LISER

LUXEMBOURG INSTITUTE OF SOCIO-ECONOMIC RESEARCH

n° 2017-07 Mai 2017

WORKING PAPERS

Generating a located synthetic population of individuals, households, and dwellings

Jean-Philippe **ANTONI** [1]
Gilles **VUIDEL** [2]
Olivier **KLEIN** [3]

[1] University of Burgundy, France
[2] University of Franche-Comté, France
[3] LISER, Luxembourg

Generating a located synthetic population of individuals, households, and dwellings*

Jean-Philippe Antoni
ThéMA Laboratory, UMR 6049 CNRS – University of Burgundy, France

Gilles Vuidel
ThéMA Laboratory, UMR 6049 CNRS – University of Franche-Comté, France

Olivier Klein
Department Urban Development and Mobility, LISER, Luxembourg

February 2017

Abstract

Some of the new approaches in urban modeling, such as multi-agent systems (MAS) or activity-based models (ABM), require inputs in the form of disaggregated individual data. But for privacy protection reasons, such data is seldom available at this level. One way to get around this obstacle is to generate a synthetic population. This paper presents a method for generating a population from fully aggregated socio-demographic and geographic data. Based on French examples, this method can be reproduced anywhere in the country providing a relevant linkage between the characteristics of agents and those of urban spaces. The proposed method is subdivided into two steps. First, a population of agents belonging to households, as well as of households ascribed to housing units, is generated from the socio-demographic data. Second, this population is located by assignment to the buildings generated from the geographic data. Testing and validating the method on three French cities (Besançon, Strasbourg and Lille) generates useful results but also some difficulties, particularly for certain population categories. Ultimately, we obtain a realistic three-dimensional database of the study area where agents and spaces are represented and realistic individual information can be mapped and used to model the behavior of agents through MASs or ABMs.

Keywords: Multi-agent systems; Agent-based modeling; Microdata; Synthetic population; Agent/space framework

* This research is part of an INTER-PICS project supported by the Luxembourg 'Fonds National de la Recherche' (contract FNR/INTER/CNRS/12/02) and by CNRS (France). It was initiated in the framework of the LUTI MOBISIM modeling project (http://thema.univ-fcomte.fr/mobisim/).

1 Introduction

The high degree of urbanization in many parts of the world and its impact on the sustainability of urban systems has made it a critical policy and planning issue. Consequently, there is currently a great demand for planning tools able to identify development strategies that minimize the potential problems of unbridled urbanization. The first such tools were prescriptive and based on geographic information systems (GIS) with little predictive capability (Webster, 1994). More recent ones are based on dynamic modeling approaches and try to describe urban processes before predicting urban growth. These models are usually based on complex systems theory (Batty, 2007) and promote the use of tools such as cellular automata (CA) to simulate the evolution of land-use or agent-based approaches to consider the behavior of the actors involved in the urban development process. But although land-use modeling has quite a long tradition in geography and planning, approaches based on agent behavior are less operational because their actual implementation is more complex. Even so, researchers have maintained their faith in the value and capacity of these models to provide a better understanding of individual processes and their urban development (Kirman, 1992). The description of actors' behavior is central to understanding the choices and decisions people make about their residential location, their mobility strategies, and the resulting land use.

Since the early 1990s, behavioral approaches have led researchers to model urban dynamics using micro-simulation, multi-agents systems (MASs), or activity-based models (ABMs). Several experiments and research programs have shown that agent-based approaches, i.e. considering individuals as agents able to communicate and to make decisions in the environment in which they live (Phan and Amblard, 2007), sometimes offer a better solution than traditional aggregated approaches to understanding the real world because they furnish a better description of this world and its underlying processes (Benenson, 1998; Wu *et al.*, 2008). But wherever it has been developed, such an approach also leads to the identification of a crucial difficulty during the construction phase of agent-based modeling compared with CA approaches: there is often a wide gap between the data required and that available, varying with national contexts and legislation.

1.1 Data and confidentiality

In social sciences, MAS or ABM approaches must be fed with individual data, i.e. data disaggregated at the level of the individuals who make up the study population. Moreover, this data must be precisely located in space so as to integrate the behavioral dimension of the movements and trips that individuals might make in the areas in which they live. The modeling should allow to consider short pedestrian trips or segregation questions, for instance, the location of individuals at the fine scale of buildings containing housing and economic activities (shops, services, and employment) in 2D (Benenson *et al.*, 2002) and possibly 3D (Crooks *et al.*, 2011).

But depending on national legislation or privacy rules about statistical surveys or censuses, such data is rarely available at this level. Access to individual data usually poses a serious challenge because their disclosure may lead survey respondents to become distrustful. Researchers are therefore sometimes constrained in their thematic studies and forced to make their own additional surveys involving additional public costs (Ahmad *et al.*, 2010). To tackle this challenge, different practices are found around the world to protect individuals from disclosure and allow researchers to work in better conditions. As described in Figure 1, several approaches to individual data access can be observed in European OECD-member countries, depending on the nature of the data:(1) population censuses, (2) household surveys, (3) administrative registries data, (4) business data; and on the applied confidentiality method: (A) public use files, (B) extracts or samples, (C) scientific use, (D) public tabulations. The conclusion is that fully disaggregated individual information is never available for the whole range of data required to understand an individual's behavior. Consequently, inputs needed for agent-based modeling are not available in sufficient quantities to automatically constitute a complete database about agents considered as individuals.

One solution to this problem is to generate a synthetic population, i.e. an artificial population composed of individuals associated with individual characteristics (level n) and constructed from the data known at the aggregated level (level $n + 1$) of the census. The advantage of this artificial population is then that social characteristics can be correctly analyzed at a fine scale while personal data remain anonymous and confidential (Hermes and Poulsen, 2012).

