Gen*: an integrated tool for realistic agent population synthesis

Kevin Chapuis, Patrick Taillandier, Benoit Gaudou, Frédéric Amblard, Samuel Thiriot

Gen*: an integrated tool for realistic agent population synthesis

Conference Paper · September 2019

CITATIONS
0

READS
352

5 authors, including:

Kevin Chapuis
Institute of Research for Development
17 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Samuel Thiriot
EIFER, European Institute for Energy Research
25 PUBLICATIONS **117** CITATIONS
SEE PROFILE

Patrick Taillandier
French National Institute for Agriculture, Food, and Environment (INRAE)
148 PUBLICATIONS **1,399** CITATIONS
SEE PROFILE

Benoit Gaudou
Toulouse 1 Capitole University
162 PUBLICATIONS **1,323** CITATIONS
SEE PROFILE

Gen*: an integrated tool for realistic agent population synthesis

Kevin Chapuis[1,2][0000−0003−1743−2224], Patrick Taillandier[3,4][0000−0003−2939−4827], Benoit Gaudou[1,2,4][0000−0002−9005−3004], Frdric Amblard[4][0000−0002−2653−0857], and Samuel Thiriot

[1] Sorbonne University, IRD, UMMISCO
[2] USTH, ICTLab, Hanoi
kevin.chapuis@gmail.com,benoit.gaudou@gmail.com
[3] MIAT, University of Toulouse, INRA, Castanet-Tolosan, France
patrick.taillandier@inra.fr
[4] IRIT, University of Toulouse, France
frederic.amblard@ut-capitole.fr

Abstract. In recent years, the use of agent-based modeling to tackle complex societal issue has led to the massive use of data to better represent the targeted system. A key question in the development of such models is the definition of the initial population. If many tools and methods already exist to generate a synthetic population from global and sample data, very few are really used in the social simulation field. One of the major reason for this fact is the difficulty of use of the existing tools and the lack of integrated tools in the modeling platforms used by modelers. To tackle this issue, we present in this paper a new generic tool, called Gen*, allowing to generate, localize and structure by a social network a synthetic population, directly usable in the GAMA agent-based modeling and simulation platform through its modeling language. The paper presents in details the three components of Gen* (generation, localization, structuring) as well as their use in the GAMA platform.

Keywords: synthetic population generation · population localization · social network generation · GAMA platform.

1 Introduction

Agent-based modeling has become a major approach to study complex social systems and is now used in more and more domains (*e.g.* geography, ecology, sociology, economy). This boom has also led to the development of more grounded models, passing from a KISS [2] to a KIDS [12] approach.This transition was also favored by the availability of data. Indeed, we are now in the Era of Big Open Data where the quantity and the diversity of data like demographic data, GIS data and social network data is quickly growing. It is now possible to follow a Data-driven simulation approach [18] which aims at building the model from the available data. However, using such an approach requires to have tools to ease the creation, replication and exploration of descriptive social system models.

The paper focuses on a specific aspect of the integration of data in simulations that concerns the initialization of the simulation, and more particularly the generation of the initial population of agents. In order to be more realistic, agent-based models need to integrate agent population that more precisely represent target population. The question of the representativeness of the synthetic population (SP) does not only concern the distribution of the entities attributes, but also the entity localization as well as their connection between each other. Unfortunately, generating a structured and spatialized population of agents from diverse sources of data is a complex task that is out of scope of the main modeling platforms. We aim thus at filling this lack by proposing a new tool, called Gen*, integrated in the GAMA platform, that makes it possible to easily generate a synthetic population that reflect the distribution of attributes, spatial distribution and connection between entities.

The paper is organized as follows: Section 2 presents the context of this work, *i.e.* the generation of synthetic populations. Section 3 is dedicated to the presentation of the general principles of the Gen* tools and Section 4 to its integration in the GAMA platform. Lastly, Section 5 concludes and presents perspectives.

2 Realistic synthetic population for agent-based simulation

2.1 Synthetic population generation

The approach that have been most studied is based on the idea of Synthetic Reconstruction (SR) [40] and consists in building populations through the random generation of individual characteristics. The created entity is then seen as a vector of values – *e.g.* $\{blue, 1.75, 35, male\}$ – that represents the individual characteristics for each attribute – *e.g.* color of the eye, height, age and gender respectively. This process is usually conducted by drawing attribute values either from the available distributions [16] or from an estimated joint-distribution based on techniques such as the Iterative Proportional Fitting (IPF) algorithm [28], simulated Markove-Chain [13] or bayesian-based estimation techniques [29]. The available data to based distribution upon, as well as the algorithm used to sample characteristics, will decide whether individual vectors are drawn at once or by gathering separately drawn characteristics.

