Chapter 44

Microsimulation

Mark Birkin

Abstract From origins in economics and financial analysis, microsimulation has become an important technique for spatial analysis. The method relies on conversion of aggregate census tables, sometimes complemented by sample data at the individual level, to synthetic lists of people and households. The individual records generated by the microsimulation can be aggregated flexibly to small areas, linked to create new attributes, and projected forward in time under stable conditions, or in the context of ‘what-if’ policy scenarios. The chapter outlines the basic building blocks of microsimulation and shows how these are combined within a representative practical application. It is argued that further progress can be expected through advances in computation, assimilation of data into models, and greater capacity to handle uncertainty and dynamics. We also expect the creation of more sophisticated architectures to reflect the interdependence between population structures at the micro-scale, and the supply-side infrastructures and urban environments in which they evolve.

44.1 Background to Microsimulation

Microsimulation models were introduced to the literature by Guy Orcutt in the 1950s. The approach was initially conceived as a powerful way to evaluate the distributional impact of economic and financial policies. The essence and distinctive feature of the method is that it proceeds through the specification and analysis of discrete entities which typically represent persons or households, in contrast to array-based representations which count the number of occurrences of a particular type. Consider for example an appraisal of the consequences of a series of changes in taxation which depend on the age, marital status, and income of the subject. A microsimulation approach would specify the population as a list of individuals, including age, marital status, and income as characteristics, to which an updated set of taxation rules can easily be applied. The notion of applying one or more discrete rules to a list of elements in order to determine an outcome (“list processing,” see below) is a

M. Birkin (✉)
School of Geography, University of Leeds, Leeds, UK
e-mail: m.h.birkin@leeds.ac.uk

central feature of the microsimulation modeling approach. The individual elements may then be combined into groups for cross-sectional analysis as required (“flexible aggregation,” see below).

The addition of a spatial label to the list of population characteristics provides a straightforward means to introduce a geographical element. Spatial microsimulation approaches have been popular in the analysis of health-care systems, education, transport and mobility, labor markets, retailing, and demographic analysis. Often the spatial disaggregation of the model rules (or parameters) can add further value, for example by specifying place-based variations in migration rates within a demographic model, but this need not necessarily be a fundamental element of the approach. Just as economic microsimulation models were originally established to investigate the effect of changing rules, spatial microsimulation models (MSM) are equally well suited to the assessment of scenarios involving changing parameters (e.g. future demographic change) or in the provision of infrastructure or services. Hence, the models can be powerful components within spatial decision-support systems for city planning.

Another important feature of spatial MSM is that they can be used to determine the impacts of policy or scenarios across a population even when detailed profiles for individuals or households are not available. The relevant methods usually involve synthetic estimation of individual records, typically using iterative proportional fitting from aggregate data or equivalent methods. Aggregate data are often easily accessible from sources such as neighborhood-level census tables, and MSM can prove to be a very efficient means to leverage these data. However, the methods can also be adapted to exploit real individual records which are increasingly available in the age of big data, for example through government departments, service operators, and consumer-facing organizations. Since individual databases of this type are rarely comprehensive or completely representative, in this case a major interest is in reweighting samples in order to maximize their value.

In this chapter, we will provide an introduction to fundamental issues and concepts in microsimulation modeling. Through an idealized but meaningful example, the major features and techniques will be described. Against this background, a more practical and powerful implementation will be outlined, concentrating on a specific but wide-ranging program of MSM for infrastructure assessment. We will discuss—in relation to both the main case study, and other relevant applications—some of the major areas of interest and further development potential for MSM at the present time. Conclusions and reflections on the evidence will be presented.

44.2 Overview of Methods and Concepts

44.2.1 *Population Synthesis*

When dealing with spatial data, it is typically the case that a range of counts will be known for various attributes across an array of small areas. Consider the example in Table 44.1, where distributions are presented across four typical areas in a region. These are the kinds of data which have been available to researchers from population censuses and surveys for many years. The five dimensions of variation displayed are lifestyle, household size, tenure, car ownership, and socio-economic status, and these vary in a natural way across area types. For example, there are more people living in flats (apartments) in urban areas, a heavy concentration of young adults in student areas, and the highest rates of car ownership in the countryside.

The essence of the microsimulation is to substitute synthetic individuals for the cell counts in each area. So for example, in Area 1, we will move to a list showing 1000 people, each with five attributes, rather than counts for every possible attribute of each state summing to 1000. In early applications (e.g. Birkin and Clarke 1988, 1989), a straightforward sequential estimation process is adopted. Let us suppose that the first attribute to be estimated is lifestyle, and then, we would proceed immediately by creating 500 individuals in Area 1 who are young adults, 300 as family members, 100 as empty nesters and 100 as retired. In Area 2 there are 100 young adults, and so forth.