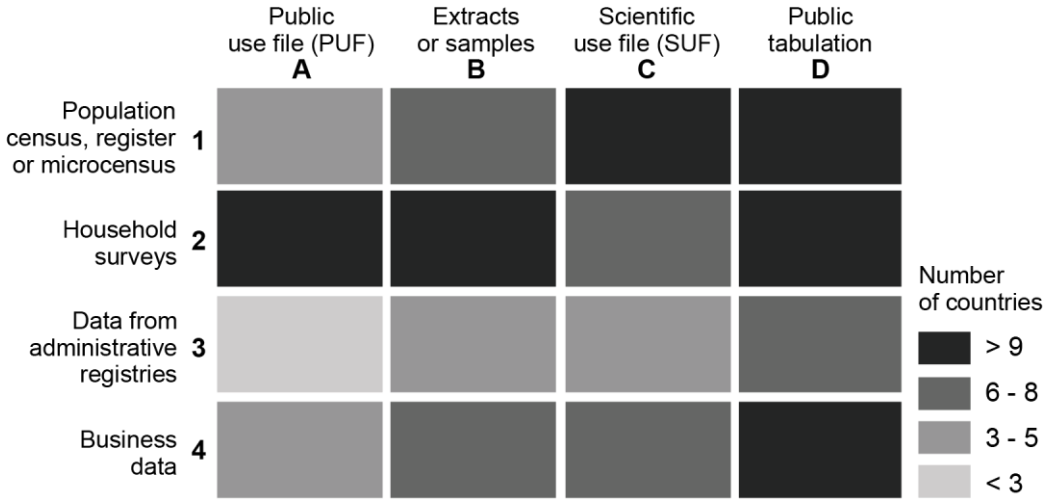

Figure 1: Microdata access in OECD countries (Eurostat, 2008)

1.2 Synthetic populations

Several methods have been developed worldwide for generating synthetic populations (Miller, 2003; Hermes and Poulsen, 2012; Farooq *et al.*, 2013; Namazi-Rad *et al.*, 2014) and a list of the most common population synthesizer software, including models and approaches such as ILUTE or Albatross, have been published by Müller and Axhausen (2010). The best known method is certainly iterative proportional fitting (IPF) introduced by Beckman *et al.* (1996) to create a collection of agents based on two datasets available in the United States: aggregated data given in the standard census file (case D1 in Figure 1) and a public use microdata sample (case B1). This method has been adapted to transportation problems by Arentze and Timmermans (2007) using Dutch datasets composed of the same kind of data: an aggregated population dataset (case D1), and a travel survey dataset (B2). In Germany, Münnich and Schürle (2003) developed another methodological framework to reach the same goal but from different kinds of data. They used Monte-Carlo simulations to generate categorical variables from a specific survey (C2). This approach was modified by Alfons *et al.* (2010) and applied at the scale of the European Union (EU) using the statistics on income and living conditions (SILC) provided by the EU. All these methods can be considered effective for the different contexts they study.

Nevertheless, two shortcomings can be identified: (i) these methods are not usable when the initial datasets do not correspond to the appropriate cases summarized in Figure 1; (ii) they ignore the spatial context at an appropriately fine scale and usually

fail to associate the generated synthetic population with specific geographic information. However, such precise geographic information is increasingly available worldwide through the survey and the dissemination of large-scale topographic vector databases, also called digital landscape models that cover many European countries[1].

1.3 Objective

Developing a new methodology by which to integrate this geographic information with aggregated population data (case D1 in Figure 1) is therefore a useful objective for all the countries that dispose of such data alone. Such a situation, where large individual samples are unavailable, is common in many European countries but also in developing countries like China, or in Southeast Asia, South America, and Africa (Long and Shen, 2013). France is also in this position and serves as a good illustration of this methodology. Moreover, only a few French studies have so far investigated the question of synthetic population generation and they usually ignore the spatial dimension of the territories studied (Mathis, 2006), or are based on highly specific surveys that are not available for the whole territory (Banos, 2010).

From the French example, the aim of this paper is to present the MobiSim Population Synthesizer, a model and software for generating a synthetic population ensuring: (i) the compatibility of the method with aggregated demographic data (case D1) and precise geographic information; (ii) the reproducibility of the method in every part of the studied territory; and (iii) the definition of a relevant linkage between the characteristics of the households (and the associated agents) and the characteristics of the spatial dimension of the city's residential structure (housing and buildings). The MobiSim Population Synthesizer[2] is tested on the French urban areas of Besançon, Strasbourg, and Lille[3], where a number of agents *A*, households *M*, and housing units

1 For example, Meridian 2 has been developed in the UK by the British Ordnance Survey, the Amtliche Topographisch-Kartographische Informationssystem (Atkis) database in Germany by national surveying agencies, and BDTopo in France by the French Geographic Institute (IGN) and is available for most French-speaking countries in Europe, etc. Other similar databases are currently collected in the OpenStreetMap project (http://openstreetmap.org) that should soon allow everyone free access to detailed geographic information worldwide.

2 MobiSim Population Synthesizer is part of the MobiSim project. MobiSim is an agent-based LUTI simulation platform dedicated to the geographical analysis of daily and residential mobility interactions, developed by J.P. Antoni, C. Tannier, and G. Vuidel (Antoni and Vuidel, 2010).

3 These three urban agglomerations differ greatly in size, functions, and structure. They have been selected empirically to provide contrasting cases through which to validate the method.

L is generated according to the aggregated available data, so that one housing unit contains one household composed of one or more agents. Agents a, households m, housing units l, and buildings b are defined as:

$$a = \{gender,\ age,\ role,\ education,\ activity,\ income\} \quad (1)$$

$$m = \{\cup a\} \subseteq A \quad (2)$$

$$l = \{type, occupation, \#rooms, m\} \quad (3)$$

$$b = \{\cup l\} \subseteq L \quad (4)$$

This structure allows the micro-level of agents to be linked to the macro-level of city shape, assuming a close link between the social and the spatial dimensions of the city. Its construction obviously depends on the structure of the available datasets described in section 2.

