Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

PROBABILISTIC POPULATION PROJECTION WITH JAMES II

Christina Bohk

Institute of Sociology and Demography
University of Rostock
18057 Rostock, Germany

Roland Ewald
Adelinde M. Uhrmacher

Institute of Computer Science
University of Rostock
18059 Rostock, Germany

ABSTRACT

Predicting future populations and their structure is a central theme in demography. It is related to public health issues, political decision-making, or urban planning. Since these predictions are concerned with the evolution of a complex system, they exhibit a considerable uncertainty. Accounting for this inherent uncertainty is crucial for subsequent decision processes, as it reveals the range of possible outcomes and their likelihood. Consequently, probabilistic prediction approaches emerged over the past decades. This paper describes the probabilistic population projection model (PPPM), a recently developed method that allows detailed projections, but has a complex structure and requires much input data. We discuss the development of P3J, a tool that helps users in managing and executing projections and is built on top of the simulation system JAMES II. We outline how even specific tools like P3J profit from general-purpose simulation frameworks like JAMES II, and illustrate its usage by a simple example.

1 INTRODUCTION

This paper is concerned with population projections that compute the future population size and composition by age and sex. They serve as a basis for various societal decision-making processes and provide data for further scientific analyses. For example, projection outputs can be used as input for other studies in the domain of public health, a discipline that is concerned with the impact factors of diseases and longevity. The main goals of public health studies are to prevent diseases by eliminating the identified risk factors, to improve the health of humans, and to prolong their lives (Lamb and Siegel 2004).

There are two major relations between public health and demographic population projections. Firstly, one may use the outcomes of population projections for public health inquiries. Future population size and structure is relevant for many simulation studies, e. g., the future proportion of disabled or elderly people might be important in the context of hospital management, home care, or pension policies (e. g., cf. (Satyabudhi and Onggo 2008)). This could motivate further investigations and might draw attention to the most important issues in public health, epidemiology, and medical research.

Secondly, population projections can be used to explore the population dynamic effect of a cause elimination, i. e., they quantify the impact of eliminating a disease, a cause of death, or a risk factor on future population size and composition. The population dynamic effect of a new disease or risk factor can be analyzed in the same way. However, creating such projections involves some serious challenges, as there is almost no reliable information about the interdependency among diseases or death causes. Another important source of uncertainty concerns the awareness of relevant influencing factors on the future population and their evolution over time.

Considering these uncertainties requires probabilistic approaches that, e. g., assign probabilities to possible trends of key parameters. These could reflect the likelihood of different future developments, e. g., economical crises or breakthroughs in health care, and should be set by experts. Probabilistic population projections take these probabilities into account by associating probability distributions with their results, or by providing confidence intervals.

Being aware of the societal importance of accurate population projections, it seems indispensable to improve existing population projection methods as much as possible. But not only the projection methods itself have to be improved – it is also important to develop software tools that support users in conducting projections and analyzing their outcome. If probabilistic projection methods base on Monte-Carlo simulation, it seems natural to build such tools on top of simulation frameworks.

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

In the following, we describe the probabilistic population projection model, PPPM (Bohk and Salzmann 2006, Salzmann and Bohk 2006), as well as the re-implementation of a tool that allows to conveniently create and execute PPPM-based projection models. It is based on the open source simulation framework JAMES II (Himmelspach and Uhrmacher 2007).

2 BACKGROUND AND RELATED WORK

Population projections predict the future population size and its composition by age and sex with assumptions on demographic events in fertility, mortality, and migration. The cohort-component method is a classical demographic method that was introduced by (Cannan 1895), reinvented independently by (Bowley 1924) and (Whelpton 1928, Whelpton 1936), and finally formalized in matrix algebra by (Leslie 1945). It allows to divide a population into subpopulations that share certain characteristics, e. g., location, religion, or race. Each subpopulation has specific *vital rates*, such as fertility or death rates, that also vary with age and sex. Basically, the cohort-component method combines a recent subpopulation count with assumed births, deaths, and migrants (in accordance to the subgroup-specific vital rates) to a future subpopulation count with the balancing equation of population change. All projected subpopulation counts can be aggregated to a total projected population count in each projection interval. The number of considered subpopulations depends on the objectives of a projection and the available data (Smith, Tayman, and Swanson 2001, Preston, Heuveline, and Guillot 2001). In public health studies, characteristics like the exposure to a risk factor or disease can be used to distinguish further subpopulations. This would, for instance, allow to compare future survivals of ill and healthy people.

As already mentioned, population projections are uncertain because the future vital rates are a result of complex and unforeseeable individual behavior. Hence, a single best guess of just one possible evolution path per vital rate is insufficient – particularly with regard to each vital rate's complex potential evolution spectrum, due to more or less (un)known and (un)expected future determinants. Probabilistic population projections solve this problem by accounting for more than one possible evolution path per model input parameter (e. g., vital rates). Their results augment each output quantity, such as population count or proportion of the elderly, with an occurrence probability distribution for each projection interval (e. g., years).

The way of generating and entering the evolution paths for a vital rate in a projection model is the main difference between existing probabilistic projection approaches. They can be generated via extrapolation (e. g., by time series methods, such as (Lee and Carter 1992)), expert-judgment (e. g., (Lutz, Sanderson, and Scherbov 1998)), regression techniques (e. g., (Swanson and Beck 1994)), simulation techniques (e. g., Monte-Carlo simulation, such as (Pflaumer 1988)), ex-post error methods (e. g., (Keyfitz 1981, Stoto 1983)), error propagation methods (e. g., (Alho and Spencer 1997)), or a mixture of several methods (e. g., (Alho and Spencer 2005, Alders, Keilman, and Cruijsen 2007)). These *assumptions* on the future evolution of vital rates can be either generated by predetermined model-based methods, or by arbitrary external methods. Even though common probabilistic approaches have the advantage to capture a projection's uncertainty with an occurrence probability distribution for each output quantity, they also have some shortcomings. Common approaches are often limited to one or more model-based methods to generate assumptions for the vital rates. Therefore, they might not be able to appropriately capture the full spectrum of a vital rate's potential future evolutions. The assumptions of a predetermined model-based method are limited to certain value ranges, as well as to certain patterns. Instead, a mixture of methods should be used to represent a projection's uncertainty more realistically. This is done in recent probabilistic population projection projects, such as UPE (Statistics Netherlands 2005). Other probabilistic approaches, particularly those with an extrapolative model-based assumption generation method (e. g., (Lee and Carter 1992, Lee and Tuljapurkar 1994)), require long historical data series – but only few countries provide long historical data series for vital rates. Consequently, missing or erroneous data can distort the assumptions for them. An accurate projection also has to differentiate between relevant subpopulations with differing vital rates. Hence, not only natives and migrants should be projected separately, but also the descendant migrant generations in the target country (as they successively adopt the behavior of the natives). Conventional probabilistic projection approaches seldom distinguish between vital rates of natives and migrants. (Alho, Cruijsen, and Keilman 2008, Alders, Keilman, and Cruijsen 2007), for example, are aware of that problem, but argue that the lack of data may prohibit more accurate predictions. Therefore, most methods treat migrants just like natives, i. e., both subpopulations have equivalent vital rates. In addition, common probabilistic projection approaches often use net migration instead of gross migration in absolute numbers. Thus, they miss important population dynamic effects concerning structural shifts in age and sex structure of the migrants, and may therefore generate rather inaccurate predictions.

