Outlining some requirements for synthetic populations to initialise agent-based models

Roxburgh, Nick; Paolillo, Rocco; Filatova, T.; Cottineau, C.; Paolucci, Mario; Polhill, J. Gareth

Publication date
2025

Document Version
Final published version

Published in
Review of Artificial Societies and Social Simulation

Citation (APA)
Roxburgh, N., Paolillo, R., Filatova, T., Cottineau, C., Paolucci, M., & Polhill, J. G. (2025). Outlining some requirements for synthetic populations to initialise agent-based models. *Review of Artificial Societies and Social Simulation*.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Review of Artificial Societies and Social Simulation

CONTENT

Outlining some requirements for synthetic populations to initialise agent-based models

JANUARY 29, 2025 | THESUBMISSIONAUTHOR | LEAVE A COMMENT

By Nick Roxburgh (https://www.hutton.ac.uk/people/nick-roxburgh/)[1], Rocco Paolillo (https://www.irpps.cnr.it/en/rocco-paolillo/)[2], Tatiana Filatova (https://www.tudelft.nl/tbm/resiliencelab/people/tatiana-filatova)[3], Clémentine Cottineau (https://www.tudelft.nl/en/staff/c.cottineau/)[3], Mario Paolucci (https://www.istc.cnr.it/en/people/mario-paolucci)[2] and Gary Polhill (https://www.hutton.ac.uk/people/gary-polhill/)[1]

[1] The James Hutton Institute, Aberdeen AB15 8QH, United Kingdom
{nick.roxburgh,gary.polhill}@hutton.ac.uk

[2] Institute for Research on Population and Social Policies, Rome, Italy
{rocco.paolillo,mario.paolucci}@cnr.it

[3] Delft University of Technology, Delft, The Netherlands {c.cottineau,t.filatova}@tudelft.nl

Abstract. We propose a wish list of features that would greatly enhance population synthesis methods from the perspective of agent-based modelling. The challenge of synthesising appropriate populations is heightened in agent-based modelling by the emphasis on complexity, which requires accounting for a wide array of features. These often include, but are not limited to: attributes of agents, their location in space, the ways they make decisions and their behavioural dynamics. In the real-world, these aspects of everyday human life can be deeply interconnected, with these associations being highly consequential in shaping outcomes. Initialising synthetic populations in ways that fail to respect these covariances can therefore compromise model efficacy, potentially leading to biased and inaccurate simulation outcomes.

1 Introduction

With agent-based models (ABMs), the rationale for creating ever more empirically informed, attribute-rich synthetic populations is clear: the closer agents and their collectives mimic their real-world counterparts, the more accurate the models can be and the wider the range of questions they can be used to address (Zhou et al., 2022). However, while many ABMs would benefit from synthetic populations that more fully capture the complexity and richness of real-world populations – including their demographic and psychological attributes, social networks, spatial realms, decision making, and behavioural dynamics – most efforts are stymied by methodological and data limitations. One reason for this is that population synthesis methods have predominantly been developed with microsimulation applications in mind (see review by Chapuis et al. (2022)), rather than ABM. We therefore argue that there is a need for improved population synthesis methods, attuned to support the specific requirements of the ABM community, as well as commonly encountered data constraints. We propose a wish list of features for population synthesis methods that could significantly enhance the capability and performance of ABMs across a wide range of application domains, and we highlight several promising approaches that could help realise these ambitions. Particular attention is paid to methods that prioritise accounting for covariance of characteristics and attributes.

2 The interrelationships among aspects of daily life

2.1 Demographic and psychological attributes

To effectively replicate real-world dynamics, ABMs must realistically depict demographic and psychological attributes at both individual and collective levels. A critical aspect of this realism is accounting for the covariance of such attributes. For instance, interactions between race and income levels significantly influence spatial segregation patterns in the USA, as demonstrated in studies like Bruch (2014).

