A different Understanding of Probability in a Probabilistic Population Projection Model and its Outcomes

Christina Bohk and Thomas Salzmann

Introduction

In general, population projections have a high relevance for societal aspects, because they compute the future population and its composition by age, sex and perhaps more features. Therefore, they are the base for important, e.g. political, decisions that influence the future of a society.

Some scientists differentiate between forecasts and projections. For them, the most important difference between a forecast and a projection is the likelihood of their outcomes. In that meaning, a projection is the numerical outcome of a specific set of assumptions. Consequently, the likelihood of the outcomes of a projection does not need to be very high, because they only want to show the population dynamic consequences of some assumptions. In comparison, forecasts are projections, too, but they want to make an accurate prediction of the future population and its composition. So, the likelihood of the outcomes of a forecast is assumed to be very high.

These definitions imply that population projections only can have a computation mistake, while forecasts even can be designated as false when the predicted population values differ from the actually observed population values in the future.

In this paper, we will describe the creation of a new designed probabilistic population projection model. With this projection model, we want to make projections and forecasts, although we will only speak of population projections.

Again, population projections are often used as a rational basis for decision-making. Changes in population size and its composition concerning the age- and sex structure have social, economic, environmental, and political effects. Therefore, population projections are often used as a basis for further projections like projections for households, unemployment, (energy or other goods) consumption, and diseases. A consequence of projected trends can be for instance the reconstruction of the health care system, the pension system, the range of a company's articles, and so on.

Being aware of the societal importance of accurate population projections, it is indispensable to improve existing population projection methods.

Progress from deterministic to probabilistic projection methods

During the last decades, a progress from deterministic to probabilistic (or stochastic) population projection methods took place. This was an important enhancement to capture the forecast uncertainty of the future evolution of vital rates, and in consequence of future population size and its composition by age and sex.

Deterministic population projections only have one assumption matrix for each input parameter. Therefore, the outcome of a deterministic population projection has no occurrence probability. For this reason, producers of deterministic population projections, like the Federal Statistical Bureau of Germany, often provide alternative scenarios as an indicator for forecast uncertainty. But the problem of this procedure is that users of these deterministic population projections can be easily confused when deciding which one of the alternative scenarios is the most likely one.

Since the deterministic population projections still capture the forecast uncertainty, even with alternative scenarios, imperfectly, probabilistic or stochastic population projections were established.

Probabilistic population projections have several different assumption matrices for each input parameter. These assumption matrices can be generated with different methods, e.g. with complex extrapolation methods like stochastic processes, as it is for example an ARIMA time series model. After generating these assumption matrices, they are initiated in the computation process, which is often an application of the cohort-component method, for n trials. Consequently, n result path matrices arise and the outcomes can then be assigned to occurrence probabilities.

Well-known probabilistic population projection methods

There are several different widely applied probabilistic population projection methods that produce so called prediction intervals for the projected future total population and other characteristic output quantities. Prediction intervals, or confidence intervals as they are often called, too, assert an occurrence probability to a projected future population size. Assuming that the underlying assumptions of the population projection will hold, a prediction interval of 80 per cent means that the projected future population size will range with a probability of 80 per cent between the two values a and b.

These prediction intervals can be computed with different well-known probabilistic population projection methods or approaches: e.g. with models that use time series models to project the total population size or vital rates (e.g. Lee and Carter 1992, Lee and Tuljapurkar 1994, Alho 1990), with projection models that base on expert-judgement (Lutz, Sanderson and Scherbov 1999), with projection models that base on

regression techniques (Swanson and Beck 1994), with models that base on Monte Carlo simulations, e.g., for fertility and migration rates (Pflaumer 1988), and with models that investigate errors of past population projections to construct prediction intervals for present population projections (e.g. Keyfitz 1981, Stoto (1983)).