Population projection models can also be distinguished by the level of detail they consider, i. e., micro vs. macro perspective. Micro-level projection models usually work on a sample population. For each individual in the sample, they project a life course with occurrence-exposure rates for certain demographic events (e. g., immigration, emigration, childbirth, etc.); depending on age, sex, and perhaps some additional characteristics. Macroscopic output quantities, like future population

size, are generated by aggregating the individual results from the sample and extrapolating them proportionally to the total population. The major advantage of micro-level projection models is the incorporation of individual demographic behavior. But this advantage might also be considered as a drawback, due to the lack of input data to model individual behavior correctly (van Imhoff and Post 1998, Willekens 2006, Andreev and Vaupel 2006).

Macro-level models project a total population by age and sex, and also may include other characteristics. Usually, some variant of the cohort-component method is applied to the respective age-sex subgroups. No individual demographic behavior can be considered: people of the same age-sex subgroup are assumed to experience the same vital rates of mortality, fertility, and migration. An advantage of macro models is the availability of data to generate realistic assumptions for future vital rates. Still, assuming homogeneous instead of heterogeneous, i. e., individual, behavior is a strong simplification of reality.

There is a long history of performance comparisons between simple and complex projection models in demography. Applications show that complex models do not always outperform simple models (Stoto 1983, Rogers 1995). Micro-level population projections are a promising approach, but the validity of their results might be hampered by the lack of information that is needed to model individual behavior properly. This problem is aggravated by the fact that, so far, not even all determinants of complex individual behavior are completely known; neither can they all be measured with enough detail to put them into a projection model. Still, micro-level projections can be very useful when the available data suffices for their construction.

All in all, micro- and macro-level approaches should be regarded as complementary methods. How to use them in conjunction is subject of current research. For example, the MICMAC project tries to combine micro and macro perspective from two separate projection models, MIC and MAC (MicMac). MIC employs a micro-level perspective, whereas MAC is a macro-level approach. Both project a population by certain characteristics and are intended to generate consistent results on macro level because they rely on the same transition intensities. MIC additionally projects individual life courses in dependence of personal characteristics (Willekens 2005, Willekens 2006, Gampe and Zinn 2007).

Micro-level projections can be realized by micro-simulations, i. e., the simulation of individual model entities on a single level. Their requirements are rather similar to other applications of simulation, e. g., in terms of scalability or output observation. It is also the level of detail where sophisticated simulation techniques, e. g., from parallel and distributed simulation, can be applied (Satyabudhi and Onggo 2008). PPPM, however, works on a macro-level and therefore has differently weighted requirements, e. g., when it comes to model storage or the management of input parameters. Overall, the requirements of PPPM and tools that also use micro-simulation, e. g., MICMAC, are quite different. However, both MicMac (Gampe and Zinn 2007) and PPPM are supported by JAMES II plug-ins, the latter of which will be described in section 4.

3 THE PROBABILISTIC POPULATION PROJECTION MODEL (PPPM)

3.1 Basic Model

The PPPM is a macro-level projection model that uses an extension of the classical cohort-component method. Demographic research literature provides several examples of dividing a population in natives and migrants as subpopulations (Espenshade, Bouvier, and Arthur 1982, Mitra 1983, Cerone 1987, Schmertmann 1992). In PPPM, this concept is extended in accordance to (Edmonston and Passel 1992) and (Dinkel 1989), so that natives, immigrants, emigrants, and the descendant generations of immigrants and emigrants can be considered and projected separately (see figure 1). The number of considered descendant generations depends on the length of the *projection horizon*, i. e., the time span of the projection, and the assumed fertility pattern. In industrial countries, it is appropriate to consider five to six generations of descendants for a projection horizon of 100 years. Consequently, different demographic behavior can be modeled by subpopulation-specific vital rates. This very detailed consideration of migration is a distinctive feature of the PPPM when compared to other probabilistic projection approaches.

The PPPM considers a large number of input parameters, which is due to the various subpopulations that are distinguished, as well as the specific modeling of the demographic components, in particular that of mortality. Here, input parameters are the survivors at age x, the survival probability of persons in the open end age interval (which cannot be computed from the survivors at age x), and infant death in the first half-year (i. e., the proportion to all infants that die in their first year of their life). All these mortality parameters have to be defined for *each* subpopulation.

The age-specific fertility rates belong to the input parameters of all female subpopulations, and the sexual proportion at birth to both male and female subpopulations. The age-specific population at the end of a projection's jump-off year is another input parameter for the native female and male subpopulation. Finally, the total number of immigrants and emigrants by age and sex are input parameters of the corresponding subpopulations (see figure 1). The assigned input parameters

