9. Microsimulation

Nik Lomax

1. INTRODUCTION

Social systems are made up of individuals, for example: people, households, cars or firms, and the distribution, behaviour and changes exhibited by these individuals add up to trends seen at the macro level. Using models which deal with individuals, rather than groups or aggregated data, allows for the analysis of different behaviours or outcomes which takes in to account the heterogeneity of these individuals. This allows researchers to assess distributional patterns and other impacts in more detail than is possible in macro models.

Much contemporary microsimulation modelling stems from the work of Orcutt (1957) who was frustrated with the limitation of macroeconomic models for assessing the impacts of policy simulations and recognised that macro approaches largely ignore any distributional effects. He argued that theoretical models of socio-economic systems are best applied at the individual level because it is individuals who make decisions within the system. Today, microsimulation is applied in a wide range of different contexts by researchers from across many different fields of study.

1.1 When to Use Microsimulation

When deciding if microsimulation is an appropriate approach to take, Spielauer (2011) provides guidance for researchers under three headings. First is population heterogeneity, whereby, if there are differences in the attributes of the population, and there are too many combinations of attributes to split the population into a manageable number of groups, it makes more sense to update a dataset of individuals than an increasing number of tables representing counts of cross-tabulations. Second, if there is a problem with aggregate analysis because the behaviours being modelled are better understood at the individual level, then microsimulation is appropriate. Spielauer (2011, p. 15) points to the example of the tax system and argues that '*to calculate total tax revenues, we need to know the composition of the population by income (progressive taxes), family characteristics (dependent children and spouses), and all other characteristics that affect the calculation of the individual tax liability.*' Third, in the case of dynamic microsimulation (defined and explained later), if individual histories matter, that is an individual's modelled outcome is dependent on their past, then microsimulation is the approach to use.

1.2 Broad Appeal of Microsimulation Methods

Microsimulation is a (relatively) computationally hungry method which requires good data inputs. In their introduction to the documentation for the modelling framework, Modgen, used by Statistics Canada to produce population projections, Bélanger et al. (2017) note that, thanks to improvements in computing power and better availability of microdata, microsimulation is

an approach which has become increasingly popular over the last few decades. Ballas et al. (2005) cite a range of studies from the 1990s which used some form of microsimulation to provide insight into society, although, as they note, the authors of these studies did not necessarily realise they were undertaking microsimulation and so there was a degree of 'reinventing the wheel' as the methods evolved. Contemporary applications of microsimulation make use of a wide range of survey microdata and utilise computational resources to leverage the benefits that microsimulation methods offer. Today, microsimulation is used in a wide range of applications, for example, to produce population projections (Lomax and Smith, 2017), estimate future elderly healthcare demand (Clark et al., 2017), model tax benefit systems (Immervoll et al., 2006), estimate household expenditure patterns (James et al., 2019) and to model health behaviours at a small geographical scale (Smith et al., 2011).

1.3 Chapter Outline

This chapter provides an overview of a number of techniques, broadly grouped together under the heading of microsimulation, which are used to estimate the distribution of phenomena (both spatial and between population groups) and evolution of individual units over time to produce projections or to test out policy interventions. Discussion is broken down in to three key areas of research: (1) the generation of synthetic, attribute rich, individual level population data which are often used to assess distributional differences or as an input to other models; (2) static models which assess short term changes to a system, for example, the immediate distributional effect of policy change; and (3) dynamic models, where time is incorporated to assess the longer-term impact of changes on the individuals as they transition through the system of interest. These microsimulation models become spatial when the individuals have a geographical identifier and results can be compared across different areas. In this context the methods are often termed *spatial microsimulation.*

2. CREATING SYNTHETIC POPULATIONS: SPATIAL MICROSIMULATION

In many different contexts, populations of individual units which have the required attributes needed for analysis (for example, age, sex, ethnicity, health and economic status, income and expenditure behaviour) are needed to model social systems. As such, there is a large and growing literature focused on the creation of synthetic, individual level populations which draw information from different data sources; for example, a complete population base from a census can be supplemented by the more attribute rich information about individuals from a sample dataset such as a survey. This process itself is often referred to as microsimulation – the simulation of the individuals within the synthetic population. Adding geographical attributes to these units (and often using geography as an additional constraint in the generation process) is referred to as spatial microsimulation.

A comprehensive review of different application areas which utilise spatial microsimulation including health, transport, agriculture and land use planning is provided by O'Donoghue et al. (2014). Examples include the combination of census and household panel survey data to assess health inequalities (Ballas et al., 2006), combining census data with household expenditure data to create local area expenditure profiles (James et al., 2019) and combining census data with

longitudinal survey data to analyse commuting patterns (Lovelace et al., 2014). Consideration of the broad range of application areas where spatial microsimulation has been applied, as well as broader discussion about the utility of spatial microdata, can be found in Hermes and Poulsen (2012). These microdata are often used as the input to dynamic microsimulation models, discussed later, or as input to other types of model, for example, Agent-Based Models (Crooks and Heppenstall, 2012). For a detailed discussion of Agent-Based Models, see the *Simulating Geographical Systems Using Cellular Automata and Agent-Based Models* chapter of this handbook.

