Microsimulation in Demographic Research

Emilio Zagheni

Abstract

Demographic microsimulation is an individual-based and computationally intensive tool used by population scientists to model demographic processes, to gain insights on life course transitions and to make projections. Several microsimulators have been developed, since the 1960s, to address questions for which standard techniques or datasets cannot provide answers. Examples of applications include historical studies of demographic constraints on household formation, the impact of the HIV/AIDS epidemic on kinship resources for orphans and the elderly, the study of interaction and feedback mechanisms in demographic behavior. From the methodological point of view, calibration of stochastic microsimulators is an area of active research.

Keywords: Agent-based models; Calibration; Demographic methods; Demography; Heterogeneity; Historical demography; HIV/AIDS; Individual-based models; Kinship; Microsimulation; Monte Carlo methods; Social interaction.

Demographic microsimulation is a computationally-intensive tool used by population scientists to model demographic processes, to gain insights on life course transitions and to make projections. Two important features distinguish microsimulation from other models. First, the unit of analysis is the individual (hence "micro"). Second, the sequences of events that individuals experience over time are the result of stochastic experiments with predetermined probabilistic rules. Transitions between states are typically generated using computer algorithms and techniques also known as Monte Carlo methods (hence "simulation").

In a standard situation, each individual in the simulation is an observation in a rectangular data file, which contains records of demographic characteristics and other key variables of interest. The simulator takes as input population files and demographic rates, and updates the population accordingly. Each simulated person is subject to a set of rates, conditional on certain demographic characteristics such as age, sex, marital status, etc. For every predefined time interval, each individual faces the risk of a number of events including death, marriage, childbirth, and migration. The selection of the event and the waiting time until the event occurs are determined stochastically, often using a competing risk model. Some constraints may be included in the simulation program in order to restrict the range of potential events for particular subgroups of the population (e.g., to avoid social taboos such as incest, to allow for a minimum interval of time between births from the same mother, etc.). Each event for which the individual is at risk is often modeled as a piecewise exponential distribution. The waiting time until each event occurs is randomly generated according to the input demographic rates. The individual's next event is the one with the shortest waiting time.

Major modeling options

A number of microsimulators have been developed over the course of several decades to address different types of research questions. Microsimulation models can be classified into various categories based on some distinctive features.

Continuous vs discrete time

The scheduling of events for simulated individuals is an important feature of each microsimulator. The algorithms that determine the next event for each agent may treat time as a continuous or as a discrete variable.

In continuous-time microsimulations, the timing and sequence of events that individuals experience over the life course are the result of a competing risk model. For every event for which an individual is potentially at risk, the simulator randomly generates a waiting time until the event occurs. The most common waiting time distribution is the piecewise exponential. Other standard distributions include the Weibull and the Gompertz distributions (Willekens 2009). The parameters for the distributions of waiting times are chosen according to the input demographic rates for each subgroup in the population. The individual's next event is the one with the shortest waiting time. Every individual is always at risk of death; only women in reproductive age are at risk of giving birth; only married people are at risk of divorce, etc. The relative risk of each event happening next depends on input demographic rates, which typically vary by age, sex, marital status, parity, etc. Once the scheduled event occurs, a new set of waiting times is generated and a new event is scheduled accordingly.

In discrete-time simulations, time intervals are modeled, instead of the exact times before events occur. Each time period is considered separately from the others, and each event is executed only once within the time period. Different events may occur within the same interval of time. As time periods get shorter and shorter, results from discrete-time simulations get more and more similar to the ones of continuous models. Continuous models have some logical and practical advantages. In continuous microsimulations, it is the sequence of events and transitions that is modeled, and not state occupancies at different points in time. In discrete-time models, the duration of events can be modeled only approximately and multiple transitions within a period of time require assumptions about the ordering and the timing of events. Moreover discrete-time models cannot handle complex and interdependent sequences of events (Willekens 2009).

