Computers, Environment and Urban Systems 63 (2017) 38–49

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata

Amy N. Rose [a,*], Nicholas N. Nagle [a,b]

[a] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, United States
[b] Geography Department, University of Tennessee, Knoxville, United States

ARTICLE INFO

Article history:
Received 5 October 2015
Received in revised form 19 July 2016
Accepted 22 July 2016
Available online 1 August 2016

Keywords:
Validation
Small area estimation
Reweighting
IPF
P-MEDM
Census
DHS

ABSTRACT

Despite the increasing availability of current national censuses, these datasets are limited by their lack of small area demographic depth. At the same time, spatial microdata that include detailed demographic information are only available for limited geographies, thus limiting the complex analysis of population subgroups within and between small areas. Techniques such as Iterative Proportional Fitting have been previously suggested as a means to generate new data with the demographic granularity of individual surveys and the spatial granularity of small area tabulations of censuses and surveys. This article explores internal and external validation approaches for synthetic, small area, household- and individual-level microdata using a case study for Bangladesh. Using data from the Bangladesh Census 2011 and the Demographic and Health Survey, we produce estimates of infant mortality rate and other household attributes for small areas using a variation of an iterative proportional fitting method called P-MEDM. We conduct an internal validation to determine: whether the model accurately recreates the spatial variation of the input data, how each of the variables performed overall, and how the estimates compare to the published population totals. We conduct an external validation by comparing the estimates with indicators from the 2009 Multiple Indicator Cluster Survey (MICS) for Bangladesh to benchmark how well the estimates compared to a known dataset which was not used in the original model. The results indicate that the estimation process is viable for regions that are better represented in the microdata sample, but also revealed the possibility of strong overfitting in sparsely sampled sub-populations.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Demographic information from censuses and surveys are used to support a wide range of decisions for public and private planning. For example, knowledge of the characteristics of a population in an area is critical to determine the need and feasibility of new programs including schools or community centers. Furthermore, changes in the size, distribution, and composition of a population will directly impact future planning of housing and infrastructure such as roads, water supply, and energy.

Users must choose between using publicly available tabulations from large scale, national censuses and surveys, or collecting individual-level data from custom surveys. National censuses and surveys offer a large sample size, and tabulations of relatively small areas, such as neighborhoods or communities, are often publicly available. Such small area estimates are important for understanding local variations in the distribution of population. Unfortunately, these tabulations may not contain the variables that are most relevant to a particular use, nor do they provide individual- and household-level detail that is necessary to understand human behaviors. In contrast, users may construct custom surveys to collect information about the relevant variables and to understand individual- and household-level behaviors. It is usually too expensive however, to construct surveys with a large enough sample size to understand small area variations.

Synthetic spatial microdata can be developed to fuse together information from census tabulations and individual survey microdata. Synthetic spatial microdata are unit record data that represent individuals or households at a small area level, and thus the methods to generate these data are part of the broader category of small area estimation techniques. The importance of the development of synthetic spatial microdata is two-fold: they allow for analysis of estimates of variables that are not available at a small area level, while simultaneously eliminating confidentiality concerns that are typical when dealing with microdata that reflects personal data. Furthermore, generating synthetic microdata is a way to create cross-tabulations that do not already exist in summary statistics.

Despite the existence of techniques to create such synthetic spatial microdata, the difficulty of validating their outputs limits their potential for use. Model outputs are useless to researchers, planners, and policymakers if those outputs are not reasonable representations of

* Corresponding author at: Computational Sciences and Engineering Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831-6017, United States.
E-mail address: rosean@ornl.gov (A.N. Rose).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.006
0198-9715/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the real world. Recent literature dealing with synthetic microdata highlights that validation of these data is still a shortcoming (Ballas & Clarke, 2001; Birkin, 2013; Edwards & Tanton, 2013; Morrissey & O'Donoghue, 2013; Ruther, Maclaurin, Leyk, Buttenfield, & Nagle, 2013; Williamson, Birkin, & Rees, 1998). The lack of finer spatial and demographic detail in census data is one of the primary motivators for creating the synthetic microdata in the first place, but is also the reason why validation is a difficult problem. There are rarely confirmatory data by which to validate against.

Simply describing the estimating method and reporting the inputs and outputs of the model are not good enough. Rigorous interrogation of the results must be attempted as to give the community of practice some confidence that the estimates are reliable. Voas and Williamson (2001) provide an excellent discussion of the many ways to test the fit of synthetic microdata estimates. Their point, which should be well taken by the larger community is that there is not one "best" method for measuring fit, but rather a give and take with regard to a variety of criteria including validity, ease of calculation, a known distribution, and familiarity to the user community.

There are two chief ways to approach the validation of small area estimation results. In internal validation, some of the input data are withheld from the model, and reserved from comparison with the outputs. In reality, these data would not be withheld, and the concern with this approach is that the errors of estimation may be different when these data are withheld versus when they are included. In external validation, the modeled estimates are compared to a data source that was not used in the model. In many cases, depending on the model and available data, it's only possible to perform internal validation. However, attempts should be made to also externally validate modeled estimates if possible. This study examines methods by which to perform both internal and external validation, and considers issues associated with these validation measures, both in a general sense and specific to our case study. In the study we develop new microdata estimates for infant mortality at the District level, which currently do not exist. We do this using household and population characteristics from the 2011 Bangladesh Census as margins for which to scale data from the Bangladesh Demographic and Health Survey (DHS).