Next, we add car ownership as an attribute, and since the rate of car ownership in Area 1 is 40%, then 200 young adults become owners of a car, and 300 are not. We continue this process for tenure, household size, and socio-economic status. The number of simulated individuals adhering to each attribute combination can be

Table 44.1 Population distributions in four idealized urban areas

		1: City	2: Country	3: Students	4: Suburbs
Lifestage	Young	500	100	400	100
	Family	100	200	300	500
	Empty-nest	100	300	200	300
	Retired	300	400	100	100
Household	Single	600	200	750	200
	Multi-person	400	800	250	800
Tenure	House	400	800	200	800
	Apartment	600	200	800	200
Car-owners	Car	400	800	200	600
	No car	600	200	800	400
Socio-economic status	Managerial	250	600	200	800
	Manual	750	400	800	200

expressed as:

$$X_i^{km} = \prod_k (p_i^{km}) X_i^{**}$$

for characteristics m relating to attribute k in area i , where X is a count and p is a probability.

For example, the most numerous group in Area 1 (City) within the simulation will have a profile reflecting the most numerous characteristics for each attribute, that is, young non-car-owners, living alone in apartments, with manual occupations. Members of this group will appear 81 times ($= 0.5 \times 0.6 \times 0.6 \times 0.6 \times 0.75 \times 1000$). A natural way to represent members of this group is simply as a list (11222)—lifestage is 1 (young), household is 1 (single), tenure is 2 (apartment), car ownership is 2 (does not have a car), and occupation is 2 (manual worker; see Table 44.1). The reader should be easily satisfied that the most numerous grouping in Area 2 is (42111); in Area 3, it would be (11222); and in Area 4 (22111).

Among many objections to this excessively simplified, presentation of the method is that the value in converting a small number of counts ($N = 12$) for each area into a list of 1000 people with 5 attributes ($N = 5000$) is not immediately apparent—but this should be more obvious by the end of this short exposition. Another problem is that it is unlikely a simple integer value will result from the product of a number of residents in an area (rarely likely to be as convenient a number as 1000 in practice) multiplied by a number of probabilities. This issue is usually addressed in MSM using Monte Carlo sampling—if there is a 60% chance that an individual lives alone then we draw lots, or random numbers, to assign household size. If that number is less than 0.6, then a single person household is the result (Lovelace and Ballas 2013 is one instance of a more sophisticated presentation and discussion of using integer weights to avoid any problems which might result from the assignment of fractions of individuals or households in spatial MSM).

44.2.2 Iterative Proportional Fitting

A third obvious objection to the simplified example in 2.1 is that independence between characteristics will rarely be a useful assumption. Thus, affluent white-collar workers are much more likely to be car owners than the unemployed, regardless of geographical location. Young people are more likely to be apartment dwellers, and so on.

This problem is usually handled using iterative proportional fitting (IPF). In the example above, it has in effect been assumed that compound probabilities for five attributes can be created as a linear combination of five independent constraint vectors, that is:

$$p(x_i^{k1}, x_i^{k2}, x_i^{k3}, x_i^{k4}, x_i^{k5}) = p(x_i^{k1})p(x_i^{k2})p(x_i^{k3})p(x_i^{k4})p(x_i^{k5})$$

In practice, more complex tables will allow much better estimates to be generated. For example, in the UK Census 2011, it is possible to utilize tables of car ownership by age (V1, V4), socio-economic status by age (V1, V5), household size by age and tenure (V1, V2, V3), and household size by age and socioeconomic status (V1, V2, V5). IPF provides the means to assemble such multidimensional constraints into a single set of estimates of the combined probability distribution:

$$p(x_i^{k1}, x_i^{k2}, x_i^{k3}, x_i^{k4}, x_i^{k5}) = f^{IPF}[p(x_i^{k123})p(x_i^{k125})p(x_i^{k14})p(x_i^{k15})]$$

As the name implies, the mechanics of this procedure involve successive adjustment of the combined probability distribution for consistency with each probability subset. This iterative procedure is known to be robust and convergent for the great majority of relevant problems (Fienberg 1970; Lomax and Norman 2016). Furthermore, IPF can be extended to accommodate large numbers of constraints with complex interactions.

44.2.3 *Reweighting*

Thus, IPF provides a robust and effective way for creating combined probability distributions across attribute sets. Ultimately, however, the method relies on the statistical estimation of individual data from aggregate totals. An alternative approach is to use data which are directly generated at the individual level. For example, suppose that a local authority holds data on claimants of housing benefits, then it may be possible to make a direct estimate of the impact of changing benefits rules on that population. Even in this situation, however, a common situation would be that changing brings a new target population into view—hence, to identify those affected, some more comprehensive simulation of the population will be required. MSM provides the means for extensive assessment of this kind.

A more typical situation is that some sample of individual data may be accessible (e.g. a Sample of Anonymized Records in the UK Census, or the Public Use Micro-Sample or PUMS in its U.S. equivalent). Provided that the sampling is robust, then data of this kind can be relied on to preserve cross-attribute relationships in the underlying population. The task for microsimulation is now to reweight the sample data in order to represent the nature of small areas: So in our example above, one would wish to apply higher weights to young people still in education when reconstructing the population of a student area; in the countryside, one oversamples for car-owners; and so on. Now, the procedure must ensure that weights are generated in such a way that when the data are aggregated all known constraints are still observed. In practice, the common approach to this problem is to select at random from a sample population and then switch individual records in order to improve the fit to known constraints. Simulated-annealing algorithms which allow backward steps have been found to be particularly effective (Harland et al. 2012), although genetic algorithms

and other heuristics such as tabu search have also been applied (Williamson et al. 1998; Zhu et al. 2015; Lidbe et al. 2017).

