Check for updates

B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2023, Vol. 50(8) 2309–2316
© The Author(s) 2023

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23998083231203066
journals.sagepub.com/home/epb
Sage

Synthetic population Catalyst: A micro-simulated population of England with circadian activities

Hadrien Salat and **Dustin Carlino**
The Alan Turing Institute, UK

Fernando Benitez-Paez
University of St Andrews, UK
The Alan Turing Institute, UK

Anna Zanchetta
The Alan Turing Institute, UK

Daniel Arribas-Bel
University of Liverpool, UK
The Alan Turing Institute, UK

Mark Birkin
University of Leeds, UK
The Alan Turing Institute, UK

Abstract

The Synthetic Population Catalyst (SPC) is an open-source tool for the simulation of populations. Building on previous efforts, synthetic populations can be created for any area in England, from a small geographical unit to the entire country, and linked to geolocalised daily activities. In contrast to most transport models, the output is focussed on the population itself and the way people socially interact together, rather than on a precise modelling of the volume of transport trips from one area to another. SPC is therefore particularly well suited, for example, to study the spread of a pandemic within a population. Other applications include identifying segregation patterns and potential causes of inequality of opportunity amongst individuals. It is fast, thanks to its Rust codebase. The outputs for each lieutenancy area in England are directly available without having to run the code.

Corresponding author:
Fernando Benitez-Paez, University of St Andrews, St Andrews, Scotland, UK.
Email: Fernando.Benitez@st-andrews.ac.uk

Keywords
Population micro-simulation, social interactions, transport flows, synthetic data

Introduction

Governments expend significant resources capturing demographic data from censuses and social surveys. Censuses typically provide small area counts for a limited range of attributes. By contrast, surveys provide numerous variables for representative individuals, but are usually not geographically aggregated and limited by small sample sizes. However, both research and applied policy analyses require the combination of these qualities into individual level data that are rich in attributes and locally specified. This can be achieved by creating high resolution synthetic data that are representative of the real population, its sociodemographic characteristics, and its interactions.

Our approach is rooted in spatial micro-simulation, which integrates census data with individual level surveys, providing a synthetic list of individuals or households allocated to geographical locations (Arentze et al., 2007; Farooq et al., 2013). This approach has been applied to multiple fields, such as health (Smith et al., 2011; Wu et al., 2022), transport (Horl and Balac, 2021; Lovelace et al., 2014), policy (O'Donoghue et al., 2013), and social inequality (He et al., 2020). Recently, it has been used to feed complex dynamic models where time and age of the population are integrated (Birkin, 2021; Lomax and Smith, 2017), in particular to study the spread of a pandemic (Spooner et al., 2021).

There are potential challenges to generating synthetic population datasets: extensive computational requirements, excessive complexity in the integration of data sources, or limited area coverage (Horl and Balac, 2021; Lovelace et al., 2014; O'Donoghue et al., 2013). Tanton (2017) describes four areas of improvement: variability estimation; reusability by other models; explicit results for policymakers; and integration methods for big data streams to promote 'what-if' scenario modelling. The lack of reproducibility, reusability, and the scarcity of data sources and tools create significant impediments for researchers looking for synthetic populations that can easily be integrated into other models.

The tool introduced in this article is focussed on the socio-economic characteristics and daily activities representing complex social interactions within the population at national level. It is easy-to-read, computationally efficient and encapsulates multiple data sources. In addition, we provide pre-compiled areas, matching all the lieutenancy areas of England, that can be used directly as input by other models. This paper is therefore the description of both open-source software and an open data product.

The methods and software originate from the Dynamic Micro-simulation Model for Epidemics (DyME). This project arose from the Royal Society's Rapid Assistance in Modelling the Pandemic (RAMP) initiative (Spooner et al., 2021). DyME was designed to study the spread of the COVID-19 pandemic in Devon. The synthetic population generation phase of DyME was extracted, extended, and significantly enhanced. Then it was ported to a highly efficient implementation using the Rust language, and renamed as the Synthetic Population Catalyst (SPC) (Carlino et al., 2022b). It produces a synthetic population of individuals covering all England at Middle-layer Super Output Area (MSOA – census area of about 8,000 people) scale for the year 2020.

