

Computers, Environment and Urban Systems

我的一篇文章被折磨的历史，目前已经发表的三篇几乎偏偏如此，不得不承认审稿人厉害，要求严格，文章品质从头到尾一直在改善。

R1

The response letter to the reviewers

We authors of this manuscript would like to thank the anonymous reviewers' comments and the editor's suggestion on the earlier manuscript of this paper. We have carefully revised this manuscript according to all comments as follows (sentences in bode are our replies).

The Reviewer 3

Some very good ideas but your model validation is problematic. The paper also needs a serious editing by an English speaker/writer.

A: This manuscript has been polished by an English speaker to improve its readability.

Check out EpiSim to see how they disaggregated data for Portland, Oregon(S Eubank, H Guclu, VS Anil Kumar, MV Marathe. - Nature, 2004)

A: We have referred this paper in the introduction section as follows.

The TRANSIMS population synthesizer also uses IPF for generation of synthetic households with demographic characteristics in addition to placement of each synthetic household on a link in transportation network and assignment of vehicles to each household (Eubank *et al.*, 2004)

Abstract

Page 1 (PDF page numbers)

Line 8: re-write

Line 10: "restraining" to "constraining"

Line 13: "have" to "has"

A: The abstract has been revised according to the above comments.

Main
Page 2
Line 39: rewrite
Line 57: rewrite

A: The two lines have been rewritten accordingly in the modified manuscript.

Page 5
Line 20-4 rewrite "obeying"?

A: These lines have been rewritten accordingly in the modified manuscript.

Line 27 section 2.2, are you losing information by putting things in ranges? Does it assume a uniform distribution over the age class? Individual characteristic can change drastically over a range (i.e. 16-21 ages: in high school vs in college)

A: The categories of a continuous attribute, e.g. age, are set according to the available information on the distribution of this attribute or its relationship with other attribute(s). For the prior condition, if we know the count of each age category from a survey (namely the distribution of age), e.g. 50 persons are 0-3 years old, and 60 persons are 4-7 years old, then we set the age categories as 0-3 and 4-7 yrs. For the later condition, if we have the cross table of age with marriage condition, the age class in this cross table will be referred in discretizing the age attribute. For instance, there are counts for marriage status for persons in 18-30 and 30-40 years old, 18-30 and 30-40 will be regarded as the age categories in the discretizing process.

Accordingly, the age class is not assumed to uniformly distribute as commented by the reviewer. We acknowledge that the suggested education status may change dramatically over an age range 16-21. We suggest that the discretizing process considers both the available information and the common sense like what the reviewer commented.

We have revised the example as follows in the revised manuscript.

“For example, the attribute MARRIAGE has two categories, married, and unmarried, while the attribute AGE can be divided into a variety of categories such as 0 to 4 years old, 4 to 7 years old, and 7 to 12 years old.”

Page 8
Line 13 reference for MNL model

A: A reference with regard to the MNL model has been added into the revised manuscript.

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*, Advances in Spatial Science. Springer-Verlag, Berlin, Heidelberg.

This line has been rewritten in the revised manuscript.

Page 15

Line 39 3.4 You are using a model generated set of disaggregated individuals to test the efficacy of a set of model (same one) generated disaggregated individuals. What you are really doing is seeing how different one model run from the average of 500 model runs. This is not terribly meaningful and you will get different results with each single run (and range over the PDF of model runs). A real set of individuals may not behave as your model has predicted because your assumptions about how individuals behave may not be very good.

A: In the revised manuscript, we states that we verified the model in the hypothetical space and validated the model in the BMA experiment using observed micro samples available in a household travel survey of Beijing in the current Section 4.2 Validation using the 2005 travel survey.

The Reviewer 2

1. Introduction

The paper proposes an approach to disaggregate heterogeneous agents based on existing data sources such as aggregate data, small-scale surveys, and empirical researches. It addresses an important problem in agent-based modeling research. Ideally, agent-based models would use data disaggregated at individual level. However, such data is very hard to obtain due to privacy concerns. The authors propose a method aim to disaggregate census data into disaggregated agents, based on in terms of both location and attributes in accordance with known information under current poor condition of data availability.

2. Remarks on Section 1 – Introduction and motivation

There is a mismatch between the paper's stated goals and its scientific content. The paper deals with disaggregating census data to individuals, a technique known broadly as *geocoding*. As a scientific problem, geocoding has many more applications than agent based models, including public health, poverty, gender and race studies. Thus, the paper's title is misleading, as agent-based modeling is but one of the motivations for disaggregating census data. Since the methods presented in the paper are not specific to agent-based modeling, the paper's objective should be rewritten. Instead of "*we aim to create individual agents with disaggregated characteristics in terms of both location and attributes in accordance with known information under current poor condition of data availability*", it would be better to write something like "*this paper presents a method of disaggregating census data into individual attributes in a situation where microdata is not available. Our interest is to use individual-level data for agent-based modeling*".

A: We have broadened the potential application of our approach from ABMs to micro models like ABMs and microsimulation models for regional, urban and population studies. We mentioned ABMs are our interest in the revised manuscript as suggested by the reviewer. In addition, the sentence has been modified accordingly.

The authors are encouraged to elaborate on possible differences between disaggregation for population studies and for agent-based modeling. Highlighting the differences would improve the value of the paper. It is plausible to argue that agent-based modeling does not require the exact reconstruction of the original spatial distribution of the surveyed population. The requirements of agent-based modeling require that the inferred distribution approximates the actual distribution for the purposes of reproducing similar patterns and interactions as those found in the actual data.

A: We are appreciated for this suggestion. We have added this in the last paragraph of the discussion section to highlight the differences for improving the value of this paper. In addition, we have compare our approach with IPF for population synthesis in the field of population studies in the introduction section from various aspects, which are also expected to improve the potential value of our paper.

