STATISTICS IN TRANSITION new series, August 2020
Special Issue, pp. 201–221, DOI 10.21307/stattrans-2020-038
Submitted – 31.01.2020; paper accepted for published – 30.06.2020

A generic business process model for conducting microsimulation studies

Jan Pablo Burgard[1], Hanna Dieckmann[2], Joscha Krause[3], Hariolf Merkle[4], Ralf Münnich[5], Kristina M. Neufang[6], Simon Schmaus[7]

Abstract

Microsimulations make use of quantitative methods to analyze complex phenomena in populations. They allow modeling socioeconomic systems based on micro-level units such as individuals, households, or institutional entities. However, conducting a microsimulation study can be challenging. It often requires the choice of appropriate data sources, micro-level modeling of multivariate processes, and the sound analysis of their outcomes. These work stages have to be conducted carefully to obtain reliable results. We present a generic business process model for conducting microsimulation studies based on an international statistics process model. This simplifies the comprehensive understanding of dynamic microsimulation models. A nine-step procedure that covers all relevant work stages from data selection to output analysis is presented. Further, we address technical problems that typically occur in the process and provide sketches as well as references of solutions.

Keywords: multi-source analysis, multivariate modeling, social simulation, synthetic data generation

1. Introduction

Microsimulation studies represent a powerful tool for the multivariate analysis of populations (Merz, 1993; O'Donoghue, 2001; O'Donoghue and Dekkers, 2018; Burgard et al., 2019a). While macrosimulation methods are limited to selected population characteristics on an aggregated level, microsimulation methods are capable of considering

[1]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: burgardj@uni-trier.de. ORCID: https://orcid.org/0000-0002-5771-6179.
[2]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: dieckmann@uni-trier.de. ORCID: https://orcid.org/0000-0002-6455-1210.
[3]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: krause@uni-trier.de. ORCID: https://orcid.org/0000-0002-2473-1516.
[4]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: merkle@uni-trier.de. ORCID: https://orcid.org/0000-0001-7653-383X.
[5]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: muennich@uni-trier.de. ORCID: https://orcid.org/0000-0001-8285-5667.
[6]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: neufang@uni-trier.de. ORCID: https://orcid.org/0000-0002-8241-4818.
[7]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: schmaus@uni-trier.de. ORCID: https://orcid.org/0000-0002-2037-4312.

individual characteristics and interactions. This allows for a more comprehensive understanding of the population and sophisticated projections on its development. As a result, microsimulations are increasingly applied for the analysis of complex systems. Exemplary applications are provided by Bourguignon and Spadaro (2006), Pichon-Riviere et al. (2011), Li and O'Donoghue (2013), Markham et al. (2017), O'Donoghue and Dekkers (2018), and Burgard et al. (2020).

Microsimulation studies are often performed according to a basic procedure. First, an adequate base dataset as a representation of the target population is needed. This requires either synthetic data or empirical observations from administrative records and surveys (Li and O'Donoghue, 2013). Next, selected features that characterize the population in its initial state are altered in scenarios. The scenarios are projected into future periods and construct individual branches in the evolution of the base population. After a sufficient number of simulation periods, the branches are compared. The comparison provides insights into essential dynamics and interdependencies within the population that typically cannot be assessed otherwise (Li and O'Donoghue, 2013; Burgard et al., 2019b).

However, there is a lack in generic descriptions on how to construct, implement, and evaluate microsimulations. This makes it difficult for researchers that are new to the field to properly conduct their own studies. Microsimulations require the statistically sound combination of multiple data sets, the construction of a sophisticated simulation infrastructure, as well as the careful analysis of simulation outcomes. If these challenges are not addressed properly, microsimulation results are not reliable and may lead to false conclusions in the analysis.

In this paper, we present a generic business process model for conducting microsimulation studies. We develop a coherent framework that can be used as instruction for all relevant work stages, including data generation, population projection, and output analysis. Drawing from the generic statistical business process model by UNECE (2013), our model consists of nine sequential steps. For each step, we elaborate on data requirements, methodological challenges, as well as possible solutions. Our descriptions can be broadly used as guidance to properly perform microsimulation research for various applications.

The remainder of the paper is organized as follows. In Section 2, we cover the specification of needs, data selection and preparation. Section 3 describes the population projection. In particular, we look at the design of the microsimulation model, population dynamics, as well as the actual simulation. In Section 4, we address output analysis. Here, relevant aspects are the analysis of simulation results, dissemination strategies, and evaluation. Section 5 closes with some concluding remarks and an outlook on future research.

2. Requirements and data selection

2.1. Step 1: Specification of needs

The underlying concept of microsimulations is to model the actions and interactions of micro-level units in a population to analyze their impact on the macro-level (Spielauer, 2011). For instance, micro-level units may represent individuals in the context of social change, firms in a competitive market situation, and cars as part of traffic or transport systems. Thus, in order to conduct a microsimulation study that allows for reliable results, a suitable simulation frame and clear research questions have to be defined. This can be done by answering the following questions:

- What kind of system shall be simulated?
- What are the characteristics of interest?
- Under which scenarios shall these features be studied?
- Which hypotheses shall be investigated?
- What are the smallest relevant entities for this purpose?
- What are potentially relevant processes and interdependencies?
- What temporal frequency for projection has to be considered?

An important distinction is between static and dynamic microsimulations (Rahman and Harding, 2017; Hannappel and Troitzsch, 2015). Static microsimulations typically have fewer data requirements and demand less computational resources than dynamic microsimulations. They are suitable for applications where the immediate effect of a clearly defined external change on micro-level units is of interest. The attributes associated with micro-level units are mainly persistent over the simulation process. In this setting, the temporal change of micro-level attributes can be modeled indirectly via reweighting and uprating (inflating/deflating) of variables (Dekkers, 2015). Prominent models such as EUROMOD (Sutherland and Figari, 2013) commonly focus on the impact of possible (e.g. tax-related) policy changes.

