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

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

standing for the dependent relationship of income and education) of agents. Small-scale surveys that store samples can also be used to deduce the distribution of an attribute and the relationships among attributes. They can also be used to validate the disaggregation approach through a comparison of surveyed samples to disaggregated agents. The probability distribution of an attribute and its relationship with other attributes can also be deduced using empirical studies; for example, the height attribute of a resident obeys a normal distribution (A'Hearn et al., 2009). To convert aggregate data to individual samples, the probability distribution of the attributes and the relationship between them must be estimated. All attributes of agents are discretized in Section 2.2 and the specific distribution and relationship forms of the proposed approach are introduced in Section 2.3.

The disaggregation order among agent attributes and location should also be considered. It is a context-dependent process. For example, attribute A has a known distribution based on existing data sources. Attribute B has a known distribution and known relationship with attribute A, and attribute C has a known relationship with attribute D, according to empirical statistical correlations. The independent attributes, which do not depend on other attributes, should be disaggregated first. Next, dependent attributes, which have known relationships with other attributes, can be disaggregated. In this case, attribute A should be disaggregated first, then attribute B. This is because attribute B has a relationship with attribute A. A logic check should be conducted to guarantee the disaggregated results make sense. The authors admit that the method for determining disaggregation order is, to some degree, ad-hoc and informal, and dependent on the empirical statistical correlations that the users apply for disaggregation. Other population synthesizers cannot avoid this problem. Users are encouraged to utilize empirical statistical information to determine the order of disaggregation. The flowchart of the disaggregation approach we proposed is shown in Fig. 1, detailed in the following subsections.

2.2. Discretizing attributes of agents

There are various strategies for disaggregating agent attributes. According to Stevens, the data describing agent attributes can be divided into four different types of scales: nominal (such as marriage and education), ordinal (such as rank order), interval (such as date and temperature) and ratio (such as age and income) (1946). Nominal and ordinal types are qualitative and categorical, while interval and ratio types are quantitative and numerical. These scales could be further divided into discrete and continuous classifications. In most aggregated data, such as census reports and yearbooks, the information available for continuous attributes is presented in discrete form. For this, all continuous data were converted into discrete data to reduce the disaggregation time. For example, the attribute age can be divided into a variety of non-numeric intervals, such as 0–4 years old, 4–7 years old, and 7–12 years old. In the process of disaggregation, it is possible to generate a value randomly, such as 3 years old from 0 to 4, and this can serve as the disaggregated result. Notably, the process for discretizing a continuous attribute, such as age, should consider the existing age ranges presented in the known information (in terms of distributions and relationships) discussed in Section 2.3. Dramatic changes in individual characteristics over a range (such as the age class 16–21: in high school or in college) were avoided by using common sense and empirical research results to evaluation of the attribute.

2.3. Disaggregating non-spatial attributes of agents

2.3.1. Known distributions of information

The known distribution of information, including the categories and intervals (discretized from continuous values) of an agent attribute and their frequencies, are discussed in the description of the disaggregation process. For example, if the categories of the attribute *marriage* are married, unmarried, and divorced, and the corresponding frequencies are 45, 20, and 35, then 45 agents are married, 20 are unmarried, and 35 are divorced among every 100 individuals. The frequencies and probability density function (PDF) are two forms of known distribution information. For the former, the disaggregated values of the attribute follow the frequency distribution. For example, in the case of attribute A among 6 agents, the categories of attribute A are a, b, and c, and the frequencies of this attribute are 3, 2, and 1. Then the disaggregated values of attribute A for all agents may be as follows: {a; b; a; c; b; a}. For the disaggregation of an attribute with a known PDF (such as Gaussian or uniform), the value range of this attribute can be divided into several bins and the frequency for the middle value in each bin can be determined using the PDF. In this way, this condition can be converted in the same manner as the previous frequencies.

2.3.2. Information regarding known relationships

Two types of relationships are considered in this paper. The first is the functional relationship (RA). In this instance, the value of an attribute depends on one or more attributes and is a function of these attributes. The value of this attribute can be calculated using other attributes. The second is the conditional probability relationship (RB, also called joint probability). Under these conditions, there is a probability relationship between the attribute j and its related attribute h, expressed as $P(h|j)$. The frequencies of attribute j, such as $P(j)$, and the categories or intervals of both attributes are known. Then the probability of each combination of categories or intervals of attributes h and j can be calculated using $P(hj) = P(h|j)P(j)$; the count of each combination is $P(hj)$ multiplied by the agent total count. For every 100 persons, the attribute j (AGE) has two intervals, 18–30 years old (40%) and 31–60 years old (60%). Its conditional probability with the attribute h (marriage) is known: Out of all individuals 18–30 years old 60% are married and 40% unmarried. Out of all individuals 31–60 years old, 80% are married and 20% unmarried. This means there will be 40% * 60% * 100 = 24 married persons within the age range of 18–30 and 40% * 40% * 100 = 16 unmarried persons are in 18–30, 60% * 80% = 48 married persons are in 31–60, and 60% * 20% * 100 = 12 unmarried persons are within the age range of 31–60.

A cross-tabulation of two attributes in which the information of marginal totals in the rows and columns are known is the general form of RB. RB can also be inferred using data mining platforms, such as SPSS. For conditions in which both the frequencies ($P(h)$ and $P(j)$) and RB ($P(h|j)$) are known, IPF can be directly applied to the disaggregation procedure.

2.4. Mapping spatial location of agents

For the allocation of agents, the entire study area must be partitioned into small spatial units. In this study, parcels served as the basic spatial units for allocating agents according to the availability of spatial data and the expected application requirements. The process of disaggregating the location of agents caused the allocation of the agents into parcels as point objects. Two solutions were developed for the disaggregation of the agents' locations for the sake of different requirements. The solution used to allocate agents into space was selected based on current knowledge and data availability of the spatial distribution of agents. The first solution parcel allocates agents into parcels in accordance with the statistical information associated with the spatial distribution of those agents. For example, if the number of agents in each parcel are in a region comprised of 80 parcels, and five agents are known to

