

# From agent-based models to statistical emulators<sup>1</sup>

Jakub Bijak<sup>a,✉</sup>, Jason Hilton<sup>a,b</sup>, Eric Silverman<sup>b</sup> and Viet Dung Cao<sup>b</sup>

<sup>a</sup> Social Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom

<sup>b</sup> Electronics and Computer Sciences, University of Southampton, Southampton, SO17 1BJ, UK

✉ Corresponding author: [j.bijak@soton.ac.uk](mailto:j.bijak@soton.ac.uk), tel. +44 23 8059 7486

## Abstract

In statistical demography information about population processes is inferred from empirical data. In contrast, agent-based approaches focus on aggregate outcomes of individual-level behavioural rules. Given the non-linearities and feedbacks present in agent-based settings, their direct statistical analysis is not always feasible. Hence, in order to bridge the gap between these two perspectives, we propose to utilise Gaussian process emulators, which enable studying the outcomes of rule-based models statistically. The suggested approach includes a sensitivity analysis, assessing the relative importance of different model parameters, and a simple calibration, aimed at selecting plausible parameter values. The discussion is illustrated by presenting a Semi-Artificial Model of Population, which augments an agent-based model of partnership formation with statistical data on natural population change in the United Kingdom. The resulting multi-state model of population dynamics is better aligned with selected aspects of the demographic reality than its underpinning agent-based component alone. The analysis also illuminates important trade-offs between different parameters and outputs considered.

## Key words

Agent-based models – Gaussian process emulator – Multi-state models – Population dynamics – Sensitivity Analysis – Wedding Ring

## 1. Introduction

Contemporary demographic micro-simulations are largely concerned with populations of *statistical individuals*, whose life courses can be inferred from empirical information (Courgeau 2012). In contrast, agent-based models study *simulated individuals*, for whom certain behavioural rules are assumed. We wish to bring these two approaches closer together by coupling the rule-based explanations driving an agent-based model with observed data. Our overarching research goal is to explain the emergence of macro-level demographic patterns as a result of reasonable micro level assumptions which are explored in the model. To that effect, we propose a method to analyse selected statistical properties of agent-based models, which utilises statistical emulators (Kennedy & O'Hagan 2001; Oakley & O'Hagan 2002).

---

<sup>1</sup> This working paper presents an abridged version of the article “Reforging the Wedding Ring: Exploring a Semi-Artificial Model of Population for the United Kingdom with Gaussian Process Emulators” by Jakub Bijak, Jason Hilton, Eric Silverman and Viet Dung Cao, forthcoming in *Demographic Research* in late 2013. Readers are kindly directed to <http://www.demographic-research.org> for the full version of the paper.

In this paper, we present a Semi-Artificial Model of Population, which aims to bridge demographic micro-simulation and agent-based traditions. We extend the ‘Wedding Ring’ agent-based model of marriage formation (Billari et al. 2007) to include empirical information on the natural population change for the United Kingdom, alongside the behavioural explanations that drive the observed trends in nuptiality. The mortality and fertility rates in this population are drawn from UK population data for 1951–2009 and forecasts until 2061 obtained from Lee-Carter models. We must note that our model is illustrative rather than attempting to be fully realistic with respect to all aspects of the underlying demographics. Subsequently, we utilise Gaussian process emulators – statistical models of the base model – to analyse the impact of selected parameters on two key simulation outputs: population size and share of agents with partners. We also attempt a sensitivity analysis, aiming to assess the relative importance of different inputs.

In general, agent-based models (ABMs) are a class of computational models designed to simulate the interactions of autonomous agents which may represent individuals or groups. The goal of such models is to assess the effects of these actions on the overall system, and to replicate incidences of complex macro-level phenomena by simulating the actions of simple, micro-level agents (Epstein and Axtell 1996, Gilbert and Tierna 2000, and Silverman and Bryden 2007). As a consequence, these simulations will generally include simple behavioural rules for autonomous agents, with the goal of observing how these low-level behaviours interact to produce higher-level complexity.

The existing examples of applying agent-based models in population-related applications are scarce, yet varied (see Billari and Prskawetz 2003 and Billari et al. 2006 for contemporaneous overviews). From the classical example of the residential segregation model of Schelling (1978), other applications include marriage formation (Todd, Billari, and Simão 2005; Billari et al. 2007), family-related decisions with respect to parenthood transitions (Aparicio Diaz et al. 2011), migration (Kniveton, Smith, and Wood, 2011; Willekens 2012), as well as overall household dynamics (Geard et al. 2013). In more general terms, Entwisle (2007) discussed the potential for harnessing the power of ABMs to understand the importance of locality and space in population models. With that in mind, the current paper attempts to narrow the gap between the behavioural assumptions of agent-based models, aimed mainly at explanations and guiding intuition about phenomena, and the higher predictive power of demographic micro-simulations.