Dynamic microsimulation models such as DYNASIM (Favreault et al., 2015) allow for a more sophisticated evolution of the population on the micro-level. A given micro-level characteristic is an endogenous factor in the simulation. The probability for a specific realization depends on both the simulated time and the realizations of other micro-level characteristics. Likewise, dynamic microsimulations are characterized by stochastic transitions and direct temporal changes of micro-level unit attributes. They are suitable for applications where multidimensional dependencies between micro-level units are relevant for the simulation outcomes. For instance, Burgard et al. (2019a) used a dynamic model for investigating future long-term care demand in a city, which required the anticipation of family structures and neighborhood characteristics. Naturally, this simulation type can be very resource-intensive.

A further distinction of dynamic models is with respect to the representation of time: discrete and continuous. In discrete-time dynamic microsimulations, temporal changes occur at predefined time intervals, such as simulated months or years. In continuous-time dynamic microsimulations, temporal changes occur at any given time within the simulated time domain (simulation horizon). Conceptually, the choice between these modes depends on whether it is necessary to account for interperiodic events in light of the research questions. Methodologically, the choice should be based on the assumptions regarding transition dynamics the researcher is willing to make. Discrete dynamics require less assumptions for the modeling of a given transition, but are also less flexible in accounting for complex interdependent event sequences (Willekens, 2017). Continuous dynamics typically require far-reaching assumptions on conditional transition rates, but are generally capable of displaying highly complex temporal event dependencies. For deeper insights into dynamic microsimulation modeling, see for example Li and O'Donoghue (2013), O'Donoghue and Dekkers (2018), and Willekens (2017).

Another crucial distinction is between open and closed population microsimulations (Spielauer, 2009). It refers to the question of whether micro-level units can interact with other micro-level units that are not initially part of the system of interest. In a closed-population microsimulation model, interactions are restricted to units that are part of the base population prior to projection. In an open population microsimulation model, new units can be generated that are added to the base population during the simulation. For instance, if a demographic projection of a regional population shall be performed, then this may correspond to migration from other regions. Conceptually, the closed approach is sensible when the research focus is on the regional population in its current state. Any effect that unfolds under a particular projection scenario is exclusively intrinsic given the initial base population. The open approach can be used when the focus is on the evolution of the region in which the base population is located. Modeling the corresponding domain as an entity requires the consideration of migration in order to be realistic. Naturally, open-population microsimulations need detailed migration data for this purpose.

After a suitable variant has been determined, the researcher has to define several simulation scenarios. They should be constructed such that they meet population characteristics that are essential in light of the research questions. A key aspect of microsimulation is to examine how target variables change under various theoretical social, economic or policy-related developments. For instance, demographic scenarios or alternative policies (e.g. tax-benefit systems) might be relevant for the research context and, therefore, be integrated into the simulation process.

2.2. Step 2: Data selection & Step 3: Data preparation

After determining research objectives and the model variant, data requirements have to be specified. The methodological challenges associated with these work stages directly depend on each other. Therefore, we address these steps jointly.

We introduce some notation and a basic data setting that helps us to illustrate the relevant aspects. Let U denote a real-world population of $|U| = N$ individuals indexed

by $i = 1,...,N$. The objective is to analyze this population via microsimulation methods. Thus, in light of the comments from Section 1, it represents the system of interest. Let $\tilde{U}$ be the base population of $|\tilde{U}| = \tilde{N}$ micro-level units indexed by $u = 1,...,\tilde{N}$. It may be viewed as a digital replica of U that we can project into future periods. Further, let $D \subset U$ be a random sample of $|D| = n$ individuals indexed by $i = 1,...,n$. Denote p_i as the inclusion probability associated with $i \in U$ given the sampling design. The sample represents an exemplary data input for the microsimulation. It can be used to construct the base population $\tilde{U}$ and to obtain empirical parameters for the projection of $\tilde{U}$. In what follows, we elaborate on potential data sources that a researcher may consider as a base population directly or for the creation of such a population.

Data Type	Characteristics	Formalization	p_i known?	Example(s)
Administrative Data	All units of a population of interest available in its entirety	$i \in U$	$p_i = 1$	Register of residents, register of taxpayers
Census Data	Usually person- and household-level data	$i \in U$	$p_i = 1$	Data collected from a census
Survey Data	A random sample of the units of the population of interest is available	$i \in D \subset U$	$p_i \in (0,1]$	Survey of units of interest, e.g. households, persons, firms
Synthetic Data	A synthetic population of interest containing (partially) synthetic units	$u \in \tilde{U}$	Yes / No	Generated data based on other data sources
Big Data	Huge, complex or steadily fast generated data	$i \in D \subset U$	No	Remote sensing data or data collected using phones

Table 1: Datatypes and their properties

A crucial point for the assessment of data quality is to know about the data production process. Since data serves as input for microsimulation models, the data quality determines also the quality of the microsimulation model. Table 1 provides a generic overview of exemplary data sources and their associated properties. The most relevant data sources are administrative data, census data, household, and survey data, as well as synthetic data (Li and O'Donoghue, 2013). The use of big data sources is not yet established in the microsimulation literature, but marks a relevant option for future research (O'Donoghue and Dekkers, 2018).

We start with administrative and census data. In the best case, these data sets cover the entire population U and there is no sampling process that has to be anticipated.

Further, they are rarely subject to measurement errors, such as inaccurate reportings of sampled individuals. Therefore, they can often be used directly. If the data sets cover all relevant characteristics in light of the research questions, then the researcher can use them as base population $\tilde{U}$ for projection. However, if essential characteristics are missing, then the data sets may be extended artificially via synthetic data generation. Further, please note that there are also occasions where administrative data does not cover the entire population, but only a subset of it given the administrative purpose. A corresponding example would be administrative data on taxation, where only tax-payers are included. In these cases, issues like coverage problems have to be accounted for in order to create the base population. For further details, see for instance Smith et al. (2009).

