SSOAR

Open Access Repository
www.ssoar.info

A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model

Münnich, Ralf; Schnell, Rainer; Brenzel, Hanna; Dieckmann, Hanna; Dräger, Sebastian; Emmenegger, Jana; Höcker, Philip; Kopp, Johannes; Merkle, Hariolf; Neufang, Kristina; Obersneider, Monika; Reinhold, Julian; Schaller, Jannik; Schmaus, Simon; Stein, Petra

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Münnich, R., Schnell, R., Brenzel, H., Dieckmann, H., Dräger, S., Emmenegger, J., ... Stein, P. (2021). A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model. *Methods, data, analyses : a journal for quantitative methods and survey methodology (mda)*, *15*(2), 241-264. https://doi.org/10.12758/mda.2021.03

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

methods, data, analyses | 2021, Vol. 15(2), pp. 241-264 DOI: 10.12758/mda.2021.03

A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model

Ralf Münnich[1], *Rainer Schnell*[2], *Hanna Brenzel*[3], *Hanna Dieckmann*[1], *Sebastian Dräger*[1], *Jana Emmenegger*[3], *Philip Höcker*[2], *Johannes Kopp*[1], *Hariolf Merkle*[1], *Kristina Neufang*[1], *Monika Obersneider*[2], *Julian Reinhold*[2], *Jannik Schaller*[3], *Simon Schmaus*[1], *& Petra Stein*[2]

[1] *Trier University, Germany*

[2] *University of Duisburg-Essen, Germany*

[3] *Federal Statistical Office, Germany*

Abstract

Microsimulation models are widely used to evaluate the potential effects of different policies on social indicators. Most microsimulation models in use operate on a national level, disregarding regional variations. We describe the construction of a national microsimulation model for Germany, accounting for local variations in each of the more than 10,000 communities in Germany. The database used and the mechanisms implementing the population dynamics are described. Finally, the further development of the database and microsimulation programs are outlined, which will contribute towards a research lab that will be made available to the wider scientific community.

Keywords: microsimulation methods, spatial microsimulation, social simulation, multivariate modelling, multi-source modelling, synthetic data generation

© The Author(s) 2021. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

What are the effects of changes in the demographic profile of the population on family formation processes? How does tax legislation impact on tax revenues? Do changes in women's employment alter the way in which the elderly are cared for? How do different qualifications and work biographies influence pensions? How will integration patterns develop in the future in the context of demographic transition regarding upcoming generations of migrants? These are some examples of problems studied using microsimulation techniques in past decades (Stein & Bekalarczyk, 2016; O'Donoghue & Dekkers, 2018; Schnell & Handke, 2020; Zwick & Emmenegger, 2020).

However, due to limited data availability and high computational complexity, such simulations were mostly done at national levels of analysis. Given the fact that social and economic change is often different for some regions and subgroups of the population, the demand for detailed modeling is increasing. Therefore, highly detailed regional datasets covering the whole population are needed. A prime example is schooling: The demand for elementary schools varies at the local level. For secondary schools, increasing population heterogeneity requires complex school planning which accounts for the diversity of schooling demands. A second example is care for the elderly, where information regarding the distance between parents and their children, who are potential care providers, is essential for modeling the demand for care.

In the Netherlands or the Scandinavian countries, population covering register datasets could provide data for regional microsimulations. In Germany, however, hardly any datasets are available that foster regional microsimulation modeling. Either the sample size is too small for regional models or data protection regulations do not allow the use of low-level identifiers in population covering datasets. Therefore, modeling tasks requiring low-level information are challenging. The MikroSim model described in this paper is aiming for a dynamic microsimulation of Germany down to each municipality.[1] The model is based on a highly detailed dataset build from many different sources, such as survey, administrative, and other data. Therefore, a series of different methods of data-integration and small area

1 The corresponding MikroSim-project, in which this model was developed and that is funded by the German research association, is described in detail in Münnich, Schnell, et al., 2020.

Acknowledgements

The research unit FOR 2559 MikroSim is funded by the German Research Foundation. The principal investigators are Ralf Münnich (Speaker), Rainer Schnell (Co-Speaker), Johannes Kopp and Petra Stein. The research unit cooperates with the German Federal Statistical Office, with Hanna Brenzel as primary contact, succeeding Markus Zwick.

Direct correspondence to

Ralf Münnich, Trier University
E-mail: muennich@uni-trier.de

estimation have been used for building the dataset. Modeling the dynamic processes is achieved by estimating transition probabilities using different statistical methods. The MikroSim model consists of different modules, simulating births, deaths, marriage, education, and other dynamic processes.

The structure of the present paper is as follows: We begin with a short overview on the history and the different approaches to microsimulations in general in section *Microsimulation Modeling*. Subsequently, we present *The MikroSim Model*. The subchapters of this section include the generation of the synthetic dataset of the German population as the base dataset of the model, the construction principles, the sequence in which the simulation modules are ordered and an overview on the specific conceptualization of each module. Some examples of modules that are to be implemented are described in section *Application Modules*, as well as a short examples on the questions that can be answered using microsimulation methods. A summary and an outlook on future developments concludes the paper.

Microsimulation Modeling

The beginning of microsimulation in economics and social sciences dates back to the 1950s when Guy H. Orcutt published the paper "A New Type of Socio-Economic System" (Orcutt, 1957). He criticizes the limited usefulness of macrosimulations due to the focus on aggregates and the inability to consider nonlinearities and discontinuities in individual behavior. He advocates a new type of modeling that focuses directly on micro-units, such as individuals, households, and firms. *This new type of model consists of various sorts of interacting units which receive inputs and generate outputs. The outputs of each unit are, in part, functionally related to prior events and, in part, are the result of a series of random drawings from discrete probability distributions* (Orcutt, 1957).