2 Material and methods

2.1 Data sources

The MobiSim Population Synthesizer requires available data sources for each part of the territory under study. Local surveys or specific micro-censuses are therefore excluded and the information must be based on national censuses and databases. In France, such databases fall into two groups according to their format:

1. Socio-demographic data are presented as statistical tables aggregated into spatial units. Such data is provided for the whole country by the general population census (RGP) conducted by the national institute of statistics and economic studies (INSEE) in census areas called IRIS. The country is divided into approximately 16,000 IRIS areas containing between 1,800 and 5,000 inhabitants. On the basis of this census, several datasets can be downloaded from the INSEE's website[4] that groups the information into thematic packages. Five packages are of direct interest for our objectives: (i) population (number of inhabitants by gender, age, and nationality, population in households, etc.), (ii) family (population in households by age, inhabitants living alone by age, population by marital status), (iii) education (school

[4] www.insee.fr

population by age, diploma by gender and age, etc.), (iv) activity (working population, kind of activity by gender, age, and social category, transport mode used for commuting, etc.) and (v) housing (number of housing units and status, housing by number of rooms, type, area, year of construction, occupation, etc.). A sample of these datasets is provided in the downloadable appendix B1.

2. Geographic data is presented as precise geo-located spatial information readable in a geographical information system (GIS) and is taken from BD-Topo, supplied by the National Geographic Institute (IGN)[5]. This global vector topographic database is accurate to the nearest meter and is available in a GIS shapefile format. It contains several information layers and includes a precise description of residential and non-residential built-up areas describing the shape, height, and altitude of all buildings. These buildings can be located in the IRIS used by the INSEE to collect sociodemographic information. A sample of this database is given in the downloadable appendix B2.

These two groups of data constitute the main inputs for the MobiSim Population Synthesizer and partially determine the methodology of disaggregation and linkage between agents, households, housing units, and buildings.

2.2 Methodology

Depending on the complexity of the algorithms used and the number of arbitrary decisions they require, the methodological aspects of the data generation process for creating synthetic populations are usually published in working papers, while most of the published literature focuses exclusively on the presentation of results (Hermes and Poulsen, 2012). In this paper, we briefly present the MobiSim Population Synthesizer approach in two points. In the first point (§ 2.2.1), a population of agents *a* belonging to households *m* belonging to housing units *l* is generated from the socio-demographic data. In the second point (§ 2.2.2), this population is located by assignment to the buildings *b* generated from the geographic data. Fully detailed information about the model's operation can be found in the downloadable software package in appendix B3.

[5] http://professionnels.ign.fr/bdtopo

2.2.1 Population generation

Agents

Initially, a number *n* of undefined agents is generated, in line with the total population of the study areas: 236,070 agents for the urban area of Besançon, 641,859 for the urban area of Strasbourg, and 1,163,936 for the urban area of Lille. The age class (0-2, 3-5, 6-10, 11-17, etc.) of these agents is then determined according to the national census and their precise age *B* is randomly distributed within this class. Depending on *B*, the gender *S* of each agent is randomly determined by a conditional probability provided by the socio-demographic data. Equation (5) shows, for instance, the probability that an agent is male if aged between 0 and 14 years old. This probability is calculated from the census variable $n_{H,0\text{-}14}$ (count of males between 0 and 14 years old) $n_{0\text{-}14}$ (count of agents between 0 and 14 years old):

$$p(S_H | B_{0-14}) = \frac{n_{H,0-14}}{n_{0-14}} \quad (5)$$

Similar probabilities are used to define the education level (depending on age and gender) and occupation (depending on age, gender, and education) of the agents, to complete the agents' vector of information defined in (1).

The role of an agent in the household (child, single, couple, etc.) is more complex to define but necessary for determining household structure. Each role category is iteratively defined from households' census data in two steps: (i) the child category depending on the agents' age, followed by single adult depending on gender and age, single parent depending on gender and age, couple with or without children depending on gender; (ii) the remainder is put into a category named "other" including house-sharing, not living as a household (military personnel, homes for the elderly, religious orders, etc.) and so on.

Households

Households are formed in two steps: (i) the creation of households and the assignment of adult agents; (ii) the assignment of children to family households:

- For adult agents → household assignment, each "single agent" can be assigned directly to a "single household". For "couple" households (with or without children), we assume they are formed based on gender (a man and a woman can be combined as a couple) and by minimizing the agents' difference in age and education. Concretely, an age-distance d_a and an education-distance d_f is calculated for each couple of agents (i, j) and is used to determine a probability of assignment p_{ij} such that:

$$p_{ij} = e^{-\alpha.d_{a_{ij}} - \beta.d_{f_{ij}}} \quad (6)$$

  where α and β[6] are calibration parameters used to weight the significance of the age/education relation. Couples are then assigned according to the probability p_{ij}. Finally, the last category of households groups the remainder of agents, who are randomly assigned to house-sharing.

- For child agents → household assignment, we firstly create sibling groups according to the household number and size given by census data. To address the lack of data about the age of siblings in the census, we add parameters from the national census to define the age distribution of the youngest child and the mean age-distance between children of a same family. We assign the youngest child of each sibling group depending on age. Then we assign other children randomly according to the family size. Finally, we use the simulated annealing algorithm (Kirkpatrick, *et al.*, 1983) to optimize the age-gap between children by randomly swapping children among different families. When sibling groups are generated, each has to be assigned to a couple household. Again we use the simulated annealing algorithm to optimize assignment depending on

[6] In the results presented in section 3, α and β are set to 0.5 and this calibration will not be discussed in this paper.

the age-gap between the mother and the oldest child. The mother's age at the time of birth of her first child has to be near to the *avgAgeMother* parameter.