Most of these well-known probabilistic population projection approaches need long historical data series to allow the application of their complex model-based statistical methods. Therefore, they are exposed to data errors, and to errors that occur while estimating the values for the input parameters. Another important thing to caution is that these probabilistic population projection approaches are often restricted to a small range of values concerning the assumption matrices of each input parameter, and to an always similar pattern of these assumptions for the vital rates, because of the model-based statistical method they use. Additionally, these probabilistic approaches are often tailored to a single statistical method, although a variety of different statistical methods might capture the forecast uncertainty better. Furthermore, some of these probabilistic projection models do not consider all subpopulations[1] separately (e.g. natives, immigrants and their descendants, emigrants and their descendants), although this aspect may have a great impact on the projection results.

Despite these above mentioned disadvantages of the well-known probabilistic population projection approaches, they have the big advantage of the prediction intervals. Although these prediction intervals are often very wide, they express the forecast uncertainty very well.

Creation of the novel Probabilistic Population Projection Model (PPPM)

Considering the restrictions of popular probabilistic population projection approaches, we introduce a novel Probabilistic Population Projection Model (PPPM) that, instead of improving already existing approaches, introduces a different meaning of probability. Therefore, the PPPM also introduces a new method of how to implement this different meaning of probability in the complex computation process.

Meaning of probability in the PPPM

The method of some well-known probabilistic population projection approaches can be roughly described as a k-folded iteration process of assumption generation with time series models. They are used to obtain k random sequences for specific vital rates, for a period t to $t + n$. By inserting each sequence into cohort-component matrices, one obtains k outcomes, e.g., k total populations. To denote the occurrence probability of the k total populations, several prediction or confidence intervals can be computed.

[1] see section: "The structure of the Probabilistic Population Projection Model (PPPM)", page iv

In contrast to the well-known probabilistic population projection approaches, the PPPM´s probability for the incidence of a specific output quantity has another origin. The procedure of the PPPM consists of the use of various exogenous assumption matrices for each input parameter[2]. The exogenous assumption matrices are not generated with a predetermined statistical method. Therefore, they can be generated by, e.g., simple or complex extrapolation methods like stochastic time series processes (see:, Alho 1990, Lee and Carter 1992, Lee 1992, Pflaumer 1992, Lee and Tuljapurkar 1994, Keilman and Pham 2000), expert-judgement (see: Keyfitz 1982, Lutz, Sanderson and Scherbov 1996), regression techniques (see: Swanson and Beck 1994), Monte Carlo simulation (see: Pflaumer 1988), or a mixture of different statistical methods.

The only sufficient condition for using the PPPM is the existence of assumption matrices for the specific input parameters for all subpopulations[3] for the complete projection horizon.

Once we receive several inputs for *i* future fertility trends of a certain subpopulation (e.g., *i* = 10), we allocate to each fertility trend an occurrence probability by expert-judgement. There is no reason to restrain the number of *i,* but all allocated probabilities have to add to one. By the use of these occurrence probabilities each assumption matrix gets a special weight. The higher the occurrence probability of an assumption matrix, the more often it will be chosen by the PPPM (considering the occurrence probabilities and random numbers) in the progression of *n* trials. Similar to the above mentioned most previous approaches, the results – e.g. *n* future population sizes – range between the limits of several estimated prediction intervals.

The structure of the Probabilistic Population Projection Model (PPPM)

