Growing populations from the 'bottom-up': an ABM approach to the generation of synthetic populations

Umberto Gostoli[1 [0000−0002−2394−3613]], Martin Hinsch[1 [0000−0002−7059−7266]], and Eric Silverman[1 [0000−0003−0147−6118]]

MRC/CSO Social and Public Health Sciences Unit, School of Health and Wellbeing, University of Glasgow, Clarice Pears Building, 90 Byres Road Glasgow, G12 8TB
umberto.gostoli@glasgow.ac.uk

Abstract. In this paper, we discuss the advantages and challenges of a 'generative' approach, as opposed to the widely used 'descriptive' approaches, to the production of synthetic populations. We present an implementation of this approach, in the form of an agent-based model which grows a synthetic population, starting from the UK fertility and mortality rates. We show how this model is able to reproduce some demographic data of the UK population.

Keywords: Synthetic populations · Computational demography · Agent-based modelling · Social simulation.

1 Introduction

In the last few years, ABM and microsimulation research have seen a significant growth of studies on the generation of synthetic populations matching the empirical demographic and socioeconomic data of a target population (for a review see [5]).

Chapuis et al. [5] distinguish the synthetic population techniques between those based on the properties of the entities (known as *synthetic reconstruction*) or reproduce known real entities (known as *combinatorial optimization*). Synthetic reconstruction generates populations by creating agents with attributes randomly sampled from existing distributions or from an estimated joint distribution, typically using methods like the IPF or the MCMC algorithms. On the other hand, combinatorial optimization generates agents which are 'copies' of real individuals, trying to fit the various population-level characteristics' distributions. Notwithstanding the different statistical algorithms used by these two paradigms, both of them focus on generating a faithful micro-level picture of the real target population, so they can both be considered as part of a high-level 'descriptive' paradigm.

Here we propose an alternative approach based on the 'generative' paradigm, able in principle not just to describe but to *explain* the emergence of the demographic and socioeconomic structure of real populations. This approach is

derived from the semi-artificial populations approach proposed by Bijak et al. [4,11]. According to this generative approach, a synthetic population is generated from the 'bottom-up', starting from a random population of agents placed on a more or less defined geographical space. Then, we simulate the agents' interactions and life course events: partnership formation/dissolution; births; and deaths. The occurrence of these events can be *data-driven*, i.e., based on empirical data such as the fertility and mortality rates, *theory-driven*, i.e., based on behavioural models according to which agents take some life course decisions (e.g., the choice of their partner or the decision to relocate), or a mixture of the two, with the agents' behaviour including constraints which ensure the empirical data is reproduced at the aggregate level. Generation after generation, the process produces a synthetic population whose structure we can compare to the structure of the target real-world population.

We argue that this approach has three main advantages compared to the 'descriptive' approach, stemming from the theory-driven dynamics of the model. First, we can assign to the agents characteristics and attributes for which there is a lack of empirical data (but which we may want to consider in our ABM). In other words, by focusing on the process through which an agent happens to have a certain characteristic, we can mitigate the problem represented by the lack of data about how this characteristic is distributed in the population.

Second, the generative approach allows us to reproduce aspects of the population's structure which may be important to include but which can hardly be reproduced through statistical techniques because of their 'complex' nature. An example is the generation of kinship networks connecting agents, and household networks which had a fundamental role in an application of this model meant to simulate the provision of informal social care. See [7] and [8]).

Third and, perhaps most importantly, this approach allows us to take into account the effect of agents' behaviour and their interactions on the demographic and socioeconomic structure of the population. Adopting a faithful picture of a society at a certain period in time as the starting point of an ABM simulation implies the assumption that the dynamics of the model's variables do not affect in any way the life course events affecting the joint distribution of the population's characteristics. To the extent that this assumption does not hold, simulations based on populations generated through the 'descriptive' approach will produce biased results.

On the other hand, this approach will give us a realistic model of the population only to the extent that the behavioural theory driving its dynamics is sound. So, the inclusion of unsupported behavioural assumptions, may increase the uncertainty of the model's outcomes.

2 The model

The current model was programmed from scratch in Julia, based on an earlier version written in Python [8]. The current version as well as the release used to

generate the results presented in this paper can be found at `https://doi.org/10.5281/zenodo.8154478`.

2.1 General concepts

Agents in the model have a **socioeconomic status** (SES), encoded as a number between 0 and 4. They can require care (due to physical or mental illness) which is represented as **care need**, a number between 0 (no care needed) and 4 (unable to live alone).

