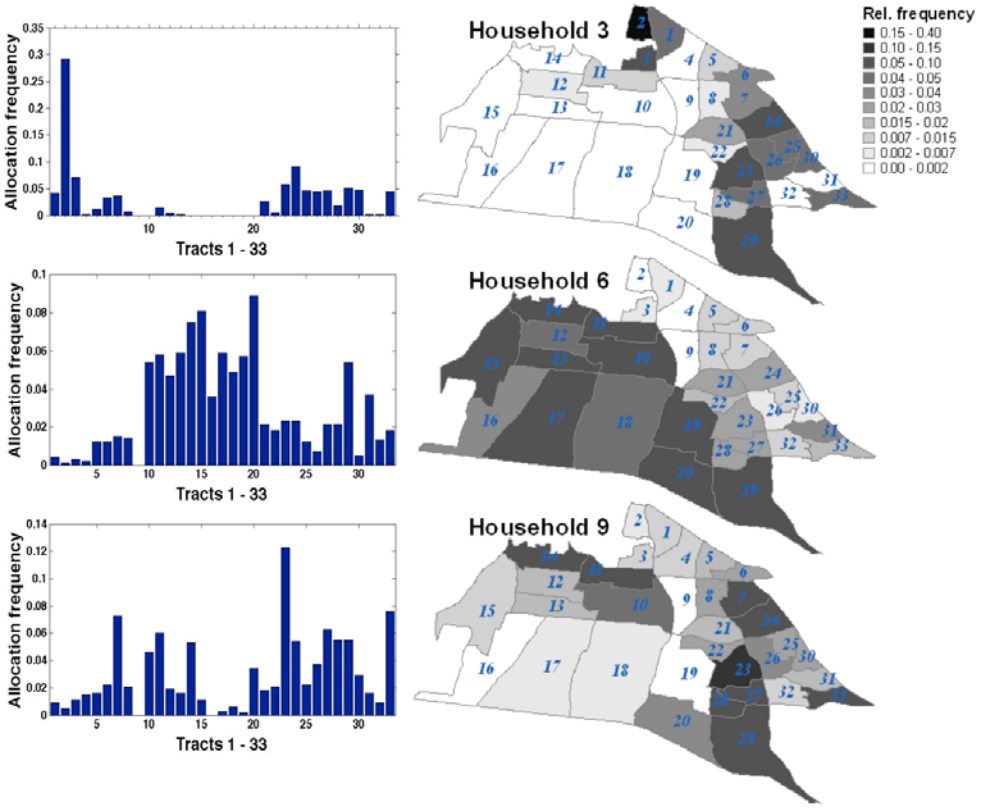

Figure 2. Allocation frequencies of three exemplary microdata households to each of the 33 census tracts based on 1000 allocation runs. On the left bar diagrams are shown allowing a purely numeric comparison and the evaluation of the allocation ambiguity. On the right probability surfaces are mapped for each household to show the inherent spatial pattern and clustering of allocation frequencies across the census tracts.

where $F_{i,j}$ is the normalized frequency in the range [0,1] with which microdata household i has been allocated to census tract j; $F_{Max,i}$ is the maximum normalized allocation frequency to a census tract observed for microdata household i and $I(x)$ is the indicator function, equal to 1 if x is true and 0 otherwise.

C. *Dasymetric mapping of allocation results*

Dasymetric mapping is applied to partition tracts into homogeneous target zones using ancillary variables; we use residential areas of varying intensity from the U.S. National Landcover Dataset (NLCD) 2001 to define areas that are used as residential land versus areas that are not. These target zones represent a spatial refinement within census tracts and commonly reflect areas of uniform value in the underlying population surface (Mennis and Hultgren, 2006). We demonstrate how the allocation of microdata households can be refined by defining such target zones, and indicate how the spatial pattern and clustering of household-specific probability surfaces would change as a result of this refinement. This step represents a demonstrative exploration to layout future endeavors in developing more advanced dasymetric modeling techniques as integrated in simulations of spatially refined household allocations.

III. Experimental Results

The allocation process of microdata households to census tracts was conducted using population weights based on race, income and tenure as constraining tract summary statistics, to derive initial experimental results. The uncertainty assessment is done for the allocation results of the 2381 PUMS microdata households (about 5% of the total population of the PUMA) to the 33 census tracts of PUMA 2203 (Fig. 1A).

As can be seen in Figs. 1B and 1C the simulation of the allocation process results in a differentiated pattern of mean frequencies of allocated microdata households to census tracts. The spatial distribution of these frequencies and their standard deviations indicate a robust allocation process, which is reflected by the summary statistics of the simulation (not shown).

Table 1 shows the computed uncertainty measures on the household level i.e., variety and confusion index, to quantify the consistency and sensitivity of the allocation for a single household. While *CI* indicates the confusion between the two highest allocation frequencies of a household, the variety measure *Var* informs about the number of tracts that have a fairly high frequency and could impede the allocation decision. For the sample of ten microdata households, both

the confusion index and the variety measure illustrate considerable differences in the quality of the allocation decisions across households.