When a sample of the target population is available, the generation process can take the form of a replication of individual records. The created population is then made of duplicated vector of characteristics taken from real individuals. This second approach is referred to as Combinatorial Optimization (CO) [39]. The process usually starts with a random initial set of individual records and then add and/or swap individual(s) drawn from the sample any number of time needed to fit required final states [35]. In fact, the replication of known real individuals is driven so to fit macroscopic descriptors, which can be the required size of the population or available aggregated data *e.g.* a certain proportion

of male and female or a certain distribution of age. Any kind of optimization algorithm can be used to manage the selection of newly drawn entities of the population, and several have already been used, *e.g.* greedy heuristics [27] or genetic algorithms [26].

2.2 Synthetic population localization

The localization of synthetic entities in agent-based simulation model is a concern as far as the goal is to put them into a realistic spatial word. To this extends, modelers must rely on GIS to create the context to locate agents in. Dealing with spatial data, several methods are classically used to desegregate population distribution of attributes to any number of scattered sub-zones. Most of them have been developed in the context of aggregated demographic data at regional level that has to be ventilated across smaller scale areas. To this purpose, census centroid data [6], dasymetric modeling [5] or a combination of those methods [20] have been mostly used.

To sum up, two main steps can be used to address the localization problem: the first one referred to as *Areal interpolation*, consists in translating data from aggregated areas to more precise ones. This type of method uses ancillary data to refine aggregated data originally given at a coarser level [22]: entities characteristics are put into correlation with spatial characteristics, making satellite imagery, land use or land cover explanatory variables of the spatial distribution of entities attributes. For example, [11] use dasymetric mapping to study U.S. wide racial segregation at a high resolution ($90m^2$ cells) combining population density at district level, and fine grained ancillary satellite imagery and land cover. The second step referred to as *Explicit localization*, consists in providing individuals with a precise location, that could be a coordinate (x,y) or a geographical object (*e.g.* a building). Very few has been done on this aspect, as mention in [8].

2.3 Synthetic population social network

According to a recent survey in the social simulation field [1], most models proposed in the domain that actually include a social network component to organize the agent population are using very simple and abstract models, *i.e.* regular, random (ER), Small-world, Scale-free. Two evident reasons for this concern are, on the one hand, the lack of available data concerning social networks on the targeted system (such data are expensive to acquire), and on the other hand such simple models are quite easy to implement and control through a limited number of parameters. Therefore, and even if several advanced social network generators, that could be appropriately used to generate synthetic social networks (such as [34, 24]) exist, they are not sufficiently solicited due mainly to the learning cost to master such tools. Therefore, whereas it is largely acknowledged that simple models are a poor approximation of real social networks [9], they are still largely used. The evident alternative to these models concerns the implementation of *ad hoc* networks including hypotheses specific to the modelled domain. Following

this trend, one of the most advanced initiative concerns contact networks in the field of epidemiology and are strongly dependent on the previous steps in the synthetic population generation: links in the network corresponds to significant properties to interact (*i.e.* propagate the virus) which is assumed to depend on household relations, schools and workplaces which are generally generated at the localization step.

3 Principles of agent population synthesis using Gen*

In this section we present the capabilities of the open-source Gen* library [5]. The main repository is made of four components: the core API that define the main concepts and overall meta-model of population, entity, attribute and value; the gospl API that encapsulate data harmonization, generation algorithms and synthetic population validation indicators; the spll API responsible of synthetic entity localization, including areal interpolation, explicit localization and spatial binding; and finally, the spin API that contains network generation algorithm between synthetic entities.

The framework have been design so that each API only rely on main concepts define in the core library, makes it possible for users to work only with generation, localization or network creation independently. Furthermore, the library contains a repository with example applications to previously generated population for the city of Rouen and Bangkok.

3.1 Generation of synthetic population

Gen* makes it possible to generate a SP from samples and/or aggregated data. Before the generation process *per se*, and because demographic data can be diversely recorded, the available input information need to be harmonized [15]. Hence, using a unified representation of data about synthetic population, algorithms can be used in order to make the best of available information: in fact, Gen* makes it possible to generate a synthetic population from any type of data, with or without a sample, only based on aggregated data or a mixed between aggregated and sample data. At the end of the generation process, Gen* provides several indicators to asses SP quality.