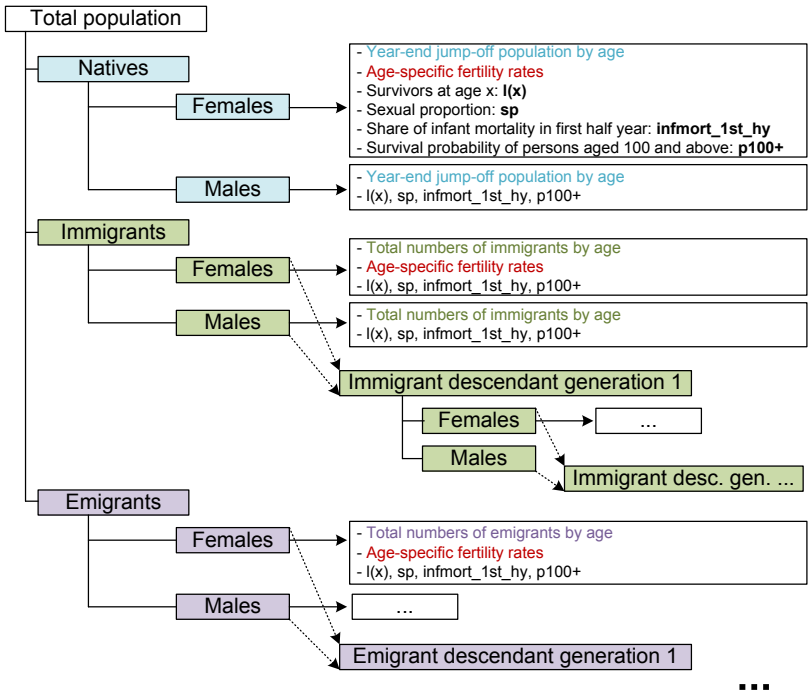

Figure 1: Overview of PPPM parameters and subpopulations, highlighting those parameters that are specific for females (red), natives (cyan), immigrants (green), or emigrants (violet)

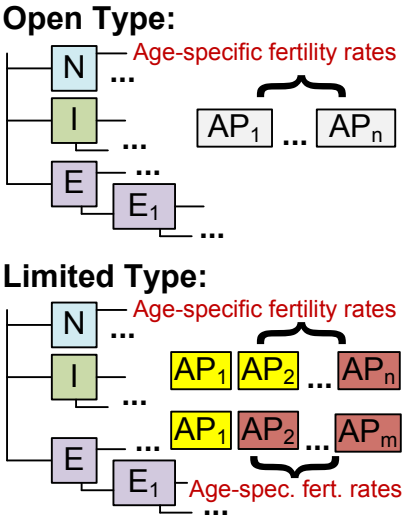

Figure 2: Open vs. limited type: While the open type allows to assign any of the defined assumption paths (AP) to a parameter, these have to be from the same *Set* (yellow, red) in the limited type.

are typically matrices referring to different subpopulations, with dimensions that depend on the projection horizon and the number of age classes.

Another special feature of the PPPM is its probabilistic projection procedure. The modeler needs to generate a suitable number of *assumptions paths*, i. e., possible evolutions during the projection horizon, for each input parameter. These should encompass the full spectrum of a parameter's potential future evolution and have to be defined by the modeler. Each assumption path contains estimated future values of a certain input parameter. Arbitrary methods, e. g., statistical or judgmental ones, can be used for generating such assumptions. In contrast to other probabilistic projection approaches, no specific model-based method has to be used for generating assumptions in the PPPM. That is due to separation of the assumption generation process and the projection process within the PPPM. Therefore, assumptions have to be generated externally, with any method the modeler might deem best. Different assumption generation methods can be used for different input parameters, or even for different assumptions regarding a single input parameter. Parameter-specific information sources like data exploration and theories can be addressed with different methods.

An occurrence probability has to be assigned to each assumption path. It should be based on expert judgment. The higher the occurrence probability of an assumption path, the higher is its expected chance of actually coming true. In each projection trial, a randomly chosen assumption path is assigned to each input parameter. This is done by considering the occurrence probabilities of all assumptions that are available for the specific parameter. All chosen assumption paths are propagated to the deterministic projection model that computes *result paths* through the projection horizon for several output quantities, e. g., the total population by age, or the number of immigrant births. Repeating this procedure n times yields therefore n possible result paths per output quantity. An output quantity's *x%-confidence interval* of $[a,b]$ indicates that $x = (1-\alpha)\cdot 100\%$ of its result paths lie between the lower bound a and the upper bound b. It can be constructed by sorting the result values in ascending order for each projection interval, and then taking the result values at index $(\frac{\alpha}{2})\cdot 100$ and $(1-\frac{\alpha}{2})\cdot 100$ as lower and upper bound, i. e., values for a and b, for each projection interval.

The procedure allows to involve strongly differing assumption paths, in value level as well as in evolution pattern, per input parameter. This is particularly important when the future evolution of a parameter is quite uncertain. Diverging assumption paths, each of which might be very improbable in itself, can now be considered together. The influence on

projection output is controlled by occurrence probabilities. Although the probabilities of rare events are usually hard to estimate (Taleb 2008), even when expert knowledge is involved, excluding them from probabilistic projections is even worse, as the true range of possible outcomes will be underestimated. Another problem of this so-called *open type* of PPPM is the possible chance of choosing assumption paths for some input parameters that do not fit each other, i. e., which are extremely unlikely *in their combination*. Ideally, each trial should represent a realistic scenario of future population development with consistent assumptions for all input parameters. In contrast, implausible assumption path combinations arise when, for instance, extremely high and low mortality paths are *both* chosen for different subpopulations within the same trial, although in reality all subpopulations still exhibit rather similar mortality levels and patterns. Such implausible combinations can be avoided by a *limited type* PPPM (Bohk and Salzmann 2006).

3.2 Limited Type

The limited type of the PPPM was developed to ensure assumption consistency for each trial. Implausible assumption paths combinations are eliminated by the introduction of *Set Types* and *Sets*. A Set Type aggregates certain input parameters over several subpopulations. For instance, a Set Type *"'Fertility"'* could subsume the age-specific fertility rates of all female subpopulations. For each Set Type, several Sets can be defined by the modeler. A Set consists of consistent assumption paths for each input parameter that is included in the corresponding Set Type.

The example in figure 2 illustrates this fundamental difference between open and limited type. While the former allows to choose *any* available assumption path for each input parameter (cf. figure 1), the latter restricts this to all *plausible* combinations of assumption paths, i. e., all paths have to be drawn from the same Set. In figure 2, there is a Set Type that subsumes the age-specific fertility rates of the natives and of the first generation of emigrant descendants (among other parameters, potentially). Two Sets are defined for this Set Type (yellow and red): one Set (yellow) defines two possible assumption paths for the natives' fertility rates, and a single one for the fertility of the emigrants. In the limited type, choosing AP_1 (yellow) for the emigrants means that the natives' fertility rate *cannot* be set to AP_n (red), since this path belongs to another Set (red). A combination of AP_n (natives) and AP_1 (emigrants) is therefore *implausible*.