2.1 A Synthetic Microdata Example

A brief example is outlined in Figure 9.1. Adapted from James et al. (2019), the objective of the work was to estimate local authority level food and drink expenditure for Great Britain by combining detailed survey microdata with local area population totals. The first box in Figure 9.1 outlines the microdata fields available in the Living Cost and Food Survey (LCFS), a detailed survey of approximately 12 000 people each year in the United Kingdom. Sociodemographic variables are available for each individual in the LCFS, along with their detailed expenditure information across 106 different categories. There is, however, not enough geographical detail to estimate this expenditure at a local authority scale from the LCFS, so spatial variation in expenditure patterns cannot be investigated using the survey alone.

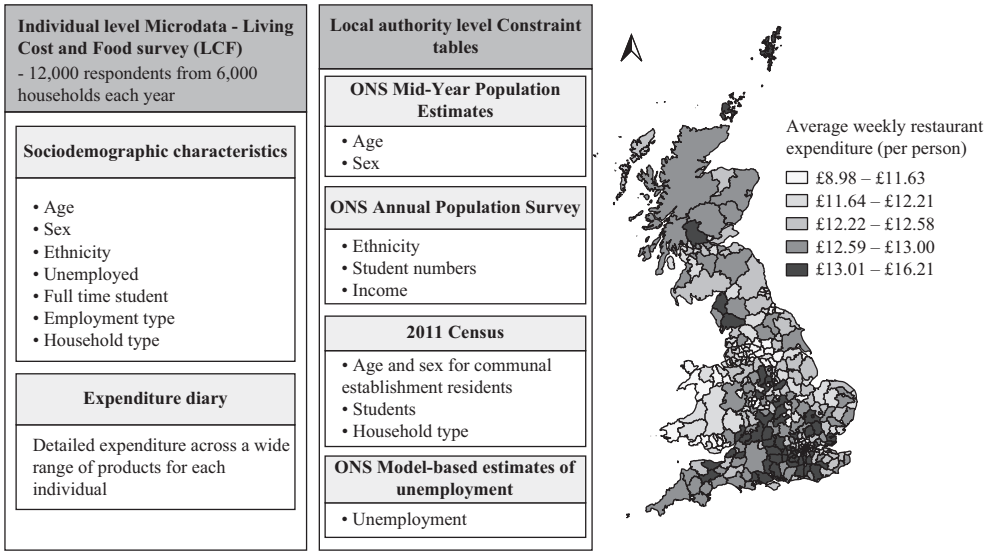

Figure 9.1 Example microdata fields, area constraints and output from a spatial microsimulation model. The figure is produced using information from James et al. (2019) and the map is created using the accompanying open-source dataset. The map output contains estimated weekly per person expenditure on eating out in restaurants in 2016

The second box in Figure 9.1 provides detail of the local authority level constraints available from a range of sources, including official mid-year estimates and census data. Note that these constraints map to the sociodemographic variables available in the LCFS, so can be used to join one dataset to the other. The survey microdata data are fitted (inflated) to the local authority level constraints iteratively so that individuals are replicated based on the distribution of their sociodemographic attributes. Iterative Proportional Fitting (IPF) is used by James et al. (2019) to undertake this estimation. IPF and other methods are discussed in more detail below. Aggregating fields in the new synthetic dataset provides estimates for the whole local authority for all expenditure categories available in the LCFS. The map in Figure 9.1 provides an example of average weekly per person expenditure in restaurants in 2016.

2.2 Methods for Creating Microdata

There are a range of methods which can be used to create these synthetic populations of individuals. Lovelace and Dumont (2016) make the distinction between *deterministic* methods, which will consistently produce the same results, and *stochastic* (probabilistic) methods, where the use of random numbers results in some variance in the outputs. Deterministic methods are described largely as reweighting algorithms, while stochastic methods randomly assign individuals from microdata and calculate goodness of fit statistics against benchmarks to see if further iteration is required. Harland et al. (2012) single out three methods which are widely used in the creation of synthetic data under the broad heading of spatial microsimulation: deterministic reweighting, and the two probabilistic methods of conditional probabilities and simulated annealing. They use univariate census tables (for example, age, gender, ethnic group, highest qualification) to create synthetic microdata at three spatial scales in the United Kingdom to test the performance of the three methods against know joint-distribution tables (for example, age by sex, sex by ethnic group, age by highest qualification). Comparing the total absolute error between the synthetic microdata and the observed joint-distribution tables, they conclude that simulated annealing was the best performing algorithm.

A further distinction is made by Hermes and Poulsen (2012) between methods referred to as *synthetic reconstruction* and those known as *reweighting*. Synthetic reconstruction includes data matching and data fusion techniques, where microdata are created from the distributions observed in aggregate tables, and further attributes are added sequentially. Reweighting uses a microdata sample and reweights this using more complete information (for example, from a census). As Hermes and Poulsen (2012) point out, the focus of research shifted from synthetic reconstruction to reweighting methods in the 1990s. For this reason, they limit their discussion to reweighting techniques, and generally the approaches discussed in this chapter fall into this category. However, IPF, discussed later, can be used as a synthetic reconstruction method, see Birkin and Clarke (1988) for a full description. A probabilistic approach to synthetic reconstruction using Monte Carlo sampling is outlined in Harland et al. (2012), which builds on the methods outlined by Birkin and Clarke (1988). Their conditional probability model sequentially builds the attributes of a population based on the distribution of the underlying dataset (rather than being reweighted from a sample).

