Computers, Environment and Urban Systems 41 (2013) 1–11

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

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

'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation

Robin Lovelace *, Dimitris Ballas

Department of Geography, The University of Sheffield, Sheffield S10 2TN, United Kingdom

ARTICLE INFO

Article history:
Received 15 June 2012
Received in revised form 20 March 2013
Accepted 20 March 2013
Available online 3 June 2013

Keywords:
Microsimulation
Integerisation
Iterative proportional fitting

ABSTRACT

Iterative proportional fitting (IPF) is a widely used method for spatial microsimulation. The technique results in non-integer weights for individual rows of data. This is problematic for certain applications and has led many researchers to favour combinatorial optimisation approaches such as simulated annealing. An alternative to this is 'integerisation' of IPF weights: the translation of the continuous weight variable into a discrete number of unique or 'cloned' individuals. We describe four existing methods of integerisation and present a new one. Our method – 'truncate, replicate, sample' (TRS) – recognises that IPF weights consist of both 'replication weights' and 'conventional weights', the effects of which need to be separated. The procedure consists of three steps: (1) separate replication and conventional weights by truncation; (2) replication of individuals with positive integer weights; and (3) probabilistic sampling. The results, which are reproducible using supplementary code and data published alongside this paper, show that TRS is fast, and more accurate than alternative approaches to integerisation.

1. Introduction

Spatial microsimulation has been widely and increasingly used as a term to describe a set of techniques used to estimate the characteristics of individuals within geographic zones about which only aggregate statistics are available (Ballas, O'Donoghue, Clarke, Hynes, & Morrissey, 2013; Tanton & Edwards, 2012). The model inputs operate on a different level from those of the outputs. To ensure that the individual-level output matches the aggregate inputs, spatial microsimulation mostly relies on one of two methods. *Combinatorial optimisation* algorithms are used to select a unique combination of individuals from a survey dataset. This approach was first demonstrated and applied by Williamson, Birkin, and Rees (1998) and there have been several applications and refinements since then. Alternatively, *deterministic reweighting* iteratively alters an array of weights, *N*, for which columns and rows correspond to zones and individuals, to optimise the fit between observed and simulated results at the aggregate level. This approach has been implemented using iterative proportional fitting (IPF) to combine national survey data with small area statistics tables (e.g. Ballas et al., 2005a; Beckman, Baggerly, & McKay, 1996). A recent review, published in this journal, highlights the advances made in methods for simulating spatial microdata (Hermes & Poulsen, 2012) since these works were published. Harland, Heppenstall, Smith, and Birkin (2012) also discuss the state of spatial microsimulation research and present a comparative critique of the performance of deterministic reweighting and combinatorial optimisation methods. Both approaches require micro-level and spatially aggregated input data and a predefined exit point: the fit between simulated and observed results improves, at a diminishing rate, with each iteration.[1]

The benefits of IPF include speed of computation, simplicity and the guarantee of convergence (Deming, 1940; Fienberg, 1970; Mosteller, 1968; Pritchard & Miller, 2012; Wong, 1992). A major potential disadvantage, however, is that non-integer weights are produced: fractions of individuals are present in a given area whereas after combinatorial optimisation, they are either present or absent. Although this is not a problem for many static spatial microsimulation applications (e.g. estimating income at the small area level, at one point in time; for example see Anderson (2013)), several applications require integer rather than fractional weights. For example, integer weights are required if a population is to be simulated dynamically into the future (e.g. Ballas et al., 2005a; Clarke, 1986; Holm, Lindgren, Malmberg, & Mäkilä, 1996; Hooimeijer, 1996) or linked to agent-based models (e.g. Birkin &

[1] In IPF, model fit improves from one iteration to the next. Due to the selection of random individuals in simulated annealing, the fit can get worse from one iteration to the next (Hynes, Morrissey, ODonoghue, & Clarke, 2009; Williamson et al., 1998). It is impossible to predict the final model fit in both cases. Therefore exit points may be somewhat arbitrary. For IPF, 20 iterations has been used as an exit point (Anderson, 2007; Lee, 2009). For simulated annealing, 5000 iterations have be used (Goffe, Ferrier, & Rogers, 1994; Hynes et al., 2009).

* Corresponding author.
E-mail address: robin.lovelace@shef.ac.uk (R. Lovelace).

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.03.004

Table 1
Criteria for reproducible research, adapted from Peng et al. (2006).

Research component	Criteria
Data	Make dataset available, either in original form or in anonymous, scrambled form if confidential
Methods	Make code available for data analysis. Use non-prohibitive software if possible
Documentation	Provide comments in code and describe how to replicate results
Distribution	Provide a mechanism for others to access data, software, and documentation

Clarke, 2011; Gilbert, 2008; Gilbert & Troitzsch, 2005; Pritchard & Miller, 2012; Wu, Birkin, & Rees, 2008).