Data harmonization Our proposal consider two types of input data: aggregated frequency or contingency and sample of population. The first one is transposed into a sparse multi-dimensional matrix with one dimension for each described attribute. The second source of data is considered to be a population of entity and can be treated as a SP, making a sample readily usable as an initial population of agent.

Both type of data are manipulated through input files in csv or excel format. In order for Gen* to read data, modelers needs to create a configuration. The

[5] `https://github.com/ANRGenstar`

library provide several features to build and store it in a JSON file that includes: input data files path, type of data for each and a description of all attributes (type of value, encoded form).

To ease the accessibility of data, Gen* provide a *dictionary* template in order to directly read data from known sources. We can find two templates that allows to ease the use of data from IPUMS (data-base of demographic data for more than 100 country around the world) and INSEE (french national institute of statistics). The schema can be extended to include any kind of pre-formated demographic data simply by defining the proper dictionary for Gen* to directly read files from any sources.

Methods Gen* has support for SR and CO based algorithms. They have classically be split into sample–free and sample–based techniques, mainly because the later needs as a stringent requirement a sample while the former only needs it when performing joint-distribution estimation.

For the first category, as default method, we provide direct sampling from known distribution, or as a refinement a hierarchical sampling that organized attributes in a graphical model with the ability to tune parameters of the Bayesian network [4]. When a sample of the population is available, modelers can use the IPF procedure to estimate the underlying joint distribution to be used in the sampling phase (see [23] for a recent critical review of the techniques and implementation issues). In all cases, we provide three sampling algorithms either linear, binary or alias search to draw individual vector of characteristic from the conditional, hierarchical or joint distribution respectively.

The alternative CO based techniques is also available with several optimization algorithms to choose from. In order to setup CO algorithms one must define several meta-parameters. First for the starting population, modelers can choose either to use the sample or to draw a defined number of records from the sample. Next, we consider a synthetic population to be a solution and searching through neighbor solutions involve swapping any number of individual. Modelers can choose either to elicit records randomly or in order to shift on one dimension of the distribution of attribute (*e.g.* swap two individual record that only differ on one attribute). At the end, modelers must define the fitness function that can rely on any quality indicators define in sub-section 3.1. Ultimately, Gen* provides several optimization algorithms that can be used to monitor the CO process: random search, hill climbing, tabu search and simulated annealing.

Indicators Basic principles of quality assessment of generated SP implies a comparison between output population aggregated data and available data about targeted real population. Gen* provides state of the art indicators that can be split into two types of distance metrics: the index focusing on categorical distance and those that target continuous distance. In the first category we can mention the Total Absolute Error (TAE) that count the number of misclassified records [39]. It is used thanks to its simplicity of computation, and also preferred in CO algorithm to compute fitness [35]. The second type of indicator features classical

average indicator, namely Relative Average Percentage Difference (RAPD) and Standard Root Mean Square Error (SRMSE) that focus on aggregating relative distances over the distribution of attributes [21]. Last, we implement an indicator that does the trade off between categorical and continuous distance metric: Relative Sum of Square modified Z score (RSSZ*) [38]. This index is based on the $\mathcal{X}^2$ and take into account both misclassified entities as well as relative difference in the distribution of attributes.

3.2 Localization of synthetic population

Gen* provides as well many tools to localize a synthetic population. We describes here the main principles of this localization, but more details can be found in [8]. The synthetic population localization tool of Gen* comprises two different while close processes: the first one, called *nesting process*, concerns the actual localization of each entity, *e.g.* assigning to each entity a "home place", *i.e.* a geographical object (building, area, etc.) referred as a *nest* and a coordinate in this *nest*; the second process, called *biding process*, enables to associate different spatial objects to the entities of an already localized population (*e.g.* a workplace).

Data requirements The only mandatory data to provide is a geographic file specifying the geographic objects on which Gen* will locate the entities (*nesting*). We have made the choice to explicitly locate the entities in spatial objects so that each user can define the location best suited to the needs of the model: the *nest* can be buildings, cells of a raster file or even simply a polygon representing the limits of the studied system.

The modeler can also provide other data to the localization process, such as "mapping" data, to create a link between the population entities and geographical objects, *e.g.* the number of people per administrative region.