After these operations, the households are defined as groups of agents, containing between 1 and *N* agents (*N* is the maximum number of people in a family, i.e. a two parents with two children).

Dwellings

Dwellings are generated using the same principle as for agents. The adequate number of individual and collective housing units is defined from census data. Then a number of rooms is chosen randomly for each housing unit, in accordance with the housing's size distribution given in the census. Other characteristics contained in the housing package are then associated with the dwelling using the same probability law as described in (5). A discrete probability law, based on the difference between the size of the household s_m (number of agents) and the size of the housing unit s_l (number of rooms), is used to assign households to housing units.

$$p_{m,l} = \frac{e^{-d_{m,l}}}{\sum_k e^{-d_{m,k}}} \tag{7}$$

where

$$\delta_{m,l} = s_l - s_m$$

$$d_{m,l} = \frac{\delta_{m,l}-1}{s_m} \quad \text{if } \delta_{m,l} > 0$$

$$d_{m,l} = -10\,\frac{\delta_{m,l}}{s_m} \quad \text{if } \delta_{m,l} < 0$$

$$d_{m,l} = 0 \quad \text{if } \delta_{m,l} = 0$$

As a result, we obtain an information vector for housing units *l* containing households *m* as defined in (3). The next section presents the assignment of these housing units to residential buildings, i.e. the linkage between sociodemographic and geographic data.

2.2.2 Population location

The precise location of the agents in the study area depends on the structure of the buildings described in the geographic data (§ 2.1): for each IRIS, a residential density can be calculated and decomposed into detached houses or apartments in which the housing units are assigned according to their respective characteristics.

- At the outset, houses and apartments are distinguished by the height of the buildings in the geographic data (§ 2.1): buildings lower than 9 m and areas between 30 and 450 m^2 are considered to be detached houses; others are considered to be apartment blocks. Then for each spatial unit, the 9 m threshold is adjusted (up or down) to obtain a number of houses that matches the census data. As the number of rooms in houses and apartments is known for each IRIS (housing package), the mean volume of a room in houses or apartments can be calculated and the number of floors and rooms for each building is inferred from this volume.
- Household → housing assignment is the last step in the method. Each individual housing unit is associated with one detached house, the buildings provided from the IGN geographic database are randomly assigned to housing units with respect to the total number of rooms in the dwellings. As a result, this assignment means households can be located by means of the location of the housing they live in. This location is also defined in a *Z* position, determined by the storey of their apartment (or associated with ground level if they live in a detached house).

3 Results validation and discussion

3.1 Results

The successive operations described in § 2.2 are currently programmed in Java language in the MobiSim Population Synthesizer (downloadable, see Appendix B3) that is used to integrate the sociodemographic and geographic data described in § 2.1, and to generate and locate agents, households and housing, using default parameters

or user-modified parameters. It provides results rapidly (1-10 minutes' computation time).

As a result, the MobiSim Population Synthesizer is able to generate the data structure of a three-dimensional map of the city where agents are associated with households, households are assigned to apartments or detached houses, apartments are distributed in apartment blocks floor by floor, and all buildings are located to the nearest meter in the urban space. These relationships are calculated at the level *n* and globally match the structure observed at the *n* + 1 level, but the agents and the households do not strictly correspond to the demographic reality. Individuals' privacy is observed and no-one can be identified or located. Nevertheless, realistic individual information can be used to model agents' behavior, in response to the main objective of this research, as presented in § 1.3.

Technically, the corresponding data structure is decomposed into three files containing information on agents *a*, households *m*, and housing *h*. These files may be interlinked and associated with the buildings' geographic information from the BD Topo IGN shapefiles in a relational database (Figure 2). As these data are harmonized and collected for the whole of France, the software is useful for generating a synthetic population located anywhere in France, independently of local specificities (rural or urban area). Conversely, it is obviously strongly constrained by this specific data but any other data can easily be converted into the required format from local aggregated census data and a digital landscape model.

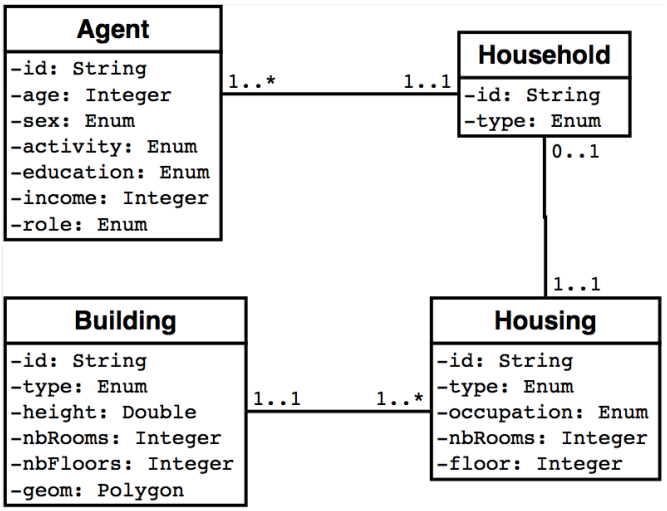

Figure 2: UML model as it can be exploited from the MobiSim Population Synthetizer results

Examples of results based on the relational database model produced by MobiSim Population Synthesizer (SQL request) are given in Figures 3 and 4, which show the characteristics of the synthetic population of Besançon aggregated by cells or by buildings in the city. These results are partially dependent on the parameters used to calibrate the software and the random algorithms included in the method. In the next section, we study how these parameters are used to validate and verify the model and ultimately to test its robustness.

