Computers, Environment and Urban Systems 42 (2013) 14–25

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Disaggregating heterogeneous agent attributes and location

Ying Long [a,*], Zhenjiang Shen [b]

[a] Beijing Institute of City Planning, Beijing 100045, China
[b] School of Environment Design, Kanazawa University, Kanazawa 920-1192, Japan

ARTICLE INFO

Article history:
Received 12 May 2011
Received in revised form 27 August 2013
Accepted 9 September 2013
Available online 8 October 2013

Keywords:
Agent-based models (ABMs)
Disaggregation
Population synthesis
Aggregate data
Agenter

ABSTRACT

The use of micro-models as supplements for macro-models has become an accepted approach into the investigation of urban dynamics. However, the widespread application of micro-models has been hindered by a dearth of individual data, due to privacy and cost constraints. A number of studies have been conducted to generate synthetic individual data by reweighting large-scale surveys. The present study focused on individual disaggregation without micro-data from any large-scale surveys. Specifically, a series of steps termed Agenter (a portmanteau of "agent producer") is proposed to disaggregate heterogeneous agent attributes and locations from aggregate data, small-scale surveys, and empirical studies. The distribution of and relationships among attributes can be inferred from three types of existing materials to disaggregate agent attributes. Two approaches to determining agent locations are proposed here to meet various data availability conditions. Agenter was initially tested in a synthetic space, then verified using the acquired individual data, which were compared to results generated using a null model. Agenter generated significantly better disaggregation results than the null model, as indicated by the proposed similarity index (SI). Agenter was then used in the Beijing Metropolitan Area to infer the attributes and location of over 10 million residential agents using a census report, a household travel survey, an empirical study, and an urban GIS database. Agenter was validated using micro-samples from the survey, with an average SI of 72.6%. These findings indicate the developed model may be suitable for using in the reproduction of individual data for feeding micro-models.

1. Introduction

Micro-models using individual-level data, such as agent-based models (ABMs) and microsimulation models, have been discussed increasingly in the context of regional, urban, and population studies as supplements to traditional macro-models (Wu, Birkin, & Rees, 2008). However, the use of micro-models has been hindered by the poor availability of individual data due to privacy and cost constraints. To rectify this hindrance, a number of studies have been conducted to generate synthetic individual data by reweighting large-scale surveys. This study focused on individual disaggregation without micro-data from large-scale surveys. This situation is common in developing countries like China, Southeast Asian countries, South American countries, and African countries. Specifically, a series of steps were proposed to disaggregate heterogeneous agent attributes and locations from aggregate data, small-scale surveys,[1] and empirical studies. These disaggregated results could be used as input for ABMs and microsimulation models. Microsimulation models tend to pay attention to micro-data based on policy evaluation (such as taxes, insurance, and health). ABMs focus more on exploring emerging phenomena at the macro-level, using interactions among agents, simple behavior rules, and interactions between agents and their environment. In this paper, the term ABMs is used, but the present approach also applies to microsimulation models.

Conditions of micro-data availability can be divided into three levels. The first level involves sufficient micro-data for ABMs. Such conditions occur in areas like Sweden, where original surveyed micro-data can be freely accessed (Holm, Lindgren, Makila, & Malmberg, 1996). The study conducted by Benenson et al. in Israel also fit the criteria for the first level (2002). Householder agents were conducted using the 1995 Population Census of Israel. The second level includes surveyed samples, such as the UK. Sample of Anonymised Records (SARs) and the U.S. Census of 2000. These samples can be used to feed agents in ABMs directly or after necessary reweighting (synthetic creation), as in studies conducted by Birkin, Turner, and Wu (2006) and Smith, Clarke, and Harland (2009). The third level is the absence of large-scale micro-data for initializing ABMs. Such conditions exist in regions in which only statistical yearbooks or census reports with aggregate information of surveys are published, such as in China and other developing countries. The ABM constructed by Li and Liu was constructed at this level (2008).

* Corresponding author. Tel.: +86 10 8807 3660.
E-mail address: longying1980@gmail.com (Y. Long).

[1] In this paper, surveys with a sampling ratio of less than 2% and incomplete attributes are defined as small-scale surveys.

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.09.002

Individual disaggregation has been discussed in the field of population studies, especially population synthesis, which is used to generate synthetic individual data for microsimulation models using aggregate data. Synthetic construction and reweighting are two dominant approaches to individual disaggregation, as demonstrated by Hermes and Poulsen reviewed current methods for reweighting (2012). Müller and Axhausen reviewed a list of population synthesizers, including PopSynWin, ILUTE, FSUMTS, CEMDAP, ALBATROSS, and PopGen (2010). The iterative proportional fitting (IPF) techniques[2] adopted by PopGen, were first proposed by Deming and Stephan (1940), and comprise one of the most widely used methods for population synthesis. IPF, which involves reweighting, can adjust tables of data cells so they add up to selected totals for both the columns and rows (in two-dimensional cases). The unadjusted data cells are referred to as seed cells, and the selected totals are referred to as marginal totals. Fienberg used IPF to combine multiple censuses into a single table (1977).