In the case of survey data, the researcher must be aware of the sampling design in order to use the data correctly (Dekkers and Cumpston, 2012). Depending on the application, it is necessary to apply weighting and imputation procedures, provided that they are not already implemented by the data producer. These steps involve the adjustment for possible nonresponse. For unit-nonresponse, the design weights (typically inverse inclusion probabilities) are altered such that relevant sample totals reproduce known population totals (Haziza and Beaumont, 2017). This is achieved via calibration methods, such as the generalized regression estimator (Deville and Särndal, 1992; Särndal, 2007) and empirical likelihood techniques (Chen and Quin, 1993). For item-nonresponse, the missing observations are imputed, for instance via multiple imputation (Schafer and Graham, 2002). Once the data set is adjusted, it can either be directly used as base population or has to be expanded by means of adding synthetic individuals.

However, often the required data might not be available due to disclosure control, as the data provider is obligated to delete regional identifiers. In this case, the generation of synthetic data is an option (Drechsler, 2011). For this, often multiple data sources (e.g. survey data and known totals) can be combined to construct a synthetic population based on real-world observations. For instance, the researcher may calculate calibration weights (Deville and Särndal, 1992; Burgard et al., 2019c) for survey observations such that (synthetic) marginal distributions reproduce known population totals for a set of relevant characteristics. The synthetic population then consists of units allocated (with replacement) to spatial regions according to their newly calculated weight (Williamson, 2013; Tanton et al., 2014; Lovelace, 2016; Rahman and Harding, 2017; Tanton, 2018). Alternatively, a synthetic population can be modeled by estimating distribution or model parameters from survey data and actually reconstruct the population (Huang and Williamson, 2001; Münnich and Schürle, 2003; Alfons et al., 2011a, Alfons et al., 2011b). This can avoid cluster effects arising from units that are repeated frequently within a region. In conclusion, there is a reweighting and an imputation approach to generate synthetic data. For the imputation approach, one considers to apply editing procedures to avoid implausible variable outcomes (Drechsler, 2011). After the synthetic population has been generated, it can be used as base population for projection.

Although not yet established, using big data for microsimulation research is an important topic. These data sets are typically very rich in detail and allow to survey complex

phenomena, such as network structures. As a result, social media data is already used in humanity fields like sociology (Murthy, 2012). Microsimulations could greatly benefit from corresponding data in order to improve the modeling of network structures or social behavior. However, big data sources also impose several methodological challenges, such as coverage problems or unknown inclusion probabilities. These issues mark a central subject for future research in the field.

3. Population projection

3.1. Step 4: Design of the microsimulation model

In Section 2.2, we stated the importance of constructing suitable scenarios given the research questions. However, not only the scenario design is crucial, but also the design of an overall functional simulation infrastructure. There are different approaches to ensure that the infrastructure works as desired. Naturally, these approaches depend on the type and complexity of the microsimulation variant chosen in Step 1. We elaborate on this aspect hereafter.

Depending on the requirements concerning performance, flexibility, additional features and costs, researchers are offered different software solutions to conduct their microsimulation study. Following Li and O'Donoghue (2013), packages to program microsimulation models can be categorized according to their development environment, having pros and cons. General-purpose programming languages (such as C/C++/C#, Python, or Java) offer high flexibility, but also require high programming skills. General-purpose statistical or mathematical packages (such as Stata, SAS, or R) might be less efficient in computing the model, but provide pre-implemented statistical operations that can be applied for simulation. There are also simulation modeling packages that focus exclusively on setting up microsimulations, such as EUROMOD (Sutherland and Figari, 2013), Modgen (Spielauer, 2006; Bélanger and Sabourin, 2017), JAMSIM (Mannion et al., 2012) or LIAM2 (de Menten et al., 2014). These packages are typically less flexible, but easier to use for applied researchers without advanced knowledge in statistical programming.

When creating microsimulations, it is recommendable to use a modular structure as basis for the implementation of population dynamics. Population dynamics are driven by multiple subprocesses that are usually organized independently. Note that an independent organization does not necessarily imply that state transitions within corresponding subprocesses are stochatistically independent. We will address that aspect later in this section. The conceptual distinction between these points can be made according to certain transitions or content groups. O'Donoghue et al. (2009, p. 20) describe modules as "the components where calculations take place, each with its own parameters, variable definitions and self-contained structure, with fixed inputs and outputs."

In a given programming language, the modules may correspond to functions that require the base population as input. Figure 1 shows a four-step process that takes place within an exemplary module for discrete-time dynamic microsimulation. In the first step, the individuals have to be selected regarding their eligibility for a change. This is

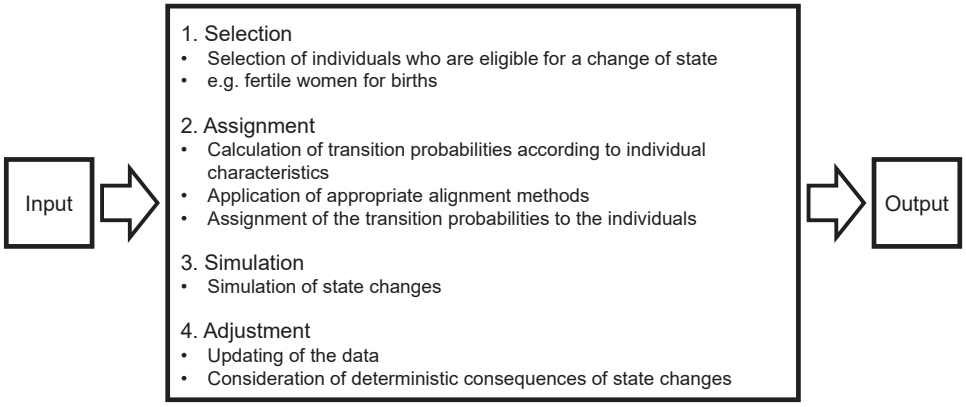

Figure 1: Module structure for a discrete-time dynamic microsimulation

to prevent implausible changes of state and to ensure the consistency of the population. The potential subpopulation for the event of birth includes, for example, women of fertile age. In the second step, the transition probabilities are calculated according to individual combinations of characteristics and linked to the individuals. If external benchmarks are not reached, calibration methods for an adjustment (so-called alignment methods) can be applied. Then, the state of the following period is simulated based on stochastic processes. This part corresponds to the simulation in the actual sense, since the concrete change of state is conducted. Finally, the population is updated according to all direct and indirect consequences of the simulated state changes. It should be noted that the exact structure of a module is individually designed in different models. Likewise, probabilities or transition matrices can serve as module output (O'Donoghue et al., 2009).