Besides biological family relations, the model explicitly includes the concept of **guardian**, i.e., a person legally responsible for a child (usually the parents), and **provider**, i.e., a person economically providing for another person.

2.2 Setup

The population is created using the age and gender distribution as measured in the 1921 UK census [2]. A proportion of **startProbMarried** of the adult population is assigned as couples (with age difference maintained between 5 and -2). Then with the probability $1 - \textbf{startProbOrphan} \cdot \text{age}$ for each individual in the population a random female agent that is between 18 and 40 years older is assigned as the mother and, if present, its partner as the father.

To build the map, houses form towns containing a number of houses in rough proportion to real-world population density. Each adult female is assigned to a randomly selected house together with her partner and minor children, if present. All remaining agents form single-person households.

The simulation starts in the year 1920 and updates in one-month time steps, until the year 2040.

The demographic and socioeconomic time course events we included in the framework are shown in Figure 1. The events are placed along the timeline to reflect their approximate, typical, timing. Some life course events (such as birth, adoption or death) are events which do not depend on the agents' decision process, whereas others (indicated with the letter D in Figure 1) involve the agents' decision process. Moreover, some events logically follow others (e.g., divorce is a probabilistic event following the formation of partnership).

2.3 Births

Married females between **minPregnancyAge** and **maxPregnancyAge** and whose youngest child is older than 1 can give birth. The probability to give birth is calculated from a base birth with a bias which depends on the woman's socio-economic status and on the number of previous children (capped at 4).

Fertility rates are computed similarly to mortality rates: data from the Eurostat Statistics Database [6] and the Office for National Statistics [10] are used from 1950–2009, with Lee-Carter projections taking over thereafter.

For years before 1951 the base birth rate is obtained by scaling the empirical overall population fertility (i.e. children per person and year) by the proportion

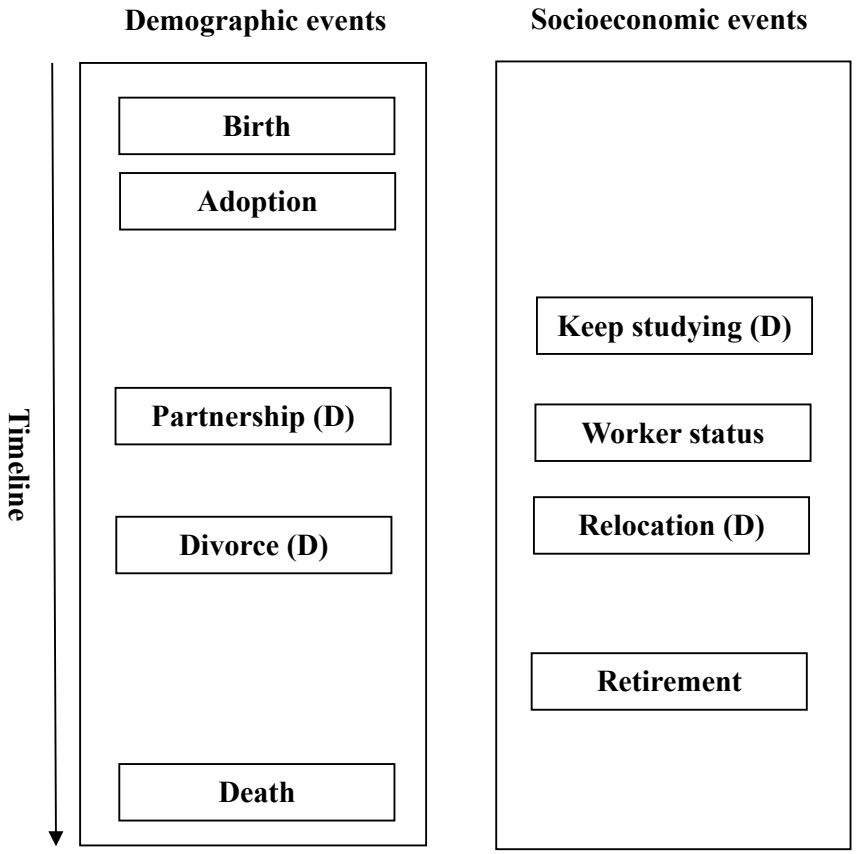

Fig. 1. Life course events.

of potential mothers in the population and an empirical age-specific fertility factor. For years after 1950 empirical age specific fertility data is scaled by the proportion of potential mothers of that age.

If a woman gives birth a new agent is created with the woman and her partner as parents and guardians, the woman as a provider and the woman's house as home.

