Transportation Research Part B 58 (2013) 243–263

Contents lists available at ScienceDirect

Transportation Research Part B

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

Simulation based population synthesis

Bilal Farooq [a,*], Michel Bierlaire [b,1], Ricardo Hurtubia [c], Gunnar Flötteröd [d]

[a] *Département des Génies Civil, Géologique et des Mines, École Polytechnique Montréal, 2500 Ch. Polytechnique Montréal, H3T 1J4 Montréal, Canada*
[b] *Transport and Mobility Laboratory, ENAC, École Polytechnique Fédérale de Lausanne, Station 18, CH-1015 Lausanne, Switzerland*
[c] *Departamento de Urbanismo, Facultad de Arquitectura y Urbanismo, Universidad de Chile, Santiago, Chile*
[d] *Division for Traffic and Logistics, Royal Institute of Technology, Teknikringen 72, 11428 Stockholm, Sweden*

ARTICLE INFO

Keywords:
Markov chain Monte Carlo simulation
Population synthesis
Agent based model
Integrated urban systems planning

ABSTRACT

Microsimulation of urban systems evolution requires synthetic population as a key input. Currently, the focus is on treating synthesis as a fitting problem and thus various techniques have been developed, including Iterative Proportional Fitting (IPF) and Combinatorial Optimization based techniques. The key shortcomings of these procedures include: (a) fitting of one contingency table, while there may be other solutions matching the available data (b) due to cloning rather than true synthesis of the population, losing the heterogeneity that may not have been captured in the microdata (c) over reliance on the accuracy of the data to determine the cloning weights (d) poor scalability with respect to the increase in number of attributes of the synthesized agents. In order to overcome these shortcomings, we propose a Markov Chain Monte Carlo (MCMC) simulation based approach. Partial views of the joint distribution of agent's attributes that are available from various data sources can be used to simulate draws from the original distribution. The real population from Swiss census is used to compare the performance of simulation based synthesis with the standard IPF. The standard root mean square error statistics indicated that even the worst case simulation based synthesis (SRMSE = 0.35) outperformed the best case IPF synthesis (SRMSE = 0.64). We also used this methodology to generate the synthetic population for Brussels, Belgium where the data availability was highly limited.

1. Introduction

Large-scale activity based travel demand and land use evolution models that take into account the individual agent decisions and interactions, are actively been developed in research and practice (Waddell, 2002; Miller and Roorda, 2003; Arentze and Timmermans, 2004; Balmer et al., 2006; Miller et al., 2011). These behavioral and market oriented models are an active tool for detailed impact forecasting of transportation, land use, environmental, and energy related policies.

Among other data, these simulations require at least a base year population of agents (households, families, and/or persons) and their attributes. These attributes are then used in various behavioral models estimated on a sample and implemented in the simulations for forecasting. The attributes are needed for not only the base year population, but also for the future years population. Agent level data on the complete population in a study area is almost never available–not including the few exceptions like Switzerland where the complete census is available for research. Instead in most cases a microsample called public use micro sample (PUMS) with out any high resolution location information is available. It

[*] Corresponding author. Tel.: +1 514 340 5121x4802.
E-mail addresses: bilal.farooq@polymtl.ca (B. Farooq), michel.bierlaire@epfl.ch (M. Bierlaire), rhurtubia@uchilefau.cl (R. Hurtubia), gunnar.floetteroed@abe.kth.se (G. Flötteröd).
[1] Tel.: +41 21 693 9327.

0191-2615/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2013.09.012

may not have information on associations of agents either. Travel surveys conducted by governmental bodies (e.g. municipality) provide a sample for use. In addition to that census and travel surveys also provide aggregate level data at various zonal systems. There might also be various other bits and pieces of information that are available to the researcher about the population. These various sources that are partial views of the population are used to reconstruct it, using synthesis techniques. Future year population can either be generated using the endogenous demographic update mechanism or by generating new population every year using a synthesis technique. Demographic update is not the topic of this paper, but is extensively covered in Farooq et al. (2009) and elsewhere.

The existing population synthesis techniques focus on fitting a single contingency table to the available data. Using that as the ground truth the microsimulations are run to produce the outputs. In existing literature there is no investigation or discussion on the error that may be propagated forward via this approach due to: (a) incompleteness of the data (b) systematic and deliberate tampering of the data at source to conserve privacy (c) differences in the definitions, aggregation levels, scale, etc. (d) assumptions and short coming of the fitting procedures. Moreover, due to unavailability of the data on real population (ground truth) in most of the cases, there rarely has been a complete and systematic analysis done on the performance of the proposed techniques. In this context we propose a new approach that uses all the partial views of the joint distribution of the real population, available through various data sources, in order to draw the synthetic populations from it. We have access to the census on the real population of Switzerland, which we used for the comprehensive performance assessment. The proposed approach is able to overcome major issues faced by the existing approaches, while maintaining at least the same level of accuracy as the leading approaches outlined in the existing literature.

The rest of the paper is organized as follow: we first describe the types of available datasets that can be used in the synthesis. Existing literature is outlined and key shortcomings are discussed. We then formally introduce the problem statement and present our methodology to address the problem. Various performance comparison experiments and a case study are presented. In the last section we discuss key features of the proposed approach and present conclusions.

2. Available data sources

Traditionally, primary data sources to construct a synthetic population have been census and travel surveys. Other sources include: household spending survey, labor force survey, statistics from revenue agency, real estate cadaster etc. Although, they are rarely used in the existing literature. The information from these sources is available in two different forms: sample of individual agents and cross-classification tables. These data are associated to one or more spatial zoning systems.

2.1. Zoning systems

The data is available at certain aggregations of space that is defined by a zoning system. The aggregation may be based on certain maximum density levels, physical obstacles (river, street etc.), and political boundaries. There may also be hierarchy of aggregations within each zoning system. For instance, in case of Canadian census the lowest level of zone is called dissemination area where 400–700 persons are living/working. One level above is the census tract where the limit is 2500 to 8000 persons. Further aggregations are census sub-division, division, and municipality, respectively. The zoning system also changes with time i.e. a discrimination area in year 2001 census, may have been divided into two in 2010 census, so as to satisfy the constraint on number of persons.

The travel surveys are available at the lowest level of aggregation called Traffic Analysis Zone (TAZ). TAZs are usually defined based on road network; its size may vary depending on the agency conducting the survey; and may not overlap with any of the census zoning system. Another zoning system that may be used is the postal code system.

2.2. Sample of individuals

Statistics bureau of a country among other surveys also conducts periodic census of the entire population. The periodicity of this census range from 5 (in case of Canada) to 10 years (USA, Switzerland etc.). While the whole dataset is almost never available for the research (with some exceptions, like that of Switzerland), bureaus do provide a representative sample for public use. In North America, this sample of individual agents is called Public Use Micro Sample (PUMS) and in the UK and few other countries, Sample of Anonymised Records (SARs). In this paper we use the term PUMS. This sample is only available at a large spatial area (for instance, City of Toronto, London etc.), so as to make sure that the privacy of the individuals is protected. The size of sample may range from 1% to 5% of the total population. The sample may contain range of demographic and socioeconomic information on households, families, and persons. The exact location, income details, and some other details may be missing due to privacy concerns. Furthermore, census bureau may hide information on certain individuals, if they deem it to be exposing individuals' identity.

Another source of the sample is the travel survey, usually conducted by the urban regions, municipalities, counties, etc. The focus of this survey is on the travel demand patterns of agents and mode shares, but it also has some information on socieconomics and demographics of agents. There is more fluctuation here in terms of the details, size, and periodicity of