The modular structure plays a major role, especially in discrete-time dynamic microsimulations, since the changes of states have to be simulated successively. In continuous-time dynamic microsimulations, however, state changes cannot be determined independently of each other. Therefore, the estimated waiting times in the individual states could be specified as module output. The state changes are then carried out in an extra step after the simulation of all waiting times. While the structure of the simulation in continuous-time variants should not influence the simulation results, it heavily influences the dynamics in discrete-time models. This is briefly demonstrated hereafter. Let Y and X be two random events that may represent state changes within the microsimulation study. There are two different approaches to obtain the joint probability $P(Y,X)$ (Schaich and Münnich, 2001):

$$P(Y,X) = P(Y) \times P(X|Y) = P(X) \times P(Y|X). \tag{1}$$

We see that $P(Y,X)$ can be obtained by means of the conditional probability $P(X|Y)$, but also via the conditional probability $P(Y|X)$. In general, discrete dynamics do not provide

an exact point of time when a given transition is realized. Thus, in the case of two interdependent or competing events, it is necessary to determine which event occurrs first in light of the simulation context. For instance, if the event of a birth is simulated prior to the event of a marriage, the probability of marriage can be conditioned on the event of birth (Burgard et al., 2020). That is to say, the order of the simulation modules has a direct impact on the simulation outcomes. This motivated van Imhoff and Post (1998) to investigate three different strategies for organizing the modular structure. The first strategy is a randomized order of events, which, however, is hardly used in practice. Another possibility is a two-stage simulation of competing events, whereby the first stage simulates the occurrence of at least one of the competing events and the second stage the concrete event. As a third way, the sequence of modules or events of the microsimulation is considered in the modeling process (van Imhoff and Post, 1998).

For the basic functionality of microsimulations, it is generally not necessary to divide the population dynamics into different modules. However, the modularization provides clear practical advantages for the handling and the transparency of the simulation. Modularization allows individual modules to be easily adapted, exchanged and compared. It creates a flexible structure that allows the model to be further developed and adapted for further research questions. Moreover, it also facilitates working on different modules individually as well as in project teams (Lawson, 2014). In addition to that, modularization allows for the inclusion of module-specific debugging devices (O'Donoghue et al., 2009). The module can be written such that potential errors are detectable and precisely displayable. Ideally, the user can be informed about the reason for termination, otherwise at least about the exact position within the module. Additionally, plausibility checks within the modules are a useful extension to ensure data consistency. These checks verify whether the status changes have occurred even in the predefined subpopulation and whether implausible combinations of characteristics occur. Naturally, a modular structure is implemented as standard in many existing microsimulation tools such as LIAM2 (O'Donoghue et al., 2009; de Menten et al., 2014).

Nevertheless, using a modularized simulation structure also has some downsides. As mentioned before, the order of modules directly influences the simulation outcomes. Thus, the segmentation of population dynamics has to be conducted carefully with suitable theoretical justification. What is more, there is an ongoing debate to what extent probability estimation methods that are applied within each of the modules induce systematic errors across modules. For instance, a regression model may produce independent error terms in a given module. Still, these errors may not be independent from the error terms of another module, which may cause inferential problems. Hence, if a modularized structure is implemented, the simulation outcomes have to be carefully investigated with respect to these issues.

3.2. Step 5: Population dynamics

After the module sequence is defined and the modules are created, the dynamics for the projection of $\tilde{U}$ have to be established. They mark the underlying processes that drive the evolution of $\tilde{U}$ over the simulation horizon S. The nature and data requirements for

the projection depend on the type of microsimulation that is chosen in Step 1. In case of static microsimulations, only a few selected population characteristics evolve over time. The projection is often deterministic and can be performed without additional data sources. Typically, a set of scenarios for the selected variables is created. The base population evolves through the interaction of the remaining variables with them (Li and O'Donoghue, 2013).

In case of dynamic microsimulations, the projection is more sophisticated. Since all population characteristics evolve over time, the initialization of multivariate stochastic processes for $\tilde{U}$ is necessary. These processes need to resemble all relevant dynamics of the real population U as closely as possible to allow for genuine simulation outcomes. An essential concept for this purpose is called state transition, which we briefly explain hereafter. Let Y be a population characteristic with J different realizations within the finite state space $\mathscr{Y} = \{Y_1, ..., Y_J\}$. For instance, if a microsimulation on long-term demand is conducted Y may correspond to micro-level care dependency and its realizations could resemble different degrees of care dependency. Based on the theoretical developments of Burgard et al. (2019b), a state transition is defined as follows:

Definition 1 *Let $y_u^{(s)}$ be the realized value of Y for a unit $u \in \tilde{U}$ in period $s \in S$. A state transition is the outcome of a stochastic process where $y_u^{(s+1)} = Y_j$ and $y_u^{(s)} = Y_k$ with $Y_j, Y_k \in \mathscr{Y}$ and $Y_j \neq Y_k$. Its probability is given by $\pi_u^{(s+1)jk} := P(y_u^{(s+1)} = Y_j | y_u^{(s)} = Y_k)$.*

Accordingly, a state transition is a change in the realized value of a population characteristic for a given unit from one simulation period to the next. Recalling the long-term care example, a state transition would then correspond to a change of micro-level care dependency. In light of the previous comments, the probability $\pi_u^{(s+1)jk}$ must be determined such that the overall evolution of $\tilde{U}$ is realistic with respect to U. This is achieved by considering suitable data sources, such as a (panel) survey sample $D \subset U$. If a corresponding data set is available, transition probabilities can be quantified based on statistical models. In the first step, the statistical relation between transition probabilities and observed auxiliary variable realizations is estimated over all sampled individuals $i \in D$. In the next step, $\pi_u^{(s+1)jk}$ is determined via model prediction by using the realized values of the auxiliary variables for $u \in \tilde{U}$ in the simulation period $s+1$.