As a refinement, it is possible to input any kind of spatialized data that will monitor the areal interpolation methods. Those data are commonly refereed as *ancillary data* and can be raw satellite imagery, land use or land cover, either raster or vector data.

Methods

Nesting process: The *nesting* process consists of connecting each entity of the population with a *nest*, and to define a location in this *nest*. In order to achieve a more realistic specialization of the entities, Gen* allows the modeler to define spatial constraints to filter possible *nests*. Gen* integrates three basic types of spatial constraints: geometric, contingency and density. In addition, Gen* enables as well to define spatial constraints on the location of entities, such as a maximum distance from roads or point of interest (POI). If the constraints do not find an appropriate *nest* (over-constrained), they can be released. More precisely, for each constraint, a relaxation process can be defined as well as a

maximum relaxation function which defines to what extent the stress can be relaxed. When multiple *nests* satisfy all the constraints for an entity, Gen* uses a distribution function to choose one. Gen* provides the modeler with three predefined distribution functions: uniform distribution, area distribution and capacity distribution.

In addition, in order to facilitate the localization process, Gen* provides the modeler with a tool to generate a contingency map or a density map from raw input data or by using statistical learning techniques such as regression. Available areal interpolation techniques make possible to infer a spatial distribution of entities and attributes from the initial mapping data and any ancillary data. Then, it can be used to allocate entities according to it, together with the aforementioned constraints.

Spatial binding: Gen* enables to link a set of geographic objects to each entity. This linking process is very close to the *nesting* process: the principle is to choose for each type of link (for example, workplace, school) a geographical object among a set of possible places. To do this, Gen* uses a combination of spatial constraints and a spatial distribution function. Since it is often necessary to link features with geographical objects close to their place of residence, Gen* offers the possibility of defining a distribution function according to the distance between the entity's location and the spatial objects such as the gravity model. Distribution function can also take into account the attributes of agent and places to bind them according to user define Bayesian rules.

3.3 Social network in synthetic population

Following the observed situation in the social simulation domain [1] that is that 69 percent of the published papers using social networks generation were actually relying on basic abstract models. As a central proposal, Gen* enables modellers to access such generic generation models. More precisely, Gen* integrates methods to generate:

- regular lattices: either 1D or 2D lattices, with the possibility to adjust the degree of the nodes in the 1D case;
- random networks: following the Erdös-Rényi model and enabling to play on the probability of connection among individuals that actually controls the network density as well as the average degree of the network;
- small-World networks: using the beta-model of Watts [37] that has two main parameters, the degree of the 1D lattice used as the starting structure and the noise beta added to the lattice. The latter corresponds to the probability of randomly rewiring a link in the network;
- scale-free networks: using the preferential attachment model [3] that allows to control the network density as well as the slope of its power-law distribution of degrees.

Apart of such solutions, Gen* facilitates the building of *ad hoc* networks. Indeed, it provides access to the properties of each generated individual and

their location (following the preceding steps in the generation process of the synthetic population) enabling therefore to build spatial networks (based on the location of individuals) or p1 social networks [19] (based on the individuals' attributes and homophilic rules).

4 Integration into the GAMA platform

The main motivation of the Gen* project has been to provide tools to generate a synthetic population for agent-based models. Reviews of the literature in the JASSS[6] journal have shown both for synthetic population [7] but also for synthetic networks [1] that synthetic populations are rarely used in agent-based models for social simulation. When it is the case, the algorithms used are often very basic. We argue that a reason is that these tools are not easily accessible for modelers, in particular when they are not computer scientists or mathematicians. We thus argue that it is necessary to integrate such a tool inside an agent-based modeling tool to ease its use. In this paper, we choose to extend the GAMA modeling and simulation platform[7] [17, 31]; we choose this platform as it manages very well the spatial data and because it is particularly well-suited to develop large-scale models, in which the synthetic population generator library Gen* is particularly relevant.

The GAMA platform is a generic agent-based modeling and simulation tools which provides a dedicated modeling language (GAML) specifically designed for any modeler to build his/her own model. Its main features are to integrate very easily spatial data (vector and raster), to be natively multi-level and designed to support large-scale models with a huge amount of data. It also provides many tools to design participative simulations [32]. It benefits for a very dynamics community of developers and users and is used worldwide.