This paper is structured into four sections. After this Introduction, in Section 2 we introduce a Semi-Artificial Model of Population based on a reimplementation of the ‘Wedding Ring’ model of Billari et al. (2007). The presentation of SAMP starts from a brief description of the Wedding Ring, followed by a discussion of empirical and projected demographic inputs, and emulator-based methods for analysing the uncertainty in complex computational models. Selected results of the simulations are shown in Section 3. Finally, Section 4 offers a brief discussion of the results, followed by suggestions for further work. The code for the current version of the model is available from the OpenABM archive (<http://www.openabm.org/model/3549/version/2>).

## **2. Semi-Artificial Model of Population**

### *2.1. Model Architecture*

Here we present a Semi-Artificial Model of Population (hereafter: SAMP), a simple multi-level and multi-state model of population dynamics, combining statistical and agent-based modelling approaches. The model follows the life courses of simulated individuals (agents), who are subject to empirical patterns of fertility and mortality. For illustration, we use time-varying data on age-specific birth and death rates for the United Kingdom (UK) for the period 1951–2010, and their further predictions yielded by Lee-Carter type models. The agent-based component is focused on the process of marriage, and thus also household formation. For this purpose, we use an adapted version of the ‘Wedding Ring’ model of Billari et al. (2007). Since SAMP is intended to be illustrative and exploratory, we have omitted other demographic processes such as migration for the sake of transparency. In terms of multi-level structure, SAMP operates at three levels: individuals (agents); households; and the whole population, with a direct bottom-up aggregation between these levels. Various technical aspects of the model are discussed in more detail in Sections 2.2 and 2.3; Section 2.4 describes a framework for analysing uncertainty in such a model, based on the concept of Gaussian process emulators.

### *2.2. Agent-Based Component: Marriage Formation on the Wedding Ring*

In order to illustrate the potential benefits and pitfalls of combining the demographic micro-simulation and agent-based approaches, we replicate and expand upon the ‘Wedding Ring’ agent-based model of marriage formation designed by Billari et al. (2007). The model attempts to explain age-at-marriage patterns seen in contemporary developed countries. In brief, the Wedding Ring represents the process of marriage formation as a consequence of social pressure. Pressure arises from contact between married- and non-married individuals within a given social network. This conceptual framework serves as a means of formalising some recent research in social influence and social learning, which has shown that these processes are highly relevant in individuals’ decisions to get married (e.g., Bernardi 2003, *idem*).

The Wedding Ring is so named due to the fact that in the original model agents live in a one-dimensional ring-shaped world (Billari et al. 2007). Each agent’s location is thus specified purely by a single coordinate (angle). The authors appear to have chosen the ring shape to avoid edge effects for agents located near a boundary. As the simulation progresses, each time-step in the simulated world is equivalent to one year. The agents are thus effectively situated in a cylindrical space, with one dimension of space and another of time (alternatively, age). Each agent’s network of ‘relevant others’ is then defined as a two-dimensional neighbourhood on that cylinder (*idem*). The size of the spatial interval for the agent’s network of relevant others is symmetric around their location, and varies according to the size of the initial population; in our reimplementations we have included a parameter for ‘spatial distance’, denoted as  $d$ , which determines the search space.

Within that neighbourhood, the proportion of married agents determines the ‘social pressure’ felt by an individual agent, which influences their decision to seek out a partner (prospective spouse). The overall level of social pressure and the agent’s age influence parameter determine the range in

which agents search for suitable partners. The age influence value is defined using a piecewise-linear function that varies with the age of the agent. As social pressure increases, agents widen their search range, and thus have a greater chance of successfully finding a partner (*idem*). However, the search is *mutual*: if one unmarried agent finds another within its acceptable range, marriage may only occur if the suitable partner has the searching agent within its acceptable range as well. Once married, agents may bear children; these children are then placed into the ring-world at a random spot in their parents' neighbourhood and begin life at age zero.

### 2.3. Demographic Components: Mortality and Fertility

To ensure that the starting structures within the simulation are reasonable, initial populations have been generated randomly, but with agent distributions by age, sex, and marital status corresponding to the breakdown observed in England and Wales in the 1951 census<sup>2</sup>. To the same end, fertility and mortality rates experienced by agents over the course of the simulation are based on empirical and projected data for the United Kingdom. For mortality, the first 59 years of the simulation are based on age-specific mortality rates for the UK for 1951–2009. The data are split by individual year and single years of age from birth to the open interval 110+, and are based on population exposure estimates and death counts from the Human Mortality Database (2011). To obtain logarithms of mortality rates  $\ln(m_{x,t})$  for the next half a century (2010–2061), predictions were produced using the well-known Lee and Carter (1992) model.