2.4 Adoptions

All individuals that can not live alone and do not have a guardian get assigned a new guardian if possible.

The list of potential guardians in this class is, in order, the individual's parents, partners of their previous guardian(s), parents and siblings of the individual's parents and parents and siblings of the previous guardian(s). The first person out of this list that is alive and an adult is selected, if available.

If no family guardian is found a random couple where both partners are adults and have worker status is selected. If a guardian is found the individual is moved to the guardian's house and the guardian and their partner (if there is one) are assigned as guardians to the individual.

2.5 Marriages

All single adult males with a care need level below 4 attempt to find a partner. Single females that are older than **minPregnancyAge** are eligible as partners.

Marriage probability Given a man's age class $c = \text{age}/10$, the basic yearly probability of that man to find a partner is calculated as

$$p_{\text{m,base,c}} = \textbf{basicMaleMarriageProb} \cdot \textbf{maleMarriageModifierByDecade}_c \cdot f_{\text{work}}.$$

Where f_{work} is defined as **notWorkingMarriageBias** if the man has a care level above one or is not working, and 1 otherwise.

If $r_{n,c}$ is the proportion of men without any children living with them in age class c then the realised probability to marry $p_{m,c}$ in that age class is defined as

$$p_{m,c} = \begin{cases} p_{\text{m,base,c}} \cdot \frac{1}{r_{n,c}+(1-r_{n,c})\cdot\textbf{manWithChildrenBias}} & \text{men without children} \\ p_{\text{m,base,c}} \cdot \frac{\textbf{manWithChildrenBias}}{r_{n,c}+(1-r_{n,c})\cdot\textbf{manWithChildrenBias}} & \text{men with children.} \end{cases}$$

Every eligible man marries with probability $p_{m,c}$.

Partner selection If a man marries, a woman is selected out of those eligible that also

- do not live in the same house as,

– and are not a relative to the first degree of

the focal man.

The selection is done by weighted random choice where a woman i's weight is calculated as a product of a number of different factors:

$$w_i = \text{geoFactor} \cdot \text{socFactor} \cdot \text{ageFactor} \cdot \text{childrenFactor} \cdot \text{studentFactor}.$$

Given the Manhattan distance (= sum of distances in x and y direction) between the man's and the woman's town d,

$$\text{geoFactor} = 1/e^{d \cdot \mathbf{betaGeoExp}}.$$

The status distance s is the absolute difference between the man's and woman's social classes r_m and r_w (the maximum of the woman's parents classes if she is a student), normalised by the number of classes. The social factor is then calculated as

$$\text{socFactor} = \begin{cases} 1/e^{s \cdot \mathbf{betaSocExp}} & r_m < r_w \\ 1/e^{s \cdot \mathbf{betaSocExp} \cdot \mathbf{rankGenderBias}} & \text{otherwise.} \end{cases}$$

The age factor is calculated from the adjusted age difference

$$d_{\text{age}} = \text{age}_{\text{man}} - \text{age}_{\text{woman}} - \mathbf{modeAgeDiff}$$

as:

$$\text{ageFactor} = \begin{cases} 1/e^{d_{\text{age}}^2 \mathbf{maleOlderFactor}} & d_{\text{age}} > 0 \\ 1/e^{d_{\text{age}}^2 \mathbf{maleYoungerFactor}} & d_{\text{age}} \leq 0 \end{cases}$$

The children factor is calulated from the number of children living in the same house as the woman, n as

$$\text{childrenFactor} = 1/e^{n \cdot \mathbf{bridesChildrenExp}}.$$

Finally, if the woman is a student, her probability of being selected is reduced by a factor which is a parameter of the model.

The couple are set as each others' partners and all dependents of either individual become dependents of both.

With probability **probApartWillMoveTogether** both individuals in the couple as well as their dependent children or other dependent members of the households move in together. With probability **coupleMovesToExistingHousehold** the house they move to is the house with the least occupants out of the two houses of the new couple. Otherwise they move into a randomly selected empty house in the same or an adjacent town to one of the two households.

2.6 Divorces

Each couple has the potential to divorce. The probability to divorce is calculated from a base divorce rate p_{div} biased by SES, using parameter **divorceBias** with

$$p_{\text{div}} = \begin{cases} \textbf{basicDivorceRate} \cdot \textbf{divorceRateModifier}(\frac{age(m)}{10}) & \text{year} < 2012 \\ \textbf{divorceVariable} \cdot \textbf{divorceRateModifier}(\frac{age(m)}{10}) & \textit{otherwise} \end{cases}$$

If the woman's status is 'student', she then starts working. The man moves out together with each of the man's children who are not the woman's children, as well as with a probability of **probChildrenWithFather** each of the man's dependents who have the same relationship status with both the man and woman. The new home is a randomly selected house from either the same or an adjacent town or the entire country.