Thus, the main focus of microsimulation is to look at the smallest unit of a system. Li and O'Donoghue (2013) describe microsimulations as *a tool to generate synthetic micro-unit based data, which can then be used to answer many "what-if" questions that, otherwise, cannot be answered.* These questions are usually understood as the investigation of different scenarios, such as different social and tax systems and behavioral assumptions. Contrary to macro simulations, not only single target values but complex interrelations and distributions within the system can be investigated. According to Li and O'Donoghue (2013), microsimulations can be split into two tasks. The first step is to generate a high-quality dataset with the relevant variables of interest in the necessary geographic depth. In a second step, a set of scenarios is performed on this dataset in order to answer the what-if questions.

The basis of any microsimulation is a dataset – the so-called base population – that contains micro-level information about the system of interest. Socioeconomic questions usually require information on individuals and households. Due to the easy availability and large amount of information, survey datasets are mainly used as base populations (Li & O'Donoghue, 2013; Burgard, Dieckmann et al., 2020). However, this kind of data contains only a relatively small number of individuals and allows only very limited regionalized analyses. Larger datasets from administrative sources and census data often contain a limited set of variables. Hence, (partially) synthetic base populations have been increasingly used for regionalized models recently. The methods used to create small-scale datasets are often described as small-area or spatial microsimulation techniques (Tanton, 2014; Rahman & Harding, 2016).

Another distinction in microsimulations relates to the temporal component. In static microsimulations, there are usually no changes in individual states during the course of the simulation. The immediate distributional impact of (political) changes is evaluated without reference to the time dimension. In this case, it is assumed that the characteristics of the population of interest do not change rapidly (Merz, 1991). Thus, this kind of modeling is primarily suitable for short- and medium-range predictions. To implement a temporal component, re-weighting and uprating/deflating techniques can be implemented. In a re-weighting process, the survey weights are calibrated to exogenously given aggregate data of another time period while uprating/deflating changes the specific variables (for example specific income components) directly (Merz, 1991; Sutherland, 2018).

In dynamic models, micro-units interact and evolve over a temporal horizon. This type of simulation can account for micro-level dependencies and complex interaction allowing long-term projections and time-dependent behavior simulations. The focus is on sophisticated ceteris paribus analyzes over time under an approximation to real-world complexity. The so-called ageing process can either be continuous or discrete. In discrete time models, the base population is aged considering discrete – mainly annual or monthly – time intervals and events are realized in each period using transition probabilities. Continuous microsimulations, on the other hand, allow events to occur at any point in time until the simulation horizon is reached. Instead of transition probabilities, the simulation is usually based on survival analysis (Li & O'Donoghue, 2013; Burgard, Dieckmann et al. 2020).

For a more detailed methodological differentiation of dynamic microsimulation models, we refer to Li and O'Donoghue (2013) and Hannappel and Kopp (2020).

The focus in this paper is on discrete time dynamic microsimulations. The simulated events can either be deterministic or stochastic. Deterministic changes of states are, for example, the ageing of individuals in each period or the loss of income after the termination of employment. However, dynamic microsimulations

are usually characterized by the fact that all events depend, directly or indirectly, on one or more stochastic processes. The simplest way to simulate changes is based on first-order Markov processes, where the occurrence of an event depends exclusively on the state of the previous period. The probabilities are organized in transition matrices, which are usually differentiated according to socio-demographic and socio-economic characteristics. Transition matrices can be easily estimated using conditional distributions. However, the most common way to obtain individual transition probabilities is to estimate logit (multinomial) regression models based on panel data. In the models, the dependent variable can either be conditioned to the state of the previous period or the lagged variable can be included directly. The individual transition probabilities are predicted in the simulation process using the estimated model parameters. The simulation of state changes within the simulation is usually organized in modules. In each period, all individuals run through each module in a fixed order. Within each module, specific events are simulated. A module can be understood as a function which uses the base population as input and returns the updated population (Burgard, Krause, Merkle et al., 2019). A more detailed explanation of the estimation process of transition probabilities and the simulation process can be found in Burgard, Krause, Merkle et al. (2019) and Burgard, Krause, and Schmaus (2020).

The MikroSim Model

Base Data

A synthetic dataset is used as the base dataset for the microsimulation model MikroSim. In general, the purpose of synthetic datasets is to mimic a non-accessible or non-existing dataset so that the relevant characteristics of a synthetic dataset matches the characteristics of the underlying population as close as possible (cf. Münnich & Schürle, 2003; Münnich, Gabler et al., 2012; Kolb, 2013; Alfons, Kraft et al., 2011). The characteristics to be matched are distributional parameters, correlations, cluster effects, and totals.

The generated dataset is based on an anonymized national register of residents that has been used for methodological research for the census 2011. The German population with respect to all 11,339 municipalities is modeled.[2]

Since the register of residents contains only a few variables, additional data is generated. Since most data stems from German official statistics, data evaluation methods are performed within the statistical office to ensure confidentiality and privacy or, alternatively, using scientific use files. No record linkage between

2 A previous project (REMIKIS) modeled the region Trier using a similar modeling strategy (Burgard, Krause, Merkle et al., 2019).

official microdata and the synthetic dataset or microdata from other sources takes place. The generation of additional variables is formally defined recursively as

$$f(x, y, z) = f(x) \cdot f(y|x) \cdot f(z|x, y)$$

The variable set x contains the basic demographic variables such as age, gender and marital status. Further sets of variables are included in the German Microcensus including x to provide the information on the conditional distribution $f(y|x)$. A second block of variables y comprises variables related to education and activity status. Further variable sets z include variables of special interest for MikroSim, such as care or migration related topics. The conditional distributions are, in general, modeled using the multinomial logit models (Alfons, Filzmoser et al., 2011; Kolb, 2013). For the data synthesis, cluster effects resulting from a positive correlation between household members are considered. The variables for the household members are generated considering a household type variable to account for these cluster effects.