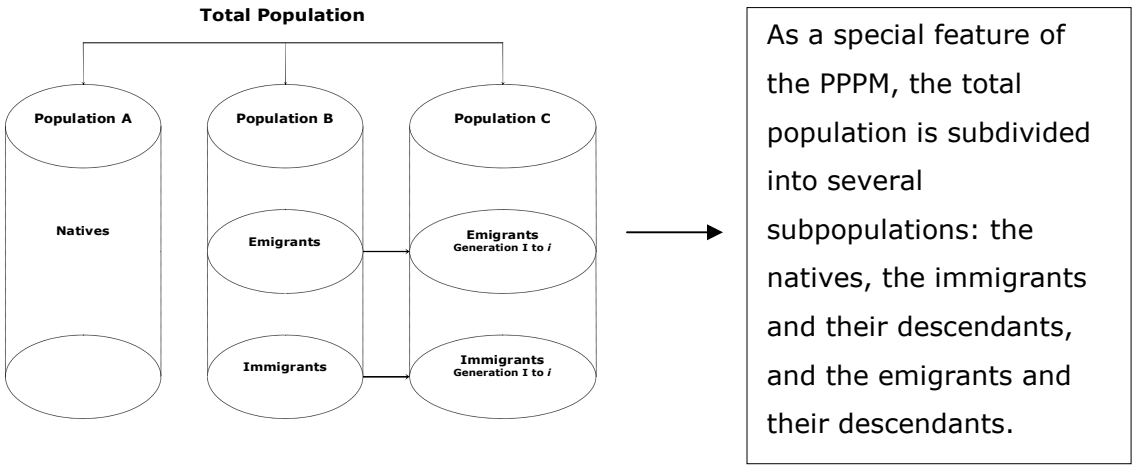
The PPPM is based on a combination of several ideas. The basic idea of the PPPM is the combination of an extended cohort-component method and the work by Espenshade, Bouvier and Arthur (1982), who divide an aggregate-population into subpopulations. But unlike to the Anglo-American literature dividing a given population into natives and migrants (see: Espenshade, Bouvier and Arthur 1982, Mitra 1983, Cerone 1987, Schmertman 1992), we use a more detailed partition.

According to Edmonston and Passel (1992) and Dinkel (2006), we will divide a given population into three main subpopulations A, B, and C. Subpopulation A comprises of people who live in a defined spatial area at a given time *t* (domestic population). The migrants in a period of time from *t* to *t* + *n* are defined as subpopulation B. Subpopulation C emanates from subpopulation B – the children, grandchildren, and great

[2] see section: "The input parameters of the PPPM", page v
[3] see section: "The structure of the Probabilistic Population Projection Model (PPPM)", page iv

grandchildren etc. of the migrants (see figure 1). Subjected to a common fertility distribution of industrialized countries, up to six generations could be expected to be computed with a projection horizon of 100 years.

Figure 1: General description of the structure of subpopulations in the PPPM



Because of its self-producing generations (see: Dinkel 2006), the assembled subpopulation C becomes even larger with continuing time when the Net Reproduction Rate is greater or equal to one. Dealing with the classification of A, B, and C, it must be pointed out that the development of the total number of subpopulation A is independent from the other two subpopulations B and C. Additionally, population C is unidirectional dependent from population B.

For both sexes, all subpopulations include single age calculations for every year as matrices. Furthermore, the created model structure of the PPPM enables us to determine the projection horizon for generating short, middle, and long term projections.

A specific feature of the PPPM structure is the partition of subpopulation B and C into immigrants and emigrants. This allows us to examine all population dynamic effects of migration, e.g., to specify the effects of the reproduction value for the final net migrant population. Immigrant and emigrant populations are calculated separately.

The input parameters of the PPPM

The original cohort-component method is based on the population balance equation. It requires input data for fertility, mortality, and migration in an annual age-sex composition.

In the PPPM, this method is used as a basic framework, and it is enlarged by more detailed input parameters. In general, the age-specific fertility rates, the survivors at age

x (l(x)), the total numbers of immigrants and emigrants at age *x*, the sexual proportion at birth, and the initial population for every single age *x* up to 100+ belong to the input parameters. Moreover, the mortality is modelled in a more sophisticated way: The survival probability for persons in the open end age interval (100+) and the specific distribution of infant mortality within the first year of life belong to the input parameters as well, and the survivors at age *x* are used to compute semi-annual death probabilities.

All input parameters are generated separately for each subpopulation A, B, and C.

The Open Type and the Limited Type of the PPPM

For a better understanding of the Open and the Limited Type of the PPPM, new terms and definitions have to be introduced. In the PPPM, a *variable* denotes an input parameter for a specific subpopulation. For example, the age-specific fertility rates of subpopulation A are one variable. To conduct a population projection, each variable has to be defined by a set of assumption matrices. The matrices are in turn associated with their occurrence probabilities, which have to add to 1 for each variable. To calculate one projection, the PPPM considers the occurrence probabilities of each variable's assumption matrices and chooses one for each variable randomly. This approach is called the *Open Type* of the PPPM.