We developed an extension to the GAMA platform[8], which extends the GAML language to allow modelers to create the synthetic population, to locate it and to structure it in a social network directly using the modeling language. An example of code is provided in Fig. 1 and the Fig. 2 illustrates a more complex generated population (agents with 5 attributes, located in buildings, mapped using IRIS demographic data, and structured using two social networks). The principles of the integration are the following ones:

1. Creation of the population generator and its configuration with a generation algorithm (Direct Sampling in the example), the individual attributes of the population to be generated (age and sex in the example), the constraints in terms of localization (in buildings in the example) and the social networks that can be generated (a spatial network in the example);

[6] JASSS stands for *The Journal of Artificial Societies and Social Simulation* and is available at the address: `http://jasss.soc.surrey.ac.uk/`.

[7] `http://gama-platform.org/`

[8] The extension can be directly downloaded from GAMA 1.8 at the URL: `https://www.irit.fr/genstar/p2updatesite/`

```
// Define the population generator
gen_population_generator pop_gen;

// Set the generation algorithm
// Here it will be the Direct Sampling
pop_gen <- pop_gen with_generation_algo "IS";

// Set the individuals attributes and their possible values or range of values
// Here indivduals have an age (integer value) and a couple status (among single or couple)
pop_gen <- pop_gen add_attribute("Age", gen_range, ["0 to 17", "18 to 110"]);
pop_gen <- pop_gen add_attribute("CoupleStatus", string, ["single", "couple"]);

// Set the constraints in terms of localization
// Here agents will be located in buildings
pop_gen <- pop_gen localize_on_geometries("../data/shp/buildings.shp");

// Set the social networks linking individuals
// Here the neighbourhood network is generated as a spatial proximity graph
pop_gen <- pop_gen add_network("neighbours","spatial",500.0);

// Creation of 100 (localized) agents from this population generator
create people from: pop_gen number: 100 ;

// Generate the graph of neighbours
pop_gen <- pop_gen associate_population_agents(people);
graph<people> graph_neighbours <- pop_gen get_network("neighbours");
```

Fig. 1. Example of GAML code illustrating the three aspects of the Gen* population generation.

2. Creation of a chosen number of agents of a given species[9] generated and located using the population generator;
3. Creation of the social network(s), once this population of agents is created.

We argue that providing such a tool directly in an Agent-based platform will ease the use of SP in agent-based social simulation. Modelers will thus be able to generate a population within modeling environment and to make use of it directly in the simulations. Furthermore, it can help the modelers to run batch exploration over the synthetic populations in order to assess its impact on the simulation results: e.g. to analyze the effect of population size keeping distribution of attributes constant or tune the structure of the population according to user define hypothesis on the distribution of attribute.

5 Conclusion

In this paper, we presented Gen*, an integrated synthetic population generation tool, that can be used in the generic agent-based modeling and simulation GAMA platform. We tried with Gen* and the GAMA plug-in to provide a powerful tool, but at the same time, an easy one to use and adapt to many applications, either in terms of entities to generate or in terms of data used. A first encouraging result is that Gen* is already used in several projects dealing with different applications: evacuation of an urban area evacuation facing hazards [10], diffusion of vegetarian diets, study of urban congestion and pollution emission in the city of Dijon [14], etc.

[9] `species` is the GAML keyword representing a kind or a class of agents.

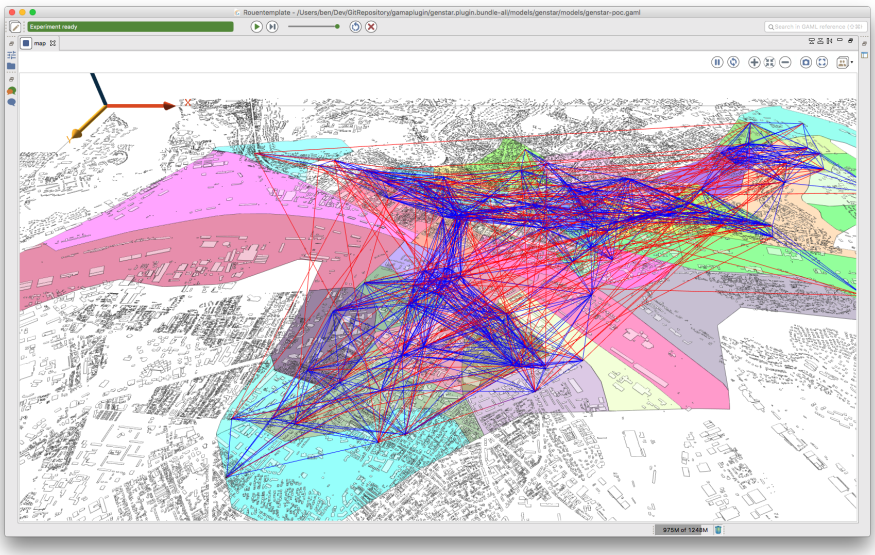

Fig. 2. Example of a population composed of 100 agents generated with 5 attributes (age, CSP, sex, couple and iris) from INSEE data, localized on IRIS inside buildings and linked through 2 social networks (a random network and a proximity network).