Table 9.1 outlines some of the most common methods used to create synthetic microdata, along with references where the methods are used. A selection of these are discussed further in the following sections.

Table 9.1 Methods used for creating synthetic populations

Method	Deterministic/ Probabilistic	Studies where method is used for microsimulation
Iterative Proportional Fitting (IPF)	Deterministic	Lomax and Norman (2016), James et al. (2019)
Generalised Regression and Weighting of sample survey results (GREGWT)	Deterministic	Tanton et al. (2011), Miranti et al. (2011)
Conditional Probabilities	Probabilistic	Harland et al. (2012)
Simulated Annealing	Probabilistic	Hynes et al. (2009), Kavroudakis et al. (2013), Harland et al. (2012), Ma et al. (2014)
Hill Climbing	Probabilistic	Williamson et al. (1998)

2.3 Deterministic Methods

Given consistent inputs, deterministic algorithms will always produce the same output. The term deterministic reweighting appears widely in the microsimulation literature (Smith et al., 2011) and of these methods, IPF and GREGWT are the most commonly used for the construction of synthetic microdata (Hermes and Poulsen, 2012).

2.3.1 Iterative proportional fitting (IPF)

One of the most well-established methods for creating synthetic populations from different data sources is IPF. In a demographic context the approach was first used by Deming and Stephan (1940) who, using 1940 U.S. Census data, found that cross-tabulated tables by geographical area were limited to samples of the population. They took these cross-tabulated tables as the starting *seed* distribution and applied IPF to produce data which matched the total population in a given area. It is worth noting that the approach is applied in a wide range of disciplines but is known by different names, for example, the RAS method in economics, Cross-Fratar or Furness in transport and raking in computer science. There are a wide range of implementations of IPF, but for a detailed guide on undertaking spatial microsimulation in the statistical language R see Lovelace and Dumont (2016). IPF is explained in detail, with worked examples by Lomax and Norman (2016) but in brief, the method requires:

(1) A seed table, which is the cross-tabulation of characteristics, often derived from a sample or survey dataset. For example, this might be microdata from a census or data about the health or consumption behaviour of individuals from a survey; and
(2) marginal population totals, often for a given geographical area if the implementation is a spatial microsimulation. These need to match those available in the microdata or survey, so for example, a count of the number of people by age and sex in a given area.

The seed table is then iteratively adjusted (that is, scaled) to match the population totals. It is adjusted to fit the row totals, then the column totals, and the process is repeated until there is either a perfect fit or the algorithm reaches a pre-defined stopping criterion or number of

iterations. In many incarnations, IPF produces fractional weights in order to reach convergence, so it is usual to end up with non-integers of individuals who meet the required constraints. If the aim is to have a count of 'whole people,' this can be dealt with by post calculation rounding, but an adapted version of IPF described by Lovelace and Ballas (2013) as *integerised IPF*, produces integer weights rather than the non-integers produced by regular IPF. The downside to this integerised method is that it reduces model fit. This approach is used by Ballas et al. (2005) in the SimBritain model.

Further examples of where IPF is used to create microdata include Beckman et al. (1996), who combine the 1990 U.S. Census data with the Public Use Microdata Sample (PUMS) to estimate households within census tracts which have certain demographic characteristics. They find that the joint distribution of size of household and number of vehicles created using IPF do not differ substantially from observed values. In the United Kingdom, Simpson and Tranmer (2005) use IPF to estimate a cross-tabulation of car ownership and tenure type using 1991 census data. As discussed in the example earlier, James et al. (2019) use IPF to create estimates of food and drink expenditure for Great Britain. IPF is used to create an individual level population as an input to the dynamic SMILE (Simulation Model for the Irish Local Economy) model (Ballas et al., 2005).

2.3.2 Generalised regression (GREGWT)

Generalised Regression and Weighting of sample survey results (GREGWT) is a method which has been widely applied to produce small area synthetic data in Australia. The geographical focus is largely because the primary implementation of GREGWT for microsimulation comes from a series of SAS statistical software macros written by the Australian Bureau of Statistics. Rahman et al. (2010) describe the algorithm in detail. Tanton et al. (2011) outline how the method is used to reweight Australian sample surveys, which are representative at the Territory level, to small area units using a worked example. Following this example, the data requirements for using GREGWT in a spatial microsimulation context are:

(1) An initial weight, such as a survey weight, scaled to a small area population. As Tanton et al. (2011) specify, these initial weights need to be reasonable but not perfect as they will be rescaled using the GREGWT algorithm.
(2) Benchmark tables (for example, income by rent paid).
(3) Benchmark variables (for example, income).
(4) Grouped benchmark classes within the tables (for example, income bands).

