Computers, Environment and Urban Systems 36 (2012) 281–290

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

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

A review of current methods to generate synthetic spatial microdata using reweighting and future directions

Kerstin Hermes *, Michael Poulsen

Department of Environment and Geography, Faculty of Science, Macquarie University, NSW 2109, Australia

ARTICLE INFO

Article history:
Received 13 July 2011
Received in revised form 25 March 2012
Accepted 26 March 2012
Available online 21 April 2012

Keywords:
Synthetic microdata
Small area estimation
Microsimulation
Neighbourhoods

ABSTRACT

Synthetic spatial microdata enable analyses of artificial populations in the form of individual unit record files at a small area level. They allow analyses of estimates of variables that are otherwise not available at this small area level, while preserving the confidentiality of personal data. This type of data has mainly been used to provide more detailed census data and for spatial microsimulation modelling: for example to analyse social policy and population changes, transportation, marketing strategies or health outcomes. We argue that many potential applications for synthetic spatial microdata remain to be developed. One reason for this is the lack of information about and confidence in this type of data. Introductory literature about creating synthetic spatial microdata and discussions on the decisions that need to be taken during the data generation process are rare. In this paper, we therefore review currently existing methods to generate synthetic spatial microdata in a manner which will support most readers who are considering this approach, and we address the main issues of the data generation process with regards to analyses of neighbourhood level data. We discuss further possible applications of these data and the importance of synthetic spatial microdata.

1. Introduction

Census data available at a small area level are restricted to a limited number of variables. By contrast, social surveys include more variables but are restricted in their potential use at the small area level because of the limited number of observations. The generation of synthetic spatial microdata provides a way which can resolve this issue. Synthetic spatial microdata or artificial populations are record files of individuals that represent the actual population at a small area level. How these data can be generated and what issues should be considered to obtain reliable synthetic spatial microdata are subject of this paper. This type of data has the advantage that variables included in social survey microdata can be analysed at a small area level while preserving the anonymity and confidentiality of personal data. Synthetic spatial microdata are also referred to as synthetic population data, (micro-) simulation or simulated data since these data are used in spatial microsimulation modelling.

Spatial microsimulation models that apply synthetic spatial microdata have been widely used across multiple projects. For example, social policy issues such as healthcare, housing or taxes (for example Ballas, Rossiter, Thomas, Clarke, & Dorling, 2005; Taylor, Harding, Lloyd, & Blake, 2004; Williamson, 2007c), population dynamics (Ballas, Clarke, Dorling, & Rossiter, 2005), and marketing strategies (Hanaoka & Clarke, 2007). In addition to scenario modelling – analyses of what-if situations and projecting data – this type of data has also been used to obtain small area estimates of poverty rates (Tanton, 2011), obesity (Edwards, Clarke, Thomas, & Forman, 2010), or smoking prevalence (Smith, Pearce, & Harland, 2011). Overviews about the variety of fields where spatial microsimulation models are employed can be found in Ballas and Clarke (2009), Zaidi, Harding, and Williamson (2009) or Birkin and Clarke (2011).

Synthetic spatial microdata and methods to generate these data are part of the broad family of small area estimation techniques. Marshall (2010), for example, distinguishes three small area estimation approaches: demographic models which mainly engage with age-related estimates, synthetic estimations which include model-based approaches like indirect standardisation and different types of individual, area, and multilevel modelling, and microsimulation which refers to the generation of synthetic spatial microdata as discussed in this paper. Rahman (2009) makes the distinction first between direct and indirect or model-based small area estimations and second between statistical (e.g. synthetic estimators) and geographical approaches (spatial microsimulation modelling) among the indirect or model-based estimations. As he emphasises, the advantage of the spatial microsimulation modelling approach over statistical approaches such as those described

* Corresponding author. Tel.: +61 (0)2 9850 9672; fax: +61 (0)2 9850 6052.
E-mail addresses: kerstin.hermes@mq.edu.au (K. Hermes), mike.poulsen@mq.edu.au (M. Poulsen).

0198-9715/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2012.03.005

by, for example, Pfeffermann (2002) or Rao (2003) is the generation of a "base data file" (Rahman, 2009, p. 15). Hence, synthetic spatial microdata enable a wider range of analyses by not only providing aggregated values for small areas but also individual level microdata. Because spatial microsimulation modelling does commonly not only include the generation of synthetic spatial microdata but also static or dynamic what-if simulations, the term synthetic spatial microdata is used throughout this paper.

While synthetic spatial microdata are currently only used by a small – though growing – research community (Ballas, Rossiter et al., 2005; Harding, 2003), we believe that more researchers could benefit from these data. However, recent introductory literature about methods to generate this type of data and discussion about issues and decisions involved in the data generation process and the reliability of the outcome are rare (Smith, Clarke, & Harland, 2009; van Leeuwen, Clarke, & Rietveld, 2009). Most of the published literature focuses on the presentation of results. Methodological aspects of the data generation process are mostly published in working papers.[1]

The process of generating synthetic spatial microdata requires a number of arbitrary decisions. These include the decision on the method for creating the data, what input data can and should be used, at what scale the data need to be represented, and how the resulting estimates for small areas can be tested for reliability. Additionally, concerns exist regarding the reliability and validity of synthetic spatial microdata because there is not a set of standards that approve or certify their quality. Even in computing the data, to date there is still no generally accepted method of assessing the goodness-of-fit between synthetic spatial microdata and census data on common values (Rahman, 2009; Voas & Williamson, 2001). Different researchers use different methods to test the reliability of their results. This makes it more difficult for 'outsiders' to evaluate the value of a model or set of artificial population data.