In the following sections, we describe the new data schema and the additional modelling methods we implemented. The intended usage of the tool is presented and three applications of the SPC outputs are mentioned.

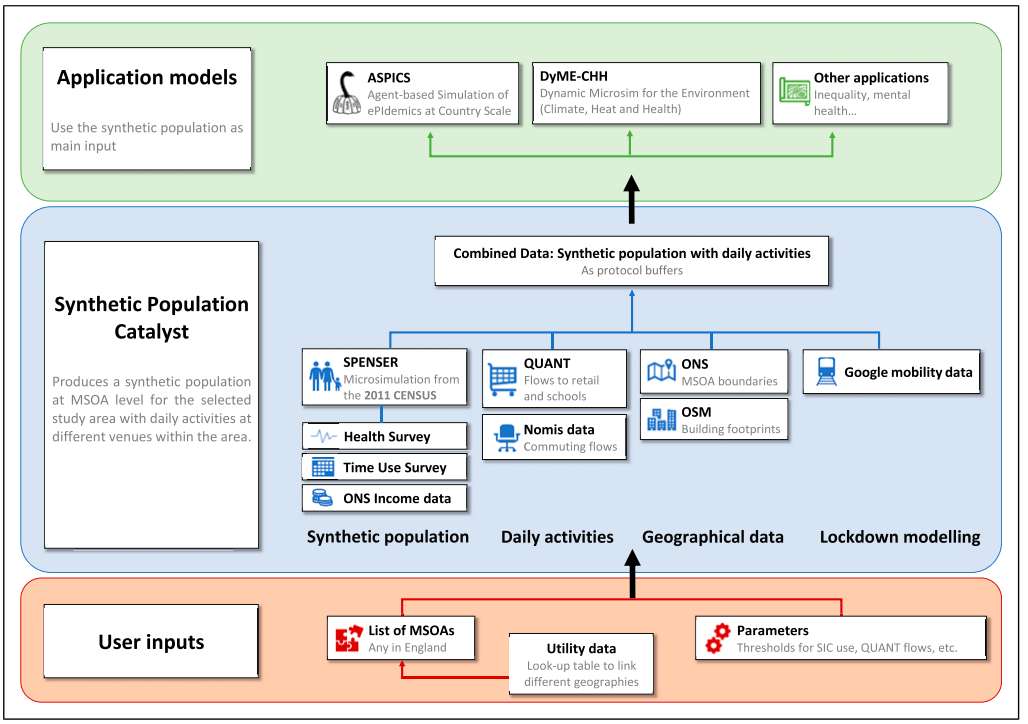

Figure 1. The SPC framework.

Data sources and methods

The full framework of the SPC is summarised in Figure 1. The SPC uses a mix of population, mobility, and geographical data, best representing the year 2020 in England. All that is necessary to run the model is stored in a pre-processed form inside a dedicated repository, excepted building footprints that are downloaded on the fly from OpenStreetMap. Specific results and descriptions of the code which are sufficient to replicate the applications can be found in the supplementary material (Carlino et al., 2022b).

The population data are based on the SPENSER microsimulation model that disaggregates the 2011 census (Office for National Statistics, 2020) to the MSOA scale and updates it to the year 2020 (Lomax et al., 2022). Propensity score matching is used to join these data to the Health Survey for England 2017 (University College London, 2021) and the Time Use Survey 2014–15 (Gershuny and Sullivan, 2021), see Morrissey et al. (2015). An hourly wage and annual salary is added for each individual classified as an employee from the Office for National Statistics (ONS) salary data (Office for National Statistics, 2022a, 2022b), based on their standard occupational classification category, home region, hours worked, and sex. The resulting data contain a range of identifiers making it possible to add to the output other fields from the source data that were not retained by the SPC.

The modelling of trips to schools and retail is taken from the QUANT project (Batty and Milton, 2021). It is based on a probability matrix of flows between any pair of MSOAs over Britain, derived from a spatial interaction model along the shortest paths of the road network (see Spooner et al., 2021). The modelling of commuting flows uses a breakdown of all single workplaces in England. These are described by size and industry classifications, at Lower-layer Super Output Area scale,

estimated from two Nomis datasets (Nomis, 2015, 2020). The employees for each workplace are drawn according to Schlapfer et al. (2021).