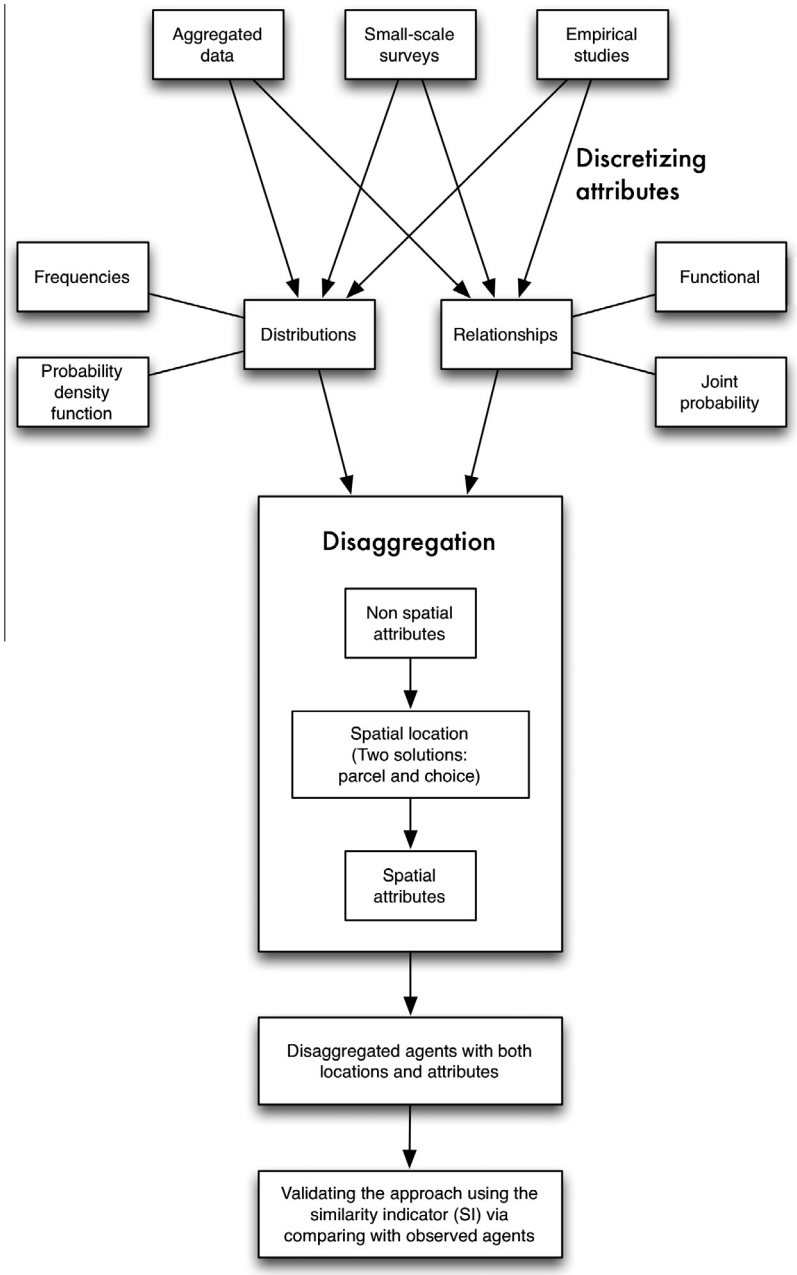

Fig. 1. Flowchart of the disaggregation approach.

be in parcel A, then parcel A will randomly select five agents to occupy, and five points as agents will be randomly created within the parcel. In this way, the agent location can be disaggregated using the same approach for the disaggregation of non-spatial attributes as discussed in Section 2.3.

The second solution allocates agents based on the residential location choice theory. The most common residential location choice model used in practice, the Multinomial Logit Model (MNL) (see Pagliara & Wilson, 2010 for a review), was used to map disaggregated agents. The basic logic of the MNL model is that households are evaluated based on their own attributes, such as income and household members. The sampling of available, vacant housing units and their characteristics, such as price, density, and accessibility to service facilities were considered. The relative attractiveness of these alternatives was measured by their utility. The model then computed the probability that a given household would select a given location from the available alternatives, defined as vacant housing units, given the preferences and budget constraints of the households seeking housing. This idea was borrowed and used to allocate agents into spaces while considering each agent as a resident and each geographical space as a housing market for residents to select. The agent location then depends on both its non-spatial attributes and related spatial layers in its environmental context. For example, a residential agent's socio-economic attributes can influence its preference for each type of spatial layer, such as the accessibility, amenities, and landscape. Parcels have distinguished spatial attributes, and residential agents with different preferences for spatial layers will select the parcel with the greatest preference as their place of residence. This solution is expressed as follows:

$$P_{ij} = \sum_k w_{ik} * F_{kj} + r_{ij} \quad (1)$$

Here, P_{ij} is the preference of agent i for parcel j, F_{kj} is the value of the spatial layer k at parcel j, which can be calculated by overlaying the parcel with the spatial layer in GIS; w_{ik} is the preference coefficient of agent i for spatial layer k, and r_{ij} is the random item of agent i for parcel j. P_{ij} is standardized to range from 0 to 1.

An updated form of choice, the constrained choice solution allocates agents using a residential location choice theory that obeys the statistical information of agent spatial distribution. It differs from choice in that the number of agents with the highest preference selected by a parcel is constrained by the statistical information. For example, if the aggregate data indicate there are six agents in parcel B, then parcel B can be used to select the top six agents with the highest preference for this parcel, after evaluating preferences for all parcels by all agents. From a conceptual point of view, the constrained choice solution is the most useful because it uses additional information. This solution may produce better results than other solutions.

2.5. Validation of the disaggregation approach

The disaggregation approach was validated by calculating the similarity between disaggregated and observed agents. The following similarity indicator was proposed for comparison of agents:

$$\begin{aligned} \mathrm{SI} &= \frac{\sum_{ui} A_{ui} + \sum_{vi} B_{vi} + \sum_{wi} C_{wi}}{(U+V+W)*I} \\ A_{ui} &= 1 - \left| \frac{s_{ui}^{dis} - s_{ui}^{obs}}{s_{u,\max} - s_{u,\min}} \right| \\ B_{vi} &= \begin{cases} 1, & \text{if } s_{vi}^{dis} = s_{vi}^{obs} \\ 0, & \text{otherwise} \end{cases} \\ C_{wi} &= 1 - \left| \frac{the\_rank\_of\_s_{wi}^{dis} - the\_rank\_of\_s_{wi}^{obs}}{\#ordinals - 1} \right| \end{aligned} \tag{2}$$

Here SI is the similarity indicator; A_{ui} is the similarity between the disaggregated value and the observed value of the ratio and interval attribute u of agent i. B_{vi} is the similarity between the disaggregated value and the observed value of the nominal attribute v of agent i; C_{wi} is the similarity between the disaggregated value and the observed value of the ordinal attribute w of agent i; U, V, and W are the number of ratio and interval, nominal, and ordinal attributes to be compared, respectively; I is the total number of agents; s_{ui}^{dis} and s_{ui}^{obs} are the disaggregated and observed values of the ratio and interval attribute u of agent i, respectively; $s_{u,max}$ and $s_{u,min}$ are the maximum and minimum values of the ratio and interval attribute u, respectively; s_{vi}^{dis} and s_{vi}^{obs} are the disaggregated and observed values of the nominal attribute v of agent i, respectively; s_{wi}^{dis} and s_{wi}^{obs} are the disaggregated and observed values of the ordinal attribute w of agent i, respectively; *#ordinals* is the count of ordinals in an ordinal attribute. The similarity index SI is 100% for two sets with the same agent attribute values. To calculate SI, both sets must be sorted by the same rule. The location attribute should be sorted first, followed by sorting the other attributes in increasing order. It is also necessary to disaggregate the same number of agents as observed agents.