3. Remarks on Section 2 – Methods

Section 2 of the paper presents a method for disaggregating individual from census data using the relationships between the census variables. Section 2.1 contains many assumptions that are used later in the paper. Many of these assumptions are described in the following paragraphs:

“Attributes of agents are divided into two types, non-spatial attributes (e.g. age, income and education for a residential agent) and spatial attributes (e.g. subway accessibility, land use type and building height this residential agent occupies). (...) Since the spatial attributes of an agent depend on its location and environment context, the agent disaggregation order is non-spatial attributes, location, and spatial attributes.”

“The disaggregation order among agent attributes and location should be taken into account. For instance, from existing data sources, the attribute A has known distribution. The attribute B has known distribution and known relationship with the attribute A. The attribute C has known relationship with the attribute D. We should firstly disaggregate the independent attributes, which do not depend on other attributes. Following with this, the dependent attributes with known relationships with other attribute(s) can be disaggregated” (page 5, line 6ff).

Although used throughout the paper, the authors do not justify the assumptions on the above paragraph. Their assumptions are a case of “ecological fallacy”. To quote from the Wikipedia entry:

An ecological fallacy is an interpretation of statistical data from census or ecological data, whereby inferences about the nature of specific individuals are based solely upon aggregate statistics collected for the group to which

those individuals belong. This fallacy assumes that individual members of a group have the average characteristics of the group at large. However, statistics that accurately describe group characteristics do not necessarily apply to individuals within that group. The variability of individuals is much greater than the variability of their mean.

A: We have validated our approach in Beijing via comparing the disaggregated results with the observed samples (about 200,000 micro samples) in a household travel survey. The similarity index between them is over 70%, further indicating the applicability of our approach. In addition, besides the statistic indicator like the average value, we mostly use an attribute's distribution list to disaggregate agents, thus guaranteeing the variability of disaggregated individual members. The mentioned assumptions are the precondition of our approach and they come from empirical works and researches as well as researchers' common sense. We hope this could justify the assumptions. More, the most common approach for population synthesis also uses similar assumptions with our approach.

The authors assume implicitly the correlation structure of the disaggregated data is the same as the one of the aggregated data. There is extensive literature on census data analysis that undermines these assumptions. See, for example, the classic paper by Stan Openshaw, "Ecological fallacies and the analysis of areal census data" (Environment and Planning A, 1984) or the book "The Census Data System" by Rees, Martin and Williamson.

A: To our knowledge, the heavily used population synthesis approach IPF also uses the correlation structure (namely cross tables between two attributes) for population synthesis. There are not areal problems in our approach thus avoiding this problem. In addition, our approach will benefit the avoiding of ecological fallacies and MAUP problems since we generate micro data instead of aggregated data.

The authors also assume there are "independent attributes" which will be disaggregated first, and then used to disaggregate the dependent attributes. However, the relation between attribute such as MARRIAGE and AGE is rather arbitrary and culture dependent. They authors cite two papers as evidences of the relations between MARRIAGE and AGE, one related to Jews in Israel, and another comparing pre-industrial Germany with Japan. It is doubtful that one can generalize based on these two case studies.

A: This example and the two references have been deleted in the revised manuscript. In addition, the relationship between MARRIAGE and AGE from the two references is only an example to indicate that empirical researches could contribute to the disaggregation process. The actual relationship may vary from the application case of our model.

Thus, the approach taken by the authors should be considered an *ad hoc* solution to the problem of disaggregating census data. In section 3.1, they consider that MARRIAGE depends on AGE, and that TRAVEL depends both on AGE and MARRIAGE. The concept of dependence used by the authors is thus informal. It would be better to use more empirical statistical correlations, which could support the

authors' assumptions.

A: The experiment in Section 3 is conducted in a hypothetical space aiming to demonstrate the applicability of our approach. We also verifying our approach in Section 3 using assumed inputs. Thus, the dependence and relationship in this virtual space are not totally practical in the real world. There is more empirical statistical information in the BMA experiment as the reviewer suggested.

In section 2.2, the authors propose to convert all attributes to categorical (ordinal) scales of measurement. They state: “*we convert all continuous data into categorical type for simplification of the disaggregation process and reduce the computing time*”. This argument is questionable. It is not clear to the reader why simplification is needed.

A: We have revised this sentence as follows.

Since the information provided by most continuous attributes is presented in categorical form, we converted all continuous data into categorical data to conform to this characteristic and reduce the disaggregation time.

In section 2.3.1, the formula and its explanation are not strictly needed. The authors could state that the disaggregated values follow the frequency distribution. Also, the author's description of the procedure to be applied in the case of a probability density function could be described in a simpler way.

A: We have deleted Formula 1 and related explanation according to this comment. In addition, the description of the procedure to be applied in the case of a probability density function is also simplified in the revised manuscript.

In section 2.3.2, formula (2) is not necessary, since it adds no new information. The term “*probability relationship*” is a little confusing. From the description, it appears that the authors refer to conditional probability, which would express the relationship between two attributes (A and B). Given the use of conditional probability, it would be natural to use Bayes' rule. Instead, the authors provide a solution that is not easy to follow. Finally, they say they use the “Monte Carlo” method but do not provide details.

A: Formula 2 in the former manuscript is deleted according to this comment. We use the term “conditional probability” in the revised manuscript. We have revised this part as follows.