However, the exact methodology for estimation and projection depends on the concept of time that is chosen in Step 1. Recall that we distinguish between discrete-time and continuous-time dynamic simulations. For discrete-time, the simulation horizon $S := \{1, ..., T\}$ is a finite set of periods, such as months within a year. State transitions can only occur from one period to the next. In this setting, common approaches are generalized linear (mixed) models for the quantification of odds, such as logit models (McCullagh and Nelder, 1989; Greene, 2003). For continuous-time, the simulation horizon $S := [1, T]$ is a closed interval. State transitions may occur at any given point within this interval. In that case, estimation and prediction are performed using survival analysis, for instance via proportional hazard models (Cox, 1972; McCullagh and Nelder, 1989). Further, note that there are also models whose dynamics rely on Markovian processes with infinite state spaces, such as random walks for income simulation (Muennig et al., 2016).

Another important aspect of population projection is the consistency of simulated transition rates in $\tilde{U}$ to observed real-world realization frequencies in U. Let

$$\tau^{(t)k} := \sum_{i \in U} y_i^{(t)k}, \quad y_i^{(t)k} = \begin{cases} 1 & \text{if } y_i^{(t)} = Y_k \\ 0 & \text{else} \end{cases} \tag{2}$$

be the absolute frequency of Y_k in the real population U for a point of time t related to the simulation period $s+1$. Revisiting the long-term care example again, this figure may correspond to the number of individuals in a population that have a specific degree of care dependency. A corresponding figure could be known, for instance, from administrative records. In dynamic microsimulations, it is often the case that

$$\sum_{u \in \tilde{U}} \sum_{j \in \mathscr{Y}} \pi_u^{(s+1)jk} \neq \tau^{(t)k}. \tag{3}$$

The formula indicates that the simulated transition dynamics do not reproduce the empirically observed frequency for Y_k properly. This inconsistency may intensify over subsequent simulation periods and can lead to an implausible evolution of $\tilde{U}$. The latter ultimately causes the simulation outcomes to be not reliable for U, which is the main purpose of microsimulation studies. In order to ensure consistency in this case, so-called alignment methods are often applied (Li and O'Donoghue, 2014). These are (algorithmic) procedures that modify the transition probabilities such that they fit external benchmarks. Recently, several methodologies to achieve this have been proposed. Exemplary approaches are ex-post alignment via logit scaling (Stephensen, 2016) and parameter alignment via constrained maximum likelihood estimation (Burgard et al., 2019b).

3.3. Step 6: Performing the simulation

As dynamic microsimulations are based on stochastic processes, new populations are generated in each simulation run. Especially, if there are many individuals in the base population, it is not often possible to save them separately for each period and simulation run. Still, it is necessary to be able to reproduce the simulation results at any time. When conducting simulation studies, it is common practice to set seeds in order to repeat the random processes. In the sense of open and reproducible research, it is desirable to publish the seeds with the simulation code (Kleiber and Zeileis, 2012). In the case of error messages during the simulation, setting seeds enables the subsequent replication and analysis of the whole process.

Checking for plausibility and possible errors plays an important role not only within modules but also during the entire simulation. In order to identify potential causes in a targeted manner, predefined queries should be implemented at several points during the simulation process. The focus is on the functionality of the combination and interaction of different modules.

4. Output analysis

4.1. Step 7: Analysis of results

A big advantage of microsimulation models is providing information about possible impacts on a population, given the implemented scenarios. These advantages can only be converted into practical use if the analysis of the produced information is done properly. Several aspects have to be taken into account. First, the data has to be analyzed to prevent programming errors or logical errors in the simulation. Second, an uncertainty analysis should be performed to identify different sources of variation. And finally, the output of the simulation has to be analyzed concerning the research question. This includes both, the analysis of the final simulation states but also the processes that lead to the final results. Fourth, the analysis results have to be visualized. The visualization helps to understand the output and provides a good basis for the dissemination of the results.

4.1.1 Programming and logical errors

Even though Step 3 and Step 4 already include several plausibility checks, oftentimes problems in the coding or setup of the simulation only become apparent after a full simulation run. Surprising results may stem from non-linear population dynamics or errors in the code or setup of the simulation. It is therefore of utmost importance to first investigate the results of the simulation to the extent that outcomes seem sensible and the inner logic of the data set are met. If this is not the case, it is necessary to revisit the code and to explore how the results may be explained by the given process.

4.1.2 Uncertainty analysis

One major challenge in microsimulation modeling is the assessment of uncertainty. Typically, when analysing estimates, confidence intervals are calculated to quantify the uncertainty. Especially in dynamic microsimulations, the degree of complexity is high making a simple determination of confidence intervals hardly possible (Lappo, 2012). First of all, the potential sources of uncertainty should be identified. These depend on the type of modeling. Different types of uncertainty in microsimulation models can be distinguished (e.g. Lappo, 2015; Godemé et al. 2013, Sharif et al., 2012):

- Monte Carlo error
- Parameter uncertainty
- Structural uncertainty
- Uncertainty from the base population

The Monte Carlo error is a result of the stochastic processes and therefore occurs especially in case of dynamic microsimulations. However, behavioral changes in static simulations can also cause Monte Carlo errors. Parameter uncertainty is directly linked to

the models on which the microsimulation processes are based. If these are estimated on sample data, they are directly related to sampling uncertainty. Even assumption-based parameters are associated with, in this case subjective, uncertainty (Sharif et al., 2012). Structural uncertainty is primarily due to the type of modeling. This can be the type of estimation of transition probabilities or survival times on the one hand, but also the type of the entire microsimulation on the other hand. Since many microsimulations use survey data as base population, the uncertainty of the sampling must be taken into account. In the case of synthetic data sets, in turn, different sources of uncertainty arise, for example, from underlying data sources, used methods, parameters and stochastic processes in the preparation.