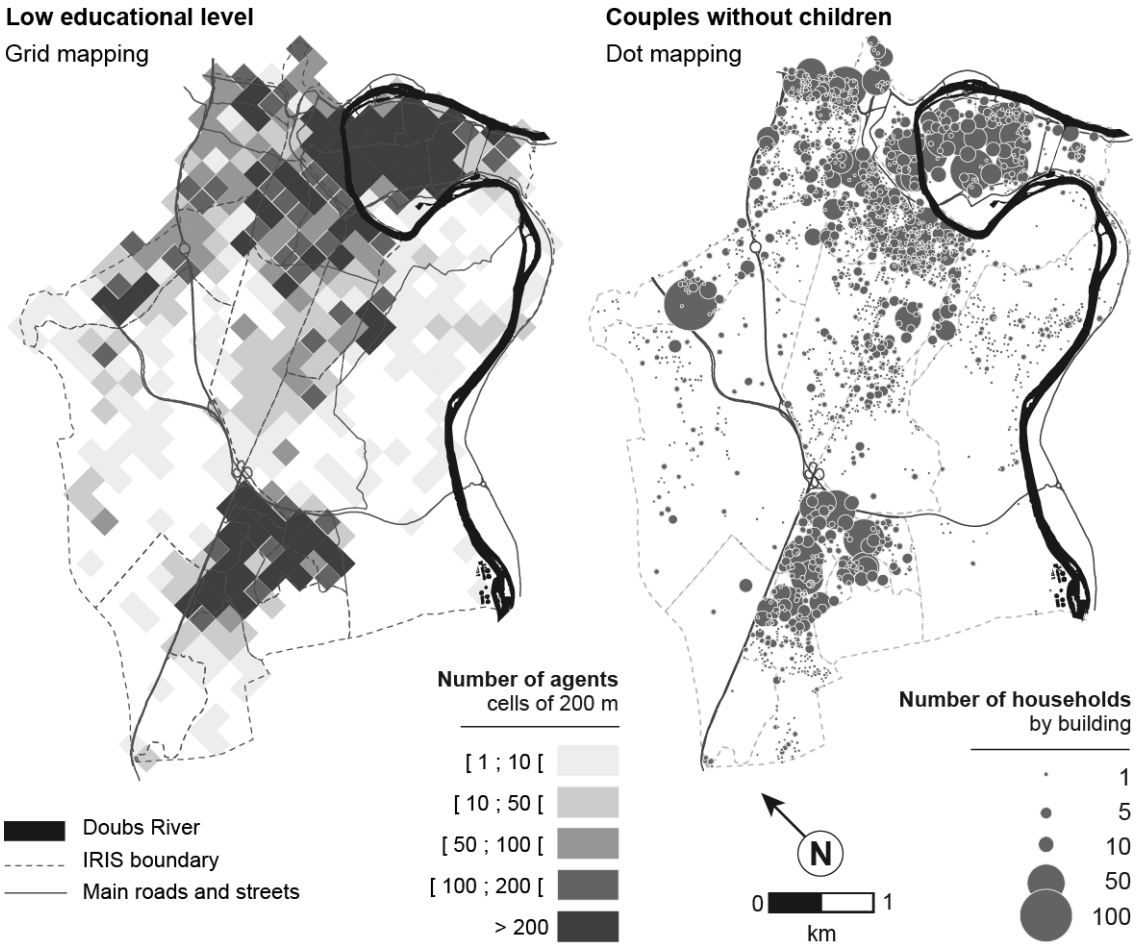

Figure 3: Some results for the city of Besançon: low educational level population aggregated in 200 m cells (left) and couples without children represented by proportional dots mapped at the scale of the buildings (right)

3.2 Validation

There is no generally accepted method for assessing the goodness-of-fit between synthetic populations and census data (Voas and Williamson, 2000, 2001). As pointed out by Li *et al.* (2008), verification and validation are two complementary processes that can be used

to create reliable models. Validation relates to the ability of the simulation results to mimic the real world; it includes a phase of calibration, i.e. of a choice of parameters leading to realistic results. Verification is about the model's correctness in terms of structure, equations, and stability and can be used to assess how successfully the model generates a coherent reality.