Finally, the synthetic population for each municipality is adjusted to published census totals. Using simulated annealing (cf. Laarhoven & Aarts, 1987; Huang & Williamson, 2001; Williamson, 2012; Tanton, 2014), households are selected randomly and entered or deleted sequentially to minimize the differences between the synthetic population and the census totals.

Construction Principles and Module Ordering

The MikroSim model is designed as a closed population simulation model. Therefore, modules simulating the paths for individuals entering or leaving is key to obtaining realistic projections of the population. The modules providing these paths are mainly the modules Mortality, Births, and Regional Mobility, which are later described in more detail. The simulation model is implemented in R.

Most of the modules in the MikroSim model are based on statistical models to estimate individual transition probabilities for the micro-units. Mainly used to determine these individual transition probabilities are for example data tables as well as multinominal and binary logit regression models (Burgard, Krause, Merkle et al., 2020). Regional differentiations as well as rural-urban disparities are modeled using adequate auxiliary variables within the models.

The characteristics of the simulated population are updated once for each simulated year. Therefore, no information about the exact time of occurrence of an event within the simulated period is available. However, the occurrence of one event might determine other events (for example, a death triggers further changes). There are different strategies available to deal with such dependencies (van Imhoff & Post, 1998). In MikoSim, probabilities for many events are only estimated for those persons who are eligible for a change of state. The eligibility is modeled by

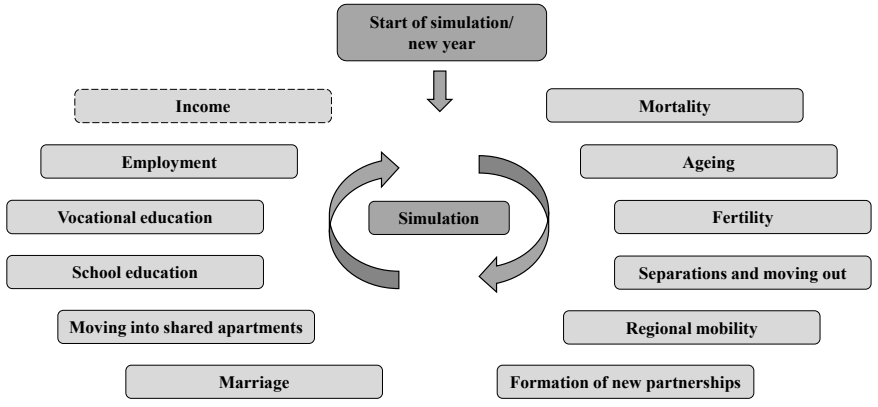

Figure 1 MikroSim Module Sequence

the ordering of the modules. For example, as the event of birth is simulated prior to the event of a marriage, the probability of marriage can be conditioned on the event of a birth (Burgard, Krause, Merkle et al., 2020). It is also possible to start with the marriage event and use this information for the prediction of the birth. From a theoretical perspective, not the order of modules but the modeling strategy is crucial for the simulation. Let $f(x, y)$ be the conditional distribution of the events birth $f(x)$ and marriages $f(y)$. There are two possibilities to reach the joint distribution:

$$f(x, y) = f(x) \cdot f(y|x) = f(y) \cdot f(x|y).$$

Nevertheless, it is always possible to take states from the previous time period into account in the estimation process. Regarding the example above, this means that the marital status in the current period can influence the transition probabilities for the birth of a child in the next period.

The sequence in which the modules are processed is shown in Figure 1.

The simulation of state transitions is conducted using random draws from the predictive distribution of the variables of interest (inversion method). First, the cumulative individual transition probabilities are calculated. Then, a uniformly distributed random number is drawn for each individual and the state is set to the value in which interval the random number lies. For example, let the transition probabilities for a full-time employed person to remain in this state be 0.70, to change into a part-time employment 0.2 and to get unemployed 0.1. The cumulative probabilities are $\{0.7, 0.9, 1\}$ and the random number is 0.83. Consequently, the person changes to a part-time job as $0.83 \in [0.7, 0.9)$.

One major challenge in dynamic microsimulation is in the fact that the transition probabilities are mainly estimated using sample data. Because of data limitations, the number of estimated events often does not match known benchmarks. In addition to the small number of observations, regional differentiations often cannot be made for data protection reasons. This problem is common in the field of microsimulation modeling since *no country has the ideal dataset for [...] estimating the parameters of all the processes in a dynamic microsimulation model* (Bækgaard, 2002). However, the application of small area methods are applied to provide accurate regional benchmarks (Rao & Molina, 2015; Münnich, Burgard, & Vogt, 2013).

To harmonize the individual transition probabilities with the known benchmark values on a macro level, alignment methods are applicable. In the context of dynamic microsimulations, various methods to adjust the transition probabilities or the number of transitions are available (Bækgaard, 2002; Li & O'Donoghue, 2014; Klevmarken, 2008; Stephensen, 2016; Burgard, Krause, & Schmaus, 2020). However, the methods differ considerably with regard to their applicability and functionality. A simple and well performing method is logit-scaling, where the transition probabilities are calibrated to a benchmark using a bi-proportional scaling algorithm. The solution corresponds in a logit framework to the adjustment of the intercept and leads to a solution which minimizes the Kullback-Leibler divergence between the estimated and calibrated probabilities (Stephensen, 2016). Klevmarken (2008) suggests a method to align the parameters by minimizing the quadratic difference between the estimated and adjusted values weighted by the inverse variance-covariance matrix. A constraint likelihood approach where the parameters are aligned while maximizing the original likelihood function aslo shows very good results (Burgard, Krause, & Schmaus, 2020). In the first version of the simulation, alignment is conducted using logit-scaling. Currently, other methods are also being implemented and can be applied via function arguments.