Integerisation solves this problem by converting the weights – a 2D array of positive real numbers ($N \in \mathbb{R}_{\geqslant 0}$) – into an array of integer values ($N' \in \mathbb{N}$) that represent whether the associated individuals are present (and how many times they are replicated) or absent. The integerisation function must perform $f(N) = N'$ whilst minimising the difference between constraint variables and the aggregated results of the simulated individuals. Integerisation has been performed on the results of the SimBritain model, based on simple rounding of the weights and two deterministic algorithms that are evaluated subsequently in this paper (see Ballas et al., 2005a). It was found that integerisation "resulted in an increase of the difference between the 'simulated' and actual cells of the target variables" (Ballas et al., 2005a, p. 26), but there was no further analysis of the amount of error introduced, or which integerisation algorithm performed best.

To the best of our knowledge, no published research has quantitatively compared the effectiveness of different integerisation strategies. We present a new method – truncate, replicate sample (TRS) – that combines probabilistic and deterministic sampling to generate representative integer results. The performance of TRS is evaluated alongside four alternative methods.

An important feature of this paper is the provision of code and data that allow the results to be tested and replicated using the statistical software R (R Core Team, 2012).[2] Reproducible research can be defined as that which allows others to conduct at least part of the analysis (Table 1). Best practice is well illustrated by Williamson (2007), an instruction manual on combinatorial optimisation algorithms described in previous work. Reproducibility is straightforward to achieve (Gentleman & Temple Lang, 2007), has a number of important benefits (Ince, Hatton, & Graham-Cumming, 2012), yet is often lacking in the field.

The next section reviews the wider context of spatial microsimulation research and explains the importance of integerisation. The need for new methods is established in Section 3, which describes increasingly sophisticated methods for integerising the results of IPF. Comparison of these five integerisation methods show TRS to be more accurate than the alternatives, across a range of measures (Section 4). The implications of these findings are discussed in Section 5.

2. Spatial microsimulation: the state of the art

2.1. What is spatial microsimulation, and why use it?

Spatial microsimulation is a modelling method that involves sampling rows of survey data (one row per individual, household, or company) to generate lists of individuals (or weights) for geographic zones that expand the survey to the population of each geographic zone considered. The problem that it overcomes is that most publicly available census datasets are aggregated, whereas individual-level data are sometimes needed. The ecological fallacy (Openshaw, 1983), for example, can be tackled using individual-level data.

Microsimulation cannot replace the 'gold standard' of real, small area microdata (Rees, Martin, & Williamson, 2002, p. 4), yet the method's practical usefulness (see Tomintz, Clarke, & Rigby, 2008) and testability (Edwards & Clarke, 2009) are beyond doubt. With this caveat in mind, the challenge can be reduced to that of optimising the fit between the aggregated results of simulated spatial microdata and aggregated census variables such as age and sex (Williamson et al., 1998). These variables are often referred to as 'constraint variables' or 'small area constraints' (Hermes & Poulsen, 2012). The term 'linking variables' can also be used, as they *link* aggregate and survey data.

The wide range of methods available for spatial microsimulation can be divided into static, dynamic, deterministic and probabilistic approaches (Table 2). Static approaches generate small area microdata for one point in time. These can be classified as either probabilistic methods which use a random number generator, and deterministic reweighting methods, which do not. The latter produce fractional weights. Dynamic approaches project small area microdata into the future. They typically involve modelling of life events such as births, deaths and migration on the basis of random sampling from known probabilities on such events (Ballas et al., 2005a; Vidyattama & Tanton, 2010); more advanced agent-based techniques, such as spatial interaction models and household-level phenomena, can be added to this basic framework (Wu et al., 2008; Wu, Birkin, & Rees, 2010). There are also 'implicitly dynamic' models, which employ a static approach to reweight an existing microdata set to match projected change in aggregate-level variables (e.g. Ballas, Clarke, & Wiemers, 2005b).

2.2. IPF-based Monte Carlo approaches for the generation of synthetic microdata

Individual-level, anonymous samples from major surveys, such as the Sample of Anonymised Records (SARs) from the UK Census have only been available since around the turn of the century (Li, 2004). Beforehand, researchers had to rely on synthetic microdata. These can be created using probabilistic methods (Birkin & Clarke, 1988). The iterative proportional fitting (IPF) technique was first described in 1940 (Deming, 1940), and has become well established for spatial microsimulation (Birkin & Clarke, 1989; Axhausen, 2010).

The first application of IPF in spatial microsimulation was presented Birkin and Clarke (1988) and Birkin and Clarke (1989) to generate synthetic individuals, and allocate them to small areas based on aggregated data. They produced spatial microdata (a list of individuals and households for each electoral ward in Leeds Metropolitan District). Their approach was to select rows of synthetic data using Monte Carlo sampling. Birkin and Clarke suggested that the microdata generation technique known as 'population synthesis' could be of great practical use (Birkin & Clarke, 2012).

2.3. Combinatorial optimisation approaches

Since the work of Birkin and Clarke (1988) and Birkin and Clarke (1989) there have been considerable advances in data availability and computer hardware and software. In particular, with the emergence of anonymous survey data, the focus of spatial microsimulation shifted towards methods for reweighting and sampling from

[2] The code, data and instructions to replicate the findings are provided in the Supplementary Information: https://dl.dropbox.com/u/15008199/ints-public.zip. A larger open-source code project, designed to test IPF and related algorithms under a range of conditions, can be found on github: https://github.com/Robinlovelace/IPF-performance-testing.