44.2.4 Data Linkage

An essential characteristic, and strength, of the MSM approach is an ability to thicken data sets, that is, to extend from a limited set of attributes into a much more extensive range of characteristics. In the simple example at Sect. 44.2.1, this is achieved by adding new characteristics from a different census table with independence. Once IPF is introduced, then the new attribute is related to the existing ones through a complex set of interrelationships. A more general approach to this problem, which is especially useful when data are reweighted from an individual sample, is to link between data sets.

Suppose we continue our example in which a population is characterized by age, socio-economic status, car ownership, etc. A lifestyle data set is made available in which respondents have declared their income based on age, car ownership, and occupation. The linkage problem is simply to add an income attribute by connecting the lifestyle data to the core demographics of the MSM. For straightforward problems, this can be achieved by creating a set of conditional probabilities for different income states in relation to the various independent variables and then using Monte Carlo sampling as above. A more general approach would be to create similarities between the individual records in each data set and then to combine the records. Where the number of records in the data is large relative to the attribute combinations, then this might result in multiple matching records in the target database. Again, this situation could be resolved by Monte Carlo sampling, that is, by selecting any of the matching records at random. Where the number of attribute combinations is very rich, or perhaps the linkage is to quite a small sample, then a perfect match may not be achievable. An alternative would be to create probabilistic linkages between the data sets, and so the linkage problem is to find a record in the target data set which has a high level of similarity to the origin record. This is tricky problem to resolve in view of the difficulty in equating (say) a situation in which two individuals are similar in every respect except they have different genders, as against two individuals who are identical except that one is a car owner and the other is not. Methods to resolve this difficulty, including a general application across ordinal, nominal, and categorical data sets, have been proposed and implemented by Burns et al. (2017). Of course, this method extends easily and naturally to the linkage of multiple attributes, either sequentially or simultaneously (e.g. if the lifestyle data set also includes expenditure, hobbies, or attitudes).

44.2.5 *Efficient Representation and Flexible Aggregation*

In Sect. 44.2.1 above, a question was raised as to why it might be advantageous to represent a city with a modest population as a list, rather than an array. Regardless of the other benefits described elsewhere, the value of this approach can quickly be seen as soon as the number of attributes and classes becomes more substantial. Van Imhoff and Post (1998) describe such an example in pure demographic terms, with a focus on a sub-model of reproduction. The likelihood of becoming pregnant might reasonably be supposed to vary substantially by single years of age in the mother, let us say in the range 15–44, but also according to marital status (married, single, widowed, or divorced), size of family (0,1,2,3,4 +), socio-economic group (6 classes), educational attainment (4 classes), employment status (3 classes), ethnicity (6 classes), and tenure (4 classes). In this situation, the number of potential unique states is evidently $30 \times 4 \times 5 \times 6 \times 4 \times 3 \times 6 \times 4 = 1.08$ million. So in any city or region with less than a million women of child-bearing age, it makes more sense to represent this population in the form of a list of individuals, rather than as a huge array with even more cells. Introduce some additional attributes (health status, socio-economic group, and educational attainment of the partner, perhaps), and the same consideration would apply across quite a large country.

This issue is doubly significant when considering small areas, especially when there are interactions, as for example in the consideration of migration, commuting, or retail flows. For example, the city of Leeds is frequently examined at a geography of more than 1000 census output areas, for example, when considering new housing developments, investments in transport infrastructure, or retail provision. Between these areas, there are evidently more than one million origin–destination pairs—many more than the number of workers, shoppers, or movers in the city. Hence, spatial MSM provides a powerful basis for efficient representation of both the structure and interaction patterns of population groups at a variety of geographical scales.

The representation of populations at the atomic level of individuals or households also permits flexible aggregation to any desired level of spatial or sectoral detail, provided only that the attributes of concern are appropriately embedded in the underlying data model. Of course, the census itself uses a complete (or almost complete) register of individual and household returns, and then aggregates these across specific topic areas for neighborhoods and regions—as we saw above, for example, in the case of car ownership or household composition by age of head. If car ownership, household composition, and age of head are included in the MSM along with a spatial identifier, then it is a straightforward matter to reproduce this logic, with the potential to cross-tabulate all three variables simultaneously if that is desirable. Should the MSM be extended to include twenty, thirty, or forty plus variables, then the potential attribute combinations become explosive, and the scope for diverse perspectives on a wide range of problems becomes very rich indeed.

44.2.6 List Processing

Another essential strength of MSM is the ability to apply rules for individual units of the population. A straightforward and common example of this would be in applying changing regimes for taxation: The impact of a new budget might be a change of income tax according to the earnings and marital status of a householder; the effect of changing fuel duty would depend on vehicle ownership and utilization; the impact of duties on cigarettes and alcohol would vary in relation to specific behaviors and habits. Each of these elements can quite easily be computed through a MSM, provided only that the determinants (i.e. income, car ownership, alcohol consumption, and so on) have already been represented in the base population. This means that not only is it possible to estimate potential benefits to the tax authorities, but also to evaluate distributional impacts on demographic sub-groups or small area populations in a city.