The first module in the MikroSim model simulates widowhood for married persons not living in the same household with their spouse. Widowhood directly influences the possibility of the respective persons to enter the trailing modules such as separations or the formation of new partnerships. Updating this relationship status at the beginning can prevent an underestimation of widowhood within separated couples.

The Mortality module is placed prior to the Aging module since also newborns face a non-zero mortality risk. The event of death depends exclusively on age and gender. Following the Aging module, births are simulated as the first way to add new individuals to the population. The position of the Birth module at this early stage of the simulation is required since birth decisions must precede the birth event. Thus, birth probabilities are estimated mainly based on the characteristics in the previous period.

The Leaving Household module then simulates (1) relocations of adults from parental households, (2) shared apartments, and (3) dissolutions of households after separations. Since these relocation events often have a direct impact on regional mobility, migration across district borders is simulated subsequently.

New households are formed by creating new partnerships and shared apartments. Thus, persons can directly form new households within a simulation period after leaving a household or immigrating. Formal changes in the relationship status (divorces and marriages) are simulated separately.

Changes in school education, vocational education, and employment status are the final modules in the simulation process. An income module will be integrated soon.

The modules are based on a variety of different models, modeling methods, and datasets. Table 1 gives a brief overview, including the choice of independent variables.[3] The following sections describe the modules in more detail.

Modules

Mortality

The Mortality module is the first step in the simulation process. The probabilities for death are assigned according to sex and age of the simulated person using the life tables published by the Federal Statistical Office of Germany (Statistisches Bundesamt, 2020a). The event of death does not only affect the size of the population but also individual and household characteristics. For instance, a partner's family status has to be updated if the husband or wife dies and underage orphans living alone after a parent dies must be assigned to new households. In addition, widowhood for married persons who do not live in the same household cannot be updated deterministically and therefore has to be simulated. This model is based on the German Microcensus (Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, 2018a; 2018b; 2018c).

After removing the deceased individuals from the simulation and updating the family status, the age of all remaining units is increased by one year.

Fertility

The Fertility module in MikroSim simulates births in two steps. In the first step, for all women of fertile age (15–49 years) a probability of giving birth is estimated. The model uses individual characteristics of the women as well as characteristics of

3 A detailed explanation of the mechanisms sketched in Table 1 will be the subject of a different publication.

Table 1 Overview of the MikroSim Modeling approaches

Modules	Method	Y	X
Mortality	RF	Death	age; sex
	LR	Widowhood (not in cohabitation	age; age²; sex; working
	D	Widowhood (in cohabitation)	death of the partner
Ageing	D	Age	$age_{t+1} = age_t + 1$
Fertility	LR	Birth (woman between 15 and 49 years)	age; age²; school education; vocational education; working; working (partner); marriage; age of children in household; Eastern Germany
	LR	Twin birth	age
Seperations and Moving out	LR	Leaving cohabitation/marriage (up to 60 years of age)	age; age difference; age of children in household; marriage; school education (both partners); vocational education (both partners); citizenship homogamy; working; Eastern Germany
	LR	Leaving parental home (18 years and older)	age; sex; age² sex; sex; school education; vocational education; marriage; citizenship; Eastern Germany
	LR	Leaving a shared apartment (no relationship within household; from 18 to 35 years)	age; sex; age² sex; sex; school education; vocational education; marriage; citizenship; Eastern Germany
Regional Mobility (between districts)	RF	Outflow	age; sex; marriage
	RF	Inflow	age; sex; marriage
Formation of New Partnerships	LR	Cohabitation (separate models for men and women living alone; from 18 years)	children in household; children under 6 in household; age; age²; separated; widowed; school education; vocational education; working; Eastern Germany
	M	Finding a partner	matching on age, education, citizenship

Modules	Method	Y	X
Marriage	LR	Marriage (not married cohabitations)	age; age² (women); family status (both partners); ISCED-Level (both partners); working (both partners); children under 1 in household; Eastern Germany
Moving into shared Apartments	LR	Moving into a shared apartment (one-person household; from 18 to 35 years)	age; age²; sex
School Education	RF	Age of enrollment (5-year-olds)	federal states; sex
	MR	No qualification; lower secondary; secondary; technical college level; university entrance level	age; sex; school education (parents); vocational education (parents); working (parents); citizenship (parents); single-parent household
Vocational Education	MR	Enrolled; no vocational training; vocational training; tertiary	working (parents); vocational education (parents)
Employment	MR	Working; unemployed; inactive (from 15 to 65 years)	age; age²; sex; school education; vocational education; mar riage; citizenship; federal states; $working_{(t\ 1)}$; children in household; children under 3 in household
	LR	Full-time employment	age; age²; sex; school education; vocational education; marriage; citizenship; federal states; $working_{(t\ 1)}$; children in household; children under 3 in household;

RF	Relative Frequencies	OS	Official Statistics
LR	Logistic Regression	GM	German Microcensus
D	Deterministic	SOEP	German Socio-Economic Panel
M	Matching	AIDA	Growing up in Germany (Aufwachsen in Deutschland: Alltagswelten, AID:A)
MR	Multinomial Regression	NEPS	National Education Panel Study

other people living in the household, such as the employment status of a potentially existing partner or the age of the youngest child (see Table 1).[4]

In the second step, conditional on the event of a birth, twin births are simulated.[5] The estimation model includes only the age of the women. The sex of a simulated child is assigned in accordance with the known sex distribution of newborns. All other variables of a simulated person are initialized to reasonable values (such as age to zero and school or vocational qualifications to missing values).

Since the model is based on sample data lacking regional details, differences between observed and simulated birth rates may result for some districts. Therefore, the model for birth is calibrated to known birth rates of the German districts up to the last available data (Statistisches Bundesamt, 2019a).