For the consideration of sampling uncertainty in static microsimulations through the application of standard variance estimation techniques, there are already useful examples (Lappo, 2012, Godemé et al. 2013). In the case of dynamic modeling the estimation of confidence intervals is much more difficult due to the complexity of the different sources. Sharif et al. (2012) and Sharif et al. (2017) propose techniques for the estimation of confidence intervals for the consideration of parameter uncertainty in dynamic disease microsimulation models. Petrik et al. (2018) estimate parameter uncertainty for an activity-based microsimulation model.

A possibility for quantifying the influence of various factors on univariate target values is variance-based sensitivity analysis as described in Burgard and Schmaus (2019). Here, the focus is not on estimating confidence intervals, but on measuring and comparing different influencing variables. The influencing variables can be selected variably, but must be pairwise independent. These factors may encompass all inputs that are to some extent wake. The goal of the sensitivity analysis is to attribute to the input factors a certain amount of variation observed in the target variable. For example different choices of scenarios or different parameter modeling strategies. Thus, sensitivity analyses are ideally suited for the selection of influential models for the later determination of confidence intervals. See Saltelli et al. (2008) for a comprehensive study of sensitivity analysis methods.

4.2. Hypothesis evaluation and result visualization

The hypotheses stated in Step 1 have to be checked. After conducting the microsimulation, it is necessary to evaluate whether the hypothesized outcome is a realistic development or not. It is possible to state the probability of the hypothesis to be true given the simulation evolves as the population will evolve. Of course, this condition is rarely possible to assume, and ex-ante, impossible to check in most cases. Especially, if the microsimulation is projecting the population for a long time horizon. The visualization of the results can considerably help the understanding of the simulation. Besides easing the simultaneous consideration of the measures used for the analysis it also helps to communicate the results to third persons and hence is also necessary for the dissemination of the results.

4.3. Step 8: Dissemination

The dissemination stage aims at disclosing knowledge acquired throughout the microsimulation study. To disseminate the study, it must be ensured that planned dissemination products, such as code, data and project reports are updated. In addition, the products must be available in such a way that they are comprehensible to outsiders and comply with legal requirements, such as publication standards.

The main focus of dissemination is to provide all interested parties with open access to resources related to the microsimulation study while respecting intellectual property rights. This includes, in particular, the provision of open access to peer-reviewed scientific publications, to research data and archival facilities for research results (European Commission, 2008). In particular in the case of microsimulation studies, however, open access to data cannot be granted for reasons of data protection and potential property rights to the data. The development of a security concept to guarantee privacy protection is to make data accessible through a research data center.

Additional dissemination strategies include the presentation of project research at conferences, organization of workshops and maintenance of a project website providing information about the project in general, conference contributions and publications related to the project. A project website also offers the possibility of setting up a mailing list to keep the interested public up to date. Furthermore, there are also associations such as the *International Microsimulation Association* that specifically aim at the dissemination of knowledge in the area of microsimulation (e.g. IMA, n.d.). For all dissemination strategies, especially when providing code and data, it is essential to have a contact person who accepts inquiries and supports users in the case of problems.

4.4. Step 9: Evaluation

The evaluation assesses all steps of the microsimulation study. It can be conducted either at the end or on an ongoing basis. The evaluation is based on the information gathered at the various steps and takes the experience from users, contributors and researchers into account. Continuously collected quality indicators are compiled to assess the quality of the individual preceding steps of the microsimulation study. Some steps, however, require specific measures such as the use of questionnaires to obtain information on the user-friendliness of the microsimulation study or to assess the effectiveness of the chosen dissemination strategies. As a result of the evaluation, an action plan is agreed upon. The implementation of the adopted actions will then again be part of the next round of evaluation (UNECE, 2013).

The complete business process model for conducting microsimulation studies is summarized in Figure 2.

5. Conclusion

Microsimulation methods play a more and more important role in policy support as well as in economic and social research. Major emphasis by now was laid on developing

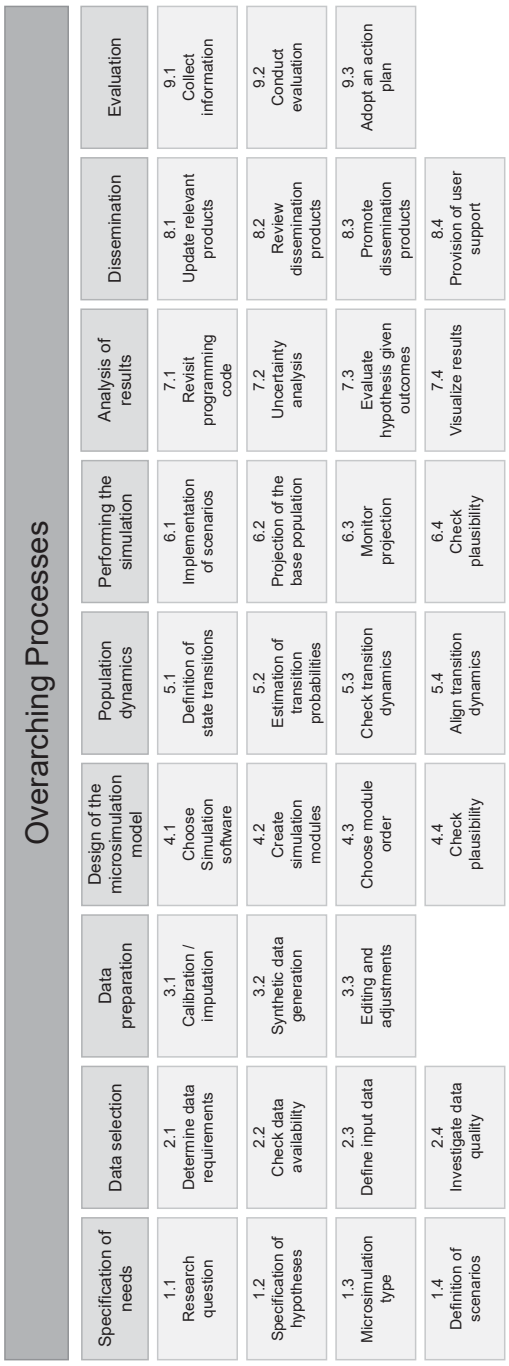

Figure 2: Generic business process model

important applications in many different areas. Less attention was put on the entire statistical production process. This becomes essentially important since the accuracy of the microsimulation heavily depends on data availability, data use, the core simulation, as well as the analysis considering all preceding steps.