In terms of perspectives, many new features are already planned. The first one concerns the possibility to define multi-level populations (*e.g.* households composed of individuals). Indeed, in its current version, Gen* does not enable to generate such populations whereas many case-studies require to take several levels into account. To face this issue, we plan to implement several new algorithms: some that rely on a layer flattening process coupled with classical SR methods (e.g. IPU [41] or HIPF [25]) or using mixed strategies that make use of CO algorithm to match up layers of SR based generated entities [30, 36].

Another perspective concerns the definition of more complex social network, mixing classic structures (small-world, scale-free, etc.) with other information such as the agent attributes and their localization. We plan for that to make a bridge between Gen* and dedicated libraries and tools like YANG [33].

References

1. Amblard, F., Bouadjio-Boulic, A., Gutiérrez, C.S., Gaudou, B.: Which models are used in social simulation to generate social networks?: a review of 17 years of publications in jasss. In: Proceedings of the 2015 Winter Simulation Conference. pp. 4021–4032. IEEE Press (2015)
2. Axelrod, R.M.: The complexity of cooperation: Agent-based models of competition and collaboration. Princeton University Press (1997)

3. Barabási, A.L., Albert, R., Jeong, H.: Scale-free characteristics of random networks: the topology of the world-wide web. Physica A: statistical mechanics and its applications **281**(1-4), 69–77 (2000)
4. Barthelemy, J., Toint, P.L.: Synthetic population generation without a sample. Transportation Science **47**(2), 266–279 (2012)
5. Bhaduri, B., Bright, E., Coleman, P., Urban, M.L.: Landscan usa: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal **69**(1-2), 103–117 (2007)
6. Bracken, I., Martin, D.: The generation of spatial population distributions from census centroid data. Environment and Planning A **21**(4), 537–543 (1989)
7. Chapuis, K., Taillandier, P.: A brief review of synthetic population generation practices in agent-based social simulation. In: submitted to SSC2019, Social Simulation Conference 2019 (2019)
8. Chapuis, K., Taillandier, P., Renaud, M., Drogoul, A.: Gen*: a generic toolkit to generate spatially explicit synthetic populations. International Journal of Geographical Information Science **32**(6), 1194–1210 (2018)
9. Cointet, J.P., Roth, C.: How realistic should knowledge diffusion models be? Journal of Artificial Societies and Social Simulation **10**(3), 5 (2007)
10. Daudé, E., Caron, C., Chapuis, K., Drogoul, A., Gaudou, B., Rey-Coyrehourq, S., Saval, A., Taillandier, P., Tranouez, P., Zucker, J.D.: ESCAPE: Exploring by Simulation Cities Awareness on Population Evacuation. In: to appear in the Proceessing of the International Conference on Information Systems and for Crisis Response and Management. Springer (2019)
11. Dmowska, A., Stepinski, T.F.: High resolution dasymetric model of us demographics with application to spatial distribution of racial diversity. Applied Geography **53**, 417–426 (2014)
12. Edmonds, B., Moss, S.: From KISS to KIDS in Multi–Agent and Multi–Agent Based Simulation. Lecture Notes in Computer Science **3415**, 130–144 (2005)
13. Farooq, B., Bierlaire, M., Hurtubia, R., Fltterd, G.: Simulation based population synthesis. Transportation Research Part B: Methodological **58**, 243–263 (2013)
14. Fosset, P., Banos, A., Beck, E., Chardonnel, S., Lang, C., Marilleau, N., Piombini, A., Leysens, T., Conesa, A., Andre-Poyaud, I., et al.: Exploring intra-urban accessibility and impacts of pollution policies with an agent-based simulation platform: Gamirod. Systems **4**(1), 5 (2016)
15. Gallagher, S., Richardson, L.F., Ventura, S.L., Eddy, W.F.: Spew: Synthetic populations and ecosystems of the world. Journal of Computational and Graphical Statistics **27**(4), 773–784 (2018)
16. Gargiulo, F., Ternes, S., Huet, S., Deffuant, G.: An Iterative Approach for Generating Statistically Realistic Populations of Households. PLoS ONE **5**(1) (2010)
17. Grignard, A., Taillandier, P., Gaudou, B., Vo, D.A., Huynh, N.Q., Drogoul, A.: GAMA 1.6: Advancing the art of complex agent-based modeling and simulation. In: International Conference on Principles and Practice of Multi-Agent Systems. pp. 117–131. Springer (2013)
18. Hassan, S., Pavon, J., Gilbert, N.: Injecting data into simulation: Can agent-based modelling learn from microsimulation. In: World Congress of Soc. Simu. (2008)
19. Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. J. of the American Statistical association **76**(373), 33–50 (1981)
20. Holm, E.: The SVERIGE spatial microsimulation model: content, validation, and example applications. Department of Social and Economic Geography, Univ. (2002)
21. Kim, J., Lee, S.: A reproducibility analysis of synthetic population generation. Transportation Research Procedia **6**, 50–63 (2015)