Separations and Moving Out

In the MikroSim model, the events of individuals leaving households are simulated by three mechanisms: (1) separating from a partner, (2) leaving of the parental household and (3) moving out of a shared apartment. Modeling these three transitions using survey data is quite challenging, as individuals are either not tracked over time at all (e.g., in the German Microcensus) or only for subgroups (e.g., in the German Socio-Economic Panel, Goebel et al., 2019). Since the person moving out is usually only observed before a change of residential status occurs, identifying the cause of a departure is difficult using available data.

The first mechanism simulates the separation from a partner as ending a cohabitation. Therefore, only persons living in a partnership with cohabitation are considered. The probabilities are estimated with a logit model based on longitudinal Microcensus data for the years 2012 to 2014 using information of the partners and their partnership (such as age difference, for details see Table 1). In the case of a separation, new households are formed. These new households are initially single or multiple-person households if children are present. Currently, children are assigned to the mother, but future versions of the simulation will include predictive models for assigning children to new households.

The second mechanism models leaving the parental home. Only persons who are at least 18 years of age and still living in the parental household are considered. The probabilities are estimated with a logit model based on the same data as the mechanism described above. As predictors, age, the current level of education or vocational training and the relationship status are used (cf. Table 1). After leaving their parental homes, the individuals initially form new single-person households.

4 We plan to extend this model to include more independent variables such as nationality and regional information.

5 The birth of triplets or more children is not simulated due to the small number of cases.

The third mechanism models moving out of shared apartments.[6] Probabilities are estimated using a similar model and the same data as for the other two mechanisms. The number of people moving out is used for modeling the number of persons moving into shared apartments in later iterations of the simulation.

To prevent overestimating the number of single person households, the moved-out persons can form new households by entering either new partnerships or shared apartments.

Regional Mobility

The module Regional Mobility simulates between municipalities. Individuals leave the simulation population when moving out of the district and are added from a copy of the base population when moving into the district. This process is based on statistics produced by the Federal Statistical Office (Destatis, 2020). These statistics contain information on regional mobility broken down by age, sex, and relationship status. Subsequently, the probabilities for regional mobility are adjusted to known margins for each district level.

To prevent minors forming a household, regional mobility is additionally considered at the household level. Probabilities are estimated by Iterative Proportional Updating (Ye et al., 2009). Iteratively, the probabilities are adjusted to the frequencies of socio-demographic characteristics at the individual level in a randomized order and then scaled to a probability between zero and one (Stephensen, 2016). This is done using the base population to assign probabilities for leaving the district and on a copy of the base population to assign probabilities for moving into the district. Households that move into a district are selected from the copy of the population, which represents the remaining part of Germany. Newly arrived households are added to the base population. Outmoving households are removed from the population.

Formation of New Partnerships

The nuclear family (cohabitation of a mother, a father, and children) is still by far the most common family type in Western Europe. However, partners within the nuclear family increasingly remain unmarried (Schneider, 2015). Therefore, the simulation requires a module simulating the formation of new partnerships independent from the official status of a relationship (which is modeled in the Marriage model).

---

6 Shared apartments are defined by us as households containing at least two people aged between 18–35 without children or partnerships.

A module for the creation of new partnerships performs two simulation tasks: The entry into the partner market and the matching of new couples considering covariates.[7] There are different modeling approaches in other simulation models (cf. Perese, 2002; Zinn, 2012). MikroSim uses stochastic matching, therefore allowing less favorable partner combinations (for example, in large age differences). The module consists of two-steps. In the first step, for persons older 18 years not living in a partnership, a model based on German Socio-Economic Panel data (Goebel et al., 2019) estimates the probability of cohabitation with a partner (for details, see Table 1). Since the model is specified separately for men and women, the estimated propensities for a relationship might yield an imbalance of men and women available on the partner market. By only considering people in the same district, regional aspects of partner markets and the importance of spatial distance for partnerships are modeled.[8]

In the second step, the selected persons are matched.[9] The probability for cohabitation is estimated with a logit model using German Microcensus data. To account for potential age difference of the partners, spline functions are used in terms of generalized additive models (Wood, 2017). Therefore, also rare but possible partnerships (for example, with large age differences) are generated. The simulated imbalances of regional partner demands models the option of better available choices in asymmetric partner markets (cf. Klein, 2000).

Marriage

Within the marriage module, only single, widowed, or divorced partners are eligible for marriage. The module starts with unmarried couples within existing households. Information on the household and individuals is used to predict the couple's probability to marry. The model is based on data from the German Socio-Economic Panel (for details of the model, cf. Table 1). Given the estimated transition probabilities, marriages are simulated by updating family status for the involved partners.

---

7 The main mechanism is homogamy with respect to age, socio-economic status, and nationality (Klein, 2015).

8 People living in different districts are not matched in the module for two reasons. On the one hand considering all potential partners would result in high computational costs. On the other hand matched partners from different districts would again lead to regional mobility and subsequently distort the marginal distribution resulting from the Regional Mobility module.

9 Due to limited information on same-sex partnerships, the partnership module is restricted to hetero-sexual partnerships only.

Moving into Shared Apartments

Since the number of apartment-sharing communities is likely to be high in large cities and university towns, neglecting this type of housing would lead to an unrealistically high number of single-person households. In the module, only people currently living alone between 18 and 35 years old are eligible to move into a shared apartment.

The probability for moving into shared apartments is based on data from the German Microcensus and is estimated with a logit model including only age and gender as independent variables (cf. Table 1). The estimated probabilities for each district are calibrated via iterative proportional fitting so that the proportion of people living in a shared apartment remains the same after the first simulation period. The proportion of people moving into shared apartments is then left constant, such that a change of the total number of people living in shared apartments is a result of changing population structures.

To form new households, all relocating persons are randomly matched, such that the average household size of three persons is created while distributional size assumptions are approximately met.