Each Set is associated with a specific occurrence probability. The occurrence probabilities of a Set Type's Sets have to be in $[0,1]$ and add up to 1. Similarly to the open type PPPM, occurrence probabilities are also assigned to all assumption paths of an input parameter, which are now associated with a specific Set. Again, a Set's assumption paths for an input parameter have to add up to 1. Calculating a limited type PPPM involves two steps: At first, a Set is randomly chosen for each Set Type. Then, an assumption path is chosen for each parameter, this time by restricting the selection to those assumption paths that are defined for the chosen Set. By restricting each Set to hold assumption paths which do not base on contradictory, i. e., inconsistent, assumptions, the user is now able to eliminate the implausible combinations. This is the main advantage of the limited type. It is bought with much additional structure that has to be managed and overseen by the modeler, which poses a considerable challenge when developing tools for a convenient usage of the PPPM.

4 P3J: A TOOL FOR PPPM CREATION AND SIMULATION

4.1 Requirements

Based on experiences with earlier prototypes (Bohk and Salzmann 2006) written in MATLAB (The MathWorks), we identified several shortcomings that had to be resolved when developing a re-implementation of the PPPM for a broader audience. First of all, the integration of new features like the limited type made the MATLAB version hard to maintain, debug, and use. Another major issue was the seamless integration of alternative calculation methods and algorithmic variants: since the PPPM is still subject to various research activities, a key requirement to the new version is its flexibility with respect to different calculation methods and different projection scopes, i. e., an arbitrary number of subpopulations should be supported. The same flexibility is also desired when it comes to the workings of the Monte-Carlo simulation as such: apart from randomly drawing parameter assignments, it might be possible to sort them by decreasing overall probability and execute the trials in that order, so that more probable outcomes are calculated before less probable ones. This would allow to stop the simulation earlier and would also focus the analysis on the most likely scenarios.

The user interface is required to make the input of all parameter assignments, as well as the definition of Sets and Set Types, as simple as possible. Ideally, the user should be able to consider other studies within the same environment and browse them with ease. The sheer amount of data that is required for a thorough probabilistic projection motivates the use of database technology here. Finally, future research might require the integration of external tools, e. g., to support the semi-automated generation of assumption paths, e. g., by using time series analysis (Box and Jenkins 1976).

We used MATLAB for prototyping because of its good performance, its easy-to-use programming language, and its powerful functions for data visualization. On the other hand, MATLAB is a commercial (and costly) software package that is predominantly used in engineering. It may not be readily available to all potential users, e. g., in the sociology departments of universities. Furthermore, its programming language is good for prototyping, but does not scale well with certain requirements, e. g., the design of complex user interfaces is rather cumbersome. While the sequential performance was satisfactory, it is also interesting to provide multi-threading capabilities in the advent of CPUs with several cores. Although MATLAB allows multi-threaded execution to some extent, it is non-trivial to write customized multi-threaded MATLAB programs. Finally, MATLAB does not inherently support the object-oriented programming paradigm, which could otherwise be used to enhance re-usability and maintainability of the original code. For these reasons we decided on a re-implementation from scratch: a Probabilistic Population Projection tool for Java (P3J). As the PPPM still requires a lot of standard functionality that most modeling and simulation tools have to provide, such as model input/output, experiment definition, and random numbers, we based our implementation on JAMES II. Newer MATLAB versions provide a Java interface, so this platform change does not eliminate the possibility of migrating the calculation-intensive parts of the Monte-Carlo simulation back to MATLAB at some point – but so far this has not been necessary.

4.2 Implementation Details

In this section we describe the three major implementation challenges we were confronted with when developing P3J:

Parameterization The PPPM allows to set various parameters. Each can be represented by a vector or matrix of real or integer values. Many parameters share the same dimensions, but these depend on the scope of the study, i. e., the number of years for which the population should be projected and also the number of age classes to be considered. For example, each subpopulation has a certain fertility that needs to be defined, and therefore all fertility parameters have the same dimensions. To avoid redundancy while still being able to support the user, e. g., by initializing all matrices with the correct size and displaying labels to guide the input, we devised a parameter management system with three main entities: `Parameter`, `ParameterInstance`, and `ParameterAssignment` (cf. fig. 4). The instances of `Parameter` provide all abstract information on parameter types, i. e., width and height of the matrices a parameter may hold as a value, its name, and the kind of population it is defined for: emigrants, immigrants, or natives. `ParameterInstance` objects are defined on this parameter information, but also include the generation they are associated with. For example, the fertility of emigrants might be a parameter with instances for each emigrant generation. This allows to manage an arbitrary number of generations without much redundancy. Finally, the class `ParameterAssignment` associates a given parameter instance with a value, i. e., a matrix, and hence represents an assumption path (see section 3). For each parameter instance, the user may define an arbitrary number of possible assignments – which will then be later chosen at random. For now, we used Colt (The Colt Project) for all matrix representations, but the corresponding classes have been wrapped, so that other implementations might be used as well. In the most basic case, just 33 parameter instances have to be configured – but nine parameters are added for each subpopulation and generation (cf. figure 1). For a projection time of 100 years, e. g., where five generations of emigrant and immigrant subpopulations are considered, this amounts to 123 parameter instances. Moreover, Sets and Set Types need to be defined as well, and probabilistic projections require to associate *multiple* assignments to parameter instances. Thus, a user is confronted with so much data and structure that this might hamper the orientation and productivity. We therefore had to take special care in creating a suitable user interface.

User Interface The user interface for editing projection data is depicted in figure 3. We designed it in the style of a file manager, with the structure of the model displayed as a tree on the left, while the right part of the window shows the content of the node that is selected in the tree. In the screenshot, a fertility assumption path from the example described in section 5 is selected. On the upper right side, the user may now edit name, description, and occurrence probability of the path. These are the properties of the corresponding `ParameterAssignment` object, which also has a reference to its `ParameterInstance` (cf. figure 4). The parameter instance of the assumption shown in figure 3 defines that this `ParameterAssignment` belongs to the fertility of third generation immigrants. This is displayed by highlighting the node in the model structure tree on the left. Set Types are displayed on the top level, followed by the Sets that are defined for each of them. Each Set has a sub-tree of the parameter instances its Set Type aggregates, and each of the parameter instances in turn has all parameter assignments as leaf nodes.