The remainder of this paper is structured as follows. Section 2 provides a brief background on techniques used for producing new small area estimates. Section 3 covers the motivation for and process of building the model for Bangladesh, including the selection of constraint variables. Model fitting and output will be described in Section 4, which will serve as a preface for the discussion of internal and external validation of the new microdata estimates in Section 5. Section 6 will conclude the paper with observations about the study and potential future work.

2. Background

Synthetic small area microdata are often calculated by methods that *reweight* a survey so as to reproduce known, aggregate data for small areas for which it was not designed to be representative. In essence, this modeling approach combines individual or household-level microdata for large spatial areas with spatially disaggregate data in order to create synthetic microdata estimates for small areas (Harding, Lloyd, Bill, & King, 2004; Taylor, Harding, Lloyd, & Blake, 2004).

A variety of techniques have been used to produce small area estimates and demographic characterizations in cases where this information was not collected as part of the national census, was collected but not reported due to privacy concerns, or was not available as cross-tabulations (Beckman, Baggerly, & McKay, 1996; Simpson & Tranmer, 2005; Williamson et al., 1998; Wong, 1992). Of these, iterative proportional fitting (IPF) approaches have a long history of use, addressing a variety of issues including: voting behavior (Johnston & Pattie, 1993), individual travel patterns (Beckman et al., 1996), rural policy analysis (Ballas, Clarke, & Wiemers, 2006; Birkin & Clarke, 1988), and small area estimation (Leyk, Nagle, and Buttenfield, 2013; Simpson & Tranmer, 2005; Wong, 1992).

2.1. 2.1 Iterative Proportional Fitting (IPF)

The Iterative Proportional Fitting (IPF) method is a well-established algorithm for aligning survey data with aggregate totals. IPF requires two datasets: one is an individual- or household-level microdataset, and the other is a dataset of known population subtotals or aggregates. Intuitively, IPF identifies weights for the microdataset so that the microdataset will be redistributed to the known totals. IPF works by iteratively adjusting an n-dimensional array until every dimension converges on the known margins. IPF can be viewed simultaneously as a mathematical scaling procedure (Deming & Stephan, 1940; Norman, 1999) as well as a procedure for creating disaggregated spatial data from spatially aggregated data (Wong, 1992). Birkin and Clarke (1988) provide an early demonstration of the utility of the IPF method in geographical research, and it is often used to overcome the lack of spatial or demographic detail in source data (Ballas, Clarke, & Turton, 1999, p. 23). IPF has been used to simulate entire national scale populations (Ballas et al., 2005), examine voting patterns (Johnston & Pattie, 1993), and to create synthetic populations in order to model the travel behavior of individuals (Beckman et al., 1996).

Wong (1992) tested the reliability of IPF results by taking a subset of his population data, treating it as the actual population, and drawing random samples from this subset. These samples were then fitted by the IPF procedure to produce population estimates. These estimates were then compared to the subset distribution and any discrepancies were attributed to random error effect. Through this process, Wong determined the method did in fact produce reliable estimates but could be improved through increased sample size. In the same paper, he argued for more extensive use of IPF in geographical research, particularly in light of studies (Fotheringham & Wong, 1991; Openshaw, 1984) that demonstrated that using areal unit data for drawing statistical inference is not justified considering the effects of the Modifiable Areal Unit Problem (MAUP).

Variations of IPF have been used in several contexts, and as clarified by Johnston and Pattie (1993), not always under the formal name of IPF. Specifically, early geographical work under the label of ***entropy maximizing procedures*** was done in the context of location-allocation (Wilson, 1971) and conducted to evaluate voting behavior (Johnston & Hay, 1983, 1984; Johnston, Hay, & Rumley, 1983, 1984; Johnston & Pattie, 1993), and small area estimation (Johnston & Pattie, 1993; Leyk, Buttenfield, & Nagle, 2013; Nagle, Buttenfield, Leyk, & Spielman, 2014; Ruther et al., 2013).

2.2. Penalized maximum entropy model (P-MEDM)

Recent work (Nagle, Buttenfield, Leyk, & Spielman, 2012; Nagle et al., 2014) formalized a *penalized* entropy maximizing approach geared toward small area estimation and particularly dasymetric mapping. Traditional maximum entropy approaches solve the model: $max - \sum_i (\frac{w_i}{d_i}) \log(w_i/d_i)$ subject to the constraints that the data reaggregate to the known margins, i.e. $\sum_{i \in k} w_i = Pop_k$, where w are the weights to be determined by IPF, d are prior survey weights and Pop_k are the known, marginal population totals. The IPF procedure estimates new weights w so that the survey estimates are now consistent with the known population totals. The *P-MEDM* adjusts that maximum entropy to account for uncertainty in the population margins, and consequently, reduces overfitting problems that commonly plague IPF applications in sparse data problems. Furthermore, by accounting for the uncertainty throughout the model, a measure of quality can be produced for the final population estimates. The penalized maximum entropy model (P–MEDM) as

Download English Version:

https://daneshyari.com/en/article/4965171

Download Persian Version:

https://daneshyari.com/article/4965171

Daneshyari.com