Several approaches to population synthesis have been developed over the years, often with a specific focus on assignment of demographic attributes. That said, where psychological attributes are collected in surveys alongside demographic data, they can be incorporated into synthetic populations just like other demographic attributes (e.g., Wu et al. (2022)). Among the most established methods is Iterative Proportional Fitting (IPF). While capable of accounting for covariances, it does have significant limitations. One of these is that it "matches distributions only at one demographic level (i.e., either household or individual)" (Zhou et al., 2022 p.2). Other approaches have sought to overcome this – such as Iterative Proportional Updating, Combinatorial Optimisation, and deep learning methods – but they invariably have their own limitations and downsides, though the extent to which these will matter depends on the application. In their overview of the existing population synthesis land-

scape, Zhou et al., (2022) suggest that deep learning methods appear particularly promising for high-dimensional cases. Such approaches tend to be data hungry, though – a potentially significant barrier to exploitation given many studies already face challenges with survey availability and sample size.

2.2 Social networks

Integrating realistic social networks into ABMs during population synthesis is crucial for effectively mimicking real-world social interactions, such as those underlying epidemic spread, opinion dynamics, and economic transactions (Amblard et al., 2015). In practice, this means generating networks that link agents by edges that represent particular associations between them. These networks may need to be weighted, directional, or multiplex, and potentially need to account for co-dependencies and correlations between layers. Real-world social networks emerge from distinct processes and tendencies. For example, homophily preferences strongly influence the likelihood of friendship formation, with connections more likely to have developed in cases where agents share attributes like age, gender, socio-economic context, and location (McPherson et al., 2001). Another example is personality which can strongly influence the size and nature of an individual's social network (Zell et al., 2014). For models where social interactions play an important role, it is therefore critical that consideration be given to the underlying factors and mechanisms that are likely to have influenced the development of social networks historically, if synthetic networks are to have any chance of reasonably depicting real world network structures.

Generating synthetic social networks is challenging due to often limited or unavailable data. Consequently, researchers tend to use simple models like regular lattices, random graphs, small-world networks, scale-free networks, and models based on spatial proximity. These models capture basic elements of real-world social networks but can fall short in complex scenarios. For instance, Jiang et al. (2022) describes a model where agents, already assigned to households and workplaces, form small-world networks based on employment or educational ties. While this approach accounts for spatial and occupational similarities, it overlooks other factors, limiting its applicability for networks like friendships that rely on personal history and intangible attributes.

To address these limitations, more sophisticated methods have been proposed, including Exponential Random Graph Models (ERGM) (Robins et al., 2007) and Yet Another Network Generator (YANG) (Amblard et al., 2015). However, they also come with their own challenges; for example, ERGMs sometimes misrepresent the likelihood of certain network structures, deviating from real-world observations.

2.3 Spatial locations

The places where people live, work, take their leisure and go to school are critically interlinked and interrelated with social networks and demographics. Spatial location also affects options open to people, including transport, access to services, job opportunities and social encounters. ABMs' capabilities in representing space explicitly and naturally is a key attraction for geographers interested in social simulation and population synthesis (Cottineau et al., 2018). Ignoring the spatial concentration of

agents with common traits, or failing to account for the effects that space has on other aspects of everyday human existence, risks overlooking a critical factor that influences a wide range of social dynamics and outcomes.

Spatial microsimulation generates synthetic populations tailored to defined geographic zones, such as census tracts (Lovelace and Dumont, 2017). However, many ABM applications require agents to be assigned to specific dwellings and workplaces, not just aggregated zones. While approaches to dealing with this have been proposed, agreement on best practice is yet to cohere. Certain agent-location assignments can be implemented using straightforward heuristic methods without greatly compromising fidelity, if heuristics align well with real-world practices. For example, children might be allocated to schools simply based on proximity, such as in Jiang et al., (2022). Others use rule-based or stochastic methods to account for observed nuances and random variability, though these often take the form of crude approximations. One of the more well-rounded examples is detailed by Zhou et al. (2022). They start by generating a synthetic population, which they then assign to specific dwellings and jobs using a combination of rule-based matching heuristic and probabilistic models. Dwellings are assigned to households by considering factors like household size, income, and dwelling type jointly. Meanwhile, jobs are assigned to workers using a destination choice model that predicts the probability of selecting locations based on factors such as sector-specific employment opportunities, commuting costs, and interactions between commuting costs and individual worker attributes. In this way, spatial location choices are more closely aligned with the diverse attributes of agents. The challenge with such an approach is to obtain sufficient microdata to inform the rules and probabilities.