On the lower right side, the user may choose between viewing the data in a plot (as shown), or directly editing the data. Alternatively, an additional editing window can be opened. The data itself is stored in a `Matrix` object (cg. figure 4). To easily overview and edit the matrices, we integrated the plotting components of `JMathTools` (JMathTools) and enhanced the `JGrid` component of JEPPERS (Jeppers). The tree view on the left has an additional tab to give the user an overview of all PPPM instances in the database, and another one to view simulation results.

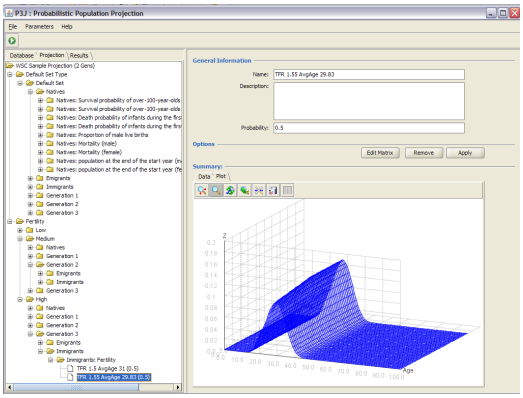

Figure 3: A screenshot of the PPPM editing interface

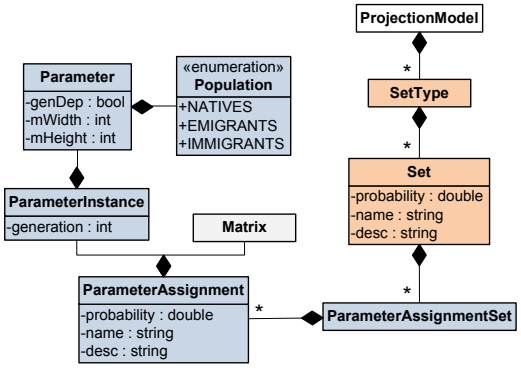

Figure 4: UML diagram of the important PPPM classes: parameters (blue) and Sets (orange). To emulate the open type, a model always has a default Set and Set Type.

Model Storage The size and the intricate structure of the models motivate their storage in a database. We use HIBERNATE to persist Java objects in a relational database (Hibernate , Elliott, Fowler, and O'Brien 2008), so that the table structure of the model database reflects the entity relations shown in figure 4. Users work directly on the database when editing the models. By this, we hope to support collaborative modeling in the future, since usually more than one expert is involved in a well-founded population projection. At the same time, our tool offers the user to export a projection to an XML file or to a binary format, so that exchanging projection set-ups can also be done via files.

4.3 Simulation Modes: Monte-Carlo vs. Analytical

The simulation algorithm for the PPPM basically consists of two components: one is the calculation of a single projection outcome, which is deterministic and reflects the underlying demographic semantics of the given parameter assignments. The other component generates the actual assignment mapping from parameter instances to matrix values, by choosing an assumption path for each instance. Consequently, these components are called *assignment generators*. The default assignment generator chooses randomly from all assignments of an instance. Since a probability $p_i \in [0,1]$ has to be defined for each assignment i, three cases may occur when the generator has to choose one of the n assignments and the sum of probabilities is defined as $p = \sum_{i=1}^{n} p_i$:

- $p = 1$. Correct setting. All probabilities will be used accordingly.
- $p \neq 1 \wedge p > 0$. The probability sum does not equal one but is still nonzero. All probabilities will be interpolated.
- $p = 0$. Invalid setting. No probabilities have been set, so that probabilities are defined uniformly.

In the two latter cases, the user is warned about the problem encountered with the probabilities. The simulator then adjusts them automatically. The necessary control methods are implemented in the `RandomNumberChecks` class, and are used for both open and limited type. While the default assignment generator implements a Monte-Carlo simulation, there might be other, more effective ways of analyzing a given model. We are currently developing an additional assignment generator that focuses on the analysis of the *most probable* assignments that can be generated. To enumerate them is non-trivial, since there are too many combinations to test them all at once. Instead, the algorithm sorts all Sets per Set Type and all assumption paths per parameter instance by decreasing probability. This allows to calculate the most probable assignment, which is generated and handed over to the calculation component. Then, a breadth-first search is executed within the combinatorial space of all possible assignments, which ensures that the next-probable assignment can be found and generated as well.

4.4 Integration into JAMES II

We used JAMES II as a simulation backbone for P3J, as it offers various features that are very useful in this context. Most importantly, its plug-in system (Himmelspach and Uhrmacher 2007) allows us to implement separate solutions for specific tasks, such as generating the actual assignment of matrices to parameter instances (see section 4.3). Moreover, we rely on its sub-system for random number generation (Ewald et al. 2008), as this is a non-trivial problem for which various solutions

exist (L'Ecuyer 1997, Matsumoto, Wada, Kuramoto, and Ashihara 2007). JAMES II also supports a flexible definition of experiments (Himmelspach, Ewald, and Uhrmacher 2008), so that the impact of, e. g., certain assumption matrices on the overall results of a study, or the performance differences between random and analytical assignment generation can now be investigated conveniently. Its experimentation layer was specifically designed for such studies, and provides both flexibility (w.r.t. to studies) and repeatability (w.r.t. defined experiments). JAMES II allows to execute several sequential simulation runs in parallel (Leye et al. 2008) – a feature that might come in handy when our tool is running on a multi-core CPU and the random assignment generator shall be used, as all generated assignments can be regarded as sampled from the same space, and hence the results of the parallel runs can be merged later. This lets us exploit parallel computing without having to implement any additional code.

On the implementation side, the integration of P3J with JAMES II was fairly straightforward: We implemented plug-ins for model creation, for reading models from the database, and for simulation. Additionally, we defined a new *plug-in type* that represents assignment generators, i. e., an extension point for the plug-in system. Both available assignment generators are implemented as plug-ins for this new plug-in type. Users can now easily configure JAMES II to use any of them when simulating PPPM models. The simulator extends the default implementation `RunnableProcessor`, which is provided by JAMES II and enables it to control and monitor the simulation process. When implementing the `nextStep()` method of the simulator, we re-interpreted its semantics (which is to eventually progress the simulation time) as the calculation of a single trial. Apart from that, various auxiliary data structures, e. g., to manage the connection information of databases, could be re-used.

5 EXAMPLE

5.1 A Sample Projection

To demonstrate the procedure of open and limited PPPM types, we now give a simplified example. Real-world data from the Federal Statistical Office of Germany was used to estimate assumption paths for all input parameters. The projection horizon starts in 2007 and ends 2048, while the base period, i. e., the period we analyzed to extrapolate current trends, starts 1990 and ends with the jump-off year 2006. A projection period of 41 years requires to consider at least two generations of migrant descendants. Single age classes are specified for age 0 to age 99, with an additional class for all persons that are over 99 years old.