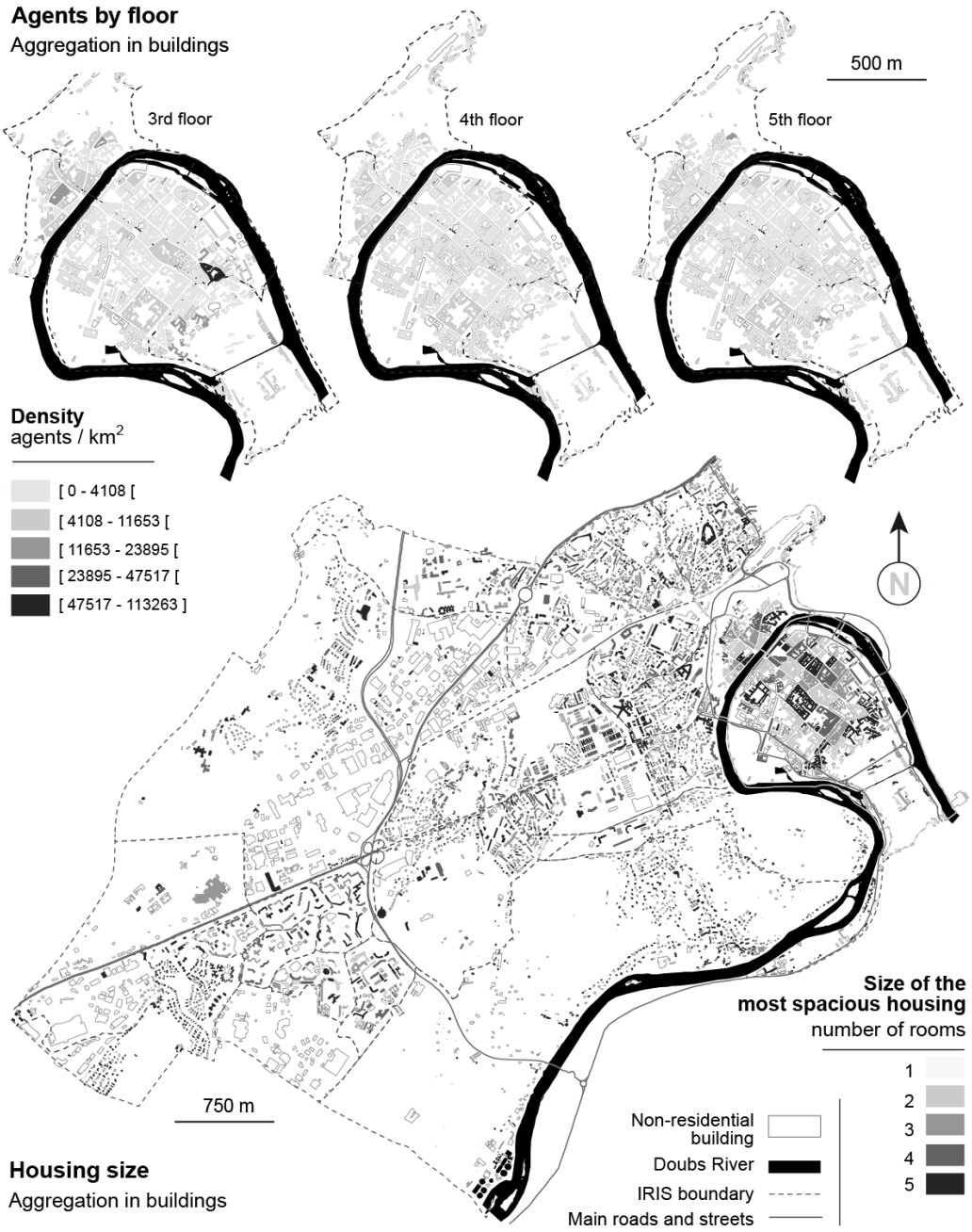

Figure 4: Other results for the city of Besançon: agent density floor by floor in the city center (top) and dwelling size (number of rooms) in the apartment blocks (bottom)

In a first step, the stability test consists in assessing how random probabilities affect the variability of population generation. This is done by running the program 100 times

with the same parameters and comparing the results in order to identify any convergence or divergence. For each test, the results are re-aggregated at the IRIS scale and compared with national statistics from the INSEE census[7]. This comparison is shown in Table 1 and expressed in terms of variation rate (> 1%) for 101 variables associated with individuals, households, and housing units within the 100 runs of the test. In a second step, the validation test consists in comparing the MobiSim Population Synthesizer results to the census data in order to observe their ability to mimic reality and to reproduce the information given by official surveys. Table 2 indicates the mean error rates of a selection of variables (variation > 1%) compared between MobiSim results and census data for the three urban areas studied. It shows relevant results: all the categories of agents are quite well reproduced, except for very specific and small categories of population like persons living alone or in retirement homes (i.e. elderly), large families, etc. (see Appendix A for a full description of the variables contained in Tables 1 and 2). The recurring errors with the number of rooms in apartments is due to the fact that the number of households and housing units do not properly match in the source data, but this has no repercussion overall on the structure of the synthetic population.

	Variable	Besançon	Lille	Strasbourg	**Mean**
Agent	All variables (58)	<1	<1	<1	<1
Household	PMEN_MENFAMMONO	1.74	1.12	1.18	1.35
	POP1524_PSEUL	<1	1.30	<1	<1
	POP80P_PSEUL	1.10	<1	<1	<1
	Other variables (25)	<1	<1	<1	<1
Housing	RP_2P	2.24	3.01	1.63	2.29
	RP_1P	2.05	2.95	1.86	2.29
	RP_3P	1.86	2.89	1.15	1.97
	RP_4P	1.91	2.52	1.44	1.96
	RP_5PP	1.41	1.30	1.27	1.32
	Other variables (10)	<1	<1	<1	<1

Table 1: Variation rate (%) of the model at IRIS level (level *n*). This table contains only the variables with a variation rate > 1 among a total of 101 variables produced by the MobiSim Population Synthesizer (for variable definitions, see Glossary in Appendix A).

[7] It seems impossible to fully validate the results at the disaggregated level (level *n*) as this would require a massive survey of individuals, households, and housing units questioning people about every aspect of their private lives. The present case study confirms the difficulties in validating synthetic population results (Voas and Williamson, 2000, 2001).

	Variable	Besançon	Lille	Strasbourg	**Mean**
Agent	AINACT1564	15.74	5.23	4.79	8.58
	INACT1564	3.55	1.60	1.34	2.17
	ACTOCC1564	1.96	<1	<1	1.19
	SCOL30P	<1	1.11	<1	<1
	All variables (54)	<1	<1	<1	<1
Household	PHORMEN	12.17	24.46	18.81	18.48
	POP1524_PSEUL	1.35	2.74	2.50	2.20
	PMEN_MENCOUPSENF	1.31	2.48	2.43	2.07
	POP80P_PSEUL	1.93	1.97	1.51	1.81
	NE24F0	1.02	1.74	1.75	1.50
	PMEN_MENFAMMONO	3.05	1.66	1.97	2.23
	POP2554_PSEUL	<1	1.31	<1	<1
	POP5579_PSEUL	1.08	1.03	<1	<1
	MENHSEUL	1.01	<1	<1	<1
	PMEN_MENHSEUL	1.01	<1	<1	<1
	PMEN_MENSFAM	1.20	<1	<1	<1
	NE24F4P	1.48	<1	<1	<1
	Other variables (16)	<1	<1	<1	<1
Housing	APPART	1.57	6.77	1.41	3.25
	MAISON	1.31	3.92	1.17	2.13
	RP_2P	1.42	1.99	<1	1.46
	RP_1P	1.26	1.87	1.04	1.39
	RP_3P	1.20	1.87	<1	1.26
	RP_4P	1.22	1.64	<1	1.23
	RP_5PP	1.14	1.03	<1	<1
	Other variables (10)	<1	<1	<1	<1