Closed vs open population

Microsimulations where individuals can enter the sample only through 'birth' from an existing woman, and can exit only through 'death', are referred to

as *closed* models. Whenever new people in the simulated population are generated other than through birth, then the simulation is defined *open*.

The distinction between closed and open populations mainly refers to the modeling technique used to match partners. In closed-population models, the choice of the spouse is restricted to individuals existing in the population. In open-population models, partners with suitable demographic characteristics are created in order to satisfy the demand for spouses at any given time.

Closed-population models require complex matching algorithms and fairly large population sizes in order for the marriage market to clear. Open-population models do not require sophisticated matching algorithms. However, the history of newly-generated individuals is typically missing. One of the main advantages of closed models is that they allow kinship ties to be tracked over time.

A standard example of closed-population kinship simulator is SOCSIM, a computer software developed at UC Berkeley in the 1970s. The computer program uses a two-stage process to pair eligible males and females from within the simulated population. When the next scheduled event for an individual is 'marriage', then the person is placed in a pool of eligible members to form an union. If a member of the opposite sex with appropriate demographic characteristics is available in the pool, then the two individuals are paired. Otherwise, the person stays in the pool until an appropriate mate 'picks' him or her, based on a random process with probabilities dependent on demographic characteristics of the two potential spouses.

In closed-population simulations, migrations are often modeled by introducing a group of people representative of 'the rest of world' that evolve independently of the population of interest. Migrants, or entire families of migrants, are drawn from the pool of people that represent the rest of the world.

Starting population

The initial population for a simulation can come from a cross-sectional sample of the entire population, like the group of respondents to a sampling survey, or can be a synthetic population generated in a way that is consistent with the expected age structure and other relevant characteristics of the population under study.

When the initial starting population comes from a cross-sectional dataset representative of the entire population, then individual life histories are sim-

ulated for the future. This approach is relatively simple and guarantees that the starting population has demographic characteristics consistent with the ones of the population under study. However, no information on the history of individuals before the starting point of the simulation is then available. Therefore, most kinship ties cannot be reconstructed. In addition, when the initial sample size comes from a survey, it may be relatively small and thus there could be high levels of stochasticity in the outputs of the simulation.

Often, the starting population is produced using a synthetic procedure. A population of a given size, composed of unrelated individuals, is randomly generated. Then the population is projected forward for a long period of time (say at least 100 years) using demographic rates that generate a population consistent with the population under study at the beginning of the simulation (in terms of age structure, age at marriage, age at childbirth, divorce rates, etc.). In the synthetic population, kinship ties have been tracked and can thus be projected for the future.

Top-down vs bottom-up

Microsimulation models simulate demographic events for individuals in a way that is consistent with macro-demographic rates. One of the goals of microsimulation is to evaluate the consequences of demographic change for a number of quantities of interest. The methodology can be seen as a top-down approach. An alternative line of individual-based simulations, often referred to as agent-based models, pursues a bottom-up approach. Agent-based models simulate agents with built-in behavioral rules of action and interaction with other individuals and their environments. The main goal of agent-based models is to study the emergence of global patterns from simple behavioral rules of autonomous agents.

Standard microsimulation models tend to emphasize the modeling of macro-to-micro interactions, and are used to evaluate the impact of policy changes, long-term demographic trends, and demographic shocks, on various quantities of interest. Agent-based models focus on the micro-to-macro direction and are often used to test theories and to evaluate the emergence of complex phenomena.

Although standard microsimulation and agent-based models are built on different assumptions and used for different goals, in practice the boundaries between the two approaches are not clearly defined. Agent-based models often include some macro-demographic rates that serve as macro-controls for

the dynamics of the simulated population. Microsimulation, on the other hand, include important behavioral rules that regulate, for instance, preferences and interaction in the marriage market.