A projection interval length of one year is used. Our exemplary projection contains just one assumption path for all input parameters, except for the age-specific fertility rates. Six assumption paths shall be provided for the age-specific fertility rates of each female subpopulation. A Set Type *Fertility* is defined to do so. It contains the age-specific fertility rates of all female subpopulations. Three Sets of Set Type *Fertility* are constructed, each of which subsumes consistent (i. e., rather similar) assumption paths: *Low Fertility*, *Medium Fertility*, and *High Fertility*. Varying the fertility and keeping every other parameter instance constant makes it possible to study the population dynamic effects of fertility.

Past and recent trends in age-specific fertility rates have been explored to estimate the six assumption paths. The shape of the age-specific fertility rate distribution determines the mean age at childbearing, which gradually increased over the last decades. Additionally, the age-specific fertility rates f_x (for age $x \in [15,49]$) are summed up to the so-called *total fertility rate* (TFR), which expresses the average number of children per woman. Age-specific fertility trends are then examined by considering $\frac{f_x}{TFR}$, i. e., the ratio of each fertility rate to the TFR. Therefore, the average growth of each ratio $\frac{f_x}{TFR}$ over the base period is assumed to continue, until a mean age at childbearing of 31 is reached. These ratios are assumed to be true for the year 2030. Then, we assumed TFR levels of 1.3, 1.4, and 1.5. Two other paths have assumed a TFR level of 1.55 and 1.3, respectively, as well as a mean age at childbearing of 29.8 in 2030. To generate a realistic assumption path between the fertility distribution of the jump-off year and one of the five aforementioned fertility rate distributions of 2030, we use linear interpolation. Constant values of 2030 are assumed from 2031 to 2048. Finally, a sixth assumption path contains just the constant age-specific fertility rates of the year 2007 through the projection horizon, i. e., a TFR level of 1.37 and a mean age at childbearing of 29.8 years.

The two assumption paths with a TFR level of 1.3 are assigned to the *Low Fertility* Set, the assumption paths with TFR levels of 1.5 and 1.55 are assigned to the *High Fertility* Set, and the remaining two assumption paths are added to the Set *Medium Fertility*. To increase the variation in this still very simple and small projection example, we chose a rather optimistic scenario by setting the probability of *High Fertility* to 40%, whereas the probabilities of *Medium Fertility* and *Low Fertility* were set to 30% each.

The assumption paths for all other input parameters are very simplistic, because they suppose constancy of a certain level through the projection horizon. Thus, the total numbers of male and female immigrants and emigrants by age are assumed

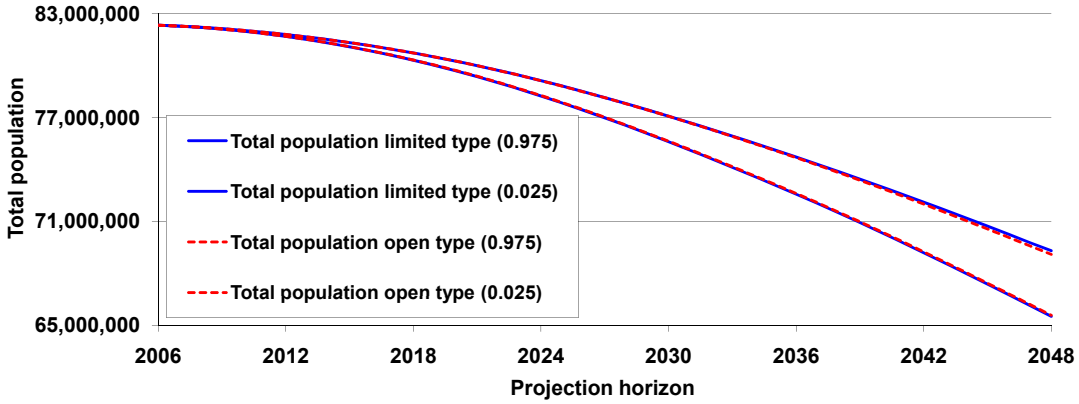

Figure 5: Comparing the 95%-confidence intervals of open and limited type

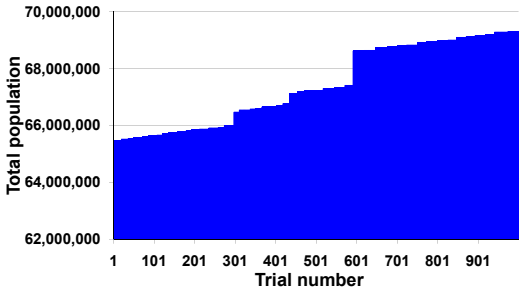

Figure 6: Final year-end populations with the limited type

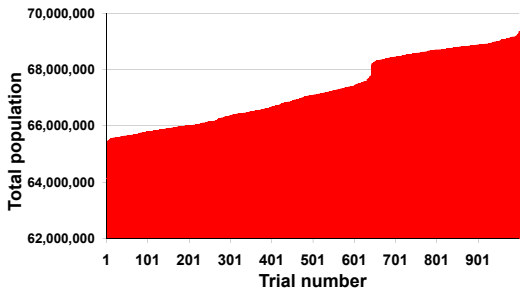

Figure 7: Final year-end populations with the open type

to remain at their average level, calculated over the last 10 years. The survivors at age x are assumed to stay constant at the level of 2006. This holds true for all male and female subpopulations. The sexual proportion of female and male births is assumed to be 0.486 and 0.514, a relatively stable level in recent years. The ratio of infant deaths in the first half-year to overall infant deaths in the first year is supposed to be constant at 0.9. The survival probability of persons aged 100 and above is assumed to be $\sqrt{0.6}$. The recent age-specific female and male population is taken from the end of the year 2006 for the natives.

5.2 Results

For each of the PPPM types, 1.000 trials have been simulated to achieve a statistically significant distribution of the output quantities of interest. In our case, this is just the total population.