2.7 Age transition

In each iteration/month, the age of all the agents born in that month is increased. Agents that are 18 years old become independent. That means all guardian-dependent relationships are dissolved.

2.8 Work

For every eligible agent changes in life stage are checked:
ageTeenagers → teenager
ageOfAdulthood → student
ageOfRetirement → retired

Students New students get a class rank of 0 and become out of town students with probability **probOutOfTownStudent**.

Retirement Newly retired agents' wage and working hours are set to 0. Their pension is calculated as

$$\text{pension} = \text{lastIncome} \cdot \text{shareWorkingTime} \cdot e^{\text{dk}}$$

with

$$\text{shareWorkingTime} = \text{workingPeriods}/\textbf{minContributionPeriods}$$

and

$$\text{dk} \in \mathcal{N}[0, \textbf{wageVar}].$$

Workers For all workers that are not in maternity leave, working period and work experience are increased by 1 and wage and income calculated.

With a worker's initial and final incomes I_i and I_f (see social transition) the base wage w_b is calculated based on the agent's SES and work experience as

$$w_b = I_f \frac{I_i}{I_f}^{e^{-\textbf{incomeGrowthRate}(\text{class})\cdot \text{workExperience}}}$$

Using the wage stochasticity

$$\text{dk} \in \mathcal{N}[0, \textbf{wageVar}],$$

the wage is then simply $w = w_b \cdot \text{dk}$

This wage is used to determine the agent's income, which is the product of the wage and the care need-dependent number of working hours.

2.9 Work status

Check transitions for all agents born in the current month whose age is equal to the **workingAge** of their class rank and whose current status is student.

Study Agents with a SES class rank lower than 4 begin to study or keep studying with a probability p_s. This increases their SES class by 1.

Probability to study p_s is 0 if both parents of the agent are dead, if the agent has no provider or if the household disposable income is 0.

Otherwise p_s is the product of income and education effects:

$$p_s = \text{incomeEffect} \cdot \text{educationEffect}$$

The income effect is calculated as:

$$\text{incomeEffect} = \frac{\textbf{constantIncomeParam} + 1}{e^{\textbf{eduWageSensitivity}\cdot \text{relCost}}} + \textbf{constantIncomeParam}$$

Where *relCost* is the ratio of forgone salary to *perCapitaDisposableIncome*:

$$\text{forgoneSalary} = \textbf{incomeInitialLevels}(\text{classRank}) \cdot \textbf{weeklyHours}(\text{careNeedLevel})$$
$$\text{relCost} = \text{forgoneSalary}/\text{perCapitaDisposableIncome}$$

With d_E as the difference between an agent's parents' maximum class rank and that of the the agent itself we obtain the education effect as:

$$E = e^{\textbf{eduRankSensitivity}\cdot d_E}$$
$$\text{educationEffect} = \frac{E}{E + \textbf{constantEduParam}}$$

Work Agents that do not study or stop studying, start working instead. Their status is set to worker and their initial income is set to

$$I_i = \textbf{incomeInitialLevels}(\text{classRank}) \cdot e^{\text{dk}}.$$

Where *dk* is again the wage stochasticity factor:

$$\text{dk} \in \mathcal{N}[0, \textbf{wageVar}]$$

The agent's final wage I_f is drawn from the class-specific income distribution.

2.10 Relocation

Single agents that can live alone, are workers and share their house with at least one other person who is neither their dependent or guardian, move into their own house with probability **moveOutProb**. They move into a randomly selected empty house (either in the same or an adjacent town or anywhere) together with their dependents.

2.11 Death

Mortality rates in the model follow Noble et al. [9] and use a Gompertz-Makeham mortality model until 1951. From that point we use mortality rates drawn from the Human Mortality Database [1]. Lee-Carter projections generate agent mortality rates from 2009.

3 Calibration

We used population Monte Carlo ABC [3] to calibrate the model against a number of empirical distributions, which, briefly, works as follows: A population of random points in prior parameter space is generated. Then, on each iteration the "quality" of a point is calculated and a proportion of the worst points is replaced with new points generated by perturbing a proportion of the best points. To obtain the quality of a point the simulation is run with that parameter combination and the sum of the relative mean square differences between the empirical data sets and the respective model outputs (from the corresponding time step) is calculated.