School Education

The School Education module consists of two sub-modules, (1) a School Enrollment module and (2) a module for assigning educational qualifications.

The Enrollment module assigns the time of enrollment based on relative frequencies from official data (Statistisches Bundesamt, 2020b). A child is either enrolled early (at age 5), regularly (at age 6) or late (at age 7).[10] The probabilities are available by federal state and sex for different types of primary schools.[11]

The School Education module assigns the duration of school attendance and the resulting certificate. Grade levels of students are promoted yearly. To take repetitions of classes into account, a pragmatic approach was chosen: The duration of primary schooling is estimated for each child. Based on the time of enrollment and the estimated duration, the age at which the students will complete the fourth grade is calculated. At this age, the child will be promoted to the fifth grade. Before this, they are simply in primary school.

10 In Germany, there are different key dates for school enrollment depending on the federal state. In MikroSim we do not take them into account, since we do not model birthdays and we use a yearly time framework to update characteristics.

11 An initial attempt to estimate the probabilities by a model using the National Education Panel Study (NEPS) (Blossfeld, Roßbach, and Maurice, 2011) was discarded due lack of predictive power: The model including education, partner education, age, age of partner, work, marital status, and body length at birth yielded a McFaddens Pseudo-R^2 of 0.03.

Children at the beginning of the simulation must be assigned a grade level. They are assigned a grade level according to their age, not considering any covariates such as repeated classes or late enrollments.

The NEPS (Starting Cohort 2) is used to estimate the probabilities for the duration of primary schooling. The current module promotes people yearly until they leave school. A model predicting re-attendance of a class from fifth grade onward will be implemented in later stages of the project.

Since school qualifications in Germany can be obtained after grade 9, degrees are tentatively assigned at this grade. As soon as a qualification is assigned, the school career is continued depending on this qualification. The probabilities for grades are estimated using a multinomial logit model based on NEPS (Starting Cohort 4) data (see Table 1 for details).

Vocational Training

In the Vocational Training module, people who left school are assigned a job qualification. Possible values are no vocational training, vocational training degree, or university degree.[12] All model estimates are based on the dataset "Growing up in Germany" (AID:A II) produced by the German Youth Institute.

The time until graduation (1–6 years) of people with an assigned Bachelor's degree is taken from official statistics (Statistisches Bundesamt, 2019b). Persons not obtaining a degree receive a vocational training which after 3 years ends with a professional qualification. Future versions of the module will use distributions from official data for the duration of training.

At the start of the simulation, people already attending a university need an estimate of the duration of their attendance, which is estimated using a regression model based on a student survey data (Georg, Ramm, & Bundesministerium für Bildung und Forschung, 2016). Age is used to predict the stage of the academic career of the student.

Employment

The Employment module consists of three models: First, the employment status is assigned to the individuals in the base dataset. Second, full-time and part-time employment are estimated for the working population. Hereby, only individuals between the age of 15 to 66 years are considered. Third, the simulated objects are sent into retirement.

12 Degrees from universities of applied sciences and PhDs are not modeled because the proportions in the population are small and are often coded as university degree in survey data.

The required transition probabilities are estimated by a multinomial logistic model using German Microcensus data. The model uses individual-level variables (such as age, gender, education, and employment status last year), household characteristics (children and type of relationship), and considers the regional variation in labor market status to predict employment status.

For the working subpopulation the probabilities for part-time or full-time employment is estimated using the same predictors (see Table 1).

In the module, currently all people passing the age-threshold of 67 get retired. In a later version we will account for early retirement.

Income

A module of central importance is the Income module. This module provides necessary information for policy analyses, such as tax or family policy reforms. Moreover, income is an important explanatory variable for other models, such as fertility decisions, internal migration, or education opportunities of children.

The Income module is based on the Taxpayer Panel, an administrative data source covering the entire population of taxpayers in Germany from 2001 to 2014. Since the tax data does not contain detailed socio-demographic variables, the dataset will be enhanced by Microcensus data using statistical matching methods (predictive mean matching and nearest-neighbor random hotdeck).

The purpose of the module is the regional prediction of income and its changes over time. Estimates of individual incomes based on a mixed model will be calibrated with published regional data, such as income, poverty, and inequality indicators. To model changes, a two step approach is used. First, the probability of a change in income is predicted, then the differences to the previous year is modeled.

Application Modules

The core simulation model can be extended easily. Currently, two extensions are implemented: (1) labor market outcomes of migrants in Germany and (2) elderly care in the family. Other modules will be added in the future. We describe the current non-core modules briefly.

Labor Market Outcomes for Migrants

The second subject-matter topic is the development of labor market qualifications in the migrant population in Germany. The aim of the module is to supplement the projection of labor market integration with a regional perspective since both the

allocation of migrants and the labor market outcomes of individuals differ regionally.

Membership to different migrant groups (e.g., EU/non-EU) and migrant generations (i.e. born elsewhere/second-generation) are modeled regionally. A specific citizenship model within the module generates naturalization probabilities for all migrants in the simulation. The models specified in Table 1 will be estimated for the migrant population. To estimate the integration of migrants in the labor market, the Employment module and its first sub-module concerning the labor market status of the simulated individuals are of interest.

The relative labor market positioning of migrants in comparison to the majority population is predicted dependent on the assignment of outcomes in the School and Vocational Education modules run previously. The Income module will allow the study of income disparities and their potential change over time for the migrant population. Through the planned Citizenship module and the extensions in the Employment module, we can then estimate different scenarios of regional labor market integration for different ethnic minorities.

Elderly Care

A central subject matter problem in the MikroSim project is the increase of informal care for the elderly depending on demographic changes and new family structures. The current aim of the module is the study of the development of intra-family care. Therefore, household structures have to be updated. For example, children and grandchildren can be potential informal care providers for the elderly. Since the same function can be performed by family members outside the household, a mechanism to add these external families to a household is needed. The details of this mechanism are the subject of ongoing research.