The concept of list processing can be applied in a different form, but with similar power and impact, to problems involving projection or forecasting of the population over time. For example, in relation to the attribute of age (in years), if we wish to project a population in time at single-year intervals, then age also increments by one at each interval. Other demographic processes, such as marriage, migration, or transitions within the labor market, may be subject to transition rates between classes. In this situation, changing states may be handled by Monte Carlo sampling of conditional probabilities (e.g. likelihood of marriage according to age, gender, and economic activity) as before.

44.3 An Example: Models of National Infrastructure

44.3.1 Overview

In 2010, partners from seven UK universities began working together on a Research Council program to explore future infrastructure options, requirements, and future scenarios. The Infrastructure Transitions Research Consortium (ITRC) considers the five sectors of transport, energy, water, wastewater, and IT, working in partnership with utilities, engineers, and regional and local providers, and acts as a trusted adviser to government through the National Infrastructure Commission. A second phase of funding with a focus on multi-scale infrastructure systems analytics (MISTRAL), including the translation of experience to international contexts, will continue until 2020.

Infrastructure projects are expensive and return on investment takes place over long-term horizons, regardless of whether these returns are measured in financial, social, or environmental terms. ITRC has a temporal framework which looks forward as far as possible toward the end of the twenty-first century. In order to create a more detailed understanding of the demand for infrastructure and its spatial and

Fig. 44.1 Model structure for infrastructure assessment

sectoral composition, ITRC requires highly disaggregate estimates of future population in relation to individual attributes, household groupings, and the character of neighborhoods and small areas.

The overall structure of the ITRC assessment process is shown in Fig. 44.1 below. ITRC uses a spatial microsimulation model to provide demographic inputs to the demand-estimation process for each of the five infrastructure sectors. The MSM is specified to the level of individuals with rich attributes, including demographics, social and economic profiles, housing, health, and labor market characteristics. Working with domain specialists in the research team, a consensus is established on the attributes representing the most important direct or proxy measures for the major drivers of infrastructure demand. Linking to consumption data from market-research surveys or direct measures of service use, for example from smart meters, sensors, or utility bills, makes it easy to translate population estimates into demand for infrastructure. Each of the demand sub-models which are driven from the MSM is linked to supply-side representations and policy options in order to drive a rich decision-support structure for infrastructure assessment. In the next sub-section, we explore the detail and a specific example.

44.3.2 *An Application of Spatial MSM to Energy Modeling*

44.3.2.1 **Population Reconstruction**

In the first phase of development of the ITRC, the UK population was recreated from the Sample of Anonymized Records (SAR; Thoung et al. 2016). Each element of the

SAR represents a real individual or household from the 2011 census from which small area labels and other potential identifiers have been removed in order to maintain the privacy of the subjects. The SAR therefore contains all of the demographic and socio-economic identifiers of the census including age, marital status, ethnicity, general health, education, occupation, car ownership, household composition, tenure, dwelling type, and a number of others.

The SARs are reweighted to reflect the composition of each census output area (a neighborhood with a typical size of no more than 200 households) using a simulated-annealing algorithm developed at Leeds (Harland 2013).

An approach to creating demand estimates for an indicative sector (energy) is described by Zuo and Birkin (2014). The English Housing Survey (EHS) contains in-depth household interviews and physical surveys for 17,000 households. EHS facilitates profiles of energy consumption and expenditure by fuel type and purpose for a rich selection of population and housing characteristics. The MSM used a CHAID (chi-square automatic interaction detection) approach to cluster households in both the MSM and the EHS into 41 categories based on a combination of dwelling type, household size, age and occupation of the household head, lifecycle, and household composition. A simple probabilistic match was applied to link records from the MSM and the EHS (i.e. records from the EHS were selected at random from the relevant cluster). Some contrasting energy-consumption profiles for different household types are shown in Fig. 44.2.

Fig. 44.2 Outputs from a microsimulation of energy consumption by household

44.3.2.2 Population Projection

The base populations within the ITRC MSM are projected forward in time using inputs from both the Office for National Statistics (ONS) National and Sub-National Population Projections (SNPP). The national projections provide the basis for estimation of aging, fertility, and mortality (“natural change”) within the population, whereas the SNPP allows the introduction of migration and the calibration of the natural change parameters to local areas. The essence of this process is therefore to list-process the base populations using a combination of demographic change rates (for fertility, mortality, and migration). The parameter estimates are managed in order to ensure consistency of the simulation outputs with the ONS regional and population profiles. For more detail, see Zuo and Birkin (2014) and Thoung et al. (2016).