We used the following data sets for calibration: the UK population age pyramid in 2020; the distribution of UK households by size in 2021; the distribution of ages of women giving birth in 2020; the distribution of the differences between the age of the men and the women forming partnerships (in 2017, using the data for France); the distribution of new mothers by number of previous children in 2020; the distribution of births by mothers' age and SES in 2020; the share of lone parents' households in 2021; the distribution of people by SES and age in 2011; the income decile distribution in 2020.

4 Results

We performed a calibration of the parameters of the model to assess its capacity of replication some features of the real UK population. Then, we performed 128 simulations sampling the parameters' combinations according to the posterior distribution resulting from the calibration. The figures below show the mean over 128 simulations and the 95% CI.

As we can see from Figure 2 the model replicated quite well the population's age distribution (i.e., the population pyramid) of the UK for the year 2020.

Figure 3 shows the share of households by size in the year 2021. We can see that, in general, the simulated distribution replicates the empirical distribution quite well, apart from a few classes (such as size 4, which is too low, and size 6, which is too high).

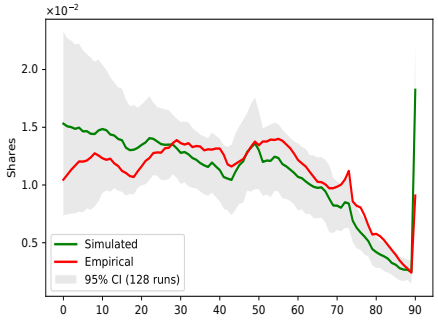

Fig. 2. Population age distribution.

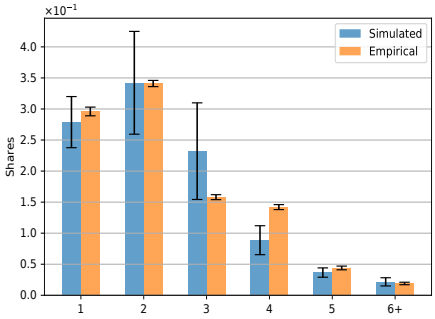

Fig. 3. Share of households by size.

From Figure 4 we can see that the model can reproduce very well the distribution of the ages of women giving birth in the year 2020.

Figure 5 shows the distribution of the differences between the ages of partners. We can see that although the model can replicate the mode of the real distribution quite well, it produces too many couples with men much older than women, so this is a section of the model which requires further attention in the future.

Figure 6 shows the dynamics of the minimum and maximum distance from the 'target' function, for the first 50 iterations.

5 Conclusions

In this paper we have demonstrated a 'generative' approach to the production of synthetic populations, which provides an alternative to current approaches. Synthetic populations are highly valuable for social simulation work, as they can produce realistic populations without the time and expense of collecting and

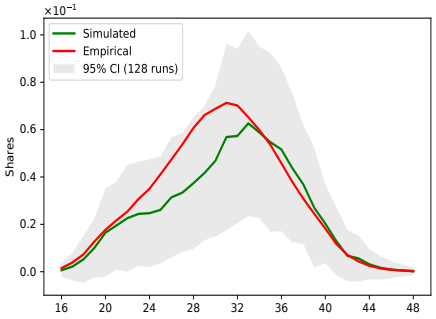

Fig. 4. Distribution of mothers' age when giving birth.

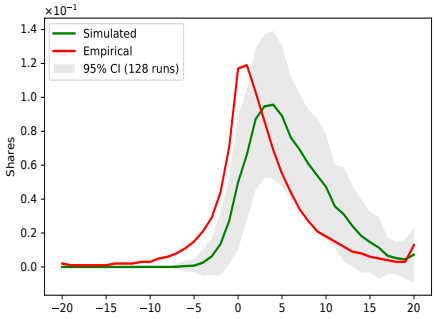

Fig. 5. Distribution of age difference between couples.