The present article provided a general view of implementing a statistics business model that includes the different steps that have to be considered to establish an accurate microsimulation. The proposed model is based on UNECE (2013) on behalf of the international statistical community aiming at providing a general procedure that is widely accepted in the international statistical system. Further, it furnished the implementation of open and reproducible microsimulations as research and policy tool.

Acknowledgements

This research was conducted within the research group FOR 2559 *Sektorenübergreifendes kleinräumiges Mikrosimulationsmodell (MikroSim)*, which is funded by the German Research Foundation. We kindly thank for the financial support.

REFERENCES

ALFONS, A., FILZMOSER, P., HULLIGER, B., KOLB, J. P., KRAFT, S., MÜNNICH, R., TEMPL, M., (2011a). Synthetic data generation of SILC data. AMELI Research Project Report WP6-D6, 2.

ALFONS, A., KRAFT, S., TEMPL, M., FILZMOSER, P., (2011b). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20(3), pp. 383–407.

BÉLANGER, A., SABOURIN, P., (2017). Microsimulation and Population Dynamics: An Introduction to Modgen 12. Springer.

BOURGUIGNON, F., SPADARO, A., (2006). Microsimulation as a tool for evaluating redistribution policies. *The Journal of Economic Inequality*, Vol. 4, pp. 77–106.

BURGARD, J. P., KRAUSE, J., MERKLE, H., MÜNNICH, R., SCHMAUS, S., (2019a). Conducting a dynamic microsimulation for care research: Data generation, transition probabilities and sensitivity analysis. In *Stochastic Models, Statistics and Their Applications*. A. Steland, E. Rafajłowicz and O. Okhrin (eds.) Cham: Springer International Publishing, pp. 269–290.

BURGARD, J. P., KRAUSE, J., SCHMAUS, S., (2019b). Estimation of regional transition probabilities for spatial dynamic microsimulations from survey data lacking in regional detail. *Research Papers in Economics*, No. 12/19, Trier University.

BURGARD, J. P., KRAUSE, J., MERKLE, H., MÜNNICH, R., SCHMAUS, S., (2020). Dynamische Mikrosimulationen zur Analyse und Planung regionaler Versorgungsstrukturen in der Pflege. In *Mikrosimulationen - Methodische Grundlagen und ausgewählte Anwendungsfelder*. M. Hannappel and J. Kopp (eds.) Wiesbaden: Springer VS, pp. 283–313.

BURGARD, J. P., MÜNNICH, R. T., RUPP, M., (2019c). A generalized calibration approach ensuring coherent estimates with small area constraints (No. 10/19). *Research Papers in Economics*.

BURGARD, J. P., SCHMAUS, S., (2019). Sensitivity analysis for dynamic microsimulation models (No. 15/19). Research Papers in Economics, Trier University.

CHEN, J., QIN, J., (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, Vol. 80, pp. 107–116.

COX, D. R., (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 9, pp. 439–455.

DE MENTEN, G., DEKKERS, G., BRYON, G., LIÉGEOIS, P., O'DONOGHUE, C., (2014). Liam2: a new open source development tool for discrete-time dynamic microsimulation models. *Journal of Artificial Societies and Social Simulation*, Vol. 17, p. 9.

DEKKERS, G., (2015). The simulation properties of microsimulation models with static and dynamic ageing – a brief guide into choosing one type of model over the other.*International Journal of Microsimulation*, Vol. 8, pp. 97–109.

DEKKERS, G., CUMPSTON, R., (2012). On weights in dynamic-ageing microsimulation models. *The International Journal of Microsimulation*, Vol. 5(2), pp. 59–65.

DEVILLE, J., SÄRNDAL, C., (1992). Calibration estimators in survey sampling.*Journal ofthe American Statistical Association*, 87, pp. 376–382.

DRECHSLER, J., (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Vol. 201. Springer Science & Business Media.

EUROPEAN COMMISSION, (2008). Commission Recommendation of 10 April 2008 on the management of intellectual property in knowledge transfer activities and Code of Practice for universities and other public research organisations (notified under document number C(2008) 1329) (Text with EEA relevance). *Official Journal of the European Union (L 146)*, Vol. 51, pp. 19–24.

FAVREAULT, M. M., SMITH, K. E., JOHNSON, R. W., (2015). The dynamic simulation of income model (DYNASIM). Research Report at Urban Institute, Washington DC.

GOEDEME, T., VAN DEN BOSCH, K., SALANAUSKAITE, L., VERBIST, G., (2013). Testing the statistical significance of microsimulation results: A plea. International Journal of Microsimulation, 6(3), pp. 50–77.

GREENE, W. H., (2003). *Econometric analysis* (5 ed.) New Jersey: Prentice Hall.

HANNAPPEL, M., TROITZSCH, K. G., (2015). Mikrosimulationsmodelle. In N.Braun, N.J.Saam (eds): *Modellbildung und Simulation in den Sozialwissenschaften,* (pp. 455–489). Springer VS, Wiesbaden.

HAZIZA, D., BEAUMONT, J. F., (2017). Construction of weights in surveys: A review. *Statistical Science*, Vol. 32, pp. 206–226.

HUANG, Z.; WILLIAMSON, P., (2001). A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata. University of Liverpool. Department of Geography. Working Paper 2001/2.

KLEIBER, C., ZEILEIS, A., (2013). Reproducible econometric simulations. *Journal of Econometric Methods*, Vol. 2, pp. 89–99.

LAPPO, S., (2015). Uncertainty in microsimulation, Master's Thesis, University of Helsinki.

LI, J., O'DONOGHUE, C., (2013). A survey of dynamic microsimulation models. Uses, model structure and methodology. *International Journal of Microsimulation*, Vol. 6, pp. 3–55.

LI, J., O'DONOGHUE, C., (2014). Evaluating binary alignment methods in microsimulation models. *Journal of Artificial Societies and Social Simulation*, Vol. 17, pp. 1–15.