This simulation process adds considerable richness to the ONS estimates by permitting detailed spatial disaggregation on the sub-national projections—which are only available over a 25 year planning horizon—and by their extrapolation alongside the national medium (50 year) and long-term projections (75 years). The flexibility of MSM is also fully exploited in ITRC through the use of variant population projections. For much of the work which has been presented to policy-makers, eight scenarios are presented which illustrate the impact of future changes in technology, affluence, and political circumstances on the population (Thoung et al. 2016).

44.3.2.3 Scenarios

The spatial detail of the MSM is particularly important when considering future infrastructure investments which have strong local dependencies, including renewable energy, personal mobility, and the supply of water. In the outline above, it has been seen that energy consumption is expected to grow in relation to expansion of the population, and be subject to compositional shifts in relation to changes in supply. One of the major motivations of ITRC is to consider the potential impacts of climate change on infrastructure (Jenkins et al. 2014). In one published application from the ITRC, climate-change projections from the Met Office Hadley Center were combined with the spatial MSM, with modified energy consumption rules relating variations in energy use to regional and seasonal variations in the climate within the EHS. This scenario was extended to 2100. A significant reduction in household energy use was expected due to global warming (see Fig. 44.3). The authors note that the potential to counterbalance due to increased use of air conditioning was not examined because of limitations in the base data. However, a variety of other behavioral shifts were also considered, with evidence drawn from extant published studies. These included adoption of solar power, insulation, double glazing, adoption of low energy lighting, and shifts to more efficient central heating systems. Behavioral change was not expected to affect cooking or the use of electrical appliances (Zuo and Birkin 2014).

Fig. 44.3 Reductions in energy consumption from a behavioral simulation

44.3.3 Extensions

The architecture of spatial microsimulation which underpins the ITRC project has recently been completely overhauled. A technology platform for Synthetic Population Estimation and Scenario Projection (SPENSER) now services the infrastructure sub-models. It is also designed to support extensions to sectors such as education and health. The capability of the new system to represent diverse behavioral components has already been demonstrated through a flexible application to consumer spending across a full range of expenditure categories (James et al. 2019). This implementation is specifically aligned to the study of future meat consumption under various alternative scenarios for production, sustainability, affluence, and lifestyle preferences.

SPENSER has a more modular design than the previous deployment within ITRC, with separate routines for data mobilization, population recreation, forecasting, and scenario building. It is hoped that a more robust design will make SPENSER amenable to a wider range of substantive improvements in the underlying scientific approach. In the next section, some key elements of the agenda for future development are discussed.

44.4 Priorities for Spatial Microsimulation

44.4.1 *Computation*

The computational burden attached to spatial microsimulation models is often quite considerable. This need arises from a desire to represent the population with significant variety (i.e. many attributes) at a fine level of spatial resolution (i.e. a lot of zones), and potentially with complex spatial or behavioral interactions to model or represent. Significant computation is needed in both the generation of the initial population, including both reconstruction and linkage, and in projections of the model forward in time.

Simple approaches to reweighting baseline populations, or using conditional probabilities from iterative proportional fitting, are not especially expensive in computational terms when they are based on one-shot estimates of the parameters. Iterative approaches including genetic algorithms (GA) and especially simulated annealing (SA) have persistently yielded better results, but are often slow to converge. These techniques depend on complex evaluations of the fitness of a model: in principle a single step of either GA or SA involves exchanging the position of two elements in the simulation (e.g. moving and replacement of an individual from one zone to another), then reaggregating the population at zone level, calculating the fit to multiple constraining totals, and then applying an evaluation function to assess the utility of the switch. This activity can be repeated multiple times for each member of a population of millions, within a loop which could itself be executed hundreds of times within the algorithm. The dynamics of the modeling also involve complex processing across a large population size, often with small time steps and multiple scenario combinations. The impacts could become explosive if adopting methods such as ensemble modeling as a means for exploring sensitivities or robustness in the model outcomes. There is no doubt that the difficulty in accessing adequate computational resources has been an impediment to exploration of some potentially fertile approaches, such as the use of ensembles.

More intense applications of spatial MSM are being permitted to some degree by the availability of high-performance computing. For example, SPENSER has access to the Data Analytics Facility for National Infrastructure (DAFNI) as a platform for executing complex model runs. Similar capability exists within the Integrated Research Campus at the Leeds Institute for Data Analytics. Nevertheless, data-services infrastructures remain scarce, difficult, and expensive to access.

Rather than the provision of enhanced computational power, simplification of the models themselves is clearly an alternative to consider. A natural strategy would be to reduce the population size, for example by sampling, or the representation of subsets rather than individuals (Parker and Epstein 2011). This approach seems more feasible for national applications than those involving small spatial zones in which the full variety of the population must be retained. A more promising method which has been adopted in dynamic microsimulation is to lengthen the time interval between processing steps. When considering discrete events such as birth, migration

or death, the usual method is to apply transition probabilities (or hazards; Clark and Rees 2017) to a population at risk at regular intervals, generally annualized. If the occurrence of such events is on average significantly less than once a year, then an option would instead be to process the time to next event and save the trouble of repeated assessments for change of state in the intervening period. This technique has been successfully introduced within the Canadian MSM DynaCan (Morrison 2007), and adopted elsewhere.