The fertility rates were obtained in a similar way to those for mortality. Age-specific rates from 1973–2009 for UK woman of childbearing age were obtained from the Eurostat database (Eurostat 2011), while earlier data for the period 1951–1972 were taken from the Office of National Statistics data for England and Wales<sup>3</sup>. A Lee-Carter model for logarithms of age-specific fertility rates,  $\ln(f_{x,t})$ , was again fitted to the data, but, in contrast to the mortality predictions, two bi-linear terms  $b_x k_t$  were required to best capture the trends in fertility. Formally, the forecasting equations for mortality and fertility have the form:

$$\begin{aligned} (1) \quad & \ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}, \\ (2) \quad & \ln(f_{x,t}) = a_x + b_{1x} k_{1t} + b_{2x} k_{2t} + \xi_{x,t}. \end{aligned}$$

where  $\varepsilon_{x,t}$  and  $\xi_{x,t}$  are normally distributed age-and-time-specific errors. For mortality,  $k_t$  was projected forwards to 2061 using a random walk with drift, while for fertility the ARIMA(1,1,1) model has been then selected for each time-variant parameter  $k_{*t}$  in the above equation using standard selection procedures, as implemented in the R package *forecast* (Hyndman 2011).

In order to ensure that fertility rates remain close to empirical values, we also utilise empirical and projected values for the proportion of births to married mothers by year and age of mother,

<sup>2</sup> Source: Table 26 of the census output: [*Population by*] *ages (quinary) by marital condition*, by courtesy of the Office for National Statistics (ONS), Titchfield (personal communication on 29/11/2011).

<sup>3</sup> Source: <http://www.ons.gov.uk/ons/rel/vsob1/birth-statistics--england-and-wales--series-fm1-/no--27--1998/birth-statistics-series-fm1.pdf> (retrieved on 15/11/2011).

denoted here as  $r_{x,t}$ . The rate of childbearing for a simulated married woman is then calculated by taking the product  $r_{x,t}f_{x,t}$  and multiplying it by the ratio of total to married women in that age group:

$$(3) \quad f_{x,t}^M = r_{x,t}f_{x,t} \times \frac{P_{x,t}}{P_{x,t}^M},$$

where superscripts  $M$  denote the population of married agents, and  $P_{x,t}$  refers to the total simulated female population at age  $x$  and time  $t$ . Similar calculations are made for unmarried women's fertility using the value  $(1 - r_{x,t})$  and the ratio of total to unmarried women. The data for  $r_{x,t}$  come from the Eurostat database (2011) for 1982–2010, and the remaining years are obtain by back- and forward-prediction for the periods 1951–1981 and 2011–2061 from another Lee-Carter model (4):

$$(4) \quad \text{logit}(r_{x,t}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t}.$$

The time varying element of this model  $\kappa_t$  is considered to be approximately proportional to the values of  $r_t$ , the proportion of the births to married women irrespective of age. Eurostat data for  $r_t$  prior to 1982 could therefore be transformed in order to continue the times series for  $\kappa_t$  by subtracting the mean value of  $r_t$  between 1982–2010 and multiplying by the ratio of the standard deviations  $\sigma_\kappa/\sigma_r$ . The `auto.arima` method of the `forecast` package (Hyndman 2011) was used to select a ARIMA(3,2,0) model for the backward, and ARIMA(1,1,2) model for the forward prediction.

#### 2.4. Framework for Analysing Uncertainty: From Monte Carlo to Gaussian Process Emulators

Due to the inherent non-linearities of relationships within agent-based models such as SAMP, and the presence of various feedback loops, the uncertainty of model outputs may not be easily (if at all) assessable analytically. Instead, a Monte Carlo simulation can be performed, where the model based on a pre-defined set of parameters is run many times, and the empirical realisations analysed in the form of statistical distributions. This solution is appropriate for assessing the code uncertainty, related to variation in the realisations of the model itself (cf. O'Hagan 2006). An example of applying the Monte Carlo approach to SAMP is presented in Section 3.3.

However, the code uncertainty is not everything. Considerable uncertainty is also associated with the unknown parameters driving the model assumptions. In principle, this issue could be also addressed using a Monte Carlo approach, although given the potentially high dimensionality of the problem, the number of required iterations, coupled with the computational complexity of the models and the time required to run them, this quickly becomes prohibitive (Kennedy and O'Hagan 2001). An alternative approach is to construct an *emulator* – effectively, a statistical model of the underlying complex computational model, reduced to the inputs and outputs of immediate interest – and to examine its properties (Oakley and O'Hagan 2002). In order for the uncertainty of the emulator to be described coherently and correctly, the preferred underlying statistical framework is the one of Bayesian inference (*idem*).

Amongst methods that have been proposed for building emulators, the one that is argued to be relatively simple, yet very flexible for applications to complex computational models, is based on Gaussian processes. A succinct introduction to Gaussian process emulators is provided below. In

general, the theoretical foundations have been laid out in the work of Kennedy and O'Hagan (2001), Oakley and O'Hagan (2002), Kennedy (2004), O'Hagan (2006), and on the website of the research community *Managing Uncertainty in Complex Models* (<http://www.mucm.ac.uk>).