A null model was proposed here to further investigate the effectiveness of the proposed approach. The null model can disaggregate the attributes and locations of agents randomly (that is, agents are randomized within the spatial units), assuming that neither distribution nor relationship information is available for the disaggregation process. The null model randomly allocates agent locations and randomly sets agent attributes. The advantage of the present approach over the null model can be determined by comparing disaggregated results obtained using the present approach to those generated using the null model.

Wong used the absolute relative error (ARE) index, which is the total absolute error divided by the sum of all cell values, to compare the disaggregated agents with the observed ones (1992). The present approach and the approach used by Wong differ in several ways. First, his approach works well for categorical attributes but not numerical attributes, whereas the present approach is applicable to both numerical and categorical attributes because it covers all four types of values. Second, Wong's approach becomes increasingly complex in cases with high dimensions and too many attributes to disaggregate. Third, the spatial location is included in the present approach. Fourth, Wong's approach consists of validation at the group level, while validation is conducted at the individual level in the present approach. The process of the present approach can also guarantee precision at the group level, whereas Wong's approach cannot.

3. Experiments in synthetic space

The Agenter (agent producer) model was developed based on the Geoprocessing module of ESRI ArcGIS using Python, and was used to disaggregate heterogeneous agents using the proposed approach. This model was applied in a synthetic space in the Beijing Metropolitan Area (BMA) to disaggregate resident locations and attributes. First, the applicability of this model was verified and tested in a synthetic space. To accomplish this, all disaggregation conditions (including spatial solutions, distribution types, and relationship types) discussed in the research approach section are included in the synthetic space experiments. The current model was verified by comparing the results of disaggregation obtained using each spatial solution to those obtained using the null model. Because there is no known observed set, the model cannot be validated in the synthetic space. For this reason, this process was done with the BMA experiment. This model was verified using a null model.

3.1. Synthetic space and agents within it

The datasets in the synthetic space, including the geometry of the space and known information regarding the residents within it, were not based on real-world data, but were assumed by the authors. It was assumed there were 100 residents living in the synthetic space, composed of 100 parcels. Each parcel was assumed to have a size of 250 m × 200 m, which is a common parcel size in Beijing. The space contains one school, as shown in Fig. 2. Amenity 1 and Amenity 2 (which represent facilities such as convenience stores) are for disaggregating agent locations in cases in which the choice and constrained choice solutions are used, as described in Section 2.4.

The residential agents in the synthetic space have the following attributes: age, marital status, income, travel, parcel (this attribute records the spatial unit ID), and school factor (Table 1). The disaggregation order of attributes was determined using known information. Age was assumed from a census survey report with aggregated information and was disaggregated first. Marital status, which empirical studies show to be related to age, was disaggregated next. Income was then disaggregated with the known PDF inferred from a small-scale survey. Travel was found to depend on age and income and was disaggregated using relationships identified in a small-scale survey. Parcel, which denotes the agent location, was disaggregated using known frequencies obtained from another census survey. After disaggregating the agents' locations, the distance to the parcel's only school was disaggregated using the school factor and the locations of the agents (Fig. 2).

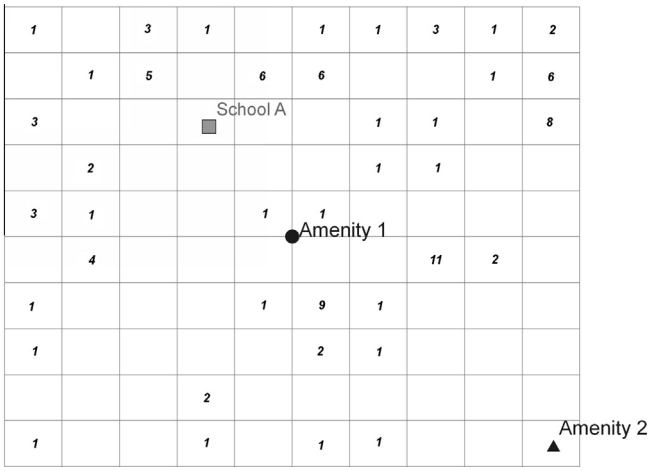

Fig. 2. The synthetic space. *Note*: The number in each parcel denotes the number of known residential agents within the parcel, and parcels with no figures indicate unoccupied parcels.

3.2. Input data

The input data of the Agenter model are presented in this subsection. It should be noted that these data were used to test our proposed approach and some may not match the practices well. The intervals and frequencies of AGE are shown in Table 2.

The dependent relationships between age and marital status are shown in Table 3. The column AGE is the age interval h, as shown in Table 2. The column MARRIAGE gives the probability of every type of marital status for the corresponding age category. For example, for people aged 18–30 ($h = 3$), the probability of residents being married ($j = 1$) is 70%, and the probability of being divorced ($j = 3$) is 10%.

INCOME is assumed to have a Gaussian distribution. It has a mean value of 6000 and standard deviation of 1500.

TRAVEL is assumed to depend on both AGE and INCOME. This relationship was simplified based on the assumption that these data could be inferred from a small-scale survey conducted within the synthetic space using the following decision tree:

$$\text{TRAVEL} = \begin{cases} \text{“No trip”, if AGE} \geqslant 0 \text{ and AGE} \leqslant 4 \\ \text{“Car”, if INCOME} \geqslant 6000 \\ \text{“Bus”, if INCOME} \leqslant 2000 \text{ and AGE} \geqslant 55 \text{ and AGE} \leqslant 70 \\ \text{“Non-mobile”, otherwise} \end{cases} \quad (3)$$

Frequencies of the attribute PARCEL are shown in Fig. 2.

Agenter also requires a table for disaggregating agents (Table 4). Specifically, this table was used to check for obvious inconsistencies among attributes of disaggregated results based on common knowledge or empirical studies. As shown in the second row of this table, in most cases the income of a resident aged 0–18 will be 0. The third row of this table shows that most residents who reside less than 1000 m from a school will travel either by bus, bicycle, or on foot. The rules in the table improved the disaggregation precision and rendered the results more consistent with reality.

Table 2
Frequencies of the age attribute.

ID (h)	Age interval (h)	Percent
1	0–10	5
2	10–18	10
3	18–30	20
4	30–55	35
5	55–70	25
6	70–100	5

Table 3
Dependent relationship between MARRIAGE and AGE.

AGE (h)	MARRIAGE (j)
1	2, 100
2	1, 0.5; 2, 99.5
3	1, 70; 2, 10; 3, 10; 4, 5; 5, 5
4	1, 50; 2, 5; 3, 10; 4, 10; 5, 25
5	1, 30; 3, 20; 4, 5; 5, 45
6	1, 15; 4, 5; 5, 80

Table 4
Evaluation of inconsistencies among disaggregated results.