Hence, this paper provides a relatively simple and concise introductory review of currently available methods to generate synthetic spatial microdata, then discusses issues and decisions that need to be made during the creation of the data.[2] We focus on the reweighting approach which is at the current stage our preferred method for generating synthetic spatial microdata. We then proceed with a discussion outlining some of the benefits of the data and potential fields where this type of data could be further employed. We conclude that in order for these data to be of wider use, we need to put a greater emphasis on providing information about the data generation and testing of the reliability of results. Addressing these aspects in this paper is a first step towards this direction.

2. Methods for creating synthetic spatial microdata

The origins of computer based microsimulation modelling go back to the 1950s when Orcutt built his own regression analyser to develop a micro-economic simulation model (Orcutt, 1957, 2007; Wolfson, 2009). Rising awareness of spatial effects has led to a substantial and growing number of spatial simulation models over the last years (Ballas, Rossiter et al., 2005; Birkin, Clarke, & Clarke, 1996). These spatial microsimulation models require as input a set of microdata that is representative for small geographic units. Generally, full population microdata for small areas are not easily available and need to be either gathered, estimated or generated synthetically. The term 'synthetic microdata' can include a wide range of data with a varying degree of synthesis: from simply weighted data, over imputation, to completely artificial microdata (Williamson, 2002b).

2.1. Synthetic reconstruction vs. reweighting

To generate synthetic spatial microdata, different approaches have been developed (Fig. 1). As Tanton, Harding, and McNamara (2010, p. 53) point out "both the UK and Australia have been spearheading efforts to develop synthetic small area household data using spatial microsimulation models". We focus in this review mainly on the reweighting approach because the main method used to create synthetic spatial microdata shifted from 'synthetic reconstruction' to 'reweighting' during the 1990s (Huang & Williamson, 2001; Rahman, 2009; Smith et al., 2009; Williamson, Birkin, & Rees, 1998).

Following the classification by Rahman (2009), synthetic reconstruction includes data matching and fusion and iterative proportional fitting. Data matching and data fusion refer to linking of different datasets based on common characteristics and identifiers, respectively, but is generally not applicable for generating small area microdata due to data confidentiality and legal restrictions (Williamson, 2002b). Synthetic reconstruction usually includes random or Monte Carlo sampling and iterative proportional fitting (IPF) or raking (Birkin & Clarke, 1988; Clarke, 1996; Frick & Axhausen, 2004; Huang & Williamson, 2001; Simpson & Tranmer, 2005). Basically, observations of individuals and households are generated artificially according to known distributions of characteristics from aggregated tables, for example age, sex, marital status. Further characteristics are then sequentially added (see Birkin & Clarke, 1988 for a good step-by-step guide). In a more recent publication, Simpson and Tranmer (2005) apply a multilevel model framework to generate estimates for small areas and show how IPF can be implemented in the statistical software package SPSS. However, their approach does not perform well for complex tables with small numbers of cases.

Today, reweighting is more often used than synthetic reconstruction. Here the microdata are not generated artificially; instead existing survey microdata are used as a basis. Basically, reweighting methods calculate new person or household weights for observations from the survey microdata according to how representative they are for each small area. Accordingly, one respondent may get a weight of '3' in an area which means he or she represents three people of the area, whereas another respondent is untypical for that area and is assigned a zero weight which excludes this respondent from representing the population in this specific small area. In another small area, however, the same respondent may represent two people. The weights are calculated and adjusted until the known marginal distribution of the population of the small area is matched by the newly weighted survey microdata. This process is then repeated for all small areas of the larger region of interest until a set of artificial population microdata is obtained to represent the real population (Fig. 2). Survey respondents can thereby represent residents of multiple locations. Information about the marginal distribution is usually gained from small area census tables or other administrative data. These tables are in the literature generally referred to as benchmarks, margins, or constraint tables.

Different methods have been developed to calculate new weights for the survey microdata. IPF, GREGWT and combinatorial optimisation techniques (esp. simulated annealing) are the most widely used reweighting methods. These can be differentiated between deterministic reweighting approaches that use the full or a

[1] Introductory literature comes, for example, from Ballas, Rossiter et al. (2005), though they concentrate on microsimulation applications and on their method to generate the data (deterministic reweighting using IPF). Reviews such as Ballas and Clarke (2009) and Birkin and Clarke (2011) focus mainly on different applications of spatial microsimulation modelling. A good example for a technical discussion about methodological issues of generating small area estimates can be found in Rahman (2009).

[2] We avoid the presentation of mathematical equations because we believe they form a barrier to some potential users of this type of data.