As shown in figure 5, both projections show that the future total population size will decrease if the exemplary assumptions come true. With a probability of 95%, the total population will range between 65.51 millions and 69.3 millions according to the limited type, while the open type predicts it between 65.58 millions and 69.1 millions in 2048. That is a counter-intuitive result. One would intuitively expect the variance of the open type to be larger than that of the limited type, due to its higher flexibility in combining assumption paths. Anyhow, the results show the opposite, because implausible combinations often result in average result paths (Bohk and Salzmann 2006). This is due to balancing effects of alternatively chosen high and low assumption paths in the open type, which lead to a variance reduction of the open type results. Thus, the open type PPPM *underestimates* the result variance and is therefore not able to completely capture a projections inherent uncertainty. In our example, the open type's 95%-confidence interval for the total population in 2048 is 7% *smaller* than that of the limited type (see fig. 5). In more complex population projections that contain many more assumption paths per input parameter, this underestimation error regarding result variance will become even worse.

Figures 6 and 7 show, for the limited and the open type, respectively, the profile of all 1.000 sorted total population outcomes for the target year 2048. The open type generates an almost continuous profile, due to the unrestricted assumption paths variations, which cause many intermediate outcomes. In contrast, the limited type's profile has the form of a step-wise

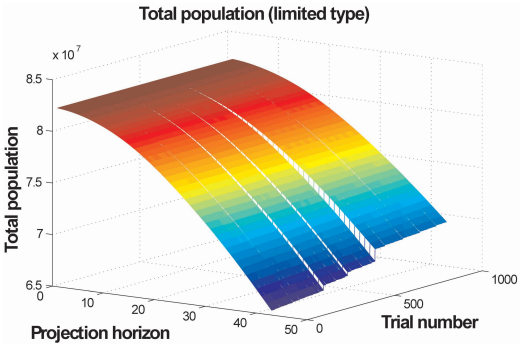

Figure 8: Total year-end population with limited type

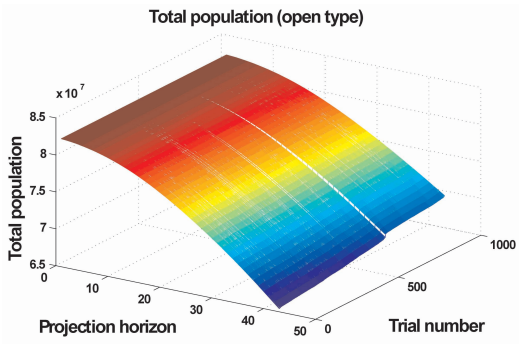

Figure 9: Total year-end population with open type

function, due to limited assumption path variations by the introduction of Set Types and Sets. Consequently, the open type generates more diverse result paths than the limited type – but these have not been intended by the modeler, and hence should be avoided. All in all, these figures illustrate that the limited PPPM will require fewer trials to generate significant result distributions than the open type PPPM. Still, many trials are needed in either case when conducting a serious population projection; the introduction of more assumption paths will allow many more combinations and therefore many more results: the now obvious difference between open and limited type cannot be seen easily anymore. Finally, figures 8 and 9 illustrate the same ordered projection outcomes as shown in figures 6 and 7, but these are now presented over time, as 3D-planes. This shows how the variance evolves differently in both types.

This simple projection example should just illustrate the intended use of P3J. Performance was not (yet) an issue, as the tool can cope very well with studies of this scale: a trial needed $\approx 0.05s$, on a Core 2 Duo system with a Java SciMark score of 440.2, using Sun's JRE 1.6.07.

6 CONCLUSIONS

This paper reviews the PPPM, a recently introduced macro-level probabilistic population projection model, and presents its implementation in JAMES II, a general-purpose framework for modeling and simulation. We discussed two variants of the PPPM, the open and the limited type, and illustrated their differences by a simple example with real data for Germany. The PPPM has several distinctive features: it does not restrict the modeler to any assumption generation method, supports an arbitrary number of subpopulations, as well as an arbitrary number of assumption paths per input parameter. Moreover, the modeler may associate each assumption path with a custom occurrence probability. This allows to take into account more improbable, but not impossible, assumption paths. Here, the limited PPPM type is particularly advantageous, as it restricts the assumption combinations to those that can be combined plausibly.

To support a broad usage of the PPPM in the future, we designed a JAMES II-based implementation, P3J. Its key features are the model storage and retrieval in a database, a user interface that is structured toward the central tasks in modeling with the PPPM (e. g., data input and management), and the flexibility to include more than one PPPM simulation method. Resolving these requirements was facilitated by the features provided by JAMES II: e. g., random number generators for our Monte-Carlo simulator, or the plug-in system to make P3J an extensible and flexible research tool in itself.

In the future, we plan to publish P3J as an open source package for the demographic community. Currently, we work on improving the result analysis capabilities of P3J and aim at further integration with JAMES II, especially on the level of the user interface.

REFERENCES

Alders, M., N. Keilman, and H. Cruijsen. 2007. Assumptions for long-term stochastic population forecasts in 18 European countries. *European Journal of Population and Development Review* 23 (1): 33–69.

Alho, J. M., H. Cruijsen, and N. Keilman. 2008. Empirically based specification of forecast uncertainty. In *Uncertain Demographics and Fiscal Sustainability*, ed. J. M. Alho, S. E. H. Jensen, and J. Lassila, 34–54. Cambridge University Press.

Alho, J. M., and B. D. Spencer. 1997. The practical specification of the expected error of population forecasts. *Journal of Official Statistics* 13 (3): 203–225.

Alho, J. M., and B. D. Spencer. 2005. *Statistical Demography and Forecasting*. Springer Science+Business Media, Inc.

Andreev, K. F., and J. W. Vaupel. 2006, May. Forecasts of cohort mortality after age 50. MPIDR Working Paper WP 2006-012.

Bohk, C., and T. Salzmann. 2006, June. A different understanding of probability in a probabilistic population projection model and its outcomes. In *Proceedings of the European Population Conference 2006*. Liverpool.

Bowley, A. L. 1924. Births and Population of Great Britain. *The Journal of the Royal Economic Society* 34:188–192.

Box, G. E. P., and G. M. Jenkins. 1976. *Time series analysis: Forecasting and Control*. Holden-Day.

Cannan, E. 1895. The Probability of a Cessation of the Growth of Population in England and Wales during the next Century. *The Economic Journal* 5 (20): 505–515.

Cerone, P. 1987. On stable population theory with immigration. *Demography* 24:431–438.

Dinkel, R. H. 1989. *Demographie. Band 1: Bevölkerungsdynamik*. München: Verlag Franz Vahlen GmbH.

Edmonston, B., and J. S. Passel. 1992. Immigration and immigrant generations in population projections. *International Journal of Forecasting* 8:459–476.