Let  $f(\cdot)$  denote the base computational model of interest – in our case, SAMP. For the purpose of building an emulator, the focus is on a pre-defined vector of  $n$  inputs,  $\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^n$ , and a single output,  $y \in \mathbf{Y} \subset \mathbb{R}$ , such that  $y = f(\mathbf{x})$ .  $\mathbf{X}$  does not have to exhaust the whole parameter space of the underlying model, but rather should relate to those inputs which are considered important from the point of view of the output studied. Following Oakley and O'Hagan (2002: 771) and Kennedy (2004: 2), we define a Gaussian process emulator, conditionally on its parameters, as a multivariate Normal distribution for  $p$  realisations of  $f$ ,  $y_1 = f(\mathbf{x}_1)$ , ...,  $y_p = f(\mathbf{x}_p)$ , denoted jointly as  $\mathbf{f}$  (*idem*):

$$(5) \quad [\mathbf{f}(\cdot) | \boldsymbol{\theta}, \sigma, \mathbf{R}] \sim N[\mathbf{m}(\cdot), \sigma^2 \mathbf{c}(\cdot, \cdot)].$$

The mean of the process,  $\mathbf{m}$ , is modelled through a vector linear regression function of  $\mathbf{x}$ ,  $\mathbf{h}(\mathbf{x})$ , with coefficients  $\boldsymbol{\theta}$ , such that for every output  $f(\mathbf{x})$ ,  $m(\cdot) = \mathbf{h}(\cdot)^T \boldsymbol{\theta}$ . Further,  $\sigma^2$  is the joint variance parameter, and  $\mathbf{c}(\cdot, \cdot)$  denotes a correlation matrix, the elements of which are here assumed as  $c_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{R} (\mathbf{x}_i - \mathbf{x}_j)\}$ . The diagonal matrix  $\mathbf{R} = \text{diag}(r_1, \dots, r_n)$  is composed of *roughness* parameters  $\{r_1, \dots, r_n\}$ , which indicate how strongly the emulator responds to particular inputs (Kennedy and O'Hagan 2001: 432–433; O'Hagan 2006).

In order to estimate the parameters of the emulator, a set of simulation data  $\mathbf{D} = [f(\boldsymbol{\delta}_1), \dots, f(\boldsymbol{\delta}_N)]$  is required for a set of  $N$  experimental points  $\Delta = \{\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_N\}$ , where  $\Delta \subset \mathbf{X}$  (Kennedy 2004: 2). Making additional assumptions on the prior distributions of the parameters of the emulator (5), allows for applying full Bayesian inferential mechanism to obtain the posterior distribution of  $\mathbf{f}$  given  $\mathbf{D}$ . In order to incorporate the code uncertainty into the emulator, an additional variance term (referred to as a *nugget*) can be subsequently included in the estimation of the mean and the covariance matrix of the posterior distribution (*idem*).

The emulator, once built, can be used for a basic *uncertainty analysis*, which looks at how much uncertainty in the output is being induced by the set of inputs  $\mathbf{X}$  under study, treated here as random variables with some assumed probability distributions (e.g. Kennedy 2006). A *sensitivity analysis*, in turn, assesses the impact of particular inputs on the output based on the reductions of the output variance due to actually observing particular inputs (Oakley and O'Hagan 2004). Output variance reductions obtained by conditioning on true observed values of single inputs are referred to as *main effects*, and the additional reductions obtained for combinations of inputs – as *joint (interaction) effects*. An illustration is provided in Section 3.

### 3. Selected Results

#### 3.1. Model Implementation

SAMP was implemented in Repast Symphony v. 2.0, a Java-based environment especially designed for agent-based modelling and simulations. Each run of the model included 110 time steps, which in our case correspond to calendar years, starting with 1600 agents in the simulated year 1951.

The starting period was chosen in order to match the initial population structure with the 1951 UK census. The results presented in this section focus on the simulated year 2011, for which empirical verification of some aspects of the simulation was possible, and on the 2061 horizon.

The summary statistics are produced every simulated year, and refer to population structures and marriage hazards. The outputs also form a basis for building statistical emulators based on  $\Delta$  consisting of  $7^3 = 343$  model runs, corresponding to seven design values for each of the three parameters. For the purpose of the Monte Carlo analysis the model was run 500 times for a selected parameter set, to assess the uncertainty resulting from the inherent randomness of SAMP.

When re-implementing the Wedding Ring model and switching to empirical and projected vital rates, the original parameter settings of Billari et al. (2007) were no longer producing results that could be considered fully plausible in the light of the empirical evidence, as discussed further in Sections 4.2 and 4.3. Most importantly, this concerned the two parameters,  $\alpha$  and  $\beta$ , related to the social pressure function  $s(r)$ , defined in the original paper as (Billari et al. 2007: 66):

$$(6) \quad s(r) = \exp\{\beta(r - \alpha)\} / [1 + \exp\{\beta(r - \alpha)\}],$$

where  $r$  denotes the proportion of agents with partners within one's network of relevant others. The parameters were originally benchmarked as  $\alpha = 0.5$  and  $\beta = 7$  (*idem*).