Calibration

In principle, perfect knowledge of demographic rates should lead to an unbiased reconstruction of population dynamics and kinship networks through demographic microsimulation. The only uncertainty associated to the simulated kinship structure would be related to the stochasticity of the microsimulation. In practice, knowledge of vital rates is far from being perfect. Kinship reconstruction and forecasting demand a level of detail for demographic rates that is often missing in available data sets. For instance, transition rates from one marital status to another one are usually not readily available and estimates may not be very accurate. Fertility rates are usually not broken down by marital status or parity, especially in the developing world. In most cases, demographic rates that are used as input to the microsimulation need to be estimated from various data sources with different sampling errors. Even when reliable data sources exist to compute demographic rates broken down by the categories of interest, the heterogeneity of the population's rates within the tabulated categories constrains the accuracy of the microsimulation (Wachter et al. 1997).

Traditionally, the problem of calibrating microsimulations has been addressed using ad hoc tuning. Input rates are adjusted on a trial and error basis in order for the output of the simulation to match key summary demographic measures obtained from a population census or sample surveys. The validity of the microsimulation has been tested by comparing kinship forecasts generated in the past with external standards provided by surveys with detailed information on numbers and ages of kin in the United States (Wachter et al. 1997).

In the past, the development of methods to calibrate microsimulations has been mainly constrained by the limitation of computer power. Traditional methods have relied heavily on minimizing the number of simulation runs. That was done, for instance, using expert judgement for adjusting the input demographic rates in a consistent and appropriate way. With increasing computer power, there has been more and more interest in the development of methods to calibrate simulation models. The Bayesian melding method

(e.g., Raftery et al. 1995; Poole and Raftery 2000), in particular, has proved useful to formalize the process of calibration and statistical inference. It is a Bayesian approach since it relies on the Bayesian machinery of combining prior distributions with likelihoods to obtain posterior distributions. It has been named 'melding' because it "provides a way of combining different kinds of information (qualitative or quantitative, fragmentary or extensive, based on expert knowledge or on data) about different quantities, as long as the quantities to which they relate can be linked using a deterministic model" (Poole and Raftery 2000). Sevcikova et al. (2007) and Zagheni (2010) extended the approach to stochastic simulations.

When using the Bayesian melding method, the researcher first has to express the available information about inputs and outputs in terms of probability distributions. For instance, this can be done by providing a prior distribution on the inputs. Then the researcher has to specify a conditional probability distribution of the data given the outputs. This yields a likelihood for the outputs, and, implicitly, produces a likelihood for the inputs. The combination of the prior on the inputs and the likelihood, using the Bayes' rule, gives the posterior distribution. In order to obtain the posterior distribution for the quantities of interest, Sevcikova et al. (2007) suggested a computational approach that is based on the sampling importance resampling (SIR) algorithm of Rubin (1987, 1988). They applied the method to urban simulations. A similar approach has been used for the calibration of demographic microsimulations (Zagheni 2010).

Some applications

The idea behind the development of microsimulation methods as a research tool dates back to the late 1950s and early 1960s (Orcutt 1957; Orcutt et al. 1961). Since then, a large number of microsimulators have been proposed and used to address demographic questions (Morand et al. 2010). It is beyond the purposes of this article to provide an exhaustive list of microsimulators and applications. Some illustrative examples that give a flavor of the scope of applications of microsimulation to demographic research will be presented in this section.

The study of kinship structure is one of the most successful applications of microsimulation models in demographic research. Two of the most widely used microsimulators are SOCSIM and CAMSIM. SOCSIM originates from a

collaboration between Peter Laslett, Eugene Hammel and Kenneth Wachter in the early 1970s. CAMSIM was developed in the late 1970s and early 1980s by Peter Laslett, James Smith and Jim Oeppen (Zhao 2006). Both simulators have been designed to study kinship networks. However, they use quite different algorithms, since SOCSIM is a closed model, whereas CAMSIM is an open-population simulator.