In addition to this modeling of care supply, the demand for care has to be simulated. Therefore, a model for the degree and duration of care required is needed. Since the degrees of care have to be consistent with other variables in the model, additional constraints have to be fulfilled.

Adding this module with complex modeling of both the need for care and the type of care allows then to answer various questions in this area. Besides the possibility to forecast certain parameters such as the number and the proportion of people in need of care under ceteris paribuis conditions, the complex interdependency structure of the MikroSim model also allows to investigate the future effects of various social processes, such as the demographic change or the differentiation of family forms.

An example of this is the discussion between the medicalization or the compression thesis. Using microsimulation methods, for example, allows to analyze whether the progress in curative medicine leads to a higher number of people in

need of care due to a higher amount of years spent in disease (medicalization thesis). In contrast, the compression thesis assumes that the number of years spent in sickness will not increase, but that more of life will be spent in health. Adding this assumption as a complex "what-if" scenario then allows to compare the effects on a variety of target values that are either directly linked to this phenomenon (e.g. number of people in need of care, burden on the health care system), or that result from the interdependence structure of the simulation model (side effects such as a potential change in women's labor force participation due to the need to care for dependents).

Achievements, current work, and further development

After the first phase of work, the datasets required for estimation have been obtained, harmonized, documented and been used to estimate parameters for cross-sectional characteristics and transition probabilities. The base dataset has been updated to new margins and enhanced by survey data estimates.

The basic structure of the simulation model was planned and implemented as separate modules as shown in Figure 1. For each module, the processes required for updating the model were specified and estimated using available data. Resulting estimates were calibrated to known (regional) totals. The program code has been documented and tested. Currently, first test runs of the schooling simulation are in progress.

Within the next period, the refinements described above will be implemented, e.g. rural-urban disparities and improved regional patterns. After that, sensitivity studies and policy scenarios will be run and analyzed. We intend to publish first simulation results 30 months after the project has started. The availability of further data will furnish additional calibration methods. The modules will be fine-tuned further, leading to improved reproduction of known regional patterns and rare subgroups. To support continuous data updates while preserving model reproducibility, a data versioning system will be implemented in the simulation environment.

During the next funding period, additional modules will be implemented to test scenarios for policy studies in housing, health service research and urban travel demand. We intend to open the simulation model for other research groups by building a research data center, operating on similar principles as other data research centers in Germany.

References

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R., & Templ, M. (2011). Synthetic data generation of SILC data (Research Project Report No. 6.2). Advanced Methodology for European Laeken Indicators.

Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20 (3), 383–407.

Bækgaard, H. (2002). Micro-macro linkage and the alignment of transition processes: Some issues, techniques and examples. University of Canberra, National Centre for Social and Economic Modelling.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (eds.). (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft 14*.

Burgard, J. P., Dieckmann, H., Krause, J., Merkle, H., Münnich, R., Neufang, K. M., Schmaus, S., et al. (2020). A generic business process model for conducting microsimulation studies. *Statistics in Transition New Series*, 21 (4), 191–211.

Burgard, J. P., Krause, J., Merkle, H., Münnich, R., & Schmaus, S. (2019). Conducting a dynamic microsimulation for care research: Data generation, transition probabilities and sensitivity analysis. In A. Steland, E. Rafajłowicz, & O. Okhrin (Eds.), *Workshop on stochastic models, statistics and their application: Dresden, Germany, March 2019* (pp. 269–290). Cham: Springer.

Burgard, J. P., Krause, J., Merkle, H., Münnich, R., & Schmaus, S. (2020). Dynamische Mikrosimulationen zur Analyse und Planung regionaler Versorgungsstrukturen in der Pflege. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 283–313). Wiesbaden: Springer VS.

Burgard, J. P., Krause, J., & Schmaus, S. (2020). Estimation of regional transition probabilities for spatial dynamic microsimulations from survey data lacking in regional detail. *Computational Statistics & Data Analysis*, 154.

Destatis. (2020). Bevölkerung: Wanderungen. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Wanderungen/_inhalt.html.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018a). Mikrozensus 2012, SUF, Version 0. doi:10.21242/12211.2012.00.00.3.1.0.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018b). Mikrozensus 2013, SUF, Version 0. doi:10.21242/12211.2013.00.00.3.1.0.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018c). Mikrozensus 2014, SUF, Version 0. doi:10.21242/12211.2014.00.00.3.1.0.

Georg, W., Ramm, M., & Bundesministerium für Bildung und Forschung. (2016). Learning conditions and student orientations 2012/13. doi:10.4232/1.12510.

German Youth Institute. (2014). Growing up in Germany: Everyday life's world (AID:A II). München. Retrieved April 7, 2020, from https://surveys.dji.de/index.php?m=msw,0&sID=107.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German socio-economic panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, 239 (2), 345–360. doi:10.1515/jbnst-2018-0022.

Hannappel, M., & Kopp, J. (Eds.). (2020). *Mikrosimulationen: Methodische Grundlagen und ausgewählte Anwendungsfelder*. Wiesbaden: Springer VS.

Huang, Z., & Williamson, P. (2001). A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata (Working Paper No. 2001/2). Department of Geography, University of Liverpool. Liverpool.

Klein, T. (2000). Partnerwahl zwischen sozialstrukturellen Vorgaben und individueller Entscheidungsautonomie. *Zeitschrift für Soziologie der Erziehung und Sozialisation*, 20 (3), 229–243.

Klein, T. (2015). Partnerwahl. In P. B. Hill & J. Kopp (Eds.), *Handbuch Familiensoziologie* (pp. 321–343). Wiesbaden: Springer.