Three distinct cases are illustrated in Fig. 2 using the bar diagrams for the relative allocation frequency over the census tracts and maps showing the resulting allocation probability surfaces for the considered household.

TABLE I. VARIETY AND CONFUSION INDICES FOR MICRODATA HOUSEHOLD ALLOCATION.

Household	Variety (Var)	Confusion Index (CI)
(1)	2	0.8700
(2)	2	0.8056
(3)	1	0.3106
(4)	3	0.7629
(5)	1	0.6950
(6)	3	0.9101
(7)	4	0.8286
(8)	1	0.7284
(9)	1	0.6179
(10)	4	0.9896

Household 3 shows a very high frequency to one tract and thus results in the smallest variety measure (*Var* = 1) and a very low confusion index (*CI* = 0.31) (Fig. 2, Table 1). Household 9 shows the same Variety but its *CI* indicates that its second-highest allocation frequency is much more similar to F_{max} than in the case of household 3. Household 6 represents a case in which the allocation decision is more ambiguous since there are 3 tracts to which the household has been allocated at least 75% of the maximum frequency. Also *CI* shows a very high value indicative of a high degree of ambiguity. The maps in Fig. 2 show the resulting probability surfaces for the single households. These spatial distributions show interesting patterns and clusters of allocation frequencies of individual microdata households.

Fig. 3 illustrates the idea of dasymetric refinement of household density estimates based on classified developed land as limiting variable. The resulting subunits of census tracts that are admitted as allocation units lead to higher spatial precision of the allocation results but also to changes in the spatial distribution of household allocations across census tracts.

IV. DISCUSSION AND CONCLUDING REMARKS

This paper focuses on the method of uncertainty analysis involved in the spatial allocation of microdata households to census tracts. The uncertainty measures allow the differentiation among varying degrees of ambiguity involved in the allocation decision. Probability surfaces for individual households represent an appropriate tool to analyze the spatial distribution and clustering effects that can be observed in the household level allocation. Thus the approach overcomes existing limitations in representing the inherent uncertainty of spatial allocations. A final validation of the accuracy of the spatial allocation process and the empirical likelihood method to impute population weights will be conducted in the near future by accessing the true microdata locations at a Census Research Data Center. This validation will help us to refine our models.

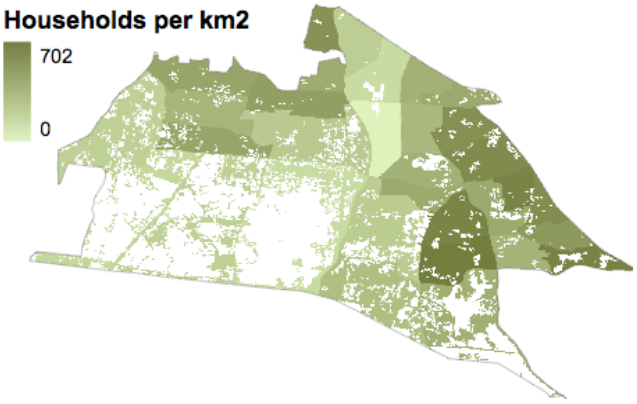

Figure 3. Areal interpolation of household densities to tract subunits of developed land cover from NLCD 2001. This spatial refinement and existing relationships between household attributes and environmental variables will allow higher precision in household allocation.

The presented experiment illustrates the potential of dasymetric modeling to improve allocation estimates of microdata households as more sophisticated rules can be formulated to characterize relationships between e.g., land cover classes or other environmental variables and certain types of households using advanced sampling methods (Mennis and Hultgren, 2006). In this context the spatial patterns of allocated households as well as the change in spatial neighborhood relationships after dasymetric refinement will be important. Three research questions will be (1) how to derive relationships between limiting or related variables defining dasymetric target zones and demographic characteristics of microdata households; (2) how to incorporate dasymetric modeling into the allocation process for simulation; and (3) How to improve the allocation decision if there is a high degree of ambiguity involved.

REFERENCES

Burrough, P.A., Wilson, J.P., van Gaans, P.F.M., and Hansen, A.J. (2001). Fuzzy k-means classification of topo-climatic data as an aid to forest mapping in the Greater Yellowstone Area, USA. *Landscape Ecology* 16(6), 523-546.

Eicher C.L. and Brewer C.A. (2001). Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science*, 28(2), 125-138.

Hynes S., Farrelly N., Murphy E., and O'Donoghue C. (2008). Modelling habitat conservation and participation in agri-environmental schemes: a spatial microsimulation approach. *Ecological Economics*. 66, 258-269.

Mennis J. and Hultgren T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science* 33(3), 179-194.

Nagle N.N., Buttenfield B.P. and Leyk S. (2010). Estimating spatially-explicit microdata by using data from small area statistics and public-use microdata: A new method. Proceedings of the 49th Annual Meeting of the Western Regional Science Association (WRSA), February 21-24, 2010, Sedona, AZ.

Owen A. B. (2001). *Empirical likelihood.* Boca Raton, FL: CRC Press.

Williamson P., Birkin M. and Rees P.H. (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.