ID	F1	N1	F2	N2
1	AGE	0–18	INCOME	0–0
2	SCH	0–1000	TRAVEL	Bus-non-mobile

3.3. Disaggregation results

3.3.1. Results produced using the parcel solution in virtual space

The first 10 of 100 disaggregated agents obtained using the parcel solution are shown in Table 5, in which the column AID shows the unique agent ID. The mapping results of all disaggregated agents are stored as the point Feature Class (Fig. 3a), and were all closely consistent with the observed frequencies of the parcels. Every point located within a certain parcel in Fig. 3a had its corresponding attributes and could be directly input into ABMs or microsimulation models. Spatial analysis and spatial statistics could be conducted based on the mapped disaggregated agents.

Table 1
Inventory of residential agent attributes in the synthetic space.

Attribute	Description	Type	Known information	Data source	Data type	Order
AGE	Age in years	Non-spatial attribute	Frequencies	Census survey	Ratio	1
MARRIAGE	Marital status	Non-spatial attribute	RB	Empirical studies	Nominal (married, unmarried, divorced, remarried, widowed)	2
INCOME	Monthly income in CNY	Non-spatial attribute	PDF	Small-scale survey	Ratio	3
TRAVEL	Means of traveling	Non-spatial attribute	RA	Small-scale survey	Nominal (car, bus, no trip, non-mobile)	4
PARCEL	Parcel in which the agent resides	Location	Frequencies	Census survey	Nominal (parcel IDs)	5
SCH	Distance to the school in meters	Spatial attribute	Location of the school	Urban GIS	Ratio	6

Table 5
Disaggregated agents (partial) in the synthetic space.

AID	AGE	MARRIAGE	INCOME	TRAVEL	PARCEL	SCH
1	46	Married	2679	Non-mobile	97	1578
2	4	Unmarried	0	No trip	0	1375
3	65	Married	4663	Non-mobile	2	1463
4	54	Married	5778	Non-mobile	3	1566
5	48	Married	4016	Non-mobile	16	1175
6	26	Married	2904	Non-mobile	16	1175
7	48	Widowed	6066	Car	29	1245
8	19	Married	3450	Car	34	1095
9	56	Widowed	4082	Non-mobile	34	1095
10	26	Married	7143	Car	35	1230

3.3.2. Results using the choice and constrained choice solutions

For the choice solution, all agents were classified into three types in terms of their residential location choice preferences (Table 6). Column w_1 indicates a preference for Amenity 1, and w_2 indicates a preference for Amenity 2, as shown in Fig. 2. In this experiment, r_{ij} in Eq. (1) was distributed uniformly from 0 to 0.1, and the sum of w_1 and w_2 was 0.9. We defined the agents with an income greater than or equal to 6000 as type A, agents with an age greater than or equal to 60 were defined as type C, and other conditions were defined as type B. In this way, a type A agent would care more about accessibility to Amenity 1 than to Amenity 2, and would therefore prefer to occupy parcels near Amenity 1. A type C agent would prefer to live in parcels close to Amenity 1, and a type B agent prefers Amenity 1 and Amenity 2 equally. The value of the spatial layer (Amenity 1 or Amenity 2) in Eq. (1) is calculated by $F = e^{-\alpha * dis}$, where α is the distance decay coefficient (set at 0.0001 in this study), and *dis* is the distance to the spatial layer. The preferences of all agents for all parcels can then be calculated as a matrix upon which the location choice process for disaggregating agent locations can be based. For the constrained choice solution, the locations of agents can be disaggregated using the approaches discussed in Section 2.4. In contrast to the choice solution, the agent count in each parcel is introduced into the disaggregation process using the constrained choice solution.

The results of various location disaggregation solutions vary significantly (Fig. 3). The agents disaggregated using the parcel solution obey the known agent number in each parcel. The agents disaggregated using the choice solution tend to approximate the two facilities, Amenity 1 and Amenity 2, due to their location preferences, and the results significantly differ from those obtained by the parcel solution. The agents using the constrained choice solution displayed similar results to those that used the parcel solution in space. However, the attributes of disaggregated agents obtained using the two approaches varied from each other greatly due to different disaggregating processes via checking agent attributes in GIS. Based on the results of various location solutions, the constrained choice is recommended here as the best solution for cases in which the spatial distribution of agents is known.

Table 6
Residential location choice preference for each type of agent.

Type	w_1	w_2
A	0.9	0
B	0.45	0.45
C	0	0.9

3.4. Model verification for all location solutions

Before applying Agenter to data synthesis, the model must first be verified by checking its response to the input, comparing two solutions for mapping agents, and comparing Agenter to the null model. Because applications in real areas are more complex than simulations, Agenter was verified in a synthetic space. The approach detailed in Section 2.5 was used to confirm the Agenter model in the synthetic space. Both attributes and locations of agents were considered in the verification process. First, since there were no observed samples in the synthetic space, one set of agents was disaggregated using the so-called best-constrained choice solution. The observed set of agents was then used to test the applicability of our model. Second, the null model was run 500 times to produce enough sets of results for comparison with those generated by Agenter (Fig. 3d). Third, for each type of location solution, Agenter was run 500 times to cover various results. Fourth, the disaggregated results produced by Agenter were compared to those produced by the null model using Eq. (2). The total

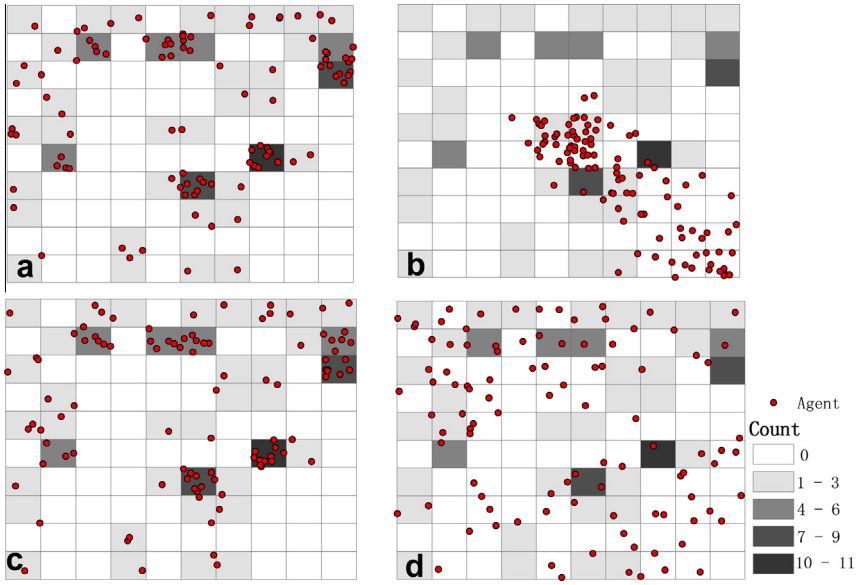

Fig. 3. Mapping disaggregated agents using various solutions: (a) parcel; (b) choice; (c) constrained choice; (d) null model. *Note*: The point for the location of each agent does not correspond to actual spatial distribution. It only shows which parcel the agent is in. The color of the parcel indicates the observed agent count, as in Fig. 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