### 3.2. Uncertainty and Sensitivity Analysis: Population Size and Marriage Rates

In this section we present two Gaussian process emulators for SAMP, with the aim of identifying areas of the parameter space that result in empirically plausible population dynamics and marriage processes. The focus here is on two features of the marriage formation mechanism: social pressure and spatial distance, both of which feed into the intensity of the partner search. In the first emulator we analyse the impact of the three underlying parameters:  $\alpha$  and  $\beta$  in equation (6), as well as the distance parameter  $d$ , on the uncertainty in the resultant overall share of population over 16 years who have entered into marriages at the simulation year 2011, denoted as  $p$ . Since  $p$  is bounded between 0 and 1, we have logit-transformed the output variable into  $u = \ln[p/(1-p)]$ .

In order to obtain the simulation data  $\mathbf{D}$  for building the emulator, we have run the model on a Cartesian product of pre-selected input values,  $\Delta = \boldsymbol{\alpha}' \times \boldsymbol{\beta}' \times \mathbf{d}'$ , where  $\boldsymbol{\alpha}' = [0, 0.333, 0.666, 1.0, 1.333, 1.666, 2.0]^T$ ,  $\boldsymbol{\beta}' = [\exp(-1), \exp(0), \exp(1), \exp(2), \exp(3), \exp(4), \exp(5)]^T$ , and  $\mathbf{d}' = [5, 10, 15, 20, 25, 30, 35]^T$ . Subsequently, a basic sensitivity analysis of the output  $u$  to the variation in the inputs has been attempted, with the aim to assess the importance of the three parameters.

The emulator was constructed, and the uncertainty and sensitivity analysis was performed in version 1.1 of the dedicated software GEM-SA (*Gaussian Emulation Machine for Sensitivity Analysis*), written by Marc Kennedy and Anthony O'Hagan (Kennedy 2004; O'Hagan 2006)<sup>4</sup>. The quality of the emulator construction was assessed by using a leave-one-out cross-validation method. The root

---

<sup>4</sup> The software is available from <http://ctcd.group.shef.ac.uk/gem.html> (retrieved on 15/07/2012).



mean-squared standardised error (RMSSE) reported by GEM-SA in this case was equal to 3.112, which indicates a fair emulator fit, in comparison with the ideal outcome of 1.

In GEM-SA, the distributions for the parameters of the Gaussian process (5) are *a priori* assumed to be vague, with  $p(\theta, \sigma^2) \propto \sigma^{-2}$  denoting limited information about the features of the process prior to observing the simulation data (inputs and outputs). The independent prior distributions for particular elements of the roughness matrix,  $r_i$ , are in turn exponential, with parameter  $\lambda = 0.01$  (Kennedy 2004: 2). For the purpose of the uncertainty and sensitivity analysis, the three input parameters are here assumed to be unknown and described by the following Normal distributions:  $\alpha \sim N(1.0, 0.25)$ ,  $\beta \sim N(2, 2.25)$  and  $d \sim N(20, 56.25)$ . The code uncertainty was handled by adding an additional error term (nugget) in calculating the posterior estimate of the covariance matrix.

The outcomes of the uncertainty analysis indicate a mean percentage of ever-married agents of  $p = 62.4\%$ , corresponding to the logit-transformed variable  $u = 0.507$ . The variance  $\sigma^2$  is estimated as 4.006, and the nugget variance as 0.092, indicating that, for  $u$ , the uncertainty in the three inputs is much more important than the code uncertainty resulting from the randomness in the model. The total output variance in  $u$  induced by input uncertainties is estimated as 2.215, of which the emulator contributed 0.0017. In terms of sensitivity, the most important variables proved to be the two parameters of the social pressure function,  $\alpha$  and  $\beta$ , accounting for 38.1% and 48.8% of the variability of the output respectively, and their interaction contributing further 9.7%. The spatial distance parameter  $d$  was responsible only for 1.7% of the variability of  $u$ .

A second emulator was constructed for population size in simulation year 2011 ( $N$ ) as an output, log-transformed as  $M = \ln(N)$ , with the same input values as before. The uncertainty analysis based on this emulator estimates the mean  $M$  as 7.57, corresponding to  $N = 1939$  agents. Proportionally, the observed mid-2011 population of 63.3 million people<sup>5</sup> corresponds therefore to 2013 agents. Knowing that cumulated net migration for the UK, since it began to be reported in 1964 until 2010, has amounted to ca. 2.1 million people<sup>6</sup>, a ball-park estimate of a corresponding closed population in mid-2011 can be put at about 61.2 million people, that is, 1945 agents.