Finally, Google mobility reports (Google LLC, 2021) at county level are used to obtain daily coefficients to reduce the duration of all activities away from home during the COVID-19 pandemic. Additional geographic data are taken from ONS boundaries (Office for National Statistics, 2021) and OpenStreetMap building footprints datasets. The modelling approach is demonstrably robust (Carlino et al., 2022a: 7), for example, hourly salaries show a correlation of 0.99 to base data (ibid, 7.2.2). BMI is in the range 0.93–0.97 (ibid, 7.3).

The pipeline in the SPC is implemented in the Rust language, a modern systems language focussed on type safety and concurrency to achieve high performance. The use of static typing prevents many types of bugs. For example, numeric and string IDs are wrapped with stronger types and the compiler prevents the code from being built when there is a mistake, giving high confidence in later refactoring the code. In addition, the performance is significantly accelerated due to the compiler laying out the data contiguously in memory, exploiting cache locality. Runtimes for all pre-compiled areas and more information about design choices are provided in the online documentation (Carlino et al., 2022a: 11). Runtimes range from about 30 seconds for West Yorkshire (2.3 million individuals) to about 13 minutes for Greater London (8.7 million individuals).

Usage and limitations

The SPC generates synthetic population files encapsulated in a protocol buffer format. Proto-buffers have an efficient binary representation, making them fast to transfer and read. Code to parse the data can be auto-generated in most programming languages.

The output files for all the lieutenancy areas of England are provided (Carlino et al., 2022a). Hence, users do not need to install the SPC to get familiar with the synthetic population characteristics the tool provides. Users can read these outputs in any supported language, and then extract and transform the parts of the data needed for their model. We include guidance to convert this file to other usual formats (JSON, numpy arrays) and plot the synthetic individuals onto a map.

Users who want to generate custom areas must install and execute the tool. A list of MSOA codes for the desired area must be provided. We include a script to get the required list of MSOAs from different geographies by names. The SPC then assembles the corresponding data to produce a single proto-buffer file. The version 1.1 of the SPC tool is openly available on GitHub under a MIT Licence (Carlino et al., 2022b).

There are limitations that users should consider before incorporating the output files into their external models.

(1) The outputs only contain individuals living within the specified area and their daily activities can only occur within that same area. Combining the synthetic populations from two areas should not be done by appending two output files. Instead, the SPC should be run with a new study area containing all MSOAs in both areas.
(2) The daily retail and school activities per person living in the same MSOA are currently taken among a fixed set of 10 and 5 venues, respectively.
(3) A venue can be randomly sampled from the flows and weighted appropriately to force people to visit specific places each day. For retail, this can be repeated every day.
(4) There are no distinctions between weekends and weekdays.

Applications

The SPC is well suited to study any phenomenon based on social interactions amongst the population. We describe here three examples.

The Agent-based Simulation of ePIdemics at National Scale (ASPICS) model is an SEIR-type model. It originates from the second stage of the DyME model, to study the spread of the COVID-19 pandemic (or any future or past epidemic if re-calibrated). It is publicly available through a GitHub repository (Greig et al., 2022). ASPICS uses the outputs of the SPC in the manner illustrated by Figure S1 in the supplementary material.

The increasing health risks associated with rising temperatures is studied by the Dynamic Microsimulation for the Environment – Climate, Heat, and Health project (DyMECHH) (Bowyer et al., 2022). DyME-CHH is a multi-level model that measures an exposure risk threshold. This threshold is based on a combination of the individual socio-demographic characteristics and health risks provided by the SPC, and the UK Climate Projection (UKCP) dataset (Met Office Hadley Centre (MOHC) 2019).

Patterns of segregation can be analysed, including those that are not the result of simple physical proximity. This could be the result of, possibly involuntary, exclusionary interactions amongst individuals. The scale of the SPC is particularly suitable for indices based on neighbourhood variation ratios (Salat et al., 2018). The absence of meaningful interactions between different social groups despite physical proximity is given by the daily activities.

Conclusions and future directions

The SPC is a solution to provide a stable, easy-to-read and scalable method to create synthetic population files that can feed other external and specialised models to study the complexity of society. The methods described in this paper provide original contributions to model the commuting flows and individual salaries. Our implementation uses a modern systems language, Rust, that provides the performance required to create the output files efficiently at the national scale.

