Computers, Environment and Urban Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

An unconstrained statistical matching algorithm for combining individual and household level geo-specific census and survey data

Mohammad-Reza Namazi-Rad[a,*], Robert Tanton[b], David Steel[a], Payam Mokhtarian[c], Sumonkanti Das[a]

[a] National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia
[b] National Centre for Social and Economic Modelling, University of Canberra, ACT 2601, Australia
[c] Domain Data Products and Insights, Domain Group, Fairfax Media, Pyrmont, NSW 2009, Australia

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form 25 July 2016
Accepted 9 November 2016
Available online xxxx

Keywords:
Imputation
Spatial microsimulation
K-nearest neighbours
Pseudo census
Small area estimation
Synthetic population

ABSTRACT

The Population Census is an important source of statistical information in most countries that is capable of producing reliable estimates of population characteristics for small geographic areas. One limitation of a census is that there are many population characteristics that cannot be collected due to respondent burden or cost. This means that statistical agencies have to conduct population based surveys to provide social, economic and demographic characteristics for a target population which are not captured by a large-scale census. These surveys are usually capable of producing direct estimates at the national level and high level regions but often cannot produce reliable estimates for smaller areas. Due to the increasing demand for comprehensive statistical information not only at the national level but also for sub-national domains, there is a wide discussion in the literature about the use of statistical techniques that combine survey with census data to provide more detailed, finer-level estimates.

Where censuses and sample surveys are based on the same reporting units, statistical matching techniques can be employed to link the records from survey and census data where exact matching of reporting units is impossible due to confidentiality restrictions. These techniques can then provide the detailed social, economic and demographic information required for small areas.

An approach is developed in this paper in which a *close-to-reality* synthetic population of individuals and households is generated from available census tables using an iterative proportional updating (IPU) method. Statistical matching using a nearest neighbour method is then used to impute survey data to the individuals and households in the synthetic population. To evaluate this approach, 2011 Bangladesh census data is used to generate a district-specific synthetic population of individuals and households. Matching is then performed by imputing the nearest possible records among the 2011 Bangladesh Demographic and Health Survey to estimate the wealth index for each household within the synthetic population. The results show that using the method presented in this paper helps with achieving more representative estimates (comparing with direct survey estimates) particularly for areas with small sample sizes where not many population units with different socio-demographic characteristics are included.

1. Introduction

The need for reliable and accurate information concerning poverty, inequality, and living conditions of people and households for geographic areas has increased substantially in recent years. Such information is a basic instrument for targeting policies and programs aimed at the reduction of poverty. Household surveys collect information on incomes, expenditures, and demographics to generate estimates of wealth and poverty at a national level and possibly for large geographic areas of a country. However, data confidentiality conditions generally restrict access to unit level survey data with small area identifiers. Even if access to such data is possible, the small sample sizes generally result in unreliable direct estimates for small areas. This is mostly the case in developing countries. Therefore, indirect estimation approaches are employed for area-level poverty mapping in different parts of the world; e.g. South Africa (Alderman, Babita, Demombynes, Makhatha, & Ozler, 2002), Ecuador (Elbers, Lanjouw, & Lanjouw, 2003), Mexico (Tarozzi & Deaton, 2009), India (Coondoo, Majumder, & Chattopadhyay, 2011), and Spain (Molina & Rao, 2010). A review of such methods is presented by Chambers and Pratesi (2014). Here, a micro-simulation technique is presented

* Corresponding author.
E-mail address: mrad@uow.edu.au (M. Namazi-Rad).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.11.003
0198-9715/

for measuring the area-specific wealth indices in different parts of Bangladesh.

In terms of estimation, *small areas* are the geographic or demographic subsets of the population whose domain-specific sample size is not large enough to produce reliable direct estimates. *Large areas*, on the other hand, are those with enough domain-specific sample information to warrant the use of direct estimators solely based on data obtained from that area. During the last few decades, different small area estimation (SAE) techniques have been developed to overcome the challenging problem of finding reliable estimates for small areas (e.g. Rao, 2003; Chambers & Tzavidis, 2006; Chambers, Chandra, Salvati, & Tzavidis, 2014; Namazi-Rad & Steel, 2015; Chandra, Salvati, & Chambers, 2015). In particular, spatial microsimulation techniques are increasingly used to derive small area estimates of many indicators using survey data (Tanton, Vidyattama, McNamara, Vu, & Harding, 2009; Tanton, Vidyattama, Nepal, & McNamara, 2011; Tanton & Clarke, 2014; Burden & Steel, 2015).