However, the Open Type allows *implausible* combinations of assumption matrices. Such combinations occur when assumption matrices, based on contradictory assumptions, are chosen to generate one projection trial. Consider the fertility rates of two subpopulations, one being set to an assumption matrix that assumes a generally high fertility, and the other one assuming the opposite. Clearly, a combination of both is not reasonable[4].

To overcome this problem, the PPPM was extended by a *Limited Type*. For the Limited Type, we introduce the notions of *Sets* and *Set Types*. A Set Type is the set of all variables that define the same input parameter for different subpopulations, e.g. the set of all age-specific fertility rates[5]. For each Set Type, several Sets can be defined by the user. A Set consists of the assumption matrices for each variable of the corresponding Set Type. Each Set is associated with a specific occurrence probability. The occurrence probabilities of a Set Type's Sets have to add to 1. Similarly to the Open Type, occurrence probabilities are also assigned to the assumption matrices of each variable, which have to add to 1 as well. To calculate a projection, the PPPM chooses a Set for

[4] Consider a population projection with a projection horizon of 45 years; a random combination of a fertility matrix for subpopulation A with an increasing trend over time from t = 1(TFR = 0.9) to t = 45 (TFR = 1.2) and a fertility matrix for the immigrants from t = 1(TFR = 1.7) to t = 45 (TFR = 2.1)

[5] As another example, a Set Type of mortality can include the survivors at age *x* (l(x)) of each subpopulation [l(x), male subpopulation A; l(x), female subpopulation A; l(x), male subpopulation B (immigrants), …])

each Set Type. The chosen Sets define all possible assumption matrices for each variable. Analogical to the Open Type, the assumption matrix to be used for the projection is chosen randomly for each variable. By restricting each Set to hold assumption matrices which do *not* base on contradictory assumptions, the user is now able to eliminate the implausible combinations. This is the main advantage of the Limited Type of the PPPM.

To illustrate the function of Set Types and Sets, an example is given. The variables, representing one input parameter of different subpopulations, are assembled to a certain Set Type – for example, the age-specific fertility rates of all subpopulations can be combined in the Set Type "Fertility" (see figure 2 and 3).