The second is the conditional probability relationship (hereafter RB, in some cases also known as joint probability). Under these conditions there is a probability relationship between the attribute j and its related attribute h as $P(h|j)$. We know the frequencies of attribute j as $P(j)$ and the categories of both attributes. Then, the probability of each combination of categories of attributes h and j can be calculated using $P(h,j) = P(h|j)P(j)$, and the count of each combination is $P(h,j)$ multiplied by the agent total count. For instance of 100 persons, the attribute j (AGE) has

two categories, 18-30 years old (40%) and 31-60 years old (60%). We know its conditional probability with the attribute h (MARRIAGE) that if a person is 18-30 years old, he/she will be 60% married and 40% unmarried, if a person is 31-60 years old, he/she will be 80% married and 20% unmarried. In this means, there will be $40\%*60\%*100=24$ persons are in 18-30 and married, $40\%*40\%*100=16$ persons are in 18-30 and unmarried, $60\%*80\%=48$ persons are in 31-60 and married, and $60\%*20\%*100=12$ persons are in 31-60 and unmarried.

In section 2.4, the authors describe two alternatives for setting up the spatial location of agents. Their first solution (“parcel”) considers the case when the number of agents in each parcel is known. The second solution (“constrained choice”) uses a multinomial logit model to allocate agents. This is arguably the most interesting part of the paper, since the problem of spatial allocation of agents is central to the proposed work. Unfortunately, the authors’ description of the method is confusing to read. They use the term “*environmental factors*” to refer to spatial attributes such as “*accessibility to education facilities, neighbourhood similarity, and landscape quality*”. The term “environmental factor” is not a usual one in this context. It would be better to say “spatial attributes”. Also, the authors do not discuss how these “environmental factors” would be computed.

A: We have revised Section 2.4 according to this comment and the term “environmental factor” has been replaced by “spatial layer”. In addition, we added a sentence to state the calculation of “environmental factors” as follows.

... which can be calculated by overlaying the parcel with the spatial layer in GIS.

From a conceptual point of view, the constrained choice approach would be the most significant result of this paper. Since the constrained choice model is using additional information than the choice model, it has the potential to produce better results than the “parcel” model. Currently, this potential contribution is limited by the shortcomings of the author’s experiments. In section 3 below, the authors point out there are differences in the results of allocating agents using the “parcel” and the “constrained choice” solutions. However, they do not explain what are the differences and how significant they are.

A: We also admit that the constrained choice solution is more appropriate for generating practical results in contrast to other solutions we proposed. In Section 3, we added more explanation regarding the differences between agents using the parcel model and the constrained choice model as follows.

The results of various location disaggregation solutions vary significantly (Fig. 2). 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 choice preferences, and the results differ

from those obtained by the “parcel” solution significantly. The agents using the “constrained choice” solution display very similar results to those that use the “parcel” solution in space. However, the attributes of disaggregated agents obtained using the two approaches vary from each other greatly due to different disaggregating processes via checking agent attributes in GIS. Based on comparison of the various location solutions, we recommend the “constrained choice” as the best solution in cases in which the spatial distribution of agents is known.

4. Remarks on Section 3 – Results

This is a confusing section of the paper. When performing the experiments, the authors make explicit assumptions that modify their model. For example, equation 7 (page 12 of the manuscript) describes a relationship between TRAVEL, INCOME and AGE. This relationship should be part of the model, described on section 2, since it describes the relationship between attributes. The presentation of such an equation on section 3 (“results”) suggests that the model for disaggregating agent proposed by the authors is mostly an ad-hoc collection of rules that are context-dependent. If that is the case, the author’s model has a limited power of generality.

A: We admit that the equation and other input data in Section 3.2 are specific for applying our Agenter model in the hypothetical space. The model provides a data and approach framework to tackle an extensive range of information as input data to disaggregate individual samples, that is, the inputted rules are context-dependent and there are no parameters embedded in the Agenter model prior to its case application. In our opinion, our model has its general power in applications. If the other user applies our model, he/she should prepare the model input in the specific form like illustrated in this paper and run the model to get desired disaggregated samples.

In section 3.4, the authors talk about “validation of the model”. They compare their “parcel” model with the results of a null model that randomly allocates of agents. Such comparison can only perform model verification, that is, it will determine if the model is plausible. It is not capable of validating the model, in the sense of finding out if the model is accurate and appropriate. The verification procedure should also have been applied to the “constrained choice” model.

A: We admit the validation process in our former manuscript is not solid. We have replaced the term validation as verification. We further verified the constrained choice model in the revised manuscript. The validation of our approach is added in Section 4.2.

5. Remarks on Section 4 – Experiment

This is the most interesting section of the paper, since it describes the author’s model when applied to real data. Unfortunately, the authors are not careful to describe the experiment with the detail required for the reader to understand what they did. In table 7, the authors state that some of the attributes (such as MARRIAGE, JOB and EDUCATION) have additional information about conditional probability distribution as (called RB by the authors and described as “*probability relationship between the attribute j and its related attribute h*”). However, Table 7 only states that these attributes have dependency relations, without saying what are the dependencies.

A: The detailed input in this experiment has been partial added (due to page limitation) in the revised manuscript for better illustrating our application to readers. More important, more detailed tables will be available for readers in the online attachments of this paper as mentioned in our revised manuscript.

Furthermore, given that the authors have access to real data, such data could be used to validate their model. Instead, they give performance indicators (“less than 1 hour” in page 19, line 51), information of limited value to the reader. Thus, the authors make limited use of an interesting dataset.

A: We have got the household travel survey data of 2005 in Beijing to validate the experiment in the BMA in the modified manuscript. Details are available in Section 4.2 Validation using the 2005 travel survey.

6. Remarks on Section 5 – Discussion

On section 5, the authors point out the need to disaggregate census data for agent based modelling and compare their approach to others in the literature on section 5.2. This reviewer thinks that section 5.2 could be moved to the start of the paper, since it contains the motivation for the author’s work. In section 5.3, the authors discuss some limitations of their approach. However, this reviewer thinks that the main limitation of the author’s work is the ad-hoc nature of the conditional relations between the attributes. Such limitation is not discussed on section 5.3.

A: We have moved the former Section 5.2 to the introduction section in the revised manuscript according to this comment. The reply regarding the ADHOC is discussed in the reply for the below comment (the last one).