2.4 Decision-making and behavioural dynamics

In practice, peoples' decision-making and behaviours are influenced by an array of factors, including their individual characteristics such as wealth, health, education, gender, and age, their social network, and their geographical circumstances. These factors shape – among other things – the information agents' are exposed to, the choices open to them, the expectations placed on them, and their personal beliefs and desires (Lobo et al., 2023). Consequently, accurately initialising such factors is important for ensuring that agents are predisposed to make decisions and take actions in ways that reflect how their real world counterparts might behave. Furthermore, the assignment of psychographic attributes to agents necessitates the prior establishment of these foundational characteristics as they are often closely entwined.

Numerous agent decision-making architectures have been proposed (see Wijermans et al. (2023)). Many suggest that a range of agent state attributes could, or even should, be taken into consideration when evaluating information and selecting behaviours. For example, the MoHub Framework (Schlüter et al., 2017) proposes four classes of attributes as potentially influential in the decision-making process: needs/goals, knowledge, assets, and social. In practice, however, the factors taken into consideration in decision-making procedures tend to be much narrower. This is understandable given the higher data demands that richer decision-making procedures entail. However, it is also regrettable given we know that decision-making often draws on many more factors than are currently accounted for, and the ABM community has worked hard to develop the tools needed to depict these richer processes.

3 Practicalities

Our wish list of features for synthetic population algorithms far exceeds their current capabilities. Perhaps the main issue today is data scarcity, especially concerning less tangible aspects of populations, such as psychological attributes and social networks, where systematic data collection is often more limited. Another significant challenge is that existing algorithms struggle to manage the numerous conditional probabilities involved in creating realistic populations, excelling on niche measures of performance but not from a holistic perspective. Moreover, there are accessibility issues with population synthesis tools. The next generation of methods need to be made more accessible to non-specialists through developing easy to use stand-alone tools or plugins for widely used platforms like NetLogo, else they risk not having their potential exploited.

Collectively, these issues may necessitate a fundamental rethink of how synthetic populations are generated. The potential benefits of successfully addressing these challenges are immense. By enhancing the capabilities of synthetic population tools to meet the wish list set out here, we can significantly improve model realism and expand the potential applications of social simulation, as well as strengthen credibility with stakeholders. More than this, though, such advancements would enhance our ability to draw meaningful insights, respecting the complexities of real-world dynamics. Most critically, better representation of the diversity of actors and circumstances reduces the risk of overlooking factors that might adversely impact segments of the population – something there is arguably a moral imperative to strive for.

Acknowledgements

MP & RP were supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union – NextGenerationEU under NPRR Grant agreement n. MUR IR0000008. CC was supported by the ERC starting Grant SEGUE (101039455).

References

Amblard, F., Bouadjio-Boulic, A., Gutiérrez, C.S. and Gaudou, B. 2015, December. Which models are used in social simulation to generate social networks? A review of 17 years of publications in JASSS. In *2015 Winter Simulation Conference (WSC)* (pp. 4021-4032). IEEE. https://doi.org/10.1109/WSC.2015.7408556 (https://doi.org/10.1109/WSC.2015.7408556)

Bruch, E.E., 2014. How population structure shapes neighborhood segregation. *American Journal of Sociology*, *119*(5), pp.1221-1278. https://doi.org/10.1086/675411 (https://doi.org/10.1086/675411)

Chapuis, K., Taillandier, P. and Drogoul, A., 2022. Generation of synthetic populations in social simulations: a review of methods and practices. *Journal of Artificial Societies and Social Simulation*, *25*(2). https://doi.org/10.18564/jasss.4762 (https://doi.org/10.18564/jasss.4762)

Cottineau, C., Perret, J., Reuillon, R., Rey-Coyrehourcq, S. and Vallée, J., 2018, March. An agent-based model to investigate the effects of social segregation around the clock on social disparities in dietary behaviour. In *CIST2018-Représenter les territoires/Representing territories* (pp. 584-589). https://hal.science/hal-01854398v1 (https://hal.science/hal-01854398v1)