The SPC is a project with ongoing extensions. The data schema presented here is versioned. Hence users can initially write their code against the current latest version of the SPC and migrate to updated versions later at their convenience. Some possible next steps include updating the population after the 2021 UK census is released and including yearly energy consumption per household and per venue. We would like to encourage other researchers wanting to enrich the synthetic population to contact the SPC team through the public repository. We aim to develop several application projects with immediate policy applications, such as the mentioned pandemic model (ASPICS) and climate change model (DyME-CHH).

Acknowledgements

We want to thank all the members of the team that developed the original DyME model, and more particularly Fiona Spooner, Nick Malleson, Alex Coleman and Karyn Morrissey for their close guidance during our work extending the model.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/W006022/1, particularly the 'Ecosystem of Digital Twin' and 'Shocks and Resilience' themes within that grant & The Alan Turing Institute. For the purpose of open access, the authors have applied a CC BY public copyright Licence to any Author Accepted Manuscript version arising.

ORCID iDs

Hadrien Salat https://orcid.org/0000-0003-0958-9715
Fernando Benitez-Paez https://orcid.org/0000-0002-9884-6471
Daniel Arribas-Bel https://orcid.org/0000-0002-6274-1619

Supplemental Material

Supplemental material for this article is available online.

References

Arentze T, Timmermans H and Hofman F (2007) Creating synthetic household populations: problems and approach. *Transportation Research Record* 2014(1): 85–91. DOI: 10.3141/2014-11.

Batty M and Milton R (2021) A new framework for very large-scale urban modelling. *Urban Studies* 58(15): 3071–3094. DOI: 10.1177/0042098020982252.

Birkin M (2021) Microsimulation. In: Shi W, Goodchild MF, Batty M, et al. (eds), *Urban Informatics*. Singapore: Springer Singapore, 845–864. DOI: 10.1007/978-981-15-8983-644.

Bowyer R, Benitez-Paez F and Ding J (2022). *DyME-CHH project [GitHub repository]*. Available at: https://github.com/alan-turing-institute/dymechh

Carlino D, Salat H and Benitez-Paez F (2022a) *Synthetic Population Catalyst (SPC) Online Documentation*. Available at: https://alan-turing-institute.github.io/uatk-spc/ (accessed on May 2022).

Carlino D, Salat H, Benitez-Paez F, et al (2022b) Synthetic Population Catalyst (SPC) v1.1. Available at: https://github.com/alan-turing-institute/uatk-spc/releases/tag/v1.1. DOI:10.5281/zenodo.6586791.

Farooq B, Bierlaire M, Hurtubia R, et al. (2013) Simulation based population synthesis. *Transportation Research Part B: Methodological* 58: 243–263. DOI: 10.1016/j.trb.2013.09.012.

Gershuny J and Sullivan O (2021) *United Kingdom Time Use Survey, 2014-2015*. [data collection]. DOI: 10.5255/UKDA-SN-8128-1.

Google LLC (2021). *Google COVID-19 Community Mobility Reports* [data collection] https://www.google.com/covid19/mobility/ (accessed on June 2021).

Greig R, Carlino D, Salat H, et al. (2022). *Urban Analytics Toolkit - Agent-Based Simulation of ePIdemics at National Scale [GitHub Repository]* Available at: https://github.com/alan-turing-institute/uatk-aspics

He BY, Zhou J, Ma Z, et al. (2020) Evaluation of city-scale built environment policies in New York City with an emerging-mobility-accessible synthetic population. *Transportation Research Part A: Policy and Practice* 141: 444–467. DOI: 10.1016/J.TRA.2020.10.006.

Horl S and Balac M (2021) Synthetic population and travel demand for Paris and Ile-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies* 130: 103291. DOI: 10.1016/j.trc.2021.103291.

Lomax N and Smith AP (2017) Microsimulation for demography. *Australian Population Studies* 1(1): 73–85. DOI: 10.37970/aps.v1i1.14.

Lomax N, Smith AP, Archer L, et al. (2022) An opensource model for projecting small area demographic and landuse change. *Geographical Analysis*. DOI: 10.1111/gean.12320.