7. Final Comments and Recommendation

This has been a difficult paper to review. Clearly, the authors are not native English speakers and the paper needs a major language review. However, the main problem of the paper is its lack of clarity. The problem that authors are trying to solve is well described in the paper. However, their approach is marred by a confusing presentation. The reviewer needed many hours and various readings of the paper to understand the paper. A normal reader of a scientific journal would have no such time or patience.

A: This manuscript has been polished by a native speaker of English.

The main drawback of the paper is the ad-hoc nature of the author’s proposed model, which mostly amounts to stating ad-hoc relationships between the variables and then using these relationships to disaggregate the census data in individuals. This arbitrary approach is not enough for the paper to be accepted in a scientific journal. The authors need to provide a sounder theoretical foundation for the disaggregation problem. A sound paper on disaggregation, describing solutions that could be valuable to CEUS readership, would be a welcome addition to the literature.

A: The model provides a data and approach framework to tackle an extensive range of information as input data to disaggregate individual samples, that is, the inputted rules are context-dependent and the model input could vary from the case application. There are no parameters embedded in the Agenter model

prior to its case application.

The inputs of the model's application in the virtual space are not totally practical in the real world since we aimed to test the applicability of our approach in Section 3. In the test in Beijing in Section 4, we use practical data coming from various official and academic channels and have successfully validate the approach using observed micro samples in a household travel survey of Beijing.

In our opinion, our model has its general power in applications. If the other user applies our model, he/she should prepare the model input in the specific form like illustrated in this paper and run the model to get desired disaggregated samples.

R2

The response letter to the reviewers

We authors of this manuscript would like to thank two anonymous reviewers' comments and the editor's suggestion on the earlier manuscript of this paper. We have carefully revised this manuscript according to all comments as follows (sentences in bode are our replies).

The Reviewer 3

This is a much-improved paper from the last review. It still needs some sharpening on a few key points. One problem I still have is with "3.4 Model verification for all location solutions". It appears that you are comparing the results of all the other models to the "constrained choice" solution? If that is the case than I'm not sure what that is really testing. I've enclosed an annotated PDF with additional comments.

A: We are verifying, rather than validating, our proposed approach in Section 3.4. Since there are no observed samples in the synthetic space, we disaggregated one set of agents using the so-called best "constrained choice" solution as "observed" existing samples. The "observed" set of agents was then used to test the applicability of our model. The purpose is to show the model response on various solutions as well as test the performance of the model. E.g., we found that *"the similarity indexes (ranging from 40% to 80%) of the Agenter and the null model do not vary greatly, indicating that both models are stable and the disaggregated agents are repeatable"*.

In addition, all minor comments embedded in the annotated PDF this referee enclosed have been addressed in the revised manuscript. We authors are appreciated for those detailed suggestions from the referee.

The Reviewer 2

1. Remarks on Section 1 – Introduction and motivation

Although the authors have improved the paper, some sentences show that they need to reflect more about what the “modifiable areal unit problem” (MAUP) means. Consider the statement:

“Analyses based on spatially aggregated data have been criticised for leading to ecological fallacies and research into the “modifiable areal unit problem” (MAUP) has demonstrated that using areal data for drawing statistical influence is not appropriate (Openshaw, 1984; Rees et al., 2002).”

This sentence is not correct. What Openshaw and co-authors have demonstrated is that the MAUP is an inherent problem of aggregated data such that produced by census. It would be better to state:

“All spatially aggregated data are known to be subject the “modifiable areal unit problem” (MAUP) (Openshaw, 1984; Rees et al., 2002). The MAUP occurs because the correlation between the variables in the aggregated data depends on the extent of the areal unit used in the aggregation. Furthermore, many authors have shown that it is incorrect to draw up correlations of individual behaviour based on aggregated variables.”

A: The authors have revised this part according to this comment.

The part of the first paragraph that begins with “Agent-based models...” (section 1, 1st paragraph, line 25) could be moved to the top of the introduction. It describes the purpose of the paper. The earlier the reader reads it, the better he will understand the author’s objectives.

A: We have revised this paragraph according to this comment. This paragraph has been moved to the top of Section 1.

The second paragraph of the introduction makes many statements about agent-based models (ABMs) that are contentious. This reviewer thinks most of them could be removed from the paper without loss of content. The readability of the paper would actually improve. Some examples: “*ABMs... can better explain issues that are less explored by macro models*” .“*The information provided by an agent includes its behaviour, attribute and spatial location, that are not well considered in existing spatial ABMs*” “*Most ABMs cannot actually reflect one agent in the model as a single micro-sample*”

These sentences are not needed. Furthermore, this reviewer considers that they are wrong and misplaced. Most spatially explicit ABMs have the capability of modeling either individuals or clusters of persons. They correctly point out the work of Benenson et al. (2002) as a good example of individual-level modeling. However, in many cases the choices of whether using ABMs as individuals or groups is based on consideration other than data availability. Removing those sentences and rewriting the paragraph would make the paper more readable.

A: We authors have removed those sentences and rewritten the second paragraph of the introduction section.

2. Remarks on Section 2 – The research approach

2.1 Remarks on Section 2.1

The authors have improved Section 2 of the paper by providing a better description of their methods. Although the section has been improved, the discussion continues to lack rigour. I will show some examples below.

In page 6, line 28 ff, the authors state: “The probability distribution of an attribute (hereafter distribution) and the dependent relationship among attributes (here after relationship) should be retrieved from existing data sources including aggregate data, small-scale surveys and empirical studies”. The verb “retrieved” is incorrect, as it implies that such probability distribution exists. In practice, the authors should use “inferred” rather than “retrieved”.

A: The verb “retrieved” has been replaced by “inferred” throughout the paper.

