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

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation techniques: Importance of the rounding method

Durán-Heras Alfonso*, García-Gutiérrez Isabel, Castilla-Alcalá Guillermo

School of Engineering, Universidad Carlos III de Madrid, Spain

ARTICLE INFO

Keywords:
Synthetic population
Iterative Proportional Fitting
Simulated annealing
Small area
IPF rounding

ABSTRACT

Approaches to space-related problems that model decision-making and interactions at the level of individuals, and thus require disaggregated population data (i.e. specifying all attributes for each individual) are increasingly being used in various research domains. Actual population data is generally unavailable due to confidentiality and cost constraints. Therefore, synthetic population generation techniques based on aggregated marginal constraints and a random sample are often used. The two sample-based techniques most frequently used are Iterative Proportional Fitting (IPF) coupled with integerization and Simulated Annealing (SA) (SA is a special case of Combinatorial Optimization, CO). Several authors have emphasized the need for further research on comparing their relative performance. Thus, a methodology encompassing statistical analysis to compare IPF and SA is presented here. Technique performance is evaluated through the percentage classification error of the generated population against the reference population. Two cases are analyzed using the 2001 census microdata in Andalusia (Spain) and the 2000 Swiss Public Use Sample as reference populations, encompassing 6 socio-demographic attributes plus geographic location (municipalities and cantons). Aggregated marginal constraints and random samples are calculated from the reference population. A set of synthetic small area populations are generated using both techniques for various scenarios within each case, corresponding to different combinations of sample sizes, number of categories and number of generated populations. Results reveal the great importance of the integerization process applied to IPF's output. IPF coupled with a marginal distributions-controlled rounding outperforms populations generated with SA in all scenarios, while as SA generally outperforms IPF coupled with the commonly used Monte Carlo rounding.

1. Introduction

There is a growing body of literature on approaches to space-related problems that model decision-making and interactions at the level of individuals, such as spatial microsimulation (Spatial microsimulation can be defined as "… an approach to the analysis of individual-level phenomena over geographical space that involves the creation, analysis and modelling of spatial microdata" (Lovelace & Dumont, 2016)) and agent-based simulation models, and thus rely on population microdata. Applications can be found in a wide variety of fields: transportation planning using travel demand models (Frick & Axhausen, 2004; MATSim, 2017; TRANSIMS, 2017); study of environmental problems linked to gas emissions in cities (Ma, Heppenstall, Harland, & Mitchell, 2014); population evolution used for demographic forecasting (Wu, Birkin, & Rees, 2008); healthcare regional planning (Morrissey, Clarke, Ballas, Hynes, & O'Donoghue, 2008); and numerous other fields, such as marketing (Hanaoka & Clarke, 2007), tourism (Van Leeuwen & Nijkamp, 2010), urban planning (Marois & Bélanger, 2015), criminology (Malleson & Birkin, 2012) or mobility (Lenormand, Huet, & Gargiulo, 2014). Ballas, Rossiter, Thomas, Clarke, and Dorling (2005), Birkin and Clarke (2011) and Ye, Wang, Chen, Lin, and Wang (2016) provide general reviews of microsimulation and its applications. These studies are typically carried out at the spatial scale level of municipalities or small urban areas such as wards or districts.

These approaches require populations of individuals ("agents"), such as households, families or individuals, each of which is characterized by the specific values assigned to a set of relevant, correlated spatial and socio-economic attributes (Farooq, Bierlaire, Hurtubia, & Flötteröd, 2013). Synthetically generated populations are generally utilized, since comprehensive, fully disaggregated data is rarely available (e.g., due to privacy issues in census-based data and due to sample size limitations in survey-based analysis) (Cho et al., 2014). A synthetic

* Corresponding author at: School of Engineering, Universidad Carlos III de Madrid, Avda. Universidad, 30, 28911, Leganés, Madrid, Spain.
E-mail address: duran@ing.uc3m.es (A. Durán-Heras).

https://doi.org/10.1016/j.compenvurbsys.2017.11.001
Received 22 March 2017; Received in revised form 6 November 2017; Accepted 8 November 2017

population synthesizes data from various sources into a common, person-centric framework, assuaging confidentiality concerns while generating realistic attributes, correlations and demographics for the synthetic individuals (Marathe & Swarup, 2013).

Several alternative techniques have been proposed to generate synthetic populations, and some comparative studies have been published. There is still, however, an ongoing debate on which techniques are more appropriate in each specific application setting. Thus, several authors have emphasized the need for further research in this area (Farooq et al., 2013; Hermes & Poulsen, 2012; Lovelace, Birkin, Ballas, & van Leeuwen, 2015).

This paper aims to contribute to this debate by proposing a methodology enabling the structured comparison among alternative approaches, relying on hypothesis testing methods to establish whether observed differences in results have statistical significance. Two of the most popular approaches for synthetic population generation, namely Iterative Proportional Fitting (IPF) and Combinatorial Optimization (CO) are therefore compared, exploiting census microdata to synthesize small area populations. Two alternative rounding procedures within the IPF approach are tested, and a sensitivity analysis exploring whether the results of the comparison are affected by changes in the scenario parameters (such as sample size or number of attribute categories) is carried out.