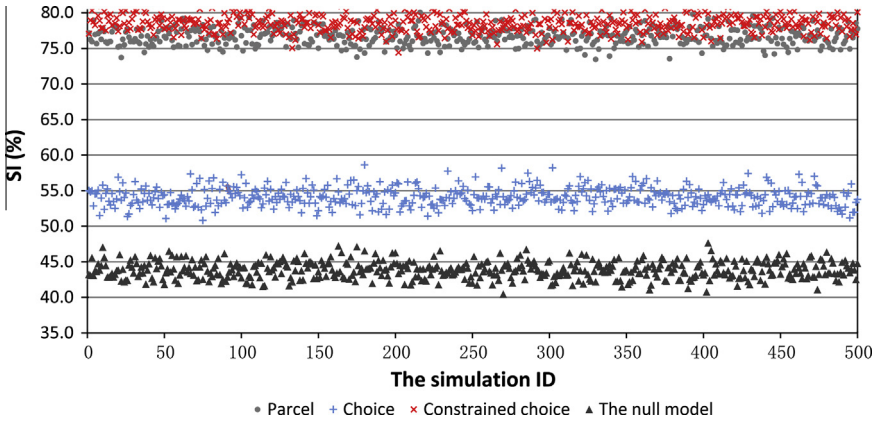

Fig. 4. Comparison of disaggregated results generated using the Agenter model and the null model.

number of agents (I) was 100. Three continuous attributes (age, income, and SCH) and three categorical attributes (marriage, travel, and parcel) were counted in this process ($U + V = 6$).

The verification results plotted in Fig. 4 showed that the similarity indexes (ranging from 40% to 80%) of the Agenter and the null model did not vary greatly, indicating that both models are stable and the disaggregated agents are repeatable. Agenter produced a significantly greater SI than the null model, with an average value of 43.9% (SD = 1.44%), demonstrating that Agenter is better suited for disaggregating agents. The better similarity index was assumed to be a result of known information being input into Agenter. Among the three location solutions, the constrained choice solution generated the greatest average SI of 78.5%, which was slightly greater than the SI of 76.7% observed for the parcel solution. The SIs of the choice solution were significantly lower than those of the constrained choice and parcel solutions, which was likely because the different patterns of agents disaggregated using different models increased the dissimilarities between the spatial location parcel and the spatial-aware attribute SCH. Overall, the results of this test suggested that the constrained choice solution is best among the three since the information introduced into Agenter was the richest.

4. Experiment in the Beijing Metropolitan Area

4.1. Input data

After the Agenter model was applied to the synthetic space, it was applied to the Beijing Metropolitan Area (BMA) to disaggregate heterogeneous residential agents. In this real-space experiment, Agenter's applicability for disaggregating residential agents was demonstrated using the Fifth Population Census Report of the BMA conducted in 2000 (the census), described in the Beijing Fifth Population Census Office and Beijing Municipal Statistical Bureau (2002), and the Household Travel Survey of Beijing conducted in 2005 (the survey). The census was conducted at the census tract level,[3] which is similar to the scale of a city block, each of which contains dozens of parcels in Beijing. Published census data were aggregated from the original census tract level to the district level (18 districts in the BMA). The total population count in the census was 13.819 million in the BMA.[4] The census data provided both the distributions and relationships among residents. Many cross-tabulations for various combinations of attributes are listed in this census report, and were used to obtain frequencies and build relationships as in Table 3.

The survey covers the entire BMA, including all 18 districts with 1118 Traffic Analysis Zones (TAZs) as the basic geographical survey unit (Beijing Municipal Commission of Transport and Beijing Transportation Research Center, 2007). The sample size was 81,760 households, housing a total of 208,290 persons. There was a 1.36% sampling ratio in contrast to the total population recorded in the census. The survey provided household information, including household size, hukou status (official residence registration),[5] residential location at the TAZ level, and personal information, including gender, age, household role and job type and location . The aforementioned information was used to further validate Agenter in the BMA. In addition, this survey included a one-day travel diary of all respondents, collected through face-to-face interviews. For each trip, the survey recorded the departure time and location, arrival time and location, trip purpose, mode of transit (both public and individual transportation), trip distance, type of destination building and transit route number.

In this experiment, the number of agents to be disaggregated was the same as the total number of residents recorded in the census (13.819 million). The attributes and locations of residential agents to be disaggregated within the BMA are listed in Table 7. The dependent relationships among attributes are illustrated in Fig. 5. To save space, all input information regarding distributions and relationships was stored online as Supplemental material for this paper. The format of the model inputs for the BMA experiment is the same as that in the synthetic space (Section 3.2 Input data).

The attribute PARCEL is the ID of the parcel used to map the disaggregated residents. All input rules are available online as tables in an MS Access file. Only typical tables are shown here. To disaggregate locations of the residential agents, the parcel GIS layer must contain agents for Agenter. The residential parcels (Fig. 6) were extracted from a land use map of the BMA from 2000, which has 133,503 polygon parcels, including 26,770 residential parcels. For each residential parcel, the floor area was obtained from aggregating buildings within each parcel. The number of residents in each parcel was allocated from each district available in the census report based on the floor area of each parcel, assuming homogeneity in residential floor areas in Beijing. This information is stored as frequencies of the attribute parcels.

[3] The spatial distribution of census tracts has never been released from the Beijing Municipal Statistical Bureau. Therefore, it is not possible to determine whether census tracts are compatible with TAZs.

[4] Both registered (with hukou) and unregistered residents (without hukou) were included.

[5] Both registered and unregistered residents were interviewed.

Table 7
Descriptions and known information for each attribute of residential agents in the BMA.

Name	Description	Type	Known information	Data source	Data type	Order
AGE	Age in years	Non-spatial attribute	Frequencies	The census	Ratio	1
SEX	Gender	Non-spatial attribute	Frequencies	The census	Nominal (male, female)	2
MARRIAGE	Marital status	Non-spatial attribute	Frequencies, RB (with AGE)	The census	Nominal (married, unmarried, divorced, remarried, widowed)	3
EDUCATION	Level of education	Non-spatial attribute	Frequencies, RB (with AGE)	The census	Ordinal (junior middle school, undergraduate, etc.)	4
JOB	Occupation	Non-spatial attribute	Frequencies, RB (with EDUCATION)	The census	Nominal	5
INCOME	Monthly income	Non-spatial attribute	Frequencies	The survey	Ratio	6
FAMILYN	Number of family members	Non-spatial attribute	Frequencies	The census	Ordinal (one person, two person, etc.)	7
PARCEL	ID of parcel at which the agent resides	Location	Frequencies	An empirical study	Nominal	8
TAM	Distance to the city center	Spatial attribute	Location of Tiananmen Square	Urban GIS	Ratio	9

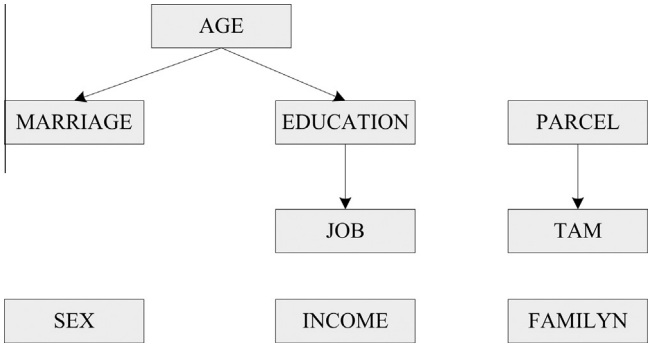

Fig. 5. Dependent relationships among attributes of residential agents.

