White paper for Specialist Meeting “Measuring People in Place”
Boulder Colorado, 4-6 October 2012

Working with Uncertainty in Small Area Estimations

Stefan Leyk, Barbara P. Buttenfield and Matt Ruther
University of Colorado Boulder
{ stefan.leyk, babs, matthew.ruther} @colorado.edu

Our research group is working on small area estimation to allocate PUMS microdata households to census tract levels. The purpose of the allocation is to refine tract-level summary statistics. The estimation process involves two stages of work. In the first stage, maximum entropy methods developed by Nicholas Nagle at Tennessee establish a set of imputed weights which in standaradized form serve as probabilities that a given household record might be encountered in each tract encompassed by the PUMA. Weights imputation is constrained by a small number of tract-level variables. Imputation guides a spatial allocation process to probabilistically derive revised tract summary statistics, for the constraining variables as well as for any other variables available in the microdata records. This is advantageous since many of the PUMS variables are not available in the tract-level SF3 summaries.

In the second stage, dasymetric modeling is utilized to refine precision of the allocations, placing households to sub-tract granularity. Dasymetric methods provide a specialized form of areal interpolation which relies on ancillary data to inform the allocation process: a commonly utilized simple ancillary variable (for example) is landuse, which informs allocation by limiting household placements to residential areas. We have developed more sophisticated dasymetric methods which utilize the imputed weights created in the first stage as ancillary variables in the dasymetric stage, and which permit meaningful refined mapping of any variables in the PUMs record for allocated households. Once allocated, the household characteristics are summarized to revise estimates of tract-level demographic summary statistics,which are compared to original tract summaries within a context of expected variation (Figure 1).

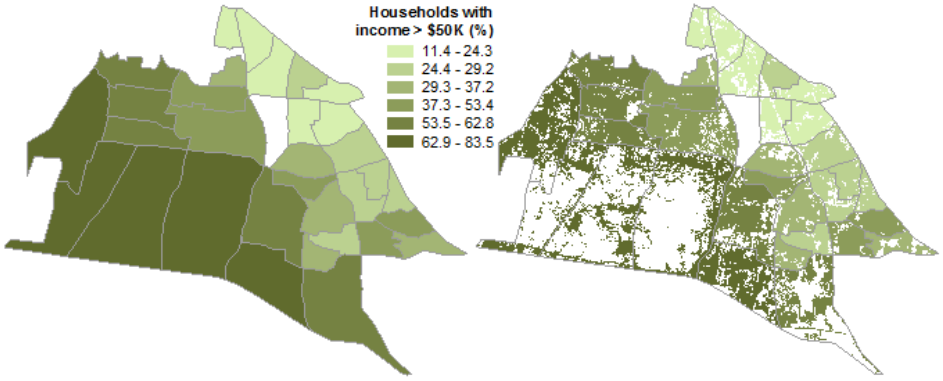


Figure 1. *Refined tract-level statistics for Davidson County TN resulting from allocation of households with income greater than $50K are shown in a choropleth map (left) and.in a dasymetrically refined map (right) using the NLCD derived limiting variable as a spatial constraint, and the imputed weights as related variable attribute relationships.*

Our current focus in the project is to develop metrics of uncertainty and integrate these metrics directly into the allocation process. While uncertainty can be characterized on many dimensions, our work emphasizes allocation ambiguity, or more specifically using Klir and Wierman’s (1999) concept of *non-specificity,* in which the assignment of an observation to any class under a given classification scheme is open to question (Fisher 1999: 192). Non-specificity is quantified for each household as a function of the distribution of imputed sample weights over all census tracts (Figure 2), and by computed metrics of confusion and variety of allocation to any census tract. The assumption in non-specificity is that some observations might fit equally well into multiple classes under any given classification scheme. Our team has also explored uncertatinty in the sequencing of allocation of households, working with Lorenz curves and Gini coefficients.

