A Dasymetric-Based Monte Carlo Simulation Approach to the Probabilistic Analysis of Spatial Variables

April Morton[1], Jesse Piburn[1], Ryan McManamay[1], Nicholas Nagle[2], Robert N. Stewart[1]

[1] Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831
Email: {mortonam; piburnjo; mcmanamayra; stewartrn}@ornl.gov

[2] University of Tennessee, Knoxville, Department of Geography, 1000 Phillip Fulmer Way, Knoxville, TN 37916
Email: nnagle@utk.edu

Abstract

Monte Carlo simulation is a popular numerical experimentation technique used in a range of scientific fields to obtain the statistics of unknown random output variables. Though Monte Carlo simulation is a powerful technique for the probabilistic understanding of many processes, it can only be applied if it is possible to infer the probability distributions describing the required input variables. This is particularly challenging when the input probability distributions are related to population counts unknown at desired spatial resolutions. To overcome this challenge, we propose a framework that uses a dasymetric model to infer the probability distributions needed for a specific class of Monte Carlo simulations dependent on population counts.

1. Introduction

Monte Carlo simulation is a numerical experimentation technique that has been widely used in a variety of scientific domains to obtain the statistics of unknown random output variables by repeatedly sampling values from a set of known input random variables and then feeding them through a computational model (Mahadevan 1997). Dasymetric mapping, on the other hand, has been widely used in the field of areal interpolation to disaggregate coarse resolution population data to a finer resolution through the use of ancillary data (Eicher and Brewer 2001).

Though Monte Carlo simulation is a powerful technique for the probabilistic understanding of many processes, it can only be applied if the probability distributions describing the required input variables can be inferred. Unfortunately, conventional inference methods cannot often be used to infer the probability distributions of population counts (i.e. counts of populations with specific characteristics) that are unknown at desired spatial resolutions. Fortunately, recent advancements in dasymetric mapping, which may not be well known to researchers utilizing Monte Carlo simulation in fields other than areal interpolation, provide novel methods for estimating the probability distributions of population counts. To highlight the potential link between dasymetric mapping and Monte Carlo simulation, we propose a framework that uses the penalized maximum entropy dasymetric model (PMEDM) proposed by Nagle *et. al* (2014) to learn the parameters of multinomial distributions describing population counts needed to complete a specific class of Monte Carlo simulations.

2. Methodology

Suppose we'd like to calculate, through Monte Carlo simulation, the sample mean $\bar{y}_t$ and sample standard deviation s_{y_t} of an output variable $y_t = f_t(a_t, x_t)$ for a set of non-overlapping regions $t \in \{1, \dots, T\}$ where $a_t = [a_{t1,\dots,}a_{tk}]$ and $x_t = [x_{t1,\dots,}x_{tk}]$ are vectors of random variables with

unknown and known probability distributions, respectively. Furthermore, assume a_{ti} represents the number of people or households with characteristic i in region t and assume x_{ti} represents some value conditioned on characteristic i in region t. For example, a_{ti} might represent the number of one bedroom households in region t while x_{ti} might represent the electricity consumption of a one bedroom household in region t.

Now, suppose that there also exist microdata, related to each of the k characteristics, for a given population survey containing n of N respondents sampled from region s, where region s is partitioned into the same $t \in \{1, \dots, T\}$ target regions. Furthermore, assume there exist summary count estimates and variances corresponding to each of the T target regions and k characteristics, and assume we know the prior probabilities q_{jt}, for all $j \in \{1, \dots, n\}$ and $t \in \{1, \dots, T\}$, that a person or household with the same k characteristics as respondent j lives in target region t. Given the preceding information, we can use the PMEDM to learn the actual probabilities p_{jt}, for all $j \in \{1, \dots, n\}$ and $t \in \{1, \dots, T\}$, that a person or household with the same k characteristics as respondent j lives in target region t. We can then simulate, for all j and t, several likely counts of each person j in target region t, from which we can compute several realizations of a_{ti}, or the total people in region t with characteristic i. Given these realizations of a_{ti}, we can complete the Monte Carlo simulation and compute the statistics of interest $\bar{y}_t$ and s_{y_t} to enhance our probabilistic understanding of the output variable y_t.