The aim is to minimise the difference between the estimated count and the sample count observed in the microdata. As such, GREGWT is an iterative approach whereby on each iteration new weights are calculated until they either match the constraint benchmarks, or a stopping criterion is reached. Once the reweighting process has finished, each individual in the survey dataset will have a weight for each small area where benchmarks were supplied (Tanton et al., 2014).

GREGWT is used by Miranti et al. (2011) to estimate poverty at small area scale in Australia, by combining the Survey of Income and Housing with benchmarks from census data. Vidyattama et al. (2015) estimate small area participation in cultural activities across Australia, using a synthetic database constructed from survey and census data, constrained to small area census benchmarks. Muñoz (2016) use GREGWT to estimate heat demand in

Hamburg, Germany. The microdata come from a 1 per cent sample of the German population and are benchmarked to small area demographics and dwelling stock information and the output is a synthetic building stock.

2.4 Probabilistic Methods

In much of the microsimulation literature, probabilistic methods are used to solve what are referred to as Combinatorial Optimisation (CO) problems (for example, Tanton 2014). The general CO approach starts with a random allocation from microdata, it then iteratively replaces these individuals and compares with the constraints. There are a number of algorithms which perform CO, and Williamson et al. (1998) investigate the utility of the three most widely used approaches for creating spatial microdata: hill climbing algorithms; simulated annealing; and genetic algorithms. They draw on the census Sample of Anonymised Records (SAR) to create population counts for small areas within the English city of Leeds. Of the three algorithms tested, they found that simulated annealing was the best performing. Harland et al. (2012), in a comparison of simulated annealing against deterministic reweighting and Monte Carlo sampling, came to similar conclusions: that when it comes to generating accurate and consistent populations, the simulated annealing algorithm provided the best performance compared with real data.

2.4.1 Simulated annealing

Openshaw and Rao (1995) discuss simulated annealing alongside other methods for reengineering the census geographies used to analyse and display 1991 census data in the United Kingdom. While not explicitly used for microsimulation by Openshaw and Rao (1995), they suggest that, at the time of writing, geographical applications of the method were limited. However, the uptake of simulated annealing for creating spatial microdata since then has been fairly strong. A clear description of the simulated annealing algorithm is offered by Ballas et al. (2007), who use it to build a small area population of the English city of Leeds, drawing on British Household Panel Survey microdata and census constraints. The method in brief:

(1) For a given small area there are census tables which contain constraint totals, for example, tenure type might come from one table and car ownership from another table. An initial random seed of individuals can be drawn from the survey data which satisfies the totals reported in the census tables.
(2) The distribution of the random allocation of households is checked against the census constraints (across all constraint tables) and the Total Absolute Error (TAE) is reported.
(3) An individual from the originally selected seed is swapped at random for a different individual in the survey dataset. The error is recomputed and if there is an improvement then the swap is accepted.

An important feature of simulated annealing is that it allows for backward steps as well as forward steps in order to avoid getting stuck in a local optimal solution where a global optimal solution might be available. Harland et al. (2012) describe how simulated annealing has a threshold at which steps which lead to deterioration in model fit are retained. This threshold is termed the annealing factor, synonymous with the cooling of metal: '*as the algorithm proceeds, the (temperature) thresholds are reduced and so backward steps become progressively*

more unlikely, so that ultimately only climbing moves are permitted towards an optimised outcome' (Harland et al., 2012, p. 2). This backward step is something not possible with other CO approaches such as hill climbing algorithms, which have a tendency to become trapped suboptimal solutions (Williamson et al., 1998).

Hynes et al. (2009) use simulated annealing to scale a national farm survey to the Irish Census of Agriculture to assess methane emissions from Irish farms. Kavroudakis et al. (2013) use simulated annealing to combine British Household Panel Survey with census data to assess social inequalities in higher educational attainment for a region of England. In Ma et al. (2014), survey data from an activity diary are combined with census data to simulate travel distance and mode choices in Beijing in order to assess CO_2 emissions. Ma et al. (2014) use a Geographical User Interface to implement simulated annealing, designed by Harland (2013) and called the Flexible Modelling Framework.

2.5 Availability of Microdata

Synthetic microdata are usually created specifically for a given project and are not routinely shared as an output, however, there are some examples where these datasets have been made available to the wider research community. See, for example, the food, drink and tobacco expenditure data which accompany the paper by James et al. (2019) and enhanced census outputs created by Morris and Clark (2018) and deposited for research use in the Consumer Data Research Centre (CDRC, 2019).

2.6 Validation of Synthetic Datasets

The validation of spatial microdata is essential, especially if they are to be used as an input to other models or used to undertake scenario analysis. The process of validation can be split in to two distinct tasks: internal validation and external validation.

2.6.1 Internal validation

Internal validation involves checking the attribute distributions of the synthetic dataset against the original input datasets. Timmins and Edwards (2016) provide a detailed assessment of the wide range of methods used to validate synthetic datasets, which is used as the starting point for Table 9.2. They advocate the Bland-Altman method for validation (see Table 9.2) as an addition to the toolbox used by microsimulation modellers for validation and calibration. Further discussion of various internal validation methods can be found in Rahman et al (2010).