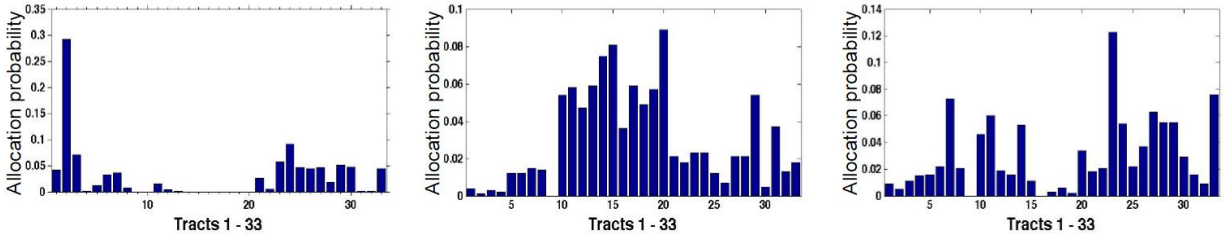
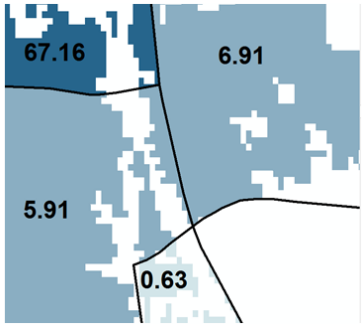


Figure 2. *Histogram distributions of imputed sample weights and probabilitiy of allocation to tracts for three PUMs households simulated for Davidson County TN (Leyk et al, 2012a). The household on the left has highest probability of allocation to Tract 2; the household on the right has high probability of allocation to Tract 23, and secondary probabilities of alloctiton to half a dozen other tracts. The middle household shows high probabilities for many tracts, indicating a household whose demographic characteristics would be encountered frequently in the PUMA.*

Our work over the past two years demonstrates the feasibility of incorporating uncertainty metrics directly into the allocation process. Differing uncertainties in the allocation sequence carry demographic and geographic meaning and assist in interpreting the resulting dasymetric patterns. These patterns are found to vary across the PUMA as well as among different variables. Incorporating uncertainty directly into the allocation and subsequent dasymetric refinement empowers the analyst to evaluate the potential of the selected set of ancillary variables to establish associations which can satisfactorily distinguish among census tracts and at the same time avoid significant departure from the original SF3 summaries.



The dasymetric refinement of allocation to sub-tract level increases spatial precision of small area estimates; and here too, uncertainty can play an informative role. It potentially can expose inconsistencies of fine scale neighborhood relations across coarser scale tract boundaries. In Figure 3, if adjacent tracts whose summary data are quite different contain a residential area spanning a shared tract boundary, one could expect residential parts to have similar demographic characteristics. If tract summaries obscure this pattern, dasymetric modeling can render uncertainties explicitly.

Figure 3. *Homogeneity within residential areas spanning tract boundaries (black lines) may be obscured if tract summary values (black text) are widely discrepant. (Leyk et al, 2012b)*

Our goal in attending this Specialist Meeting is to share insights gained from our research showing , and to engage with other meeting participants on issues of communicating uncertainty in spatial information statistically as well as by methods of visualization and visual analytics, with which we have not worked extensively. If accepted, we expect that Stefan Leyk will represent the research team at the Specialist Meeting.

References

Fisher PF 1999 Models of uncertainty in spatial data. In Longley PA, Goodchild MF, Maguire DJ, and Rhind D W (eds) *Geographical Information Systems (Volume 1):Principles and Technical Issues (Second Edition)*. New York, John Wiley and Sons: 191–205

Klir GJ and Wierman MJ 1999 *Uncertainty–Based Information – Elements of Generalized Information Theory*. Springer, Physica-Verlag.

Leyk S, Buttenfield BP and Nagle N. 2012a Modeling Ambiguity in Census Microdata Allocations to Improve Demographic Small Area Estimates. *Transactions in Geographic Information Science.* (in press)

Leyk S, Nagle N and Buttenfield, BP 2012 b Maximum Entropy Dasymetric Modeling for Demographic Small Area Estimation. *Geographical Analysis* (in press).