Table 2: Mean error rate (%) between the model's results and the national census at IRIS level (level *n*). This table contains only the variables with an error > 1

More generally, from these two test phases (Tables 1 and 2), three major conclusions can be drawn:

1. The variability of the model due to random probabilities is negligible, but large differences may appear for variables from inaccurate data sources. For example, the HM data source that was originally poorly completed by INSEE statistics cannot readily be used to match relevant and accurate results for a part of the population's characteristics. We ascribe this weakness to the data source, not the method.

2. Other kinds of error are due to the gap separating data in time. The data files needed by the software were not collected at exactly the same dates: the latest update of the IGN BD Topo (geographic information) was in 2012 whereas the latest available INSEE census is for 2009. This gap may explain some variations between the model and the real world when assigning housing units to buildings.

3. Errors may emerge from the model itself, i.e. from its structure, which relies on conditional probabilities and random assignments. In order to check the weight of this structure, complementary tests were made to verify the quality of the model and the role of the parameters introduced.

Conclusion

The MobiSim Population Synthesizer presents a results structure calculated at the level *n* that on the whole matches the structure observed at the *n* + 1 level, but the agents and the households do not strictly correspond to demographic reality. Individuals' privacy is preserved and no-one can be precisely identified or located. Nevertheless, realistic individual information can be used to model agents' behavior, in response to the main objective of this research as presented in § 1. For instance, all spatially aggregated data is usually subject to the modifiable area unit problem (MAUP; Openshaw and Taylor, 1981) and this can be avoided by the use of such a located synthetic population, as shown in Figures 3 and 4. Even so, problems associated with such fine scale representation remain because it is practically impossible to map the precision of the resulting data. Moreover, the methodology and software can be applied to the whole of French territory, or indeed to any country with the same kind of sociodemographic and geographic data. This should enable widespread use of synthetic populations in those countries where the potential of synthetic population data is currently not fully used by a broader community of researchers, planners, and stakeholders, given its benefits for microsimulation of spatial and demographic data (Hermes and Poulsen, 2012).

The main limitation of the MobiSim Population Synthesizer approach is that it is constrained by the specific data requirements provided by the national institutions listed in § 2.1, as well as the fact that users must prepare the model inputs as specified in this study by referring to the sample in the online attachment (see appendices A1 and A2).

Although the MobiSim synthetic population provides a basis for further MAS or ABM modeling, it does not integrate any dynamic dimension: the population and the residential structure is static and does not take into account any demographic or residential processes. Integrating such dynamic processes is an important task for

further developments. For example, as the agents and households are not spatially defined by an XY position but located by housing units, several permutations can be envisaged to simulate the movements of agents around the city defined by its residential structure and stimulated by the evolution of the agents' residential preferences or by particular life cycle events (birth of a child, death of a parent, etc.). Such a change obviously requires a dynamic population (Ballas *et al.*, 2005). Several methods could then be used to simulate the evolution of households, according to their possible life cycle stages. Finally, the different transport models integrated in LUTI models, and more generally in urban dynamics models (Moeckel *et al.*, 2003; Wegener, 2004), can be modified to take into account processes based on disaggregated agent and household behavior and located at the very fine 3D-scale of the buildings. The MobiSim LUTI project (Antoni and Vuidel, 2010) is currently being developed in this way to model and anticipate daily mobility and residential behaviors both of which are active parts of the ongoing process of urbanization.

5 References

Ahmad N., De Backer K., Yoon Y., 2010, An OECD perspective on micro data access: trends, opportunities and challenges, *Statistical Journal of the IAOS*, 26, 57-63.

Alfons A., Kraft S., Templ M., Filzmoser P., 2010, Simulation of synthetic population data for household surveys with application to EU-SILC. Forschungsbericht CS-2010-1, Institut f. Statistik u. Wahrscheinlichkeitstheorie, Vienna University of Technology.

Antoni J.-P., Vuidel G., 2010, MobiSim: un modèle multi-agents et multiscalaire pour simuler les mobilités urbaines. In: Antoni, J.-P. Modéliser la ville. Forme urbaine et politiques de transport, Méthodes et approches, Economica, 50-77.

Arentze T., Timmermans H., 2007, Creating synthetic household populations. Problems and approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2014, 85-91.

Ballas D., Clarke G., Dorling D., Eyre H., Bethan T., Rossiter D., 2005, SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place*, 11, 13-34.

Banos A., 2010, Miro : des trajectoires individuelles à la ville en mouvement. Modéliser la ville. In: Antoni, J.-P. Forme urbaine et politiques de transport, Méthodes et approches, Economica, 225-253.

Batty M., 2007, Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals, MIT Press, 592 p.

Beckman R. J., Baggerly K. A., McKay M. D., 1996, Creating synthetic baseline populations. *Transportation Research A*, 30 (6), 415-429.

Benenson I., 1998, Multi-agent simulations of residential dynamics in the city, *Computer, Environment and Urban Systems*, 22 (1), 25-42.

Benenson I., Omer I., Hatna E., 2002, Entity-Based Modelling of Urban Residential Dynamics: The Case of Ya_o, Tel Aviv, *Environment and Planning B*, 29 (4): 491-512.