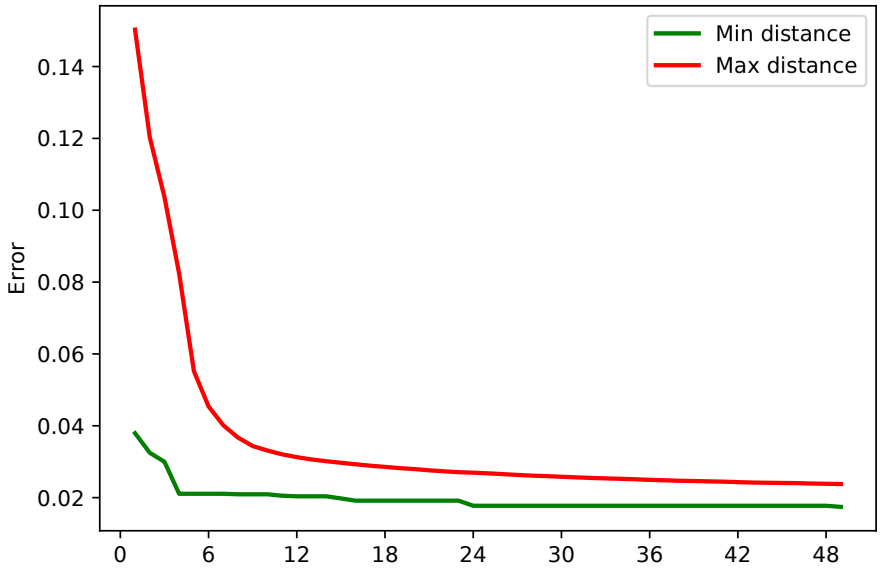

Fig. 6. Distance from 'target' function.

analysing both individual- and population-level data. We propose that the generative approach can achieve this goal while also taking into account individual-level behaviour on population dynamics. The generative approach is also less data-hungry than most current approaches to synthetic populations.

This approach also allows us to generate connections, such as kinship networks, between agents as they form partnerships, reproduce and move between households. These data can allow for the development of models examining the impact of familial connection on phenomena such as informal social care [7].

In future work, we plan to build on this generative approach to population synthesis by expanding the scope of agent behaviours, allowing the formation of realistic social networks as well. Such extensions will allow us to examine the potential effects of social network interventions on realistic synthetic populations.

Acknowledgements The authors are supported by the Medical Research Council (MC_UU_00022/1) and the Chief Scientist Office (SPHSU16)as well as the UK Prevention Research Partnership (MR/S037594/1). The authors thank Atiyah Elsheikh for help with the implementation of an earlier version of the model.

References

1. Human Mortality Database 2011. https://www.mortality.org/cgi-bin/hmd/
2. 1921 census of england and wales (1924)
3. Beaumont, M.A., Cornuet, J.M., Marin, J.M., Robert, C.P.: Adaptive approximate Bayesian computation. Biometrika **96**(4), 983–990 (Dec 2009). https://doi.org/10.1093/biomet/asp052, http://arxiv.org/abs/0805.2256, arXiv: 0805.2256
4. Bijak, J., Hilton, J., Silverman, E., Cao, V.D.: Reforging the wedding ring: Exploring a semi-artificial model of population for the united kingdom with gaussian process emulators. Demographic Research **29**, 729–766 (2013)
5. Chapuis, K., Taillandier, P., Drogoul, A.: Generation of synthetic populations in social simulations: A review of methods and practices. Journal of Artificial Societies and Social Simulation **25**(2) (2022)
6. Eurostat Statistics Database: Domain Population and Social Conditions. https://ec.europa.eu/eurostat/statistics-explained/index.php/Population_and_social_conditions (2011)
7. Gostoli, U., Silverman, E.: Modelling social care provision in an agent-based framework with kinship networks. Royal Society Open Science **6**(7) (2019). https://doi.org/10.1098/rsos.190029
8. Gostoli, U., Silverman, E.: Social and child care provision in kinship networks: An agent-based model. Plos one **15**(12), e0242779 (2020)
9. Noble, J., Silverman, E., Bijak, J., Rossiter, S., Evandrou, M., Bullock, S., Vlachantoni, A., Falkingham, J.: Linked lives: the utility of an agent-based approach to modeling partnership and household formation in the context of social care. In: Proceedings of the 2012 Winter Simulation Conference (WSC). pp. 1–12. IEEE (2012)
10. Office for National Statistics: Birth Statistics, Series FM1 (27). https://webarchive.nationalarchives.gov.uk/20160129135406/http://www.ons.gov.uk/ons/rel/vsob1/birth-statistics--england-and-wales--series-fm1-/no--27--1998/index.html (1998)
11. Silverman, E., Bijak, J., Noble, J., Cao, V., Hilton, J.: Semi-artificial models of populations: connecting demography with agent-based modelling. In: Advances in Computational Social Science: The Fourth World Congress. pp. 177–189. Springer (2014)