Microsimulation aims at fine-grained level statements about the distribution of some endogenous variables defined on a population. As such, microsimulation techniques are being used for synthesising spatial micro data based on close-to-reality-information. Having synthesised micro-level population data also enables agent-based microsimulation modelling of behaviours of individual entities or agents in different applications. The idea of having such models is motivated for having a better alliance between economic theories and behavioural research in order to improve representation of informed choices within a choice-modelling paradigm and provide an opportunity to test the what-if-scenarios. These models are being increasingly used in modelling the economy (Kokic, Chambers, & Beare, 2000; Morrissey, O'Donoghue, & Farrell, 2014), urban energy markets (Mozumder & Marathe, 2005), dynamics of regional and local labour markets (Morrissey, Clarke, Ballas, Hynes, & ODonoghue, 2008; Farrell, Morrissey, & ODonoghue, 2013), education (Wu, Birkin, & Rees, 2008), agri-environment (Hynes, Farrelly, Murphy, & O'Donoghue, 2008; Hynes, Morrissey, O'Donoghue, & Clarke, 2009), policy making (e.g. Lovelace & Ballas, 2013), public health (Tomintz, Clarke, & Rigby, 2008; Edwards, Clarke, Thomas, & Forman, 2011), population movements and traffic analysis (Lovelace, Ballas, & Watson, 2014; Treiber & Kesting, 2013), and disease control (Eubank et al., 2004; Barrett, Eubank, & Smith, 2005; Ferguson et al., 2006). Details about standard microsimulation models are discussed by Wu et al. (2008), Anderson and Hicks (2011), Birkin and Clarke (2011), and O'Donoghue (2015).

The key objective in this paper is to present a novel microsimulation approach for generating an area-specific synthesised population which can then be used for calculating indirect survey-based inferences. As such, microsimulation is solely used for generating the baseline population and agent-based models are not being used in this paper. The main objective in the first part of this paper is to simulate an artificial population from anonymous census data at the individual and household levels which realistically matches the observed population in a geographical area for a given set of table margins. The resulting area-specific synthetic population (SP) of individuals and households will then be considered as a pseudo-census for the purpose of current study. Using this approach, the identification of population units and/or their sensitive information in the generated area-specific synthetic data will be difficult (Beckman, Baggerly, & McKay, 1996; Rubin, 1987). The hybrid spatial microsimulation technique presented in this paper (for generating an area-specific SP), for the first time, has brought together the theory behind sample-based synthesis techniques (as discussed by Wilson & Pownall, 1976; Arentze, Timmermans, & Hofman, 2007; Guo & Bhat, 2007 and Namazi-Rad, Huynh, Berryman, Barthelemy, & P, 2014a) and sample-free population synthesis techniques (as discussed by Voas and Williamson, 2001). This technique is employed to achieve the highest level of accuracy in calculating area-specific population estimates.

Once a reliable SP is simulated, survey-based estimates can be projected over the entire population using the statistical matching techniques based on the same reporting units. For measuring population-specific indicators based on available census and sample data, and more specifically for measuring the poverty and wealth indicators as the main focus in this study, having close-to-reality population data helps to create a more accurate imputation of survey data based on population characteristics correctly classified within the SP. For empirical evaluation, the Bangladesh census data is used to generate a district-specific SP of individuals and households. Statistical matching is then performed by imputing the nearest possible records among the 2011 Bangladesh Demographic and Health Survey (2011 BDHS) to calculate synthesised wealth indicators for the entire population at the level of households. The wealth index is calculated for each survey individual in 2011 BDHS based on a method developed at the Bangladesh Bureau of Statistics (BBS). This method is briefly discussed in this paper.

The rest of this paper is organised as follows. Section 2 discusses the sample-based and sample-based population synthesis approach. Then, an alternative approach is presented based on which area-specific population of Bangladesh is simulated. Section 3 discusses statistical matching techniques and presents a *K*-nearest neighbours (*K*-NN) algorithm for matching population- and survey-specific data. Section 4 gives a background on how wealth index is being calculated in Bangladesh. Then, the 2011 BDHS data together with the SP simulated in Section 2 are used to calculate indirect estimators for area-based wealth indexes in Bangladesh. Section 5 provides some discussions around the findings of this paper and highlights the benefits of using the indirect approach used in this study for area-based indicators when comparing with standard direct survey estimates.

2. Population synthesis

A synthetic population aims at faithfully reproducing actual social entities, such as individuals and households, and their characteristics as described in a population census. Depending on the quality and completeness of the input datasets, as well as the number of variables of interest and hierarchical levels (usually, individual and household), a reliable SP should be able to reflect the actual physical social entities, with their characteristics and specific behavioural patterns (Namazi-Rad, Mokhtarian, & Perez, 2014b).

The sample-based synthetic reconstruction (SR) approach has been traditionally used by researchers (e.g. Namazi-Rad et al., 2014a; Farooq, Bierlaire, Hurtubia, & Flitterd, 2013) for generating SP using both disaggregated- and aggregated-level data. This method first uses available disaggregated-level data while assuming that it is a representative sample of the target population. This is generally referred to as the *seed data*. Then, population units with the required socio-demographics are randomly drawn from the representative disaggregated-level data and populated within the target area using a weighting technique so that the marginal distribution follows the aggregated-level information coming from one source covering the complete population (e.g. census data).

In order to generate a reliable SP, multi-dimensional tables of population units' socio-demographic variables are needed. When dealing with area-specific tables at the lower dimension, the iterative proportional fitting procedure (IPFP) is proposed by Deming and Stephan (1940), as an algorithm that adjusts a table of data in a way that table cells add up to given totals in all required dimensions. This application of IPFP to contingency tables with known margins is called *raking* and are discussed by Deming and Stephan (1940); Deville, Sarndal, and Sautory (1991); Fienberg (1970); Lu and Gelman (2003); Namazi-Rad et al. (2014a); Stephan (1942).

Please cite this article as: M-R. Namazi-Rad et al., An unconstrained statistical matching algorithm for combining individual and household..., Computers, Environment and Urban Systems (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.11.003