Klevmarken, A. (2008). Dynamic microsimulation for policy analysis: Problems and solutions. In A. Klevmarken & B. Lindgren (Eds.), *Simulating an ageing population: A microsimulation approach applied to Sweden*. Bingeley: Emerald Group Publishing Limited.

Kolb, J.-P. (2013). Methoden zur Erzeugung synthetischer Simulationsgesamtheiten (Doctoral dissertation, University of Trier, Trier). Retrieved from https://ubt.opus.hbz-nrw.de/files/590/Diss_Kolb_JP.pdf.

Laarhoven, P. J. M., & Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Dordrecht: Springer.

Li, J., & O'Donoghue, C. (2013). A survey of dynamic microsimulation models: Uses, model structure and methodology. *International Journal of Microsimulation*, 6 (2), 3–55. doi:10.34196/ijm.00082

Li, J., & O'Donoghue, C. (2014). Evaluating binary alignment methods in microsimulation models. *Journal of Artificial Societies and Social Simulation*, 17 (1), 15. doi:10. 18564/jasss.2334

Merz, J. (1991). Microsimulation—a survey of principles, developments and applications. *International Journal of Forecasting*, 7 (1), 77–104.

Münnich, R., Burgard, P., J., & Vogt, M. (2013). Small Area-Statistik: Methoden und Anwendungen. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 6 (3/4), 149–191. doi:10.1007/s11943-013-0126-1

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P., & Kolb, J.-P. (2012). Stichprobenoptimierung und Schätzung im Zensus 2011. *Statistik und Wissenschaft*. Wiesbaden: Statistisches Bundesamt.

Münnich, R., Schnell, R., Kopp, J., Stein, P., Zwick, M., Dräger, S., Merkle, H., Obersneider, M., Richter, N., Schmaus, S. (2020). Zur Entwicklung eines kleinräumigen und sektorenübergreifenden Mikrosimulationsmodells für Deutschland. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 109–140). Wiesbaden: Springer VS.

Münnich, R., & Schürle, J. (2003). On the simulation of complex universes in the case of applying the German Microcensus (DACSEIS Research Paper Series No. 4). Eberhard Karls University of Tübingen.

O'Donoghue, C., & Dekkers, G. (2018). Increasing the impact of dynamic microsimulation modelling. *The International Journal of Microsimulation*, 11 (1), 61–96.

Orcutt, G. H. (1957). A new type of socio-economic system. *The Review of Economics and Statistics*, 39 (2), 116–123.

Perese, K. (2002). Mate matching for microsimulation models (Technical Report No. 3). Congressional Budget Office. Washington. Retrieved from https://www.cbo.gov/publication/14211.

Rahman, A., & Harding, A. (2016). *Small area estimation and microsimulation modeling*. Boca Raton: CRC Press.

Rao, J., & Molina, I. (2015). *Small area estimation, 2nd edition*. Wiley Series in Survey Methodology. Hoboken: John Wiley and Sons, Ltd.

Schneider, N. F. (2015). Familie in Westeuropa: Von der Institution zur Lebensform. In P. B. Hill & J. Kopp (Eds.), *Handbuch Familiensoziologie* (pp. 21–54). Wiesbaden: Springer.

Schnell, R., & Handke, T. (2020). Neuere bevölkerungsbezogene Mikrosimulationen in Großbritannien und Deutschland. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 35–56). doi:10.1007/978-3-658-23702-8_3

Statistisches Bundesamt. (2019a). Bevölkerung: Eheschließungen, Geborene und Gestorbene 2018 nach Kreisen. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Eheschliessungen-Ehescheidungen-Lebenspartnerschaften/Publikationen/Downloads-Eheschliessungen/eheschliessungen-geborene-gestorbene-5126001187004.html.

Statistisches Bundesamt. (2019b). Bildung und Kultur: Prüfungen an Hochschulen. Retrieved January 21, 2020, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/pruefungen-hochschulen-2110420187004.pdf.

Statistisches Bundesamt. (2020a). Bevölkerung: Sterbefälle und Lebenserwartung. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Sterbefaelle-Lebenserwartung/_inhalt.html.

Statistisches Bundesamt. (2020b). Schulanfänger: Bundesländer, Schuljahr, Geschlecht, Einschulungsart, Schulart. Retrieved April 7, 2020, from https://www-genesis.destatis.de/genesis/online?operation=table&code=21111-0010.

Stein, P., & Bekalarczyk, D. (2016). Zur Prognose beruflicher Positionierung von Migranten der dritten Generation. In R. Bachleitner, M. Weichbold, & M. Pausch (Eds.), *Empirische Prognoseverfahren in den Sozialwissenschaften* (pp. 223–257). Wiesbaden: Springer.

Stephensen, P. (2016). Logit scaling: A general method for alignment in microsimulation models. *International Journal of Microsimulation*, 9 (3), 86–102.

Sutherland, H. (2018). Quality assessment of microsimulation models: The case of Euromod. *International Journal of Microsimulation*, 11 (1), 198–223.

Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation*, 7 (1), 4–25.

van Imhoff, E., & Post, W. (1998). Microsimulation methods for population projection. *Population: An English Selection*, 10 (1), 97–138.

Williamson, P. (2012). An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation. In R. Tanton & K. L. Edwards (Eds.), *Spatial microsimulation: A reference guide for users* (pp. 19–47). Dordrecht: Springer.

Wood, S. N. (2017). *Generalized additive models: An introduction with R (2nd ed.)*. Boca Raton: CRC Press.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In 88th annual meeting of the transportation research board.

Zinn, S. (2012). A mate-matching algorithm for continuous-time microsimulation models. *International Journal of Microsimulation*, 5 (1), 31–51.

Zwick, M., & Emmenegger, J. (2020). Mikrosimulation und Gesellschaftspolitik – ein kurzer historischer Abriss. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 17–34). doi:10.1007/978-3-658-23702-8_2