22. Li, G., Weng, Q.: Fine-scale population estimation: How landsat etm+ imagery can improve population distribution mapping. Canadian Journal of Remote Sensing **36**(3), 155–165 (2010)
23. Lovelace, R., Birkin, M., Ballas, D., van Leeuwen, E.: Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. J. of Artificial Societies and Social Simulation **18**(2), 21 (2015)
24. Menezes, T., Roth, C.: Symbolic regression of generative network models. Scientific reports **4**, 6284 (2014)
25. Mller, K., Axhausen, K.W.: Hierarchical IPF: Generating a synthetic population for switzerland
26. Otani, N., Miyamoto, K., Sugiki, N.: Goodness-of-fit evaluation method between observed and estimated sets of micro-data in land-use micro-simulation. Proceedings of CUPUM **9** (2009)
27. Srinivasan, S., Ma, L., Yathindra, K.: Procedure for forecasting household characteristics for input to travel-demand models. Tech. rep., Florida Department of Transportation (2008)
28. Stephan, F.F.: An iterative method of adjusting sample frequency tables when expected marginal totals are known. The Annals of Mathematical Statistics **13**(2), 166–178 (1942)
29. Sun, L., Erath, A.: A bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies **61**, 49–62 (2015-12)
30. Sun, L., Erath, A., Cai, M.: A hierarchical mixture modeling framework for population synthesis **114**, 199–212. https://doi.org/10.1016/j.trb.2018.06.002
31. Taillandier, P., Gaudou, B., Grignard, A., Huynh, Q.N., Marilleau, N., Caillou, P., Philippon, D., Drogoul, A.: Building, composing and experimenting complex spatial models with the gama platform. GeoInformatica (Dec 2018)
32. Taillandier, P., Grignard, A., Marilleau, N., Philippon, D., Huynh, Q.N., Gaudou, B., Drogoul, A., et al.: Participatory modeling and simulation with the gama platform. Journal of Artificial Societies and Social Simulation **22**(2), 1–3 (2019)
33. Thiriot, S.: Yang (yet another generator) (2010), `https://sourceforge.net/projects/yang-j/`
34. Thiriot, S., Kant, J.D.: Generate country-scale networks of interaction from scattered statistics. In: The 5th conference of the ESSA, Brescia, Italy. vol. 240 (2008)
35. Voas, D., Williamson, P.: An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography **6**(5), 349–366 (2000)
36. Watthanasutthi, N., Muangsin, V.: Generating synthetic population at individual and household levels with aggregate data. In: 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). pp. 1–6. IEEE (2016)
37. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. nature **393**(6684), 440 (1998)
38. Williamson, P.: An evaluation of two synthetic small-area microdata simulation methodologies: synthetic reconstruction and combinatorial optimisation. In: Spatial microsimulation: A reference guide for users, pp. 19–47. Springer
39. Williamson, P., Birkin, M., Rees, P.H.: The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A **30**(5), 785–816 (1998)
40. Wilson, A.G., Pownall, C.E.: A new representation of the urban system for modelling and for the study of micro-level interdependence. Area pp. 246–254 (1976)

41. Ye, X., Konduri, K., Pendyala, R.M., Sana, B., Waddell, P.: A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In: 88th Annual Meeting of the Transportation Research Board, Washington, DC