44.4.2 *Uncertainty*

The potential for error, and consequent uncertainty in model estimates and projections, is widespread in the microsimulation framework. While MSM are usually created from high-quality sources, including censuses and national statistics, these data are by no means free of bias and inaccuracy. For example, censuses are never completely enumerated, giving rise to errors in the imputation of missing records. Students, transient populations, and the homeless all have significant potential for misrepresentation. When these data are combined, then sophisticated models have the capability to reproduce aggregate constraints with minimal variations. However, the individual estimates are subject to unknown errors which are by definition unobservable to the extent that the purpose of the model is to simulate individual distributions which are not directly measured.

These issues become more challenging for more ambitious applications, for example if a demographic microsimulation is linked to big data for mobility, consumer spending, health, and behavior (Birkin 2018), because such data sets are themselves more variable in data quality and in view of distortions in the linkage process itself.

When the purpose of microsimulation modeling is to assess the effect of changing financial regulations, taxation, or benefits then modeling scenarios can be expected to be relatively robust. When the what-if models are reliant on changing infrastructure, uncertain behaviors, policy environments, and economic circumstances, then any attempts at projection and impact analysis are hugely uncertain. The MSM community has largely sidestepped the problems associated with uncertainty by offering single model estimates, occasionally flexed through defined scenarios with variant input assumptions. This may change if microsimulation chooses to align itself more closely with emerging disciplines in data science. A particular instance of this could be through the adoption of probabilistic programming (Improbable Research 2019). In this new style of model implementation, state variables are assigned distributions rather than discrete values, and operators may be treated in the same way. Hence, this approach lends itself naturally to the expression of outcomes in terms of likelihoods, confidence intervals, or other dimensions incorporating variability and uncertainty. A drawback of this style of research is that tools are still relatively inaccessible and in early stage of development, and experience of complex applications is limited.

44.4.3 Data Assimilation

The origins of spatial microsimulation are as a means to estimate unknown individual-level variations from aggregate data about neighborhoods and small areas. Later, applications incorporate more information by the addition of sample data, in which case the essence of the problem may be more about reweighting. In either of these cases, the ambition is to create simulations in detail from relatively restricted data, and in all circumstances, evaluation of the success of the models is a challenge, because by definition we are estimating things which are unobserved. In the age of Big Data, where increasingly more is known about the world at ever finer scales, the nature of the challenge is beginning to shift toward a view of the world in which it is possible to steer models toward more effective representations through the absorption of evidence. This could be facilitated by data assimilation.

It has been recognized for some time in the complex domain of weather forecasting that methods are needed to update models as new information becomes available. This process of data assimilation has been adopted into agent-based simulation, for example through the adaptation of pedestrian movement models to absorb movement data from street sensors (Ward et al. 2016). There seems no reason in principle that the philosophy and techniques of data assimilation might not be used to calibrate longer-term effects such as spatial diffusion or policy impacts in a microsimulation.

44.4.4 Dynamics

MSM is typically used in one of three modes, which can be characterized as static, comparative static, and dynamic. Static MSM may refer to population reconstruction processes in which aggregate data are decomposed to generate refined distributions at household or individual levels. These outputs may be valuable in their own right, for example to understand the prevalence of at risk groups, or provide inputs to agent-based models (ABM) or other policy models.

Linkage to other data sets is also a static or baseline process, for example using MSM to estimate expenditures or market potential in a retail model (James et al. 2019). As noted above, comparative static is a core mode for tax and benefits assessment (Sutherland and Figari 2013). Comparative-static applications are perhaps the most common in which some variation in the initial conditions allows the MSM to be applied in what-if mode. In SPENSER, many of the scenarios look to the future but are essentially comparative static since they start from the premise that higher level forecasts (such as ONS estimates of the future population) can be disaggregated, and then input to secondary models of demand for infrastructure or consumption of other services.

Truly dynamic models are not entirely absent (Morrison 2007; Li and O'Donoghue 2013; Rutter et al. 2011) but challenging in that they require the incorporation of longitudinal processes in relation to core demographics (e.g. fertility, mortality, and

migration) or more specific elements such as morbidity or energy consumption. Backward propagation of MSM as a basis for validating both the structure and logic of dynamic MSM is another concept that might usefully be borrowed from climate-modeling literature, but is as yet relatively unexplored.

Fast and slow dynamics are also a consideration for MSM. Much more attention has been focused on long-term or slow dynamics, and these kinds of models are important for decision making in relation to major infrastructure investment and policy making. However, fast dynamics are becoming more relevant in relation to real-time observation. This makes a connection to data assimilation, and opportunities for real-time evaluation and model enhancement. We will see increasing use of machine learning techniques like reinforcement learning for traffic lights or store promotions, and blurring of boundaries between data science, MSM, ABM, and other forms of individual-based modeling. It is surprising that these approaches are relatively unexplored in commercial applications, where personalization and precision targeting are a priority with the growing availability and fidelity of individual data.