IPF is a mathematical procedure originally developed to combine information from two or more datasets. It can be used when the values in a table of data are inconsistent, or when row and column totals have been obtained from different sources (Norman, 1999). Birkin et al. developed the Population Reconstruction Model to recreate 60 million individuals reweighted from the U.K. Sample of Anonymised Records (SARs) (2006). It provides 1% micro-data describing U.K. households. Wu et al. simulated student dynamics in Leeds, United Kingdom, based on the synthetic population using the Population Reconstruction Model and an integrated approach of microsimulation and ABM (2008). Smith et al. (2009) proposed a method for improving the process of synthetic sample generation for microsimulation models (2009). The TRANSIMS population synthesizer uses IPF for the generation of synthetic households with demographic characteristics in addition to the placement of each synthetic household on a link in a transportation network and assigning vehicles to each household (Eubank et al., 2004). However, these previous studies were primarily conducted to generate individuals based on existing large scale micro-samples, namely through reweighting, with the exception of Barthelemy and Toint, whose work was used to produce a synthetic population for Belgium at the municipality level without a sample (2013). In the present study, generating agents were investigated on a fine scale without any large-scale individual samples.

The present work focused on disaggregating agents with heterogeneous attributes and locations based on both attribute information and spatial location information stored in existing data sources. With respect to agent location, studies regarding the mapping of population distribution were considered useful (Langford & Unwin, 1994; Liao, Wang, Meng, & Li, 2010; Mennis, 2003). In these studies, population density can be interpolated using spatial factors and population census data. However, these studies did not consider the disaggregation of population attributes. Spatial attributes of agents can be probed based on the mapped agent location by overlaying the location of the agent with spatial layers, such as accessibility to educational facilities, neighborhood similarity, and landscape quality (Robinson & Brown, 2009). Spatial attributes of agents have been used in some ABMs (Crooks, 2006; Crooks, 2008; Li & Liu, 2008; Shen, Yao, Kawakami, & Koujin, 2009). With respect to disaggregation of agent attributes, Li and Liu defined agent attributes using aggregate census data (2007). However, they only considered two characteristics of the agents, while the relationships between agent characteristics and agent location were not considered.

[2] See Wong for a mathematical exploration of the IPF and see Norman for a review (1992, 1999).

The present study targets the third level of data availability, in which no large-scale micro-data are available for developing ABMs. The differences between the present study and previous IPF-based studies, such as those conducted by Birkin and Clarke (1988), Rees (1994), Birkin et al. (2006), Ryan, Maoh, and Kanaroglou (2009) and Smith et al. (2009) are as follows. First, the present synthetic reconstruction approach can generate micro-data using only aggregate data and information. This approach does not require individual samples. However, a census based IPF, which takes a reweighting approach, requires surveying large-scale individual data for the production of marginal cross-classification tables of counts and marginal tables for reweighting. IPF could be included in the present approach for cases in which large-scale samples are available. The present approach can be used to disaggregate individuals, households, and other micro-samples, such as vehicles, organizations, packages, and buildings. Accordingly, this approach is more general than micro-data synthesis studies that focus primarily on population disaggregation, such as those by Birkin et al. (2006) and Smith et al. (2009). Third, the spatial locations of samples, which are essential to spatial ABMs, receive special attention in this approach, as advocated by Birkin and Clarke (1988) and Wong (1992). Ideas are borrowed from the residential location choice approach to mapping the disaggregated individuals. Both the characteristics and location of each agent are disaggregated for the initialization of ABMs in the present paper; IPF is primarily used in microsimulation and population studies for population estimates in the years between censuses, rather than in ABMs, as advocated by Norman (1999). The present approach falls into the pool of synthetic reconstruction. It has three aforementioned advantages over existing related studies that target the disaggregation of micro-data.

The current paper presents a method of disaggregating aggregated datasets into individual attributes and locations in situations in which micro-data are not available. This paper is organized as follows: The approach to disaggregating agents is detailed in Section 2. The initial testing and verification under experimental conditions is described in Section 3. Section 4 shows the disaggregation of full-scale residents in Beijing. Discussion and concluding remarks are provided in Sections 5 and 6, respectively.

2. The research approach

2.1. Assumptions

To disaggregate agents, the approach for disaggregating attributes and location should be established separately. Attributes of agents are further divided into two types, non-spatial attributes (such as age, income, and education for a residential agent) and spatial attributes (such as access to subways and amenities, land use, and height of the building that the residential agent occupies). The approach to disaggregating spatial attributes also differs from that used for non-spatial attributes. Because an agent's spatial attributes depend on its location and environmental context, the order in which the agents are disaggregated involves non-spatial attributes, location, and spatial attributes. The disaggregating approach to each portion of the agent information varies, and these differences are elaborated on in the following subsections.

The probability distribution of an attribute (hereafter referred to as the distribution) and the dependent relationship among attributes (hereafter referred to as the relationship) can be inferred from existing data sources, including aggregate data, small-scale surveys and empirical studies. Aggregate data include the total number, distribution and relationship (such as the cross-tabulation of marriage-age standing for the dependent relationship of marriage and age, and the cross-tabulation of income-education