LOVELACE, R., DUMONT, M., (2016). Spatial microsimulation with R. Chapman and Hall/CRC.

MANNION, O., LAY-YEE, R., WRAPSON, W., DAVIS, P., PEARSON, J., (2012). JAMSIM: A microsimulation modelling policy tool. *Journal of Artificial Societies and Social Simulation*, Vol. 15, p. 8.

MARKHAM, F., YOUNG, M., DORAN, B., (2017). Improving spatial microsimulation estimates of health outcomes by including geographic indicators of health behaviour: The example of problem gambling. *Health & Place*, Vol. 46, pp. 29–36.

MCCULLAGH, P., NELDER, J. A., (1989). *Generalized linear models* (2 ed.), Vol. 37 of *Monographs on Statistics and Applied Probability* London: Chapman and Hall.

MUENNIG, P.A., MOHIT, B., WU, J., JIA, H., ROSEN, Z., (2016). Coest effectiveness of the earned income tax credit as health policy investiment. *American Journal of Preventive Medicine*, Vol. 51(6), pp. 874–881.

MÜNNICH R, SCHÜRLE J., (2003). On the simulation of complex universes in the case of applying the GermanMicrocensus. DACSEIS research paper series No. 4, University of Tübingen.

MURTHY, D., (2012). Towards a sociological understanding of social media: theorizing Twitter. *Sociology*, Vol. 46(6), pp. 1–15.

O'DONOGHUE, C., (2001). Dynamic Microsimulation: A Methodological Survey. *Brazilian Electronic Journal of Economics*, Vol. 4, p. 77.

O'DONOGHUE, C., LENNON, J., HYNES, S., (2009). The Life-cycle Income Analysis Model (LIAM): a study of a flexible dynamic microsimulation modelling computing framework. *International Journal of Microsimulation*, Vol. 2, pp. 16–31.

O'DONOGHUE, C., DEKKERS, G., (2018). Increasing the impact of dynamic microsimulation modelling. *International Journal of Microsimulation*, Vol. 11, pp. 61–96.

ORCUTT, G. H., (1957). A new type of socio-economic system. *The review of economics and statistics*, 58, pp. 116–123.

PETRIK, O., ADNAN, M., BASAK, K., BEN-AKIVA, M., (2018). Uncertainty analysis of an activity-based microsimulation model for Singapore. Future Generation Computer Systems.

PICHON-RIVIERE, A., AUGUSTOVSKI, F., BARDACH, A., COLANTONIO, L., (2011). Development and validation of a microsimulation economic model to evaluate the disease burden associated with smoking and the cost-effectiveness of tobacco control interventions in Latin America. *Value in Health*, Vol. 14, S51–S59.

RAHMAN, A., HARDING, A., (2017). Small area estimation and microsimulation modeling. Boca Raton: CRC Press, Taylor & Francis Group.

SALTELLI, A., RATTO, M., TERRY, A., CAMPOLOGNO, F., CARIBONI, J., GATELLI, D., SAISANA, M., TARANTOLA, S., (2008). Global sensitivity analysis. The Primer. Chichester: John Wiley & Sons.

SÄRNDAL, C.-E., (2007). The calibration approach in survey theory and practice. *Survey Methodology*, Vol. 33, pp. 99–119.

SCHAFER, J.L., GRAHAM, J. W., (2002). Missing data: Our view of the state of the art. *Psychological Methods*, Vol. 7, pp. 147–177.

SCHAICH, E., MÜNNICH, R., (2001). Mathematische Statistik für Ökonomen. Vahlen.

SHARIF, B., KOPEC, J. A., WONG, H., FINES, P., SAYRE, E. C., LIU, R. R., WOLFSON, M. C., (2012). Uncertainty analysis in population-based disease microsimulation models. Epidemiology Research International, 2012.

SHARIF, B., WONG, H., ANIS, A. H., KOPEC, J. A., (2017). A practical ANOVA approach for uncertainty analysis in population-based disease microsimulation models. *Value in Health*, Vol. 20(4), pp. 710–717.

SMITH, D.M., CLARKE, G.P., HARLAND, K., (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A: Economy and Space*, Vol. 41, pp. 1251–1268.

SPIELAUER, M., (2006). The "Life Course" model, a competing risk cohort microsimulation model: source code and basic concepts of the generic microsimulation programming language Modgen, MPIDR WORKING PAPER 2006–046.

SPIELAUER, M., (2009). Microsimulation approaches. Technical Report, Statistics Canada, Modeling Division.

SPIELAUER, M., (2011). What is Social Science Microsimulation? *Social Science Computer Review*, Vol. 29, pp. 9–20.

STEPHENSEN, P., (2016). Logit scaling: A general method for alignment in microsimulation models. *International Journal of Microsimulation*, Vol. 9, pp. 89–102.

SUTHERLAND, H., FIGARI, F., (2013). EUROMOD: the European Union tax-benefit microsimulation model. *International Journal of Microsimulation*, Vol. 6, pp. 4–26.

TANTON, R., (2018). Spatial microsimulation: Developments and potential future directions. *International Journal of Microsimulation*, Vol. 11(1), pp. 143–161.

TANTON, R., WILLIAMSON, P., HARDING, A., (2014). Comparing two methods of reweighting a survey file to small area data. International Journal of Microsimulation, 7(1), pp. 76–99.

UNECE, (2013). Generic statistical business process model. Version 5.0 – December 2013. The United Nations Economic Commission for Europe (UNECE). URL: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0.

VAN IMHOFF, E., POST, W., (1998). Microsimulation methods for population projection. *Population: An English Selection*, Vol. 10, pp. 97–138.

WILLEKENS, F., (2017). Continuous-time microsimulation in longitudinal analysis. In *New Frontiers in Microsimulation Modelling*. A. Zaidi, A. Harding and P. Williamson (eds.), Routledge, pp. 413–436.

WILLIAMSON, P., (2013). An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation. In In Tanton and Edwards (eds): *Spatial microsimulation: A reference guide for users*. Springer, Dordrecht.