The rest of this paper is structured as follows. Next section outlines the techniques and their reviews in the literature. In Section 3 we describe the methodology and data used to carry out the comparison and in Section 4 we discuss the results. Finally, we present the conclusions.

2. Techniques for synthetic spatial microdata generation

Synthetic populations with location attributes, also known as synthetic spatial microdata, may be generated by means of a variety of techniques. Different techniques may either be aimed at tackling different problem settings (e.g. sample-based or not) or adopt different approaches to the same problem. Hermes and Poulsen (2012) provide a thorough review of techniques used to generate synthetic spatial microdata.

IPF based techniques are among the most widely used population synthesis techniques (Farooq et al., 2013; Janssens, Yasar, & Knapen, 2014; Lovelace et al., 2015; Ryan, Maoh, & Kanaroglou, 2009; Ye et al., 2016). Rose and Nagle (2016) review IPF's evolution, highlighting that it can be conceptualized both as a mathematical scaling procedure and as a procedure for creating disaggregated spatial data from spatially aggregated data. This set of techniques are based on the algorithm proposed by Demings and Stephan (1940). This algorithm generates a multiway table, representing the synthesized population, that meets the target marginal distributions of attributes (i.e. known small area population subtotals or aggregates, such as number of males and females) while preserving the correlation structure of a given sample (Cho et al., 2014).

Later on, Williamson, Birkin, and Rees (1998) proposed an alternative approach for the generation of synthetic spatial microdata through CO techniques. They tested three CO techniques: Hill Climbing, Simulated Annealing (SA) and Genetic Algorithms, and concluded that, for that application, SA was the most promising of the three. An open access implementation of this SA algorithm is available at the author's website (Williamson, 2017). CO is commonly presented as one of the main alternatives to IPF for population synthesis (Cho et al., 2014; Farooq et al., 2013; Williamson, 2013).

In order to quantitatively compare CO and IPF, a specific setting was chosen. We have selected one of the most prevalent problems in the literature: the generation of small area microdata based on a sample and marginal information for all attributes for each small area. Examples of this type of problem may be found in Lovelace, Ballas, and Watson (2014) and Rahman (2009). In the following Sections, 2.1 and 2.2, we briefly present how IPF and CO deal with this task. In 2.3 we summarize a literature review of the reported relative performance of these techniques.

2.1. IPF based techniques

Based on Farooq et al.'s (2013) conceptualization, according to which IPF based techniques may be described as a sequence of a marginals fitting step and a population generation step, we present hereafter a simplified description of the various techniques that only differ in how the two steps are executed.

The objective of the fitting step is to create a multiway table that fits the target marginal distributions while maintaining the correlation structure found in the sample. If the synthetic population is characterized by M attributes, each cell of the M-dimensional multiway table contains the agent population count for that particular combination of attribute values (the total number of cells is the successive multiplication of the number of categories for all attributes).

The generation of this multiway table starts by using the sample as the initialization multiway table, also called seed. This multiway table is then iteratively adjusted until every dimension converges on the target margins, using the algorithm proposed by Demings and Stephan (1940). The application of this algorithm generates a new multiway table, which complies with the marginal distributions imposed, while preserving the sample's internal association structure, in terms of “conditional odd ratios”. For a thorough explanation of this concept see Rudas (1991) and Rudas (1998). This multiway table also has the property of minimizing the Kullback–Leibler divergence (relative entropy, which can be interpreted as a measure of difference) between that marginal-compliant table and the sample (Champion, 2013; Ireland & Kullback, 1968). When applied to the problem of generating small area microdata, the result of this step is one multiway table for each small area. Nevertheless, the cells in these tables contain, generally, non-integer values. Therefore, a second step is needed in order to obtain a valid number of agents for each combination of attribute values. For this second step, several alternatives have been proposed. One alternative involves Monte Carlo sampling using the normalized cell values in the small area multiway tables as probabilities; however, it entails a sampling variability, which may cause a significant deviation from the target marginal distributions (Tanton, 2014). Other methods have also been proposed, some of which incorporate mechanisms aimed at minimizing the deviation from the target marginal distributions, thus leading to a better fit. Lovelace and Ballas (2013) and Choupani and Mamdoohi (2015) present comparisons between various rounding methods.

Deterministic Reweighting is a variation of the abovementioned algorithm that produces the same result, even though in a different format. The iterative adjustments are applied in this case to the weights attached to each “record” (“agent”) within the sample, until the “reweighted” sample matches the target marginal distributions. Since it leads to the same result, a second step is likewise needed to obtain an integer population. Applications of Deterministic Reweighting to population generation may be found in Smith, Clarke, and Harland (2009) and Morrissey et al. (2008).

IPF Synthetic Reconstruction is still another variation, in which the iterative adjustments are applied to the conditional probabilities of population attributes. The resulting multiway table of conditional probabilities is then used to create the population using MonteCarlo sampling. When applied to the aforementioned problem, as in Ryan et al. (2009), it also generates the same output in the first step.

Beckman, Baggerly, and McKay (1996) identify an inconsistency that arises when IPF is applied successively to generate several small-area multiway tables. Even though each small area population table does preserve the sample's correlation structure, the total generated population that results from combining all the small areas has a slightly different correlation structure. Therefore, they propose a two-step variation. The first step generates, for the full set of small areas, an M-