Jiang, N., Crooks, A.T., Kavak, H., Burger, A. and Kennedy, W.G., 2022. A method to create a synthetic population with social networks for geographically-explicit agent-based models. *Computational Urban Science*, *2*(1), p.7. https://doi.org/10.1007/s43762-022-00034-1 (https://doi.org/10.1007/s43762-022-00034-1)

Lobo, I., Dimas, J., Mascarenhas, S., Rato, D. and Prada, R., 2023. When "I" becomes "We": Modelling dynamic identity on autonomous agents. *Journal of Artificial Societies and Social Simulation*, *26*(3). https://doi.org/10.18564/jasss.5146 (https://doi.org/10.18564/jasss.5146)

Lovelace, R. and Dumont, M., 2017. *Spatial microsimulation with R*. Chapman and Hall/CRC. https://spatial-microsim-book.robinlovelace.net (https://spatial-microsim-book.robinlovelace.net)

McPherson, M., Smith-Lovin, L. and Cook, J.M., 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), pp.415-444. https://doi.org/10.1146/annurev.soc.27.1.415 (https://doi.org/10.1146/annurev.soc.27.1.415)

Robins, G., Pattison, P., Kalish, Y. and Lusher, D., 2007. An introduction to exponential random graph (p*) models for social networks. *Social networks*, *29*(2), pp.173-191. https://doi.org/10.1016/j.socnet.2006.08.002 (https://doi.org/10.1016/j.socnet.2006.08.002)

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., Janssen, M.A., McAllister, R.R., Müller, B., Orach, K. and Schwarz, N., 2017. A framework for mapping and comparing behavioural theories in models of social-ecological systems. *Ecological economics*, *131*, pp.21-35. https://doi.org/10.1016/j.ecolecon.2016.08.008 (https://doi.org/10.1016/j.ecolecon.2016.08.008)

Wijermans, N., Scholz, G., Chappin, É., Heppenstall, A., Filatova, T., Polhill, J.G., Semeniuk, C. and Stöppler, F., 2023. Agent decision-making: The Elephant in the Room-Enabling the justification of decision model fit in social-ecological models. *Environmental Modelling & Software*, *170*, p.105850. https://doi.org/10.1016/j.envsoft.2023.105850 (https://doi.org/10.1016/j.envsoft.2023.105850)

Wu, G., Heppenstall, A., Meier, P., Purshouse, R. and Lomax, N., 2022. A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. *Scientific Data*, *9*(1), p.19. https://doi.org/10.1038/s41597-022-01124-9 (https://doi.org/10.1038/s41597-022-01124-9)

Zell, D., McGrath, C. and Vance, C.M., 2014. Examining the interaction of extroversion and network structure in the formation of effective informal support networks. *Journal of Behavioral and Applied Management*, *15*(2), pp.59-81. https://jbam.scholasticahq.com/article/17938.pdf (https://jbam.scholasticahq.com/article/17938.pdf)

Zhou, M., Li, J., Basu, R. and Ferreira, J., 2022. Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation. *Computers, Environment and Urban Systems*, *91*, p.101717. https://doi.org/10.1016/j.compenvurbsys.2021.101717 (https://doi.org/10.1016/j.compenvurbsys.2021.101717)

Roxburgh, N., Paolillo, R., Filatova, T., Cottineau, C., Paolucci, M. and Polhill, G. (2025) **Outlining some requirements for synthetic populations to initialise agent-based models.** *Review of Artificial Societies and Social Simulation*, 27 Jan 2025. https://rofasss.org/2025/01/29/popsynth (https://rofasss.org/2025/01/29/popsynth)

ABM **AGENT-BASED MODELLING** **ATTRIBUTES** **CLEMENTINECOTTINEAU** **CLIMATE** **CLIMATE-CHANGE** **COVARIANCE** **ENVIRONMENT** **GARYPOLHILL** **INITIALISATION** **MARIOPAOLUCCI** **MODELLING** **NICKROXBURGH** **POPULATION SYNTHESIS** **ROCCOPAOLILLO** **SYNTHETIC POPULATION** **TATIANAFILITOVA**