44.4.5 *Interdependence*

Applications of MSM are well-suited to the problems of demand estimation, which are typified by the uses of SPENSER as a tool within the ITRC framework for future infrastructure assessments. Similar applications can be seen in the estimation of retail expenditure (James et al. 2019), educational attainment (Kavrouidakis et al. 2013), health care (Clark and Rees 2017) and even the incidence of crime (Kongmuang 2006) and the need for jobs (Ballas and Clarke 2000). The beauties of the technique in this regard are multiple (as we have seen), providing a powerful means to connect aggregate data to individual-level modeling, introducing rich and multiple simultaneous representations of individual attributes, and a sophisticated understanding in changing drivers of consumption over time.

Nevertheless, conceptual architectures which view microsimulation purely as a foundational layer in the modeling process are often in danger of simplifying away many of the subtle and vitally important interactions which underpin real-world problems. The importance of interaction and interdependence between individuals has always been fundamental to ABM, in which the capacity for complex structures to emerge—often in unexpected ways—is a cornerstone of the method (Schelling 1969). However, while conceptually rich in this sense, ABM is typically less strongly grounded in the empirical realities of everyday life.

The benefits of linking microsimulation to meso-scale representations of land-use and service provision have been recognized in early applications to a retail market (Birkin and Clarke 1987; Nakaya et al. 2007). In this framework, a microsimulation is used to create a rich population, which in turn forms the basis for expenditure assessments across a tapestry of small areas. These expenditure estimates are then combined with networks of service provision through a spatial interaction model (SIM), hence creating revenue flows from neighborhoods to shopping centers. These

flows can then be sampled in order to create assignments of retail preferences for individual consumers, thus closing the loop from demand to supply. A similar process underlies a module within SPENSER which connects the microsimulation to migration flows through a spatial interaction model of internal migration (SIMIM; Lomax and Smith 2019). In order to fully embed microsimulation within land-use transport interaction models, however, it might be argued that the reciprocal dynamics of infrastructure systems including housing and transport must be fully incorporated within the model system.

The resulting applications would be somewhat analogous to the network planning models developed in Leeds by Geographical Modeling and Planning (GMAP) Limited the 1990s, in which service delivery was co-designed with retail demand. George et al. (1997) provide a good description of a representative problem. The broader significance, perhaps, of the GMAP experience (Birkin et al. 1996; Birkin et al. 2002, 2017) is in seeing spatial analysis approaches including MSM as elements of spatial decision-support systems (Geertman and Stillwell 2009). Robust translation of such ideas into the urban planning domain, for example through the integration of SPENSER with other models such as UCL's Quantitative Urban Analytics (QUANT) model of land-use and transport interactions, could provide stronger foundations for spatial decision support than hitherto.

While MSM is almost exclusively used to represent both individuals and households as the entities within a modeling system, there is no reason why other elements such as vehicles, houses, schools, hospitals, firms, or retail outlets might not equally be represented in a similar way, with rich characteristics and complex behavioral drivers. Indeed, one might argue whether cellular automata, in which the building blocks are land-use parcels changing in character through time, are so different to microsimulation. Hybrid models which combine MSM with SIM, land-use and transport interaction models, or even cellular automata are likely to become increasingly popular, but the absorption of more complex actors representing complementary sectors might be seen as a fully viable alternative strategy.

44.5 Conclusions

Spatial MSM has been developed as an important variant from the introduction of similar individual-based models in economics and financial policy. The technology of spatial microsimulation has progressed steadily over a period of more than thirty years, allowing population distributions in very small areas to be faithfully represented. The models benefit from increasingly detailed and diverse sources of data. This also provides underpinning for applications to a diverse range of problems.

The scope for further enrichment of spatial MSM is substantial, for example drawing on computational advances and progression of techniques in data science, machine learning, and artificial intelligence. This could help to increase the robustness of models, especially when their dynamic qualities are considered as a basis for projection and forecasting.