4.2. Disaggregating residential agents in the BMA

The disaggregated residential agents (partially listed in Table 8) are mapped in Fig. 7, and are stored as the point Feature Class in the ESRI Personal Geodatabase. This dataset embeds both the attributes and location information of residential agents, which can be regarded as a primary dataset for urban studies and initializing agent-based models. The disaggregated residential agents and parcels took up 2.9 GB; the model requires less than 1 h to accomplish the disaggregation of 1 million residential agents (every agent has 10 attributes) for the experiment in the BMA. The test was conducted using a workstation with a CPU of 3.0 GHz * 2 and memory of 4 GB. The amount of time consumed by Agenter primarily depends on the number of agents, their attributes, the distribution, and the complexity of the relationship.

5. Discussion

5.1. Validation using the 2005 travel survey

We validated the Agenter model in the BMA using agents observed in the survey. In the survey, household and personal information such as age, sex, income, and number of family members were also included. TAM can be calculated based on the TAZ location (the centroid of each TAZ is used here) for this survey. For a better comparative validation, 208,291 individuals were

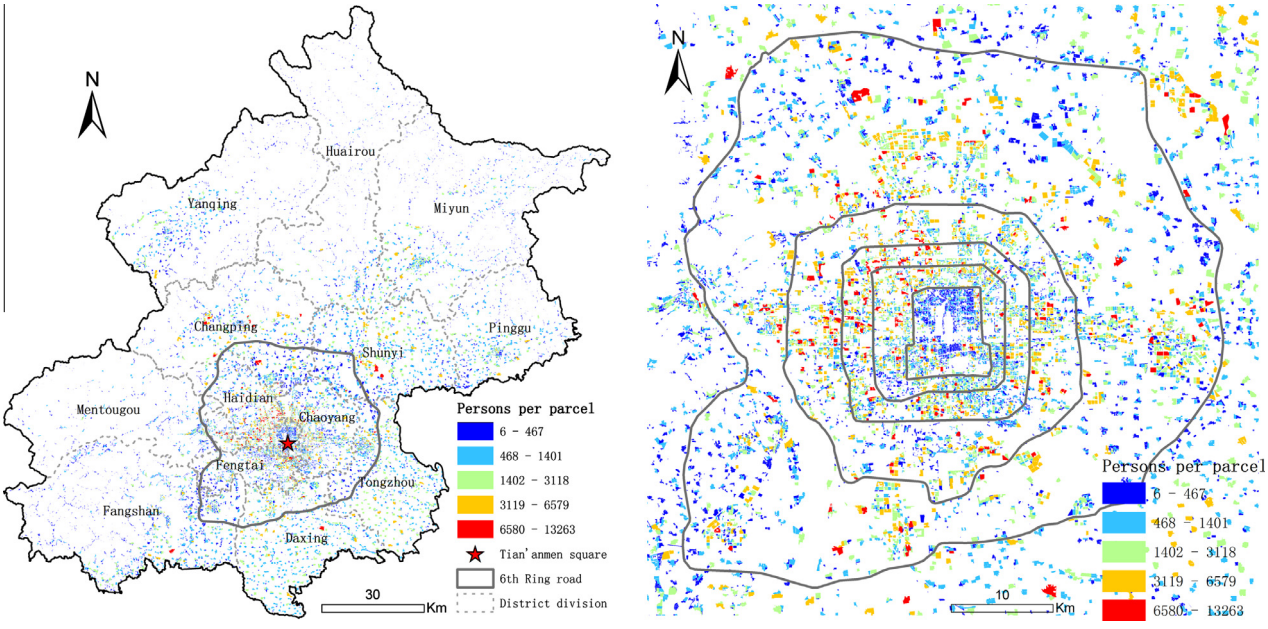

Fig. 6. Residential parcels in the BMA.

Table 8
Disaggregated residential agents (partial) in the BMA.

AID	AGE	SEX	MARRIAGE	EDUCATION	JOB	INCOME	FAMILYN	PARCEL	TAM
193392	36	Male	Married	Junior High/Middle School	Production, Transport Equipment Operator, and Related	2385	Three persons	888	2140
198316	41	Female	Married	High School	Production, Transport Equipment Operator, and Related	5966	Three persons	966	7747
37094	61	Male	Married	High School	Professional Technology Employee	4744	Three persons	523	5721
165014	27	Male	Unmarried	High School	Business and Service Employees	5559	Five persons	768	4957
2	41	Female	Married	Elementary School	Production, Transport Equipment Operator and Related	5351	Three persons	18	36739
49808	21	Male	Unmarried	Junior High/Middle School	Business and Service Employees	2684	Five persons	274	2905
189128	21	Male	Married	Junior High/Middle School	Production, Transport Equipment Operator and Related	2578	One person	878	4092
118806	8	Male	Unmarried	Elementary School	Production, Transport Equipment Operator and Related	0	Three persons	478	6949
33570	53	Female	Married	Elementary School	Production, Transport Equipment Operator and Related	1304	Five persons	929	23,760
179469	50	Male	Married	Elementary School	Farming, Forestry, Animal Husbandry and Fishery	4978	Two persons	804	2286

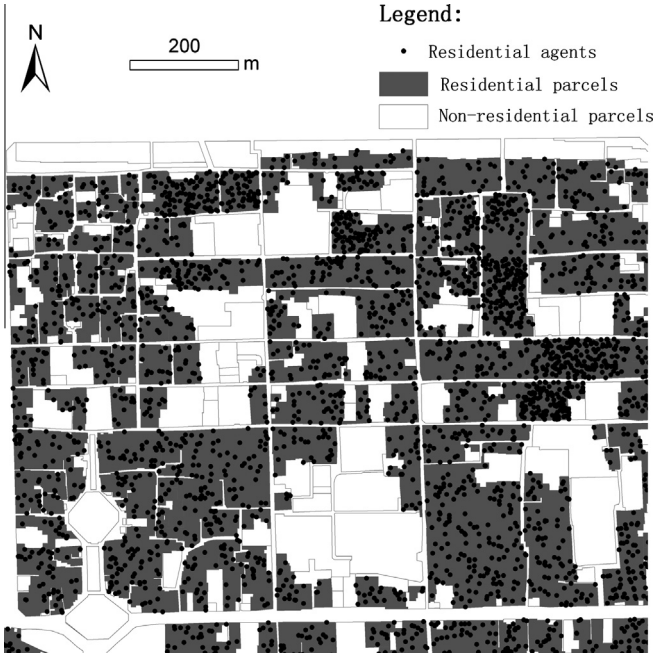

Fig. 7. Spatial distribution of disaggregated agents in the BMA (partial).

disaggregated using Agenter. Input information is given in Section 4.1. The parcel solution was adopted in Agenter. For comparison with the null model, both the Agenter and null models were run 500 times. In the disaggregated results, the attribute PARCEL was further aggregated into TAZ so the validation could be compared to that in a TAZ-scale survey. A total of six attributes were evaluated for calculation of SI. These were AGE, SEX, INCOME, FAMILYN, TAM, and TAZ. The similarity indicator SI was calculated using simulated results and observed results, as shown in Fig. 8.