Elliott, J., R. Fowler, and T. O'Brien. 2008, May. *Harnessing hibernate*. First ed. O'Reilly Media.

Espenshade, T. J., L. F. Bouvier, and W. B. Arthur. 1982. Immigration and the stable population model. *Demography* 19:125–133.

Ewald, R., J. Rössel, J. Himmelspach, and A. M. Uhrmacher. 2008. A plug-in - based architecture for random number generation in simulation systems. In *Proceedings of the 2008 Winter Simulation Conference*, 836–844. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gampe, J., and S. Zinn. 2007. Description of the microsimulation model. *MicMac Work package 2, MPIDR*.

Hibernate. http://www.hibernate.org/. Accessed 07/05/2009.

Himmelspach, J., R. Ewald, and A. M. Uhrmacher. 2008. A flexible and scalable experimentation layer. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. Mason, R. Hill, L. Moench, and O. Rose, 827–835. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Himmelspach, J., and A. M. Uhrmacher. 2007. Plug'n simulate. In *Proceedings of the 40th Annual Simulation Symposium*, 137–143: IEEE Computer Society.

Jeppers. http://sourceforge.net/projects/jeppers/. Accessed 07/05/2009.

JMathTools. http://jmathtools.berlios.de. Accessed 07/05/2009.

Keyfitz, N. 1981. The limits of population forecasting. *Population and Development Review* 7 (4): 579–593.

Lamb, V. L., and J. S. Siegel. 2004, March. Health Demography. In *The Methods and Materials of Demography* (2nd ed.)., ed. D. A. Swanson and J. S. Siegel, 341–370. Academic Press.

L'Ecuyer, P. 1997. Uniform random number generators: a review. In *Proceedings of the 1997 Winter Simulation Conference*, 127–134. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lee, R. D., and L. R. Carter. 1992. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87 (419): 659–671.

Lee, R. D., and S. Tuljapurkar. 1994. Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association* 89 (428): 1175–1189.

Leslie, P. H. 1945. On the use of matrices in certain population mathematics. *Biometrika* 33 (3): 183–212.

Leye, S., J. Himmelspach, M. Jeschke, R. Ewald, and A. M. Uhrmacher. 2008. A grid-inspired mechanism for coarse-grained experiment execution. In *DS-RT '08: Proceedings of the 2008 12th IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications*, 7–16. Washington, DC, USA: IEEE Computer Society.

Lutz, W., W. C. Sanderson, and S. Scherbov. 1998. Expert-based probabilistic population projections. *Population and Development Review, Supplement: Frontiers of Population Forecasting* 24:139–155.

Matsumoto, M., I. Wada, A. Kuramoto, and H. Ashihara. 2007, September. Common defects in initialization of pseudorandom number generators. *ACM Trans. Model. Comput. Simul.* 17 (4).

MicMac. http://www.nidi.knaw.nl/en/micmac/. Accessed 07/05/2009.

Mitra, S. 1983. Generalization of immigration and the stable population model. *Demography* 20:111–115.

Pflaumer, P. 1988. Confidence intervals for population projections based on Monte Carlo methods. *International Journal of Forecasting* 4:135–142.

Preston, S. H., P. Heuveline, and M. Guillot. 2001. *Demography. Measuring and Modeling Population Processes*. Blackwell Publishers.

Rogers, A. 1995. Population forecasting: Do simple models outperform complex models? *Mathematical Population Studies* 5 (3): 187–202.

Salzmann, T., and C. Bohk. 2006, March. The unexpected results produced by a specific design of a probabilistic population projection model. In *Proceedings of the Population Association of America 2006 Annual Meeting*. Los Angeles.

Satyabudhi, B., and S. Onggo. 2008. Parallel discrete-event simulation of population dynamics. In *Proceedings of the Winter Simulation Conference 2008*, 1047–1054.

Schmertmann, C. P. 1992. Immigrant's ages and the structure of stationary populations with below-replacement fertility. *Demography* 29:595–612.

Smith, S. K., J. Tayman, and D. A. Swanson. 2001. *State and Local Population Projections. Methodology and Analysis*. Kluwer Academic/Plenum Publishers.

Statistics Netherlands 2005. Changing population of Europe: uncertain future. *Final Report*.

Stoto, M. A. 1983. The accuracy of population projections. *Journal of the American Statistical Association* 78 (381): 13–20.

Swanson, D., and D. Beck. 1994. A new short-term county population projection method. *Journal of Economic and Social Measurement* 20:25–50.

Taleb, N. N. 2008, March. *The Black Swan. The Impact of the Highly Improbable*. Random House Inc.

The Colt Project. http://acs.lbl.gov/ hoschek/colt/. Accessed 07/05/2009.

The MathWorks. http://www.mathworks.com/products/matlab/. Accessed 07/05/2009.

van Imhoff, E., and W. Post. 1998. Microsimulation methods for population projections. *Population: An English Selection, New Methodological Approaches in the Social Sciences* 10:97–138.

Whelpton, P. K. 1928. Population in the United States, 1925-1975. *American Journal of Sociology* 34 (2): 253–270.

Whelpton, P. K. 1936. An empirical method of calculating future population. *Journal of the American Statistical Association* 31 (195): 457–473.

Willekens, F. J. 2005. Biographic forecasting: bridging the micro-macro gap in population forecasting. *New Zealand population review* 31 (1): 77–124.

Willekens, F. J. 2006. Description of the multistate projection model (multistate model for biographic analysis and projection). *MicMac Work package 1, NIDI*.

AUTHOR BIOGRAPHIES

CHRISTINA BOHK holds a diploma in Demography from the University of Rostock and pursues a PhD at the Institute of Sociology and Demography at the University of Rostock. Her main research interests are in probabilistic population projections. Her email address is <christina.bohk@uni-rostock.de>.

ROLAND EWALD holds a diploma in Computer Science from the University of Rostock and pursues a PhD at the Modeling and Simulation Group at the University of Rostock. His main research interests are in simulation algorithm selection and performance analysis. His email address is <roland.ewald@uni-rostock.de>.

ADELINDE M. UHRMACHER is an Associate Professor at the Department of Computer Science at the University of Rostock and head of the Modeling and Simulation Group. Her research interests are in modeling and simulation methodologies and their applications. Her e-mail address is <lin@informatik.uni-rostock.de> and her Web page is <www.informatik.uni-rostock.de/~lin>.