2.6.2 External validation

External validation involves comparison of synthetic outputs against other external datasets. This is more difficult than internal validation primarily because the necessity to produce synthetic data suggests that a directly comparable dataset does not exist. This means that often external validation of microsimulation models is not carried out. However, there are ways of undertaking this validation. Two methods are adopted by James et al. (2019). First, their estimates of household expenditure are compared with an independent dataset, the Index of Multiple Deprivation. They found that expenditure is higher in less deprived areas, which is consistent with expectations. Second, they aggregate their small area data up to regional level and compare with a publicly available expenditure dataset, finding similar trends exist for

Table 9.2 Summary of validation methods used to assess synthetic population outputs

Validation method	Description/application	Studies where method is used for microsimulation
Correlation	The relationship between the aggregate counts produced by simulation and observed in constraints.	James et al. (2019), Lovelace et al. (2014)
Standard Error about Identity (SEI)	Square root of the average of the sum of the squared deviations where intercept=0, slope=1 in comparison ~~of~~ to predictions against observed data.	Ballas et al. (2007)
Total Absolute Error (TAE)	The sum of error terms in a constraint category (observed minus simulated)	Voas and Williamson (2001)
Standardised Absolute Error (SAE)	TAE divided by the expected total population count in that category.	Smith et al. (2009)
Z Score	Difference in the relative size of the category between actual and simulated population.	Williamson et al. (1998), Tanton et al. (2014),
E5	In spatial microsimulation, a count of the number of areas where error is greater than 5%.	Lovelace et al. (2015)
Bland-Altman Method	A plot of the absolute difference between estimated and observed data and the mean of the observed plus estimated data.	Timmins and Edwards (2016)

Source: adapted from Timmins and Edwards (2016), Table 1.

each expenditure category and year of simulation. A similar two stage approach was taken by Edwards et al. (2011). For their estimates of obesity at small area scale they first aggregated the data to a coarser geographical level at which comparable obesity data were available, and second compared their obesity estimates with cancer data which were known to be correlated with obesity.

3. ASSESSING INTERVENTIONS IN THE SHORT TERM: STATIC MICROSIMULATION

Static microsimulation is a term used in studies which address the immediate, short-term impacts of changes or policies. Assessing these distributions is often called static microsimulation, because it does not take in to account the dynamic processes and transitions needed to assess longer term impacts. The input to these static models is usually a synthetic, attribute rich population as discussed above.

Cassells et al. (2006, p. 8) refer to these static microsimulations as '*morning after*' models, which look at the distributional effect of changes in the system, for example, assessing what the impact might be of changing the rules for calculating tax or benefits. In this example, these impacts would not, however, take into account changes in behaviour or changes to labour

supply. The value of static models is that they are able to apply any changes or scenarios to individuals within a base dataset, rather than to population groups or geographical areas, to assess the distributional effects of these scenarios. Examples of static microsimulation include the forecasting of domestic water demand in the United Kingdom (Williamson et al., 2002), the analysis of housing taxation in Italy (Pellegrino et al., 2011) and the assessment of income tax policy in Germany (Flory and Stöwhase, 2012).

One well-established static microsimulation model is EUROMOD (Sutherland and Figari, 2013), a tax-benefit model for the European Union. EUROMOD provides assessment of the distributional impacts of policy changes, but without taking into account time or behavioural adjustment. The strength of EUROMOD is that it enables cross country comparisons, focusing on harmonising input data and providing flexibility over more complex dynamic model structure.

4. SIMULATING THE FUTURE: DYNAMIC MICROSIMULATION

Dynamic models are used to simulate the trajectories of individuals over time. The early dynamic model described by Orcutt et al. (1961) evolved to become DYNASIM (The Dynamic Simulation of Income Model) (Orcutt et al., 1976), a microsimulation for forecasting and policy analysis. From this point, a wide range of dynamic models were developed for a multitude of purposes. A comprehensive review of dynamic models is provided in Li and O'Donoghue (2013) who reference 61 dynamic models developed between 1977 and 2013 used for social economic analyses.

Dynamic models take as input a starting population of individuals, who each have a number of characteristics. Often these are created using the synthetic data generation methods outlined earlier, or are taken directly from survey datasets. This population is then aged on and individuals have the opportunity to transition between different states in the model (for example, from employed to unemployed or from healthy to ill). Altering the parameters and assumptions of these transition rates or equations provides an opportunity to experiment with different outcomes, providing a framework for testing policy change. These transitions are often estimated from observed longitudinal data where they are available.

A good primer for how to develop the demographic core of a dynamic model is offered by Ballas et al. (2005). Their model, SMILE (Simulation Model for the Irish Local Economy), takes as key input a static population from a spatial microsimulation (created using IPF) with the attributes of age, sex, marital status, employment and spatial location (District Electoral Division). For these individuals, the demographic components of births, deaths and internal migration (note that international migration is excluded) are calculated and used to project the population forward. Mortality is dealt with using survival probability (as a function of age, sex and location) calculated from vital statistics and each individual, based on their demographic attributes, is assigned a survival probability. In the next step, a random number is drawn (constrained between 0 and 1) for each person. So for an individual with an 80 per cent chance of survival as reported in the vital statistics, if the random number is between 0 and 0.8 they would survive. If that random number were 0.81 or over they would die in the simulation. Fertility is similarly dealt with as a function of age, marital status and location, where an individual female had a probability of giving birth. Internal migration is also simulated based on probability, with people moved to a new location based on population size (but the authors note

this could be improved using spatial interaction models). Small area populations are projected and compared with official estimates from a more aggregate spatial scale and accuracy of the projection assessed using absolute percentage error (APE).