Figure 2: General description of a Set Type

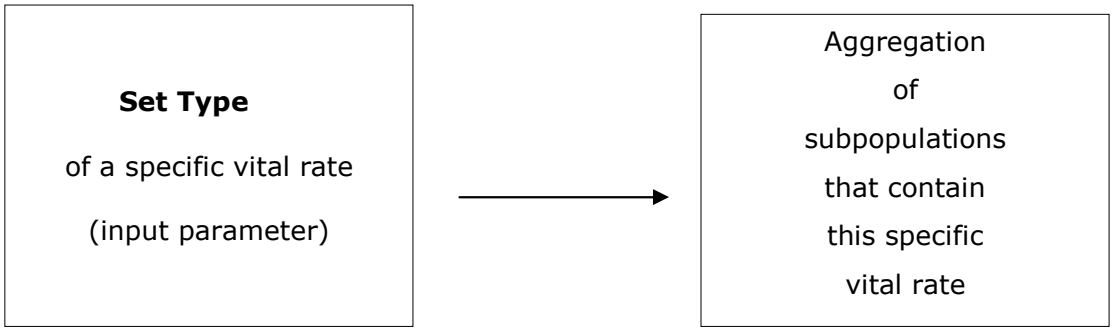
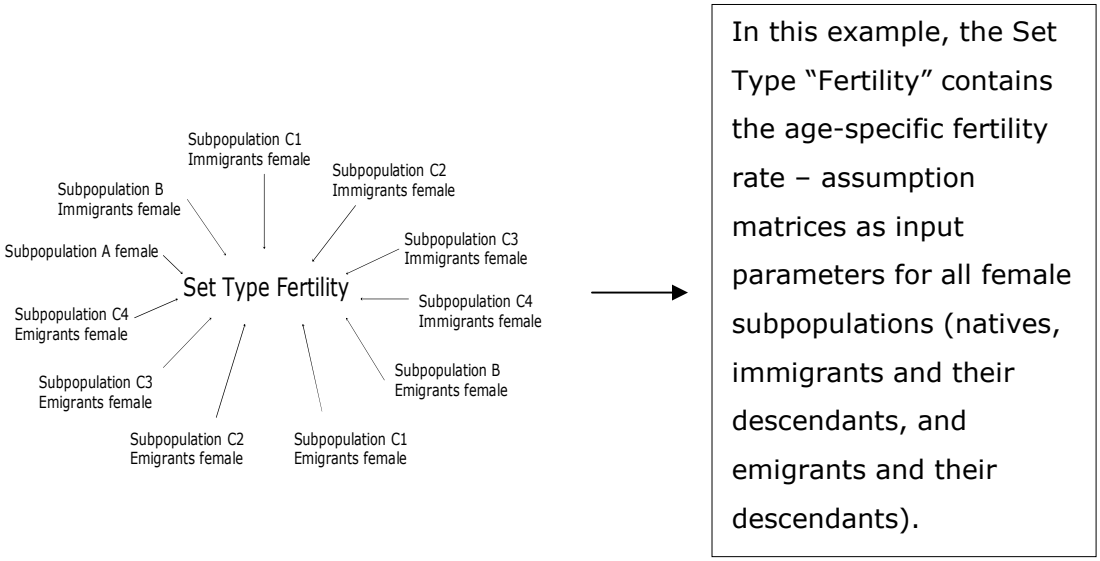


Figure 3: Example of the Set Type "Fertility"



In the next step, several Sets can be defined for this Set Type "Fertility". Each of these Sets includes predetermined combinations of assumption matrices for age-specific fertility rates, and for each subpopulation. One of these Sets could contain lower age-specific fertility rates as assumption matrices; while another Set could comprise higher age-specific fertility rate assumption matrices (see figure 4 and 5).

It is important to notice that the assumption matrices of the age-specific fertility rates in a Set for the Set Type "Fertility" are given separately for each subpopulation. Consequently, every subpopulation can have its own fertility assumption matrices. Additionally, a Set can contain more than one assumption matrix of the age-specific fertility rates for a subpopulation. There is no upper limit for the number of the assumption matrices for a subpopulation in a Set.