This time  $\sigma^2$  is estimated as 0.968, and the nugget variance as 0.797, suggesting that the code itself is almost as important as the uncertainty in the underlying marriage formation process. The variance on  $M$  is estimated as 0.00074, with 0.00002 being accounted for by the emulator. Cross-validation indicate that the fit of this emulator is worse than before, with an RMSSE of 5.00, which is not surprising given the role of code uncertainty. The sensitivity analysis reveals the proportions of the variance accounted for by  $\alpha$  as 12.0% and  $\beta$  as 32.3%, with a further 12.5% accounted for by their interaction. The spatial distance parameter  $d$  is more important than for the previous emulator, accounting for 31.5% of total variance. Figure 1 illustrates contour maps of the predicted

---

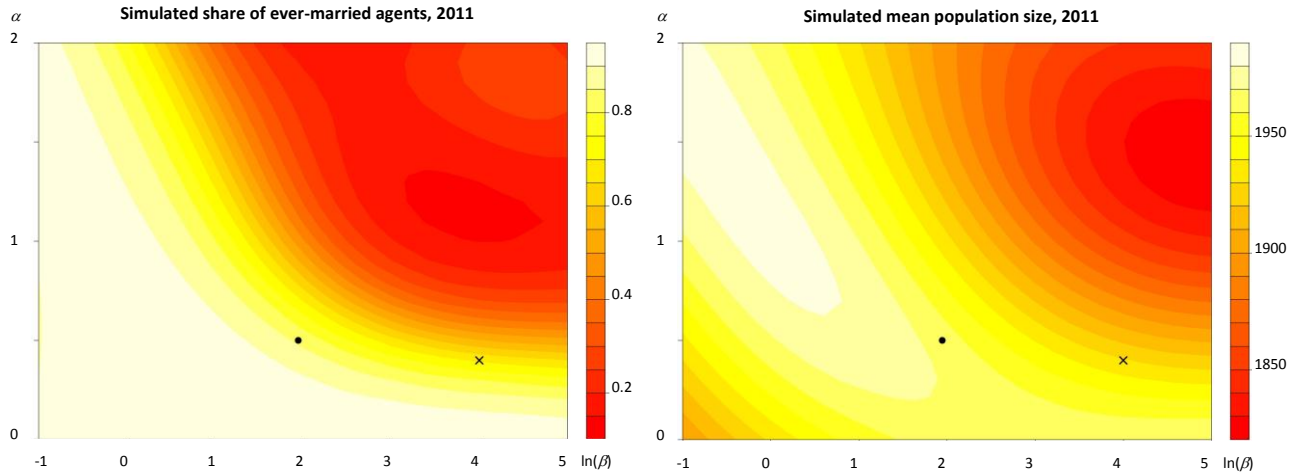
<sup>5</sup> <http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--england-and-wales--scotland-and-northern-ireland/mid-2011-and-mid-2012/stb---mid-2011---mid-2012-uk-population-estimates.html>

<sup>6</sup> Source: <http://www.ons.gov.uk/ons/interactive/theme-pages-1-1/index.html> (retrieved on 15/08/2013).



emulator means for the outputs  $p$  and  $N$ , plotted against the parameters  $\alpha$  and  $\beta$ , for  $d = 25$ .

Figure 1. Mean share of ever-married agents and mean population size by parameters  $\alpha$  and  $\beta$ , 2011



Notes: The share of ever-married people 16+ reported by the ONS for England and Wales in the 2011 census was 0.652. (Source: <http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-local-authorities-in-england-and-wales/stb-2011-census-key-statistics-for-england-and-wales.html#tab---Marital-status>; retrieved on 15/08/2013)

• - original parameter settings of Billari et al. (2007), × - default parameters used in this paper. Isolines shown for  $d = 25$ .

### 3.3. Illustration: A Scenario with Plausible Marriage Rates and Population Dynamics

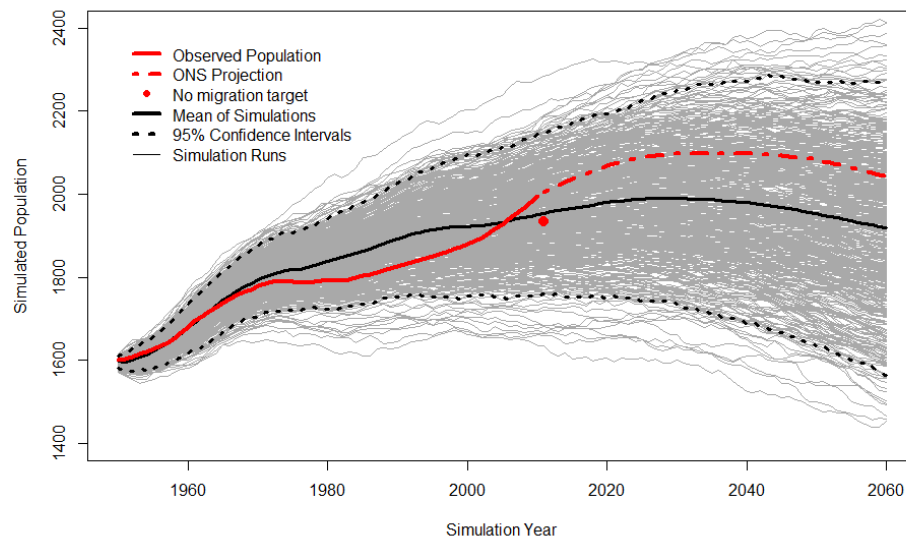
As indicated before, a comparison with the respective empirical data of the UK Office for National Statistics was conducted e.g. for the simulation year 2011. After all the changes were applied in our implementation of the original Wedding Ring model, the default parameter setting of Billari et al. (2007), with  $\alpha = 0.5$ ,  $\ln(\beta) = \ln(7)$ , would produce overall shares of ever-married agents over 80%; visibly higher than the empirical values (dots in Figure 1). In turn, parameters  $\alpha = 0.4$ ,  $\ln(\beta) = 4$  and  $d = 25$ , depicted in Figure 1 by crosses, generate plausible outputs. For these settings we present a scenario of Monte Carlo population dynamics for the overall population size.