These dynamic models can be considerably more complex, taking in to account a wide range of outcomes and attributes, although the basic objective can be thought of in a similar way to the SMILE model, with transitions for individuals between various states introduced as matrices or calculated as survival functions. For example, on top of the base demographic characteristics (age, sex, ethnicity, marital status), a wide range of health transitions, including heart disease, stroke, cancer and diabetes, are estimated for individuals aged 50 and over in the United States using the Future Elderly Model (FEM) (Goldman and Orszag, 2014). The FEM is used to produce scenarios of health care interventions and the financial cost of modelled trends in health and transitions are estimated from the longitudinal Health and Retirement Study.

4.1 Dynamic Spatial Microsimulation

Many existing dynamic models do not explicitly deal with the spatial. This is a point emphasised by Rephann and Holm (2004), who point out that space is useful in dynamic models as both an input, where geographical variables are driving forces of change and an output, where results can be interpreted for spatial regions. However, there are a growing number of dynamic spatial microsimulation models, a subset of which are reviewed by Tanton (2014). These include the SMILE model discussed above, SVERIGE for Sweden and MOSES for the United Kingdom (Birkin et al., 2017), which all produce outputs which can be analysed spatially. Of particular note is SVERIGE (Rephann and Holm, 2004), which makes use of attribute rich, longitudinal microdata covering the whole Swedish population, rather than synthetic microdata as used in other models. In addition to the core demographic components, the model is able to simulate a range of other attributes including education, employment, earnings and marital status. The treatment of internal migration is particularly well developed, involving a three-stage approach whereby an individual decides to move, chooses a labour market and then chooses a 100 m square grid to relocate to.

4.2 Treatment of Time and Sequencing of Events

Time can be dealt with differently in dynamic models, namely as discrete time steps or continuously. In a discrete-time model, transitions are estimated between two time points, t and t+1 with no consideration for the timing of events between these periods. So a person might transition from married to divorced in the period but the exact timing of that transition is not taken in to account. In continuous-time models, the event history is simulated for each individual, including the timing of these events, calculated using survival functions. Galler (1997) reviews the pros and cons of discrete vs continuous-time models. The complexities of applying a continuous-time model are substantial and to make them operational requires specific and restrictive assumptions about the conditional independence of processes and other unobserved heterogeneity. On the other hand, general model specifications and estimation procedures have been developed for discrete-time frameworks, although there is need for careful consideration of the ordering of events (for example, a person transitioning from single to married might then have a higher probability of conceiving a child). Galler (1997) concludes that discrete-time models based on short time periods are preferable.

Three options for the sequencing of events which are estimated for each individual in the simulation are offered by Rephann and Holm (2004). Time driven models implement modules (for example, mortality, migration, fertility) in a predefined order. Event driven models implement modules conditional on the outcome of other modules. Random order models implement modules in random sequence.

5. CONCLUSION

Microsimulation is a broad term which has been broken down in to three different but interlinked strands in this chapter. The first is the creation of attribute rich, synthetic datasets of individuals which are constructed from sources such as surveys which contain detailed attributes at the individual level, scaled to population totals available from censuses or other large coverage datasets. Where geography is available in the constraining dataset then individuals can be given a geographic identifier (or allocated to small areas) and these models are widely known as spatial microsimulation models.

The second strand is static microsimulation. These models typically take as an input the attribute rich individual population and look at the distribution of certain characteristics across different population groups or geographical areas. These models are used to test the immediate impact of policy intervention or other changes and to assess how these might play out across those groups and geographies. The third strand is dynamic microsimulation, where time is introduced into the models. Individuals in the microdata are aged through time and can transition between different states. Altering these transitions will produce different model outcomes, and so these models can be used to undertake scenario and what-if analysis over a long-time frame.

Applications of microsimulation are broad, ranging from population projection, to the assessment of tax and benefit policy and estimation of the impact of healthcare improvements. Methods also vary, both in the creation of the microdata and the development of dynamic models. However, all microsimulation approaches are fundamentally used because of a need to capture the behaviour of individuals who make decisions within the system and because of the need to assess the distributional effects of modelled outcomes, either between different population groups or different geographical areas.

There are limitations to all of the microsimulation approaches discussed in this chapter. Models are complex to implement properly and, especially in the case of dynamic models, require expert understanding and careful calibration. Some of the most well established models, for example, the FEM (Goldman and Orszag, 2014) have taken years to build and continue to be developed and maintained, demonstrating that it takes significant time and investment to build complex models. Microsimulation models also require good data inputs, both in terms of individual level attributes and constraints needed to build synthetic population datasets and longitudinal data needed to estimate transition probabilities for dynamic models. These data are not always readily available or accessible in the formats required. While computational power is no longer such a barrier to building these models, they still require substantial resource, especially in the case of dynamic models. Validation is also a key issue with microsimulation. While approaches used to validate synthetic microdata are discussed in this chapter, fundamentally, these synthetic data are produced because no comparable data exist. Dynamic models which provide forecasts or scenario-based outputs are even harder

to validate given their focus on future, unknown outcomes. The best way to tackle this is to produce a range of scenarios and compare results, provided the researcher is confident that the model inputs are sensible.