The average SI of Agenter is 72.6%, which is significantly greater than that of the null model (43.9%), indicating that Agenter generates sounder disaggregated agents. Because the null model represents a random disaggregation process, the rules adopted in Agenter regarding the forms of distributions and relationships are part of why the present model outperforms the null model. As more comprehensive rules are entered into Agenter, its performance in terms of precision may increase further. As shown in Fig. 8, Agenter's behavior is stable in terms of SI. This demonstrates that Agenter can be used to reproduce individuals in actual situations, which means representative disaggregated agents can be retrieved by running Agenter a few times or even only once.

For more solid validation, we also broke down the average SI 72.6% for all TAZ and attributes across space and attributes. First, the average SI for each TAZ was calculated based on existing global results of 500 simulations. Those TAZs with more samples in and around the central part of the city were found to have greater SIs, indicating that more samples lead to better-disaggregated results. This was reasonable considering that the disaggregated results did not closely match the original samples. When a TAZ has only 10 or 20 samples, the results become less likely to match. Second, the SI was calculated based on the sorted disaggregated results and observed samples, as described in Section 2.5. Location was given the highest sorting priority, followed by other attributes. Under these conditions, the location of disaggregated results was the most consistent with observed samples. The lowest priority attributes showed the least consistency. For this reason, the attribute of interest should be given higher sorting priority. This may produce more consistent results.

5.2. Discussion on the experiment results in the BMA

All of the residents were successfully disaggregated in terms of spatial distribution and socioeconomic attributes in Beijing. This is the first time that the researchers and planners of the Beijing Institute of City Planning have been able to access large-scale disaggregated micro-data regarding the city of Beijing. Agenter was evaluated and found to be a convenient tool with explicit embedded algorithms, and its users are able to easily understand the disaggregation mechanism. The disaggregation results in Beijing were applied to several small-scale detailed plans for new towns in the Beijing area for the Beijing Affordable Housing Plan. The disaggregated population was found to be effective in supporting small-scale plans and policy evaluations, which required fine-scale individual data. Before Agenter, these applications were not possible in Beijing, where the municipal governments do not release individual datasets used for census reports or yearbooks. The extensive applications of this disaggregated population in the BMA are expected in the near future. The travel survey was conducted in 2005 and the census in 2000, and there were demographic changes in Beijing from 2000 to 2005. Unfortunately, the Beijing travel survey conducted in 2000 was not accessible. If access becomes available in the near future, the travel survey in 2010 and the population census of 2010 may be used to update

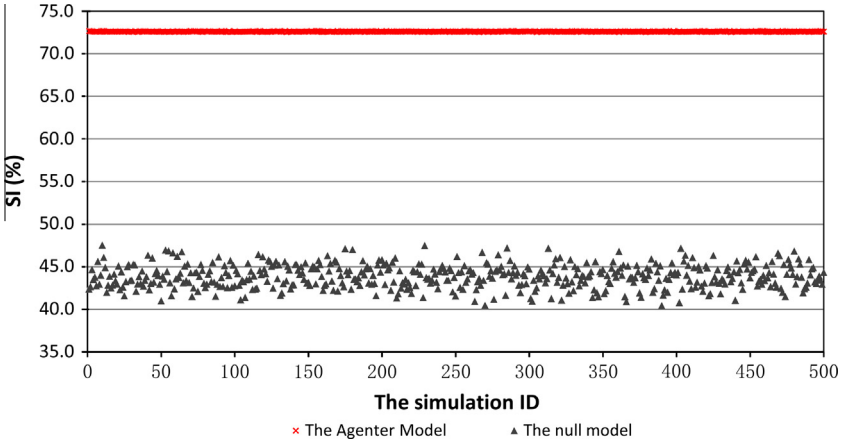

Fig. 8. Validation results of the BMA experiment and comparison with the null model.

the Agenter application in Beijing and disaggregate the 2010 population.

There is poor data availability within Beijing, as in most parts of China. Exploring the disaggregation of intensive datasets using known information provides an opportunity for micro-level simulation and analysis through agent-based models (ABMs). Crooks et al. demonstrated that ABMs focus on individual objects or categories, and thus disaggregate data are an essential determinant of their applicability (2008). Those areas for which there are insufficient micro-level data have a greater chance to develop their ABMs to simulate regional and urban dynamics supported by this approach. The approach proposed here incorporates known data and information to the greatest degree possible to initialize ABMs. This approach also sheds light on linking macro-datasets and intensive micro-datasets via disaggregation. In addition to initializing ABMs, this approach can be used in the construction of spatial population databases, which are essential to geographers and planners. The approach used herein can generate both population distributions and socio-economic attributes of the population in the form of a spatial population layer that is a key to an urban cyber infrastructure.

5.3. Limitations of the approach

There are currently several limitations to the Agenter approach. One is that its application can be constrained by the specific data requirements and assumptions of the approach itself. Specifically, the use of Agenter requires several steps. First, users must select the attributes to be disaggregated for agents based on the goals of the disaggregation. Second, the disaggregation order of attributes must be decided based on the dependent relationships among the attributes and the availability of existing information. Third, users must prepare the model input in terms of distributions and relationships as specified in this study by referring to the example provided in the online attachment. The online example is expected to solve this limitation to some degree. Regarding the second limitation, the agents disaggregated by our approach are not an exclusive set, even though they obey the same existing statistical characteristics of samples. When the disaggregated agents are used in an ABM, the user is expected to disaggregate numeric sets of agents via running Agenter repeatedly using the same input, run the ABM with each set, and then treat the mean value (or other statistical characteristics) of the results from all simulations as the final simulation result in order to reduce the uncertainty associated with applying this approach to ABMs. Because many combinations of micro-agents are generated, this process may eliminate issues associated with the ecological fallacy, a logical fallacy in the interpretation of statistical data in which inferences regarding the nature of individuals are deduced from inferences for the group to which those individuals belong. The ABM does not require an exact reconstruction of the surveyed population's original spatial distribution. Rather, it only needs an inferred distribution that approximates the actual distribution for purposes of reproducing similar patterns and interactions such as those found in the actual data. Based on this consideration, the model is acceptable for initializing ABMs, despite these limitations.

6. Conclusions

The Agenter approach is proposed in this paper as a means of disaggregating heterogeneous agent attributes and locations using known information, including aggregate data, small-scale surveys, and empirical studies. The agents to be disaggregated include non-spatial attributes, spatial locations, and spatial attributes. The known information is modeled as the distribution of an attribute and as relationships among attributes for disaggregation.