Figure 2 indicates the dynamics of the simulated population over the whole period 1951–2061. Here, the mean values are shown alongside the 2.5-th and 97.5-th percentiles from the simulated set of 500 model runs. Additionally, observed population totals for 1951–2010 are presented, as well as those projected by the ONS for 2011–2061 in the 2010 round of National Population Projections, in the zero-migration variant. The ONS projections are benchmarked to higher values, as the simulation does not take into account the positive balance of past migration into the UK. Still, the trends in the projected and simulated trajectories for the future years are very similar.

The results of this illustrative simulation indicate that the generated population trajectories and structures are plausible from the point of view of selected empirical data and official projections. Differences between the simulated and observed trajectories are in large part due to the simplifications of SAMP, in particular the exclusion of international migration, which remains a very important component of the contemporary and projected population dynamics of the UK. Further

discrepancies might result from the very basic description of the modelled marriage processes, with no explicit modelling of cohabitation, no divorce or partnership dissolution, and no re-marriage. Still, the proportions of ever-married agents averaged across the whole simulation horizon are similar to patterns observed in 2011, but with slightly higher percentages married at younger ages, and slightly lower for age 50 or above.

Figure 2. Simulated population size (black), and empirical / projected UK comparisons (red)



Notes: Scaling applied. The dotted lines correspond to 95-percent confidence bounds of the simulated population size.

Source: ONS National Population Projections, zero migration <http://www.ons.gov.uk/ons/rel/npp/national-population-projections/2010-based-projections/rft-table-j1-1-natural-change-variant---uk-summary.xls> (retrieved on 15/08/2013).

#### 4. Conclusion

The main contribution of this paper to agent-based computational demography has been to demonstrate that using Gaussian process emulators is a convenient way of identifying plausible areas within the model parameter space, and of conducting a comprehensive analysis of uncertainty in complex computational models. In our example, the sensitivity analysis shows the key role for social pressure in the marriage formation process as implemented in the model, which proved more important than the spatial distance parameter driving the partner search. We have also shown that agent-based models enhanced with selected series of real demographic data offer improved predictive capabilities when compared to agent-based scenario generation alone. By using SAMP we have obtained the simulated population characteristics that match patterns observed in the UK demography with respect to population size and share of ever-married agents.

The resulting multi-state model of population dynamics is argued to have enhanced predictive capacity as compared to the original specification of the Wedding Ring, but there are some trade-offs between the outputs considered. The sensitivity analysis indicates a key role of social pressure in the modelled partnership formation process. We posit that the presented method allows for generating coherent, multi-level agent-based scenarios aligned with selected aspects of empirical

demographic reality. Emulators permit a statistical analysis of the model properties and help select plausible parameter values. Given non-linearities in agent-based models such as the Wedding Ring, and the presence of feedback loops, the uncertainty of the model may be impossible to assess directly with traditional statistical methods. The use of statistical emulators offers a way forward.

Natural substantive extensions of models such as SAMP include the spatial dimension, and in particular, migration (see Willekens 2012), as well as partnership dissolution and heterogeneous forms of partnerships. Fertility decisions themselves can be subject to agent-based modelling, as demonstrated by Aparicio Diaz et al. (2011), with parity distribution being an explicitly-targeted emergent outcome of the underlying behavioural rules. Other innovations, such as increasing the spatial dimensionality, which relax some of the constraints on the agents' behaviour, and add further complexity to the state space by including the health status of agents, are reported elsewhere (Silverman et al. 2013).

Further important methodological extensions of the model would include learning about the input values from the benchmarking of outputs to the observed population characteristics, for example with respect to various summary measures of population structures, in a comprehensive manner. Such statistical calibration techniques could be explored by using full Bayesian inference in conjunction with emulators. This would allow for describing and propagating uncertainty stemming from different sources, not only the model code, in a coherent way. In particular, this approach could be applied to calibrating the emulator results against the series of historical data, in a process known as *history matching*<sup>7</sup>. This is especially important given the dynamic nature of the system under study. Finally, more work should be done on the design of the experimental space,  $\Delta$ , for example by using Latin Hypercube samples or randomisation (O'Hagan 2006).

Overall, the proposed methods allow for generating coherent, multi-level agent-based scenarios, whose increased predictive capacity is due to a combination of incorporating the empirical basis for selected aspects of the demographic reality, and exploring the parameter space by using emulators. Emulators are also convenient for analysing statistical properties of such models. In this way, the agent-based models can be viewed through a statistical lens, reducing the gap between 'statistical' and 'simulated individuals' (cf. Courgeau 2012). We argue that these two approaches are complementary, rather than competitive.