This chapter has discussed a snapshot of the wide range of work which has been undertaken under each of the three strands of microsimulation. Some guidance on the approaches which are available and decisions to be made when implementing microsimulation has been offered. A good starting point for the reader interested in knowing more about the creation of synthetic data using spatial microsimulation approaches is O'Donoghue et al. (2014). The recommended next step for the reader interested in dynamic microsimulation models is Li and O'Donoghue (2013).

REFERENCES

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., and Rossiter, D. (2005). SimBritain: A spatial microsimulation approach to population dynamics. *Population, Space and Place*, *11*(1), 13–34. https://doi.org/10.1002/psp.351

Ballas, D., Clarke, G., Dorling, D., Rigby, J., and Wheeler, B. (2006). Using geographical information systems and spatial microsimulation for the analysis of health inequalities. *Health Informatics Journal*, *12*(1), 65–79.

Ballas, D., Clarke, G., Dorling, D., and Rossiter, D. (2007). Using SimBritain to Model the Geographical Impact of National Government Policies. *Geographical Analysis*, *39*(1), 44–77. https://doi.org/10.1111/j.1538-4632.2006.00695.x

Ballas, D., Clarke, G. P., and Wiemers, E. (2005). Building a dynamic spatial microsimulation model for Ireland. *Population, Space and Place*, *11*(3), 157–172. https://doi.org/10.1002/psp.359

Ballas, D., Kingston, R., Stillwell, J., and Jin, J. (2007). Building a spatial microsimulation-based planning support system for local policy making. *Environment and Planning A*, *39*(10), 2482–2499.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6), 415–429.

Bélanger, A., Sabourin, P., and Bélanger, A. (2017). *Microsimulation and population dynamics*. Springer.

Birkin, M., and Clarke, M. (1988). SYNTHESIS—a synthetic spatial information system for urban and regional analysis: Methods and examples. *Environment and Planning A*, *20*(12), 1645–1671.

Birkin, M., Wu, B., and Rees, P. (2017). Moses: Dynamic spatial microsimulation with demographic interactions. In *New frontiers in microsimulation modelling* (pp. 53–77). Routledge.

Cassells, R., Harding, A., and Kelly, S. (2006). *Problems and prospects for dynamic microsimulation: A review and lessons for APPSIM*. NATSEM, University of Canberra.

CDRC. (2019). Synthetic population. Retrieved 29 October 2020, from https://data.cdrc.ac.uk/dataset/synthetic-population

Clark, S., Birkin, M., Heppenstall, A., and Rees, P. (2017). Using 2011 Census data to estimate future elderly health care demand. *The Routledge Handbook of Census Resources, Methods and Applications: Unlocking the UK 2011 Census*.

Crooks, A. T., and Heppenstall, A. J. (2012). Introduction to agent-based modelling. In *Agent-based models of geographical systems* (pp. 85–105). Springer.

Deming, W. E., and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427–444.

Edwards, K. L., Clarke, G. P., Thomas, J., and Forman, D. (2011). Internal and external validation of spatial microsimulation models: Small area estimates of adult obesity. *Applied Spatial Analysis and Policy*, *4*(4), 281–300.

Flory, J., and Stöwhase, S. (2012). MIKMOD-ESt: A static microsimulation model of personal income taxation in Germany. *International Journal of Microsimulation*, *5*(2), 66–73.

Galler, H. P. (1997). *Discrete-time and continuous-time approaches to dynamic microsimulation reconsidered*. National Centre for Social and Economic Modelling.

Goldman, D. P., and Orszag, P. R. (2014). The growing gap in life expectancy: Using the Future Elderly Model to estimate implications for Social Security and Medicare. *American Economic Review*, *104*(5), 230–233.

Harland, K. (2013). *Microsimulation Model user guide (flexible modelling framework)*. Working Paper, NCRM https://eprints.ncrm.ac.uk/id/eprint/3177

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, *15*(1).

Hermes, K., and Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, *36*(4), 281–290.