References

- Ballas D, Clarke G (2000) GIS and microsimulation for local labour market policy analysis. *Comput Environ Urban Syst* 24:305–330
- Birkin M (2018) Big data for social science research. *Ubiquity*, Jan 2018. 1–7. <https://doi.org/10.1145/3158339>
- Birkin M, Clarke M (1987) Comprehensive models and efficient accounting frameworks for urban and regional systems. In: Griffith D, Haining R (eds) *Transformations through space and time*. Martinus Nijhoff, The Hague, pp 169–195
- Birkin M, Clarke M (1988) SYNTHESIS: a SYNTHetic spatial information system for urban modeling and spatial planning. *Environ Plan A* 20:1645–1671
- Birkin M, Clarke M (1989) The generation of individual and household incomes at the small area level using synthesis. *Reg Stud* 23:535–548
- Birkin M, Clarke G, Clarke M, Wilson A (1996) *Intelligent GIS: location decisions and strategic planning*. Geoinformation International, Cambridge
- Birkin M, Clarke G, Clarke M (2002) *Retail geography and intelligent network planning*. Wiley, Chichester
- Birkin M, Clarke G, Clarke M (2017) *Retail location planning in an era of multi-channel growth*. Routledge, London
- Burns L, Heppenstall A, See L, Birkin M (2017) Developing an individual-level geodemographic classification. *Appl Spat Anal Policy* 11:417–437. <https://doi.org/10.1007/s12061-017-9233-7>
- Clark SD, Rees PH (2017) The drivers of health trends: a decomposition of projected health for local areas in England. In: Swanson DA (ed) *Frontiers in applied demography*. Applied demography series 9. Springer, Berlin, pp 21–40
- Fienberg SE (1970) An iterative procedure for estimation in contingency tables. *Ann Math Stat* 41:907–917
- Geertman S, Stillwell J (eds) (2009) *Planning support systems: best practice and new methods*. Springer, New York
- George F, Radcliffe N, Smith M, Birkin M, Clarke M (1997) Spatial interaction model optimisation on parallel computers. *Concurrency: Pract Exp* 9(8):753–780
- Harland K (2013) *Microsimulation model user guide (Flexible modeling Framework)*. NCRM working paper. National centre for research methods
- Harland K, Heppenstall A, Smith D, Birkin M (2012) Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *JASSS* 15(1):1
- Improbable Research (2019) Keanu: a probabilistic approach. <https://github.com/improbable-research/keanu>. Accessed 27 Oct 2019
- James W, Lomax N, Birkin M (2019) Local level estimates of food, drink and tobacco expenditure for Great Britain. *Sci Data*. <https://doi.org/10.1038/s41597-019-0064-z>
- Jenkins K, Hall J, Glenis V et al (2014) Probabilistic spatial risk assessment of heat impacts and adaptations for London. *Climatic Change* 124:105–117
- Kavroudakis D, Ballas D, Birkin M (2013) Using spatial microsimulation to model social and spatial inequalities in educational attainment. *Appl Spat Anal Policy* 6(1):1–23
- Kongmuang C (2006). *Modeling crime: a spatial microsimulation approach*. Ph.D. thesis. University of Leeds
- Li J, O'Donoghue C (2013) A survey of dynamic microsimulation models: uses, model structure and methodology. *Int J Microsimulation* 6(2):3–55
- Lidbe A, Hainen A, Jones S (2017) Comparative study of simulated annealing, tabu search and the genetic algorithm for calibration of the microsimulation model. *Trans Soc Model Simul Int* 93:21–33
- Lomax N, Norman P (2016) Estimating population attribute values in a table: “get me started” in iterative proportional fitting. *Prof Geogr* 68(3):451–461
- Lomax N, Smith A (2019) Spatial interaction models of internal migration. <https://github.com/nis/mod/simim>. Accessed 26 Oct 2019

- Lovelace R, Ballas D (2013) 'Truncate, replicate, sample': a method for creating integer weights for spatial microsimulation. *Comput Environ Urban Syst* 41:1–11
- Morrison R (2007) Model 6: DYNACAN (Longitudinal Dynamic Microsimulation Model). In: Gupta A, Harding A (eds) *Modeling our future: population ageing, health and aged care. International symposia in economic theory and econometrics, vol 16*. Emerald group publishing limited, Bingley, pp. 461–465
- Nakaya T, Fotheringham AS, Hanaoka K, Clarke GP, Ballas D, Yano K (2007) Combining microsimulation and spatial interaction models for retail location analysis. *J Geogr Syst* 4:345–369
- Parker J, Epstein J (2011) A distributed platform for global-scale agent-based models of disease transmission. *ACM transactions on modeling and computer simulation* 22(1): Article 2
- Rutter C, Zaslavsky A, Feuer E (2011) Dynamic microsimulation models for health outcomes: A review. *Med Decis Making* 31:10–18
- Schelling T (1969) Models of segregation. *Am Econ Rev, Pap Proc* 59:488–493
- Sutherland H, Figari F (2013) EUROMOD: the European union tax-benefit microsimulation model. *Int J Microsimulation* 6(1):4–26
- Thoung C, Beaven R, Zuo C et al (2016) Future demand for infrastructure services. In: Hall J, Tran M, Hickford A, Nicholls R (eds) *The future of national infrastructure: a system-of-systems methodology*. Cambridge University Press, Cambridge, pp 31–53
- Van Imhoff E, Post W (1998) Microsimulation methods for population projection. *Population* 10(1):97–136
- Ward JA, Evans AJ, Malleon NS (2016) Dynamic calibration of agent-based models using data assimilation. *Roy Soc Open Sci* 3:150703
- Williamson P, Birkin M, Rees P (1998) The estimation of population microdata using data from small area statistics and samples of anonymised records. *Environ Plan a* 30:785–816
- Zhu S, Tey L, Ferreira L (2015) Genetic algorithm based microscale vehicle emissions modeling. *Mathematical problems in engineering*, article ID 178490. <https://doi.org/10.1155/2015/178490>
- Zuo C, Birkin M (2014) Spatial microsimulation modeling for residential energy demand of England in an uncertain future. *GeoSpatial Inf Sci* 17(3):157–169

Mark Birkin is Professor of Spatial Analysis and Policy at the University of Leeds. He is also Programme Director and Turing Fellow at the Alan Turing Institute and Director of the Leeds Institute for Data Analytics. He is a Fellow of the Academy of Social Sciences and of the Royal Geographical Society.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.