SOCSIM has been originally developed for historical studies of demographic constraints on household formation. In particular, it was used to test the hypothesis that social norms, and not unfavorable demographic conditions, were the cause of the low proportion of stem family households in pre-industrial England (Wachter et al. 1978). Other historical analyses include the assessment of the 1698 Slavonian census, using SOCSIM (Hammel and Wachter 1996a and 1996b) and the evaluation of potential biases in genealogical data, using CAMSIM (Zhao 2001). A second line of research that has largely benefited from the use of microsimulators is the study of change of kinship availability over time. SOCSIM has been used to project kinship resources (including step-kin) for the elderly in the US (Wachter 1997) as well as long-term changes in family and kinship networks in Britain (Murphy 2011). Zhao (2006) offers a comparative review of applications of SOCISM and CAMSIM.

For countries with a generalized HIV/AIDS epidemic, microsimulation has been used to assess kinship resources for the elderly in Thailand (Wachter et al. 2002) and for orphans in Zimbabwe (Zagheni 2011). Demographic microsimulations that include modules for the transmission and progression of HIV/AIDS have been used to model population dynamics of polygynous populations in sub-Saharan Africa (Clark 2001). In the area of indirect estimation, microsimulation has proven useful to evaluate bias for methods developed to estimate mortality from sibling survival data (Masquelier 2012).

Microsimulation has been the dominant research tool to analyze the consequences of demographic change for family and kinship networks. Agent-based models have recently emerged as a complementary tool to study marriage formation, choice of partners, and the role of behavior, interaction and feedback mechanisms in demographic research (Billari and Prskawetz 2003). For instance, it has been found that relatively simple mate search rules that adjust on the basis of sequential encounters with potential partners may generate regularities in the distribution of the age at first marriage (Todd et al. 2005; Billari et al. 2007).

Strengths and limitations

Microsimulation models have been developed to address questions for which standard tools could not provide an answer. In that sense, microsimulation amplifies the power of scientific imagination of researchers, it allows population scientists to go beyond some of the simplifying assumptions made in standard macro demographic models.

Microsimulation is the most appropriate tool when population heterogeneity and interaction matter, and when the number of variables and the number of attributes that these variables can take is very large (Van Imhoff and Post 1998; Spielauer 2011). Microsimulation can accommodate interactions between individuals and between variables much better than macro models with a large state space. More specifically, "microsimulation is particularly appropriate if the results of the process are complex but the driving forces of the process are simple" (Van Imhoff and Post 1998). Anytime processes are complex at the macro level, but better understood at the micro level, or when individual histories are relevant, then microsimulation/individual-based models are the preferred choice.

Microsimulation is a very attractive tool both to gain insights on population dynamics and for policy-relevant analyses. However, there are two important limitations that need to be taken into account. First, microsimulations typically require high quality and very specific types of data that are often not available. Moreover, even if detailed demographic data are available, the calibration process may prove difficult. Second, failing to model correlations between demographic events in models of kinship may lead to less variation in the frequency of kin of any particular type than would occur in the real population (Ruggles 1993). In other words, models tend to underestimate the fraction of people with many kin and the fraction of people with few kin. In most situations the extent of the underestimation is very small and, for practical purposes, could be ignored. However, there may be cases where the size of the underestimation is not negligible and thus, in those situations, modeling of correlations is advisable.

References

[1] Billari, F. and Prskawetz, A., Eds. 2003.*Agent-based Computational Demography*. Contributions to Economics, Physica-Verlag.

[2] Billari, F., Prskawetz A., Diaz B.A., and Fent T. 2007. The "Wedding Ring": An Agent-Based Marriage Model Based on Social Interaction. *Demographic Research* 17: 59.

[3] Clark S.J. 2001. An investigation into the impact of HIV on population dynamics in Africa. PhD dissertation, University of Pennsylvania

[4] Hammel, E. and Wachter, K. 1996a. Evaluating the Slavonian Census of 1698, Part I: Structure and meaning. *European Journal of Population* 12:145-166.

[5] Hammel, E. and Wachter, K. 1996b. Evaluating the Slavonian Census of 1698, Part II: A Microsimulation Test and Extension of the Evidence. *European Journal of Population* 12:295-326.