Hynes, S., Morrissey, K., O'Donoghue, C., and Clarke, G. (2009). A spatial micro-simulation analysis of methane emissions from Irish agriculture. *Ecological Complexity*, *6*(2), 135–146.
Immervoll, H., Levy, H., Lietz, C., Mantovani, D., O'Donoghue, C., Sutherland, H., and Verbist, G. (2006). Household incomes and redistribution in the European Union: Quantifying the equalizing properties of taxes and benefits. In *The distributional effects of government spending and taxation* (pp. 135–165). Springer.
James, W. H., Lomax, N., and Birkin, M. (2019). Local level estimates of food, drink and tobacco expenditure for Great Britain. *Scientific Data*, *6*(1), 1–14.
Kavroudakis, D., Ballas, D., and Birkin, M. (2013). Using spatial microsimulation to model social and spatial inequalities in educational attainment. *Applied Spatial Analysis and Policy*, *6*(1), 1–23.
Li, J., and O'Donoghue, C. (2013). A survey of dynamic microsimulation models: Uses, model structure and methodology. *International Journal of Microsimulation*, *6*(2), 3–55.
Lomax, N., and Norman, P. (2016). Estimating population attribute values in a table: "get me started in" iterative proportional fitting. *The Professional Geographer*, *68*(3), 451–461.
Lomax, N., and Smith, A. (2017). Microsimulation for demography. *Australian Population Studies*, *1*(1), 73–85.
Lovelace, R., and Ballas, D. (2013). 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41*, 1–11. https://doi.org/10.1016/j.compenvurbsys.2013.03.004
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. https://doi.org/10.1016/j.jtrangeo.2013.07.008
Lovelace, R., Birkin, M., Ballas, D., and van Leeuwen, E. (2015). Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique. *Journal of Artificial Societies and Social Simulation*, *18*(2), 21.
Lovelace, R., and Dumont, M. (2016). *Spatial microsimulation with R*. Chapman and Hall/CRC.
Ma, J., Heppenstall, A., Harland, K., and Mitchell, G. (2014). Synthesising carbon emission for mega-cities: A static spatial microsimulation of transport CO_2 from urban travel in Beijing. *Computers, Environment and Urban Systems*, *45*, 78–88.
Miranti, R., McNamara, J., Tanton, R., and Harding, A. (2011). Poverty at the local level: National and small area poverty estimates by family type for Australia in 2006. *Applied Spatial Analysis and Policy*, *4*(3), 145–171.
Morris, M., and Clark, S. (2018). Big data application of spatial microsimulation for neighborhoods in England and Wales. In *Big data for regional science* (pp. 243–256). Abingdon: Routledge.
Muñoz H, M. E. (2016). *A spatial microsimulation model for the estimation of heat demand in Hamburg* (pp. 39–46). CORP–Competence Center of Urban and Regional Planning.
O'Donoghue, C., Morrissey, K., and Lennon, J. (2014). Spatial microsimulation modelling: A review of applications and methodological choices. *International Journal of Microsimulation, 7*(1), 26−75.
Openshaw, S., and Rao, L. (1995). Algorithms for reengineering 1991 Census geography. *Environment and Planning A*, *27*(3), 425–446.
Orcutt, G. H. (1957). A new type of socio-economic system. *Review of Economics and Statistics*, *39*(2), 116–123.
Orcutt, G. H., Greenberger, M., Korbel, J., and Rivlin, A. (1961). *Microanalysis of socioeconomic systems*: *A simulation study*. New York: Harper and Row.
Orcutt, Guy H., Caldwell, S., Wertheimer, R., and Franklin, S. (1976). *Policy exploration through microanalytic simulation*. The Urban Institute.
Pellegrino, S., Piacenza, M., and Turati, G. (2011). Developing a static microsimulation model for the analysis of housing taxation in Italy. *International Journal of Microsimulation*, *4*(2), 73–85.
Rahman, A., Harding, A., Tanton, R., and Liu, S. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation*, *3*(2), 3–22.
Rephann, T. J., and Holm, E. (2004). Economic-demographic effects of immigration: Results from a dynamic spatial microsimulation model. *International Regional Science Review*, *27*(4), 379–410.
Simpson, L., and Tranmer, M. (2005). Combining sample and census data in small area estimates: Iterative proportional fitting with standard software. *The Professional Geographer*, *57*(2), 222–234.
Smith, D., Clarke, G., and Harland, K. (2009). Improving the Synthetic Data Generation Process in Spatial Microsimulation Models. *Environment and Planning A*, *41*(5), 1251–1268.
Smith, D., Pearce, J. R., 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 and Place*, *17*(2), 618–624.
Spielauer, M. (2011). What is social science microsimulation? *Social Science Computer Review*, *29*(1), 9–20. https://doi.org/10.1177/0894439310370085
Sutherland, H., and Figari, F. (2013). EUROMOD: the European Union tax-benefit microsimulation model. *International Journal of Microsimulation*, *6*(1), 4–26.
Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation*, *7*(1), 4–25.

Tanton, R., Vidyattama, Y., Nepal, B., and McNamara, J. (2011). Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(4), 931–951.
Tanton, R., Williamson, P., and Harding, A. (2014). Comparing two methods of reweighting a survey file to small area data. *International Journal of Microsimulation*, *7*(1), 76–99.
Timmins, K. A., and Edwards, K. L. (2016). Validation of spatial microsimulation models: A proposal to adopt the Bland-Altman method. *International Journal of Microsimulation*, *9*(2), 106–122.
Vidyattama, Y., Tanton, R., and Biddle, N. (2015). Estimating small-area Indigenous cultural participation from synthetic survey data. *Environment and Planning A*, *47*(5), 1211–1228.
Voas, D., and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, *5*(2), 177–200.
Williamson, P., Birkin, M., and Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, *30*(5), 785–816.
Williamson, P., Mitchell, G., and McDonald, A. (2002). Domestic water demand forecasting: A static microsimulation approach. *Water and Environment Journal*, *16*(4), 243–248.