A similar comment applies to the sentence “Empirical studies can provide existing knowledge regarding samples”. Most likely, the authors mean “the probability distribution of an attribute can also be deduced by empirical studies”. Likewise, the sentence “To generate agent attributes and locations, the three aforementioned types of existing data should be converted into the specific forms of distribution and relationships”(page 6, line 50 ff) would better be expressed as “to convert from aggregate data into individual samples, we need to estimate the probability distribution of the attributes and the relationship between them”.

A: We have revised them according to this detailed comment.

The paragraph starting in line 57, page 6, is central to the method proposed by the authors. It reads: “*The disaggregation order among agent attributes and location should also be considered. For example, attribute A has a known distribution based on existing data sources. Additionally, attribute B has a known distribution and known relationship with attribute A, while attribute C has a known relationship with attribute D. We should first disaggregate the independent attributes, which do not depend on other attributes. Next, the dependent attributes with known relationships to other attributes can be disaggregated. For instance, we should first disaggregate attribute A, then attribute B, because attribute B has an existing relationship with attribute A. Additionally, a confliction check should be conducted to guarantee that the disaggregated results make sense.*” (pag 6, line57ff).

This paragraph needs much improvement. The authors also assume there are “independent attributes” which will be disaggregated first, and then used to disaggregate the dependent attributes. What the authors mean is that there are some attributes that can be related to each other. Given information that is available, a researcher could setup a dependency relationship between these attributes. But, as the authors show in the remainder of the paper, such dependency between attributes is arbitrary. The concept of “independent attribute” used in the paper is context-dependent. In some case, some variables are better suited to be taken as independent attributes than others. As I stated in my first review of the paper:

“The approach taken by the authors should be considered an ad hoc solution to the problem of disaggregating census data. In section 3.1, they consider that MARRIAGE depends on AGE, and that TRAVEL depends both on AGE and MARRIAGE. The concept of dependence used by the authors is thus informal. It would be better to use more empirical statistical correlations,

which could support the authors' assumptions."

Thus, the paragraph cited above needs to be rewritten so as to better express the fact that the authors use the idea of dependence in an informal way. It is an established fact in statistics that the establishment of "dependent" and "independent" variables is context-dependent. This ad-hoc relationship needs to be stated clearly in the paper.

A: The supposed relationships among attributes in this paragraph are not totally practical in the real world, and it is used as an example to express our ideas on disaggregation order. In practical application of our approach, we recommend users use empirical statistical correlations for determining the order. In addition, we have significantly rewritten the paragraph the reviewer pointed out according to all comments proposed to express this informal but applicable way.

2.2 Remarks on Section 2.2

In section 2.2, the authors state:

"The data describing agent attributes can be divided into continuous (e.g., age and income) and categorical (e.g., marriage and education) types".

As it becomes clear from the following text, both MARRIAGE and AGE attributes are categorical: *"For example, the attribute MARRIAGE has two categories, married, and unmarried, while the attribute AGE can be divided into a variety of categories such as 0 to 4 years old, 4 to 7 years old, and 7 to 12 years old."*

As the authors state, the attribute AGE has a "variety of categories". Thus, it is categorical. Thus, the distinction between "categorical" and "continuous" is not an appropriate one for the method used by the authors. It would be much better to use the distinction suggested by Stevens in his classical paper: nominal, ordinal, interval and ratio. Using Stevens' distinctions, the methods proposed by the authors could be described in a more rigorous way than what is done in the current version of the paper. Thus, in the paper, MARRIAGE, TRAVEL, PARCEL would be nominal attributes, AGE would be ordinal, INCOME would be a ratio attribute.

Thus, the authors should rewrite section 2.2 to use Stevens' description of nominal, ordinal, interval and ratio types for their attributes, instead of the problematic distinction between "categorical" and "continuous" types.

A: We authors have accepted this comment and referred to Stevens (1946)'s classification approach. This has been modified throughout the new manuscript.

2.3 Remarks on Section 2.3

The authors have improved 2.3. The text is now better to read. Some glitches remain, notably on the first paragraph, which talks about a "functional relationship" type that is no longer used in the paper (except in Table 1). I suggest the authors remove this mention.

A: The functional relationship is common in the real world. So we keep it in the revised manuscript.

2.3 Remarks on Section 2.4

The authors have improved section 2.4, but some issues still need revision. The sentence "For allocation of agents, we should first partition the entire study area into small numeric units" is most likely an inadvertent error. The authors probably mean

“partition the study area in small spatial units”.

A: We have revised this in the modified manuscript, by using “partition”.

2.3 Remarks on Section 2.5

The validation index proposed by the authors is based on the distinction between “categorical” and “continuous” attributes. However, as expressed in my comments to section 2.2, it would be more appropriate to refer to “ordinal”, “nominal”, “interval” and “ratio” types. In this case, the authors would have to improve the definition of the A_u indicator, since most of the variables they describe as continuous are in fact have “ordinal” types.

A: As the reply to the comment for Section 2.2, we have applied Stevens (1946)’s definition. The validation index has been accordingly modified.

3. Remarks on Section 3 – Results

The authors use the term “hypothetical space” throughout the paper, especially in section 3. I suggest to replace this term with “synthetic space”, since the space the authors have built for their experiment is not hypothetical, but it is a synthetic one.

A: This term has been replaced throughout this manuscript.

The results for the so-called “hypothetical space” serve to show how the model works. Therefore, the discussion on model verification (section 3) is of limited scientific value in comparison with the results on the Beijing Metropolitan Area, they have less scientific values. Thus, the paper readability and its value for the readership would be improved if the section 3 would be omitted and more details are given about the Beijing MA experiment (section 4). As it is, the experiment in the synthetic space is clearly described, but the experiment in BMA needs to be better explained.

A: In the revised version, we have paid more attention on the experiment in the Beijing Metropolitan Area with details available in the replies for the below comments as well as in the modified manuscript.