3. Application and Results

To illustrate the utility of the proposed framework, we use Monte Carlo simulation and the PMEDM to estimate the mean and standard deviation of the average aggregate monthly electricity consumption for all Census block groups t intersecting the Knoxville urbanized area defined by the Census in 2012 (Census 2012). More specifically, we compute the sample mean $\bar{y}_t$ and sample standard deviation s_{y_t}, for all t, of the average aggregate monthly electricity consumption given by

$$y_t = f_t(a_t, x_t) = \sum_{z=1}^{a_{t1r}} x_{t1z} + \sum_{z=1}^{a_{t2r}} x_{t2z} + \cdots + \sum_{z=1}^{a_{t8r}} x_{t8z} \quad (1)$$

where a_{tir} represents the r^{th} realization of the number of households with characteristic i in region t and x_{tiz} represents the z^{th} realization of the average monthly electricity consumption of a household with characteristic i. Out of the 8 characteristics i, the first 4 characteristics refer to the number of 1 through 4 or more bedroom detached houses in target region t while the last 4 characteristics represent the number of 1 through 4 or more bedroom non-detached houses in target region t. In this application "detached" house refers to all houses following the United States (US) Census' definition of "detached single-family housing units" and non-detached household refers to all other Census classifications for housing units (Census 2012). Also note that, due to limited sample sizes, studio apartments, or 0 bedroom houses, are grouped with 1 bedroom houses. Furthermore, from this point forward, the term "monthly electricity consumption" refers to the average monthly electricity consumption over a 12 month period.

3.1 Learning the Input Variable Probability Distributions

To learn the probability distributions of the random variables contained in a_t we collected all microdata variables, for all survey boundaries containing our study area block groups, matching the 8 categories defined above from the weighted 2008-2012 household-level Public Use Microdata Sample (PUMS) of the American Community Survey (ACS) (US Census 2012). Furthermore, we determined the summary count estimates and variances, related to the same characteristics, for all Census tracts and block groups, through the summary count estimates and 90% margins of error (MOEs) published in the 2008-2012 ACS summary tables (US Census 2012). In addition, we assumed each unique household had the same prior probability of belonging to each target region, and thus let $q_{jt} = \frac{w_j}{T \cdot \sum_{r=1}^{n} w_r}$, where w_j represents the weight of microdata respondent j. We then used the PMEDM to learn the probabilities p_{jt}, for all $j \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, from which we simulated several realizations of a_t.

To learn the probability distributions for the random variables in x_t we used the 2009 Residential Energy Consumption Survey (RECS) microdata, restricted to respondents living in Tennessee, published by the US Energy Information Administration (EIA) (Energy Information Administration 2009). More specifically, we assumed each of the 8 random variables x_{ti} followed a normal distribution and estimated the mean and standard deviation of the monthly electricity consumption of each of these categories using the annual kWh reported by Tennessee respondents belonging to each category.

3.2 Results and Discussion

To complete the Monte Carlo analysis we simulated, for all t, 30 sets of population counts for a_t and then computed 30 values of y_t, for each vector a_t, to obtain a total of 900 simulated values of each y_t. Figure 1 shows the mean monthly electricity consumption and standard deviation error bars for all household categories while figure 2 shows the median average monthly electricity consumption and standard deviation per household for all Census block groups intersecting the Knoxville urbanized area. As expected, the Census block groups closer to downtown Knoxville have a much lower median average consumption per household than the households in the wealthy suburban neighborhoods. This is likely due to the fact that the downtown block groups have a higher percentage of small apartments and student housing, which, according to figure 2, have a lower mean monthly electricity consumption than the wealthy suburban neighborhoods containing a higher percentage of large detached households. Though more difficult to interpret visually, the median average standard deviation per household within each block group varies according to the cumulative effect of the count and standard deviation of the mean electricity consumption coming from the mix of categories within each block group.