Crooks A.T., Hudson-Smith A., Patel A., 2011, Building 3D Agent-based Models for Urban Sytems, Working Paper, Centre for Advanced Spatial Analysis, 161, 35 p.

Eursostat, 2008, Microdata access: new developments and a way forward, Eurostats Workshop, 3-8 december 2008, Luxembourg.

Farooq B., Bierlaire M., Hurtubia R., Flötteröd G., 2013, Simulation based population synthesis, *Transportation Research B*, 58, 243-263.

Hermes K., Poulsen M., 2012, A review of current methods to generate synthetic spatial microdata using reweighing and future directions, *Computers, Environment and Urban Systems*, 36, 281-290.

Kirkpatrick S., Gelatt C.D., Vecchi M. P., 1983, Optimisation by simulated annealing. *Science*, 220, 671-680.

Kirman A., 1992, Whom of what does representative agent represent? *Journal of Economic Perspectives*, 6, 117-136.

Li Y., Brimicombe A. J., Li C., 2008, Agent based services for the validation of multi agent models, *Computers, Environment and Urban Systems*, 32, 464-473.

Long Y., Shen Z., 2013, Disaggregating heterogeneous agent attributes and location, *Computers, Environment and Urban Systems*, 42, 14-25.

Mathis P., 2006, Cohérence entre politique des transports et politique d'aménagement. PREDIT/ADEME. 44 p.

Miller E. J., 2003, Microsimulation transportation system planning: Methods and application, CRC, Chapter 12.

Moeckel R., Spiekermann K., Wegener M., 2003, Creating a synthetic population. 8th International conference on computers in urban planning and urban management, Sendai, Japan.

Müller K., Axhausen K. W., 2010, Population synthesis for microsimulation: state of the art. Proceedings of the 10th Swiss Transport Research Conference.

Münnich R., Schürle J., 2003, On the simulation of complex universes in the case of applying the German microcensus , n°4, DACSEIS Research Paper Series.

Namazi-Rad M. R, Mokhtarian P., Perez P., 2014, Generating a dynamic synthetic population using an age-structured two-sex model for household dynamics, *PLOS One*, 9 (4).

Openshaw S., Taylor P. J., 1981. The modifiable areal unit problem. In: Wrigleyand N, Bennett R.J (Eds), Quantitative Geography: A British View, Routledge and Kegan Paul: London, 60-70.

Phan D., Amblard F., 2007, Agent-Based Modeling and Simulation in the Social and Human Sciences, Bardwell Press, 448 p.

Voas D., Williamson P., 2000, An evaluation of the combinatorial optimization approach to the creation of synthetic microdata, *International Journal of Population Geography*, 6, 349-366.

Voas D., Williamson P., 2001, Evaluating the goodness to fit measures for synthetic microdata, *Geographical and Environmental modeling*, 5, 2, 177-200.

Webster C.J., 1994, GIS and the scientific inputs to planning. Part 2: prediction and prescription, *Environment and Planning B*, 21, 145-157.

Wegener, M., 2004, Overview of land-use transport models. Chapter 9 in David A. Hensher and Kenneth Button (Eds.): Transport Geography and Spatial Systems. Handbook 5 of the Handbook in Transport. Pergamon/Elsevier Science, 127-146.

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.

Appendixes

A Glossary

This glossary describes the INSEE variables listed in Tables 1 and 2. All variables are numbers located in spatial units (IRIS).

	Variable	Description
Agent	POP1524	Persons aged 15-24
	POP80P	Persons aged 80 or more
	POP1524_PSEUL	Persons aged 15-24 living alone
	POP2554_PSEUL	Persons aged 25-54 living alone
	POP5579_PSEUL	Persons aged 55-79 living alone
	POP80P_PSEUL	Persons aged 80 or more living alone
	INACT1564	Non-working people aged 15-64
	ACTOCC1564	Working people aged 15-64
	AINACT1564	Other working people aged 15-64
	SCOL30P	Full time education aged 30 or more
Households	MENHSEUL	Single men
	MENFAMMONO	Single parent household
	PMEN	Persons in households
	PMEN_MENSFAM	Persons in household without family (house-sharing)
	PMEN_MENFAMMONO	Persons in single parent household
	PMEN_MENCOUPSENF	Persons in household (couple) without children
	NE24F0	Family without children aged less than 25
	NE24F4P	Family with 4 or more children aged less than 25
Housing	MAISON	Detached houses
	APPART	Apartments
	RP_1P	Housing unit of 1 room
	RP_2P	Housing unit of 2 rooms
	RP_3P	Housing unit of 3 rooms
	RP_4P	Housing unit of 4 rooms
	RP_5PP	Housing unit of 5 rooms or more

Source: INSEE

B Downloads

MobiSim Population Synthesizer software and sample datasets can be downloaded free-of-charge from the MobiSim website:

http://thema.univ-fcomte.fr/mobisim/software-applications/mobisim-synthesizer

B.1 Sociodemographic dataset sample

This archive contains sociodemographic spreadsheet data grouped into five thematic packages, as described in § 2:1. Files are provided by the French Institute for Statistical and Economics Studies (INSEE) and describe the data so it can be read by the MobiSim Population Synthesizer.

B.2 Geographic dataset sample

This archive contains shapefile data describing the city's buildings geometry. The data structure corresponds overall to the BDTopo provided by the French National Geographic Institute (IGN) and to other common large-scale topographic vector databases or digital landscape models. File format is .shp (ESRI shapefiles).

B.3 Software package

This archive contains a basic command line version of the MobiSim Population Synthesizer that can run on any operating system after installing the latest version of Java.

LUXEMBOURG INSTITUTE OF
SOCIO-ECONOMIC RESEARCH

11, Porte des Sciences
Campus Belval
L-4366 Esch-sur-Alzette

T +352 58 58 55-1
F +352 58 58 55-700

www.liser.lu