4. Remarks on Section 4 – Results in Beijing Metropolitan Area

This section needs improvements. First of all, the authors need to improve the writing. They mix the description of the census data (which provide the attribute correlation) with the description of the Beijing travel survey. It would be better to first describe the census data available, and then in a separate paragraph to describe the travel sample survey.

A: In the revised manuscript, we have separated the descriptions for the census data with those for the travel survey. We have carefully added more details for the both data according to this and the following comments.

There are a number of assumptions in the author’s experiment that need to be better explained. These assumptions do not invalidate the experiment, but they need to be made explicit so that the readers can fully understand the potential and the limitations of the methods proposed by the authors.

A: We have revised this section according to this and the following detailed comments in the modified manuscript to make it clearer.

The authors need to provide more information on both the census and on the survey. What is the spatial resolution of the census data? Is the data available aggregated by census tract? In this case, are the census tracts compatible with the TAZs (traffic analysis zones) used in the travel survey? What does the 1.36% sample rate mean?

A: We have provided more information for both the census and the survey in this section. The spatial resolution of the census data is a census tract, which is not available in this research. The data we used in the paper is the census report aggregated in the district level published by the Beijing Municipal Statistical Bureau. The census tracts have never been released officially. In this end, we could not evaluate whether they are compatible with TAZs. 1.36% is the sampling ratio calculated via dividing all surveyed population by the total population recorded in the census report. “rate” is a typo and we have corrected in the revised manuscript.

A further point is that the travel survey was done in 2005, and the census in 2000. There was a substantial change in the city demographics in the period. How does this change affect the study?

A: Thank you for this comment. The population census in China is generally conducted in every ten years, e.g. 1990, 2000 and 2010. We did not successfully access the Beijing travel survey conducted in 2000 in preparing, as well as now revising, the manuscript. We admit the demographic changes in Beijing during 2000-2005. In our near future, with the access of the travel survey in 2010 as well as the population census in 2010, we will update our Agenter application in Beijing and disaggregate population in 2010. We have addressed this in the revised paper.

To what extent does the travel survey measures the city’s population? What is the percentage of the Beijing population that is covered in the travel survey? How is the travel survey done? Does it include public transportation (e.g., metro and bus) or also include individual transportation (e.g., cars, bicycles)? What about the non-travel residents, those with hukou permits? Are they considered in the 2000 census and in the travel survey?

A: The sample size is 81,760 households/208,290 persons, with a 1.36% sampling ratio in contrast to the total population recorded in the census (13.819 million). All travel modes are included in the travel survey we used. In addition, both registered and unregistered residents are included in the survey and census. These have been addressed in the revised manuscript.

What was the spatial resolution of the disaggregation? What does it mean to state “the count of agents to be disaggregated is the same as that in the census”? Does it mean that 13 million agents were created? How were they grouped into parcels?

A: The spatial resolution of the disaggregation in the BMA is a parcel. “the count of agents to be disaggregated is the same as that in the census” means that 13

million created and we have updated the expression of this sentence. Since we know the resident count in each parcel, we could group all disaggregated individuals into parcels.

What do the authors mean by “the attribute PARCEL indicates the ID of the parcel the resident occupies that was used for mapping the disaggregated results”? The authors state that they have had access to the land use map of the BMA, which has 133,503 parcels, of whom 26,700 are residential parcels. They state: “they have synthesized the residential count of each parcel using that of each district obtained from the census report and the floor area of each parcel”. How was that done? How was the floor area obtained?

A: For clearer expression, we have rewritten this sentence as “the attribute PARCEL indicates the ID of the parcel, which will be used for mapping the disaggregated residents”. The floor area is obtained from aggregating buildings within each parcel. In addition, the resident number in each parcel was allocated from that in each district available in the census report based on the floor area of each parcel, supposing that the residential floor area of each resident is homogenous in Beijing. These have been indicated in the revised manuscript.

5. Remarks on Section 5 – Discussion

I found the discussion section lacking in specific content. This section should be focused on discussing the results of the simulation of the agents in the BMA and what was learned from the experiment. The validation of the Agenter model (section 4.2) should be moved from the earlier section to this section and the results should be discussed in more detail.

A: In the revised manuscript, the former Section 4.2 regarding the validation of the Agenter model has been moved to Section 5. Moreover, the validation results as well as the experiment results in the BMA have been discussed with more details.

6. Final Comments and Recommendation

I stand by my earlier assessment: this has been a hard paper to review. The English revision made by the authors improved the paper, but it still lacks clarity. The authors put too much focus on secondary aspects (such as the simulated model) and less emphasis on the important results (the disaggregation of agents in Beijing). Many relevant questions are not discussed by the authors.

The problem that authors are trying to solve is well described in the paper. However, their approach is marred by a confusing presentation. The reviewer needed many hours and various readings of the paper to understand the paper. A normal reader of a scientific journal would have no such time or patience.

A: This manuscript has been polished by an English speaker to improve its readability. In addition, we hope the revisions, thanks to all comments by the reviewer, could significantly promote the readability and clarity of this manuscript.

R3

The response letter to the reviewers

We authors of this manuscript would like to thank two anonymous reviewers' comments and the editor's suggestion on the earlier manuscript of this paper. We have carefully revised this manuscript according to all comments as follows (sentences in bode are our replies). We are meanwhile submitting an attachment showing marks on our modification according to the editor's suggestion. In addition, this manuscript has been thoroughly polished by an English-editing agency.

The Reviewer 2

This is a much-improved paper from the last review. It still needs some sharpening on a few key points. One problem I still have is with "3.4 Model verification for all location solutions". It appears that you are comparing the results of all the other models to the "constrained choice" solution? If that is the case than I'm not sure what that is really testing. I've enclosed an annotated PDF with additional comments.