Lovelace R, Ballas D and Watson M (2014) A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. *Journal of Transport Geography* 34: 282–296. DOI: 10.1016/j.jtrangeo.2013.07.008.

Met Office Hadley Centre (MOHC) (2019) *UKCP Local Projections at 2.2km Resolution for 1980-2080*. [data collection] https://catalogue.ceda.ac.uk/uuid/d5822183143c4011a2bb304ee7c0baf7 (accessed on June 2022).

Morrissey K, Clarke G, Williamson P, et al. (2015) Mental illness in Ireland: simulating its geographical prevalence and the role of access to services. *Environment and Planning B: Planning and Design* 42(2): 338–353. DOI: 10.1068/b130054p.

Nomis (2015) *Business Register and Employment Survey.* Available at: https://www.nomisweb.co.uk/datasets/newbrespub (accessed on June 2021).

Nomis (2020) *UK Business Counts - Local Units by Industry and Employment Size Band*. Available at: https://www.nomisweb.co.uk/datasets/idbrlu (accessed on June 2021).

Office for National Statistics (2020) *2011 Census: Aggregate Data*. [data collection]. DOI: 10.5257/census/aggregate-2011-1.

Office for National Statistics (2021) *Middle Layer Super Output Areas (December 2011) Boundaries Full Clipped (BFC) EW V3* [data collection].https://geoportal.statistics.gov.uk/datasets/middle-layer-super-output-areas-december-2011-boundaries-full-clipped-bfc-ew-v3 (accessed on June 2021).

Office for National Statistics (2022a) *Earnings and Hours Worked, Age Group. ASHE Table 6 (2020 revised edition)* [data collection] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/agegroupashetable6 (accessed on May 2022).

Office for National Statistics (2022b). *Earnings and hours worked, region by occupation by four-digit SOC: ASHE Table 15 (2020 revised edition)* [data collection] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/regionbyoccupation4digitsoc2010ashetable15 (accessed on May 2022).

O'Donoghue C, Ballas D, Clarke G, et al. (eds) (2013) *Spatial Microsimulation for Rural Policy Analysis*. Heidelberg: Springer Berlin. Available at: https://link.springer.com/book/10.1007/978-3-642-30026-4#bibliographic-information

Salat H, Murcio R, Yano K, et al. (2018) Uncovering inequality through multifractality of land prices: 1912 and contemporary Kyoto. *PLoS One*. DOI: 10.1371/journal.pone.0196737.

Schlapfer M, Dong L, O'Keeffe K, et al. (2021) The universal visitation law of human mobility. *Nature* 593: 522–527. DOI: 10.1038/s41586-021-03480-9.

Smith DM, Pearce JR and Harland K (2011) Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health & Place* 17(2): 618–624. DOI: 10.1016/j.healthplace.2011.01.001.

Spooner F, Abrams JF, Morrissey K, et al. (2021) A dynamic microsimulation model for epidemics. *Social Science & Medicine* 291: 114461. DOI: 10.1016/j.socscimed.2021.114461.

Tanton R (2017) Spatial microsimulation: developments and potential future directions. *International Journal of Microsimulation* 11(1): 143–161. DOI: 10.34196/ijm.00176.

University College London (2021) Health Survey for England, 2017 [data collection]. 2nd edition. DOI:10.5255/UKDA-SN-8488-2.

Wu G, Heppenstall A, Meier P, et al. (2022) A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. *Scientific Data* 9(1): 1–11. DOI: 10.1038/s41597-022-01124-9.

Hadrien Salat, Dustin Carlino, Fernando Benitez-Paez and Anna Zanchetta are Research Associates at the Alan Turing Institute.

Daniel Arribas-Bel is Professor in Geographic Data Science at the Department of Geography and Planning of the University of Liverpool (UK), and Deputy Programme Director for Urban Analytics at the Alan Turing Institute, where he is also an ESRC Fellow. At Liverpool, he is a member of the Geographic Data Science Lab and directs the MSc in Geographic Data Science.

Mark Birkin is Professor of Spatial Analysis and Policy in the School of Geography at the University of Leeds, where he is co-director of both the Consumer Data Research Centre and Leeds Institute for Data Analytics. He is Programme Director for Urban Analytics at the Alan Turing Institute.