## **5. Acknowledgements**

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/H021698/1 "Care Life Cycle", funded within the 'Complexity Science in the Real World' theme. The model reported here a demographic extension of the prototype from Silverman et al. (2013). The authors are very grateful to *Demographic Research* for helpful comments that helped us improve the earlier version of the model. All the remaining errors are exclusively ours.

---

<sup>7</sup> See: <http://mucm.aston.ac.uk/MUCM/MUCMToolkit/index.php?page=ThreadGenericHistoryMatching.html> (accessed on 20 May 2013).

## References

- Aparicio Diaz, B., Fent, T., Prskawetz, A., and Bernardi, L. (2011). Transition to Parenthood: The Role of Social Interaction and Endogenous Networks. *Demography*, 48(2): 559–579.
- Bernardi, L. (2003). Channels of social influence on reproduction, *Popul Res and Policy Rev*, 22(5/6): 527–555.
- Billari, F., Aparicio Diaz, B., Fent, T., and Prskawetz, A. (2007). The “Wedding–Ring”. An agent–based marriage model based on social interaction. *Demogr Res*, 17(3): 59–82.
- Billari, F., Fent, T., Prskawetz, A., and Scheffran, J. (eds.) (2006). *Agent–Based Computational Modelling: Applications in Demography, Social, Economic and Environmental Sciences*. Heidelberg: Physica Verlag.
- Billari, F., and Prskawetz, A. (eds.) (2003). *Agent–Based Computational Demography: Using Simulation to Improve Our Understanding of Demographic Behaviour*. Heidelberg: Physica Verlag.
- Courgeau, D. (2012). *Probability and Social Science. Methodological Relationships between the two Approaches*. Methodos Series, vol. 10. Dordrecht: Springer.
- Entwisle, B. (2007). Putting people into place. *Demography*, 44(4): 687–703.
- Epstein, J.M., and Axtell, R.L. (1996). *Growing Artificial Societies: Social Science From the Bottom Up*. Cambridge, MA: MIT.
- Eurostat (2011). Eurostat Statistics Database; Domain Population and Social Conditions [electronic resource]. Luxembourg: Eurostat. <http://epp.eurostat.ec.europa.eu> (retrieved on 27/10/2011).
- Geard, N., McCaw, J.M., Dorin, A., Korb, K.B. and McVernon, J. (2013) Synthetic Population Dynamics: A Model of Household Demography. *J Artificial Societies and Social Simulation*, 16(1): 8.
- Gilbert, N., and Tierna, P. (2000). How to build and use agent-based models in social science. *Mind and Soc*, 1(1): 57–72.
- Human Mortality Database (2011). Human Mortality Database. Berkeley: University of California; Rostock: Max Planck Institute for Demographic Research. <http://www.mortality.org> (retrieved on 26/07/2011).
- Hyndman, R. (2011). *forecast: Forecasting functions for time series*. Melbourne: Monash University. <http://robjhyndman.com/software/forecast> (retrieved on 06/09/2011).
- Kennedy, M. (2004). Description of the Gaussian process model used in GEM-SA. Sheffield: Centre for Terrestrial Carbon Dynamics. <http://ctcd.group.shef.ac.uk/gem.html> (retrieved on 23/05/2012).
- Kennedy, M.C., and O’Hagan, A. (2001). Bayesian Calibration of Computer Models. *J Royal Stat Soc B*, 63(3): 425–464.
- Kniveton, D., Smith, Ch. and Wood, S. (2011). Agent-based model simulations of future changes in migration flows for Burkina Faso. *Global Env Change*, 21(S1): S34–S40.
- Lee, R.D., and Carter, L.R. (1992). Modeling and Forecasting U.S. Mortality. *J Amer Stat Assoc*, 87(419): 659–671.
- Oakley, J., and O’Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4): 769–784.
- Oakley, J., and O’Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J Royal Stat Soc B*, 66(3): 751–769.
- O’Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, 91(10–11): 1290–1300.
- Schelling, T.C. (1978). *Micromotives and Macrobehavior*. Norton, New York, USA.
- Silverman, E., Bijak, J., Hilton, J., Cao, V.D., and Noble, J. (2013) When Demography Met Social Simulation: A Tale of Two Modelling Approaches. *J Artificial Societies and Social Simulation* (forthcoming).
- Silverman, E., and Bryden, J. (2007) From artificial societies to new social science theory. In: Almeida e Costa, F., Rocha, L.M., Costa, E., Harvey, I. and Coutinho, A. (eds.) *Advances in Artificial Life, 9th European Conference, ECAL 2007*. Berlin-Heidelberg: Springer: pp. 565–574.
- Todd, P.M., Billari, F.C. and Simão, J. (2005). Aggregate age-at-marriage patterns from individual mate-search heuristics. *Demography*, 42(3): 559–574.
- Willekens, F. (2012). Migration: A perspective from complexity science. Paper presented at the Migration workshop of the Complexity Science for the Real World network, Chilworth, UK, 16 February 2012.

■