A: Thank you for the comment. We are verifying, rather than validating, our proposed approach in Section 3.4. Since there are (should be) no observed samples in the synthetic space, we disaggregated one set of agents using the so-called best solution "constrained choice" as "observed" existing samples. The "observed" set of agents was then used to test the applicability of our model. The purpose is to show the model response on various solutions as well as test the performance of the model. For instance, we found 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"*.

In addition, all minor comments embedded in the annotated PDF this referee enclosed have been addressed in the revised manuscript. We have had this manuscript edited by an English-native speaker, thoroughly. We authors are appreciated for those detailed suggestions.

The Reviewer 4

I like the notion that this paper promotes - disaggregation without microdata from large scale surveys. However, I have several issues with this paper:

(i) While I understand what the authors are trying to do, I don't fully understand why the authors have used this approach and not drawn more on microsimulation/pop syn (other methods than IPF, e.g. simulated annealing or conditional probabilities) and then linked to ABM. Fundamentally, I find this approach difficult to understand and thereby find value in (how different is it from other established methods?). Is this just a search and allocation routine (optimisation)?

A: Our approach in the manuscript focuses on disaggregation in condition of no micro-sample, as indicated in the introduction section: “*This study focused on individual disaggregation without micro data from large scale surveys.*” In cases of no large-scale micro data/samples, the traditional disaggregation approach like IPF, would not apply. The differences between our approach and existing approaches have been elaborated in the fifth paragraph of the introduction section.

The differences between the present study and previous IPF-based studies, such as those conducted by Birkin and Clarke, Rees, Birkin et al., Ryan et al., and Smith et al. are as follows (1988; 1994; 2006; 2009; 2009). First, the present synthetic reconstruction approach can generate micro data using only aggregate data and information. This approach does not require individual samples. However, census based IPF, 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 in 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. and Smith et al. (2006; 2009). Third, the spatial locations of samples, which are essential to spatial ABMs, receive special attention in this approach, as advocated by Birkin and Clark and Wong (1988; 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 initialization of ABMs in the present paper, and 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 disaggregation of micro data.

In addition, generally there are limited attributes recorded in samples, and those attributes not in samples could be disaggregated using our approach. That is, the approach proposed in this manuscript could supplement the existing approach and has potential to combine with existing approaches. This has been further highlighted in the last paragraph of the conclusion section as follows.

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 under way.

(ii) I would argue that the agents only appear at the end of this process when the method has initialised them and they will then be used to do something with. Until then, they are just empty units, to call them agents is confusing, after all, the method is an agent-producer, it could also be called a spatial-attribute unit producer.

A: We aim to use the disaggregated results generated by the approach for initializing agents in ABMs. In this regard, we term the approach as “Agenter Producer (Agenter)”. The disaggregated units have non-spatial attributes, spatial attributes (like the accessibility to the nearest school) and their spatial locations. That is, the location is not the only elements in the disaggregated results. Therefore, we keep this term in the revision and added additional explanation when it first appears in the manuscript (first paragraph of the introduction section).

(iii) I do not have confidence in the results; the Chinese census only samples a small proportion of the total population - statistics, at a small scale, are not available to allow the model to be fully tested against. As the authors note, the travel survey is only 1.36% sample. There is simply not enough detail given in the real data study or enough complexity introduced into the synthetic study.

A: Although the sampling ratio covering all of Beijing metropolitan area (not a small scale) is low in the Beijing survey, we disaggregated 208,291 persons using Agenter in Section 5.1 (validation in Beijing), rather than total population in Beijing (almost 20 million). Our logic to avoid the shortcoming of low sampling ratio is to disaggregate Beijing residents using various inputs and compare them with the surveyed ones (with the same totals).

I also found in several places the phrasing confusing - perhaps this explains my lack of understanding! There are also a number of typo's throughout.

A: This manuscript has been polished by an English-editing agency to avoid language confusing and typos.

Abstract: Last sentence: the findings show that the developed model has the potential to be applied. insert the word potential, it makes a slight change in emphasis.

A: Modified.

I would be careful saying that you have created a whole new approach; while the series of steps may have some novelty in this area, I'm not wholly convinced that it's a new method. I would suggest making it explicit that the phrase "Agenter" is a reference to the series of steps, rather than a name for a new method.

A: We have mentioned that Agenter stands for a series of steps in the first paragraph of the introduction section.

Introduction:

Set out at the beginning a clear rationale and context of your problem.

A: We have rewritten the beginning of the introduction section with indicating the context of our work as follows.

Micro models using individual-level data, such as agent-based models (ABMs) and microsimulation models, have increasingly been discussed in

the context of regional, urban, and population studies as supplements to traditional macro models (Wu et al., 2008). However, the use of micro models has been hindered by poor availability of individual data due to privacy and cost constraints. Some 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 very 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, and empirical studies. These disaggregated results could be used as input for ABMs and microsimulation models. ...

Can you clarify your sentence about "current poor condition of data availability" - application specific? "with the goal of generating individual-level data as agents for initializing the widely applied ABMs " I don't understand this sentence; are you using ABM to create new data that is then used to initialise the agents with?

A: We aim to generate micro data, using Agenter, for feeding into ABMs. That is, the disaggregated micro data could be used as input for ABMs. This sentence has been revised to avoid confusion as follows.

These disaggregated results could be used as input for ABMs and microsimulation models.

The Crooks (2008) statement was correct back in 2008; however, there has been so much data proliferation in the past 5 years that I don't think this holds as strongly. Are the authors making a statement about a particular application or geographical area?

A: We have deleted citation to Crooks et al (2008) in the introduction section. Our solution would have its potential applications in developing countries where micro data from large scale surveys are not common.

After the MAUP has been discussed, you refer to many researches - where are they references? Back up such statements. MAUP doesn't seem to be picked up on later - why not?