The Agenter model was developed based on the approach proposed here. It was used to disaggregate agents in a synthetic space and in the BMA. In the first of these experiments, several attributes were disaggregated using various types of known information, and then three types of location disaggregation solutions, including parcel, choice, and constrained choice, were tested. Agenter was verified using a similarity index (SI) to evaluate the similarities between the disaggregated and observed agents. Agenter produced significantly better disaggregation results than the null model (randomly disaggregated) in terms of SI. In the BMA experiment, the Fifth Population Census Report of Beijing in 2000, the Household Travel Survey of Beijing in 2005, an empirical study, and the Beijing urban GIS database were all used to infer frequencies of attributes and relationships among attributes for the disaggregation of all residential agents within the BMA. In this experiment, Agenter was further validated using micro-samples from the survey, and the average SI was found to be 72.6%. These findings indicate that Agenter can be applied in the real world to reproduce individuals that can then be fed into ABMs. Overall, this approach is appropriate to disaggregating agents in situations for which there are no micro-data from large-scale surveys. Specifically, the method developed here can make full use of existing statistical information, surveys, and empirical studies to disaggregate the attribute values and location of agents.

Even though this approach is best suited to preliminary exploration, it may solve the bottleneck problems associated with ABMs, like those caused by data scarcity in developing countries. Because

all spatially aggregated data are subject to the modifiable areal unit problem (MAUP), the disaggregated results may allow the user to avoid the MAUP, which occurs because the correlation between the variables in the aggregated data depends on the extent of the areal units used in the aggregation (Openshaw, 1984; Rees, Martin, & Williamson, 2002). Generally, there are few attributes recorded in samples, and unrecorded attributes can be disaggregated using this approach. The present approach could supplement conventional approaches and may be combined with traditional approaches. Studies on disaggregating Beijing populations with more attributes by incorporating Agenter and PopGen are underway.

Acknowledgments

We thank the National Natural Science Foundation of China (No. 51078213) for providing financial support. We also thank three anonymous reviewers for their valuable comments on our earlier manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compenvurbsys.2013.09.002.

References

A'Hearn, B., Peracchi, F., & Vecchi, G. (2009). Height and the normal distribution: Evidence from Italian military data. *Demography, 46*, 1–25.

Barthelemy, J, & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science, 47*, 266–279.

Beijing Fifth Population Census Office, & Beijing Municipal Statistical Bureau (2002). *Beijing population census of 2000*. Beijing: Chinese Statistic Press.

Beijing Municipal Commission of Transport, & Beijing Transportation Research Center (2007). *Beijing household travel survey of 2005*. Beijing: Internal report.

Benenson, I., Omer, I., & Hatna, E. (2002). Entity-based modeling of urban residential dynamics: The case of Yaffo, Tel Aviv. *Environment and Planning B: Planning and Design, 29*, 491–512.

Birkin, M., & Clarke, M. (1988). SYNTHESIS – A synthetic spatial information system for urban and regional analysis: Methods and explanations. *Environment and Planning A, 20*, 1645–1671.

Birkin, M., Turner, A., & Wu, B. (2006). A synthetic demographic model of the UK population: Methods, progress and problems. In *Proceedings of the second international conference on e-social science*. National Centre for ESocial Science, Manchester. <http://www.ncess.ac.uk/events/conference/2006/papers> Accessed 15.01.01.

Crooks, A. (2006). Exploring cities using agent-based models and GIS. *CASA working paper*, No. 109. Centre for Advanced Spatial Analysis, University College London, London.

Crooks, A. (2008). Constructing and implementing an agent-based model of residential segregation through vector GIS. *CASA working paper*, No. 133. Centre for Advanced Spatial Analysis, University College London, London.

Crooks, A., Castle, C., & Batty, M. (2008). Key challenges in agent-based modeling for geo-spatial simulation. *Computers, Environment and Urban Systems, 32*, 417–430.

Deming, W. E., & Stephan, F. F. (1940). On least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics, 11*, 427–444.

Eubank, S., Guclu, H., Anil Kumar, V., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature, 429*, 180–184.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge, MA: The MIT Press.

Hermes, K., & Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems, 36*(4), 281–290.

Holm, E., Lindgren, U., Makila, K., & Malmberg, G. (1996). Simulating an entire nation. In G.P. Clarke (Eds.), *Microsimulation for urban and regional policy analysis* (pp. 164–186). Pion, London.

Langford, M., & Unwin, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal, 31*, 21–26.

Li, X., & Liu, X. (2007). Defining agents' behaviors to simulate complex residential development using multicriteria evaluation. *Journal of Environmental Management, 85*, 1063–1075.

Li, X., & Liu, X. (2008). Embedding sustainable development strategies in agent-based models for use as a planning tool. *International Journal of Geographical Information Science, 22*, 21–45.

Liao, Y., Wang, J., Meng, B., & Li, X. (2010). Integration of GP and GA for mapping population distribution. *International Journal of Geographical Information Science, 24*, 47–67.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*, 31–42.

Müller, K., & Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. In *Proceedings of the 10*th *Swiss Transport research conference*. <http://e-collection.library.ethz.ch/view/eth:1623?q=microsimulation> Accessed 15.01.12.

Norman, P. (1999). Putting iterative proportional fitting on the researcher's desk. *Working paper* 99/03. School of Geography, University of Leeds, UK.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A, 16*, 17–31.

Pagliara, F., & Wilson, A. (2010). The state-of-the-art in building residential location models. In F. Pagliara et al. (Eds.), *Residential location choice: Models and applications*. *Advance in spatial science*. Berlin, Heidelberg: Springer-Verlag.

Rees, P. (1994). Estimating and projecting the population of urban communities. *Environment and Planning A, 26*, 1671–1697.

Rees, P., Martin, D., & Williamson, P. (Eds.). (2002). *The census data system*. London: Wiley.

Robinson, D. T., & Brown, D. (2009). Evaluating the effects of land-use development policies on ex-urban forest cover: An integrated agent-based GIS approach. *International Journal of Geographical Information Science, 23*, 1211–1232.

Ryan, J., Maoh, H., & Kanaroglou, P. (2009). Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis, 41*, 181–203.

Shen, Z., Yao, X., Kawakami, M., & Koujin, M. (2009). Simulating the impact on downtown of large-scale shopping centre location: Integrating GIS dataset and MAS platform as a case study in Kanazawa city in Japan. In *Proceedings of the 11th international conference on computers in urban planning and urban management*.

Smith, D. M., Clarke, G. P., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A, 41*, 1251–1268.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Wong, D. W. S. (1992). The reliability of using the iterative proportional fitting procedure. *The Professional Geographer, 44*, 340–348.

Wu, B. M., Birkin, M. H., & Rees, P. H. (2008). A spatial microsimulation model with student agents. *Computers, Environment and Urban Systems, 32*, 440–453.