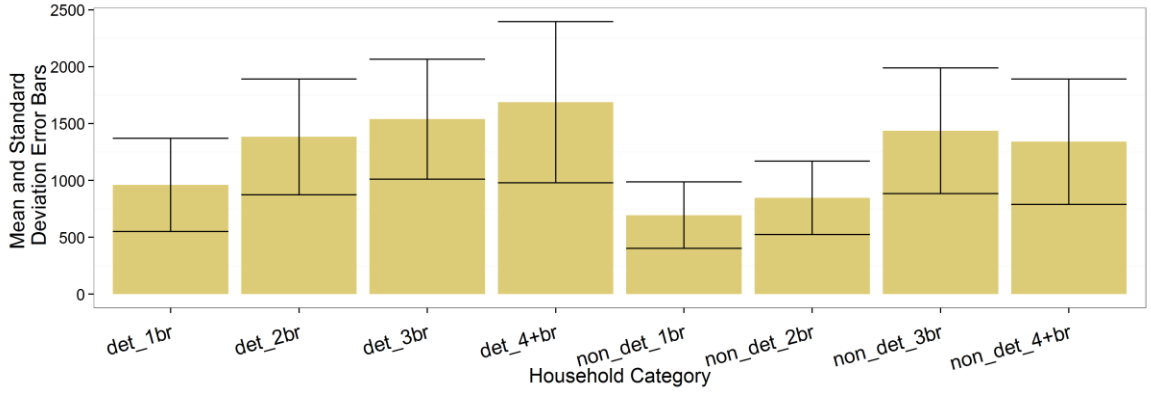

Figure 1. Mean monthly electricity consumption and standard deviation error bars (kWh) by household category

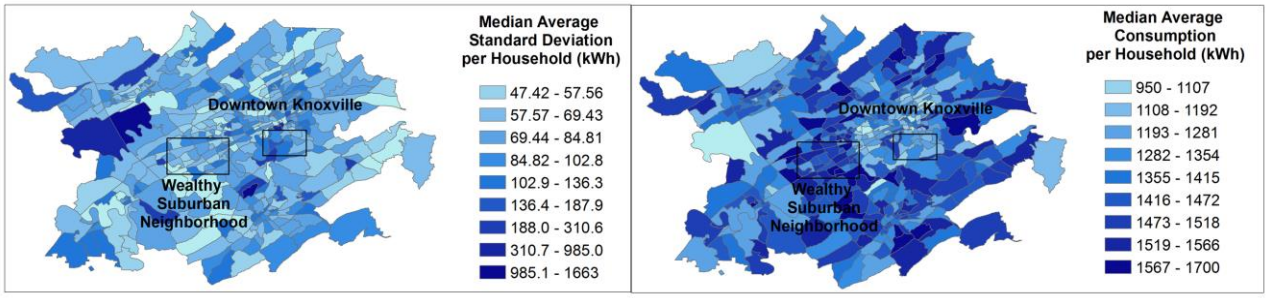

Figure 2. Median average monthly electricity consumption and standard deviation per household for all Census block groups intersecting the Knoxville urbanized area

In summary, this case study is one example of the potential usefulness of the proposed framework for completing Monte Carlo analyses which require probability distributions over population counts.

Copyright

Acknowledgements

The authors would like to acknowledge the financial support for this research from the U.S. Government for the development of a fine-resolution model of urban-energy systems' water footprint in river networks; Oak Ridge National Laboratory's Laboratory Directed Research and Development (LDRD).

References

Mahadevan, S (1997). Monte carlo simulation. Reliability-Based Mechanical Design: 123-146.

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

Nagle, N, Buttenfield, B, Leyk, S, Spielman, S, (2014) Dasymetric modeling and uncertainty. Annals of the Association of American Geographers 104(1):80–95

US Census Bureau (2012), 2008 – 2012, American Community Survey microdata. US Census Bureau. http://factfinder2.Census.gov. Accessed 6 Jan 2015

US Census Bureau (2012), 2008 – 2012, American Community Survey summary tables. US Census Bureau. http://factfinder2.Census.gov. Accessed 6 Jan 2015

Energy Information Administration (2009) Residential Energy Consumption Survey. Energy Information Administration. http://www.eia.gov/consumption/residential/data/2009/index.cfm?view=microdata. Accessed 1 April 2016