[6] Masquelier B. 2013. Adult Mortality from Sibling Survival Data: A Reappraisal of Selection Biases. *Demography* 50(1):207-228.

[7] Morand E., Toulemon L., Pennec S., Baggio R., and Billari F. 2010. Demographic Modeling: The State of the Art. *SustainCity Working Paper 2.1a*, Ined, Paris

[8] Murphy M. 2011. Long-Term Effects of the Demographic Transition on Family and Kinship Networks in Britain. *Population and Development Review* 37(1):55-80

[9] Orcutt G.H., 1957. A New Type of Socio-economic System. *Review of Economics and Statistics* 39:116-123.

[10] Orcutt G.H., Greenberger M., Korbel J., and Rivlin A. 1961. *Microanalysis of socioeconomic systems : A simulation study* Harper & Row, New York.

[11] Poole D. and A.E. Raftery. 2000. Inference for Deterministic Simulation Models: the Bayesian Melding Approach. *Journal of the American Statistical Association* 95(452):1244-1255.

[12] Raftery, A.E., G.H. Givens and J.E. Zeh. 1995. Inference from a Deterministic Population Dynamics Model for Bowhead Whales. *Journal of the American Statistical Association* 90(430):402-416.

[13] Rubin, D. 1987. Comment on "The Calculation of Posterior Distributions by Data Augmentation", by M. Tanner and W.H. Wang. *Journal of the American Statistical Association* 82:543-546.

[14] Rubin, D. 1988. Using the SIR Algorithm to Simulate Posterior Distributions. *Bayesian Statistics* 3 (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.) 395-402. Clarendon Press, Oxford, U.K.

[15] Ruggles S. 1993. Confessions of a Microsimulator. *Historical Methods* 26:161-169.

[16] Sevcikova H., A.E. Raftery and P.A. Waddell. 2007. Assessing Uncertainty in Urban Simulations Using Bayesian Melding. *Transportation Research*. Part B 41:652-669.

[17] Spielauer M. 2011. What is Social Science Microsimulation? *Social Science Computer Review* 29(1):9-20.

[18] Todd P.M., Billari F., and Simao J. 2005. Aggregate Age-at-Marriage Patterns from Individual Mate-Search Heuristics. *Demography* 42(3):559-574

[19] Van Imhoff E. and W. Post. 1998. "Microsimulation Methods for Population Projection". *Population: An English Selection* 10(1): 97-138.

[20] Wachter, K.W., E.A. Hammel and P. Laslett. 1978. *Statistical Studies of Historical Social Structure*. New York, Academic Press.

[21] Wachter K.W. 1997. Kinship resources for the elderly. *Philosophical transactions of the royal society of London - Series B: Biological sciences* 352(29).

[22] Wachter K.W., D. Blackwell and E.A. Hammel. 1997. Testing the Validity of Kinship Microsimulation. *Journal of Mathematical and Computer Modeling* 26:89-104.

[23] Wachter K.W., J.E. Knodel and M. VanLandingham. 2002. AIDS and the elderly of Thailand. *Demography* 39(1):25-41.

[24] Willekens, F. 2009. Continuous-time Microsimulation in Longitudinal Analysis. *New Frontiers in microsimulation modeling* 353-376.

[25] Zagheni E. 2010. The Impact of the HIV/AIDS Epidemic on Orphanhood Probabilities and Kinship Structure in Zimbabwe. PhD Dissertation. University of California, Berkeley.

[26] Zagheni E. 2011. The Impact of the HIV/AIDS Epidemic on Kinship Resources for Orphans in Zimbabwe. *Population and Development Review* 37(4):761-783.

[27] Zhao, Z. 2001. Chinese Genealogies as a Source for Demographic Research: A further Assessment of their Reliabilities and Biases. *Population Studies* 55:181-193

[28] Zhao Z., 2006 Computer Microsimulation and Historical Study of Social Structure: A Comparative Review of SOCSIM and CAMSIM. *Revista de Demografia Historica* XXIV, II, pp. 59-88