A: We have deleted "many researches" in the modified manuscript. We mention MAUP here just to say, micro models using individual data are possible to avoid MAUP.

Considering the emphasis on ABM and broadly speaking, microsimulation techniques in the introduction - a short paragraph discussing the differences and synergies between the two methods seems appropriate. I'm particularly keen on this as I cannot fully understand why the authors don't use other microsimulation/population synthesis techniques to do this. I can't clearly see what the gain is from using an ABM.

A: We aim to use our solution Agenter for disaggregating/synthesizing individual agents in condition of with no micro samples. The disaggregated agents (the output of Agenter) could be used for initializing an ABM or microsimulation

model. We are not using ABM for disaggregating samples in Agenter.

Although we are not using microsimulation or ABM in our approach, we have added a short paragraph in the first paragraph of the introduction section on discussing the differences and synergies between the two methods, as follows.

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 phenomenon at the macro level, using on 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.

p.6; first paragraph: (e.g. the cross-tabulation of marriage-age.)

A: Revised throughout the manuscript.

2.1 : why chose these characteristics?

A: Three types of information of disaggregated results are spatial attributes, non-spatial attributes and location. They are selected according to the general requirements of ABMs and the current data availability.

2.3: Present a flow chart or a series of equations - the text is a little difficult to get through. This all sounds like classic microsimulation. In addition, this is an important part of the methodology; it should be far more precise.

A: We have added a flow chart for Section 2.2 to 2.4 in the end of Section 2.1 for illustrating the basic ideas of our disaggregation solution.

2.4: How did you decide on the size of the spatial unit - is your method transferable? Optimisation algorithm?

A: In this study, we used the parcel as the basic spatial unit for allocating agents according to the spatial data availability and the expected application requirement. This has been indicated in the modifications. In this regards, the approach we developed could be applied in other spatial units with various sizes.

So agents are put into an area based on numbers that should be there (known from surveys)? The use of the MNL with this really does sound like conditional probabilities (okay, it's not exactly the same, but very similar). These are empty units put into an area - are they really agents?

A: In our “constrained choice” solution, agents are allocated into an area based on two types of information. The first is numbers that should be there according to known data sources e.g. surveys. The second is MNL coefficients used in the “choice” solution. Empty spatial units will not be allocated any agent in the “constrained choice” solution.

3.1 I think it's a good idea to use a synthetic data set. The authors could have

increased the levels of complexity in the space/attributes through a few different experiments. This would have tested out the approach more rigorously. The synthetic experiment is very simplistic and the results not too surprising.

A: We thank the reviewer for this comment. All types of spatial allocation solutions are included in the synthetic experiment, so are various types of distributions and relationships. More, the synthetic test is used to verify the model and check the response of the model to the model input. In this regard, the synthetic test is sufficient to verify and check our proposed approach. More practical and complex conditions have been considered in the Beijing experiment. The proposed approach was validated in the Beijing experiment, rather the synthetic test.

3.4: Can you be more explicit what exactly you are verifying against?

A: Before applying Agenter for data synthesis, we need verify the model via checking the model response to input, comparing two spatial solutions for mapping agents, as well as comparing Agenter with the null model. Since an application in a real area is more complex, we verified Agenter in the synthetic space. This has been highlighted in the revised manuscript.

4: The results of 72.6%; can these be broken down? What type of "agents" is it getting correct? Which types are being misclassified? What sort of areas are they? A more detailed breakdown of the results is needed, far more than a single SI figure.

A: Thanks for the helpful comment. We have broken down the global SI in space and attributes. A new paragraph has been added as follows.

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 match the original samples closely. 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.

Discussion is general, and does take account of the main points. Presumably, this creates a snapshot of the disaggregate data? If you have census data from 2005 and a travel survey from 2007, how do you reconcile these? Why use this approach rather than a traditional disaggregation approach such as provided by microsimulation (SA, CP or IPF)? I'm not satisfied that I have an answer.

A: The samples from a survey were for 2005 and we disaggregate residents for 2000. The reason lies in lacking of samples in 2000 for validating the results. This has been explained in the first paragraph of Section 5.2.

Our approach in the manuscript focuses on disaggregation in condition of no micro-sample, as indicated in the very beginning of the introduction section. In cases of no large-scale micro data/samples, the traditional disaggregation approach like IPF, would not apply. The differences between our approach and existing approaches have been elaborated in the fifth paragraph of the introduction section as follows.

The present study targets the third level of 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, Rees, Birkin et al., Ryan et al., and Smith et al. are as follows (1988; 1994; 2006; 2009; 2009). First, the present synthetic reconstruction approach can generate micro data using only aggregate data and information. This approach does not require individual samples. However, census based IPF, 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 in 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. and Smith et al. (2006; 2009). Third, the spatial locations of samples, which are essential to spatial ABMs, receive special attention in this approach, as advocated by Birkin and Clark and Wong (1988; 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 initialization of ABMs in the present paper, and 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 disaggregation of micro data.

In addition, there are limited attributes recorded in samples, and those attributes not in samples could be disaggregated using our approach. That is, the approach proposed in this manuscript could supplement the conventional approach and has potential to combine with traditional approaches. This has been further highlighted in the discussion section.

I found several of the images very hard to read - this is possibly due to the resolution in the version I had. Certainly Figures 2 and 5 were difficult to interpret. I would also recommend they have some accompanying figures/statistics.

A: Some of images were compressed and their resolution in PDF was low. The

original figures are with high resolution. In this revision, we have redrawn the original Figure 2 and 5 to increase their readability.

R4

The response letter to the reviewers

We authors of this manuscript would like to thank the editor's suggestion on the earlier manuscript of this paper. This manuscript has been thoroughly polished by an English-editing agency for the second time. We are meanwhile submitting an attachment showing marks on our modification according to the editor's suggestion.