Figure 4: General description of a Set

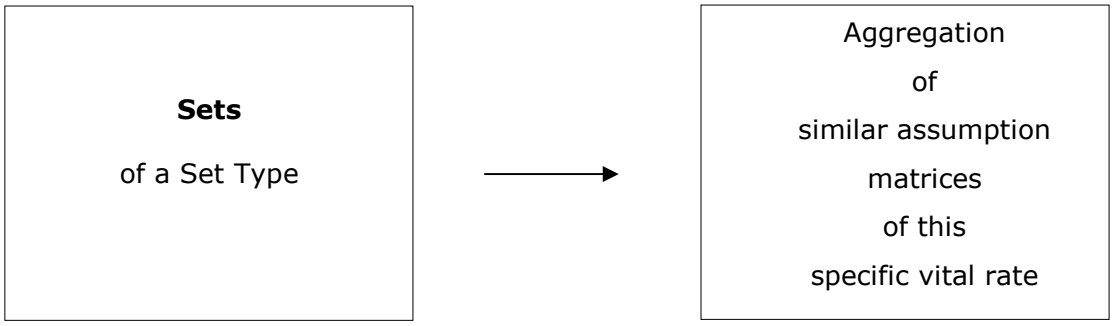
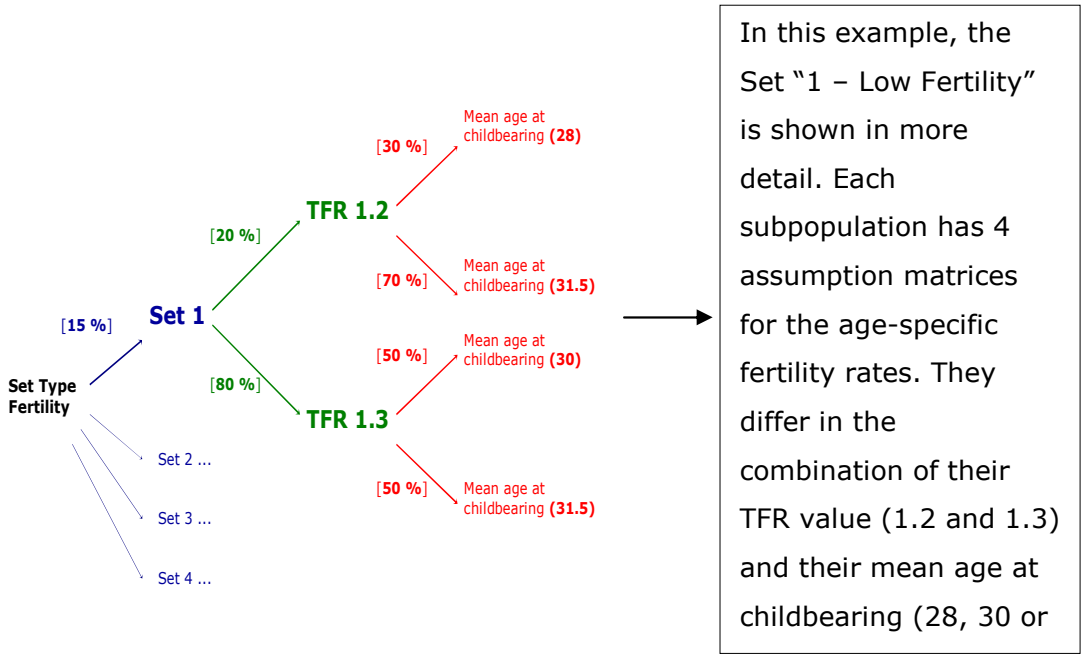


Figure 5: Example of the Set "1 - Low Fertility", based on the Set Type "Fertility". For simplicity, subpopulation distinctions are left out.



In the example depicted in figure 5, the presented subpopulation has four age-specific fertility-rate assumption matrices. They differ in their TFR value and in their mean age at childbearing. Moreover, the Set "1 – Low Fertility" has an occurrence probability of 0.15, and the age-specific fertility rate assumption matrix with a TFR value of 1.2 has an occurrence probability of 0.2.

One characteristic of the Open Type of the PPPM is the freedom in the combination of all assumption matrices. Therefore, the results are not influenced by predetermined combinations, but only by occurrence probabilities. The distribution of the occurrence probabilities implicitly regulates that the unrealistic, the realistic and the more realistic combinations have a low, high and higher occurrence probability, respectively. By using this procedure and this knowledge, the implausible combinations can be reduced to a minimum in the Open Type of the PPPM, but they still exist.

In contrast to the Open Type, the Limited Type of the PPPM eliminates the implausible combinations by using Set Types and Sets. However, there are predetermined combinations of the assumption matrices which could lead to a restriction of the probability in the PPPM. The degree of restriction can be defined by the user, by creating Sets with more or less similar assumption matrices.

The PPPM is implemented in MatLab (version 6.5) and uses MatLab's random number generator. Further improvements concerning the theoretical model and the computer program will be subject of future research.

Findings of the application of both types of the PPPM

In the master thesis of Bohk 2004, the Open Type of the PPPM was developed, on the base of Dinkel's deterministic population projection model (Dinkel 2006). An application of this Open Type has shown that the prediction intervals of the results were very wide. Consequently, the theory and the method of the PPPM were revised to reduce the range of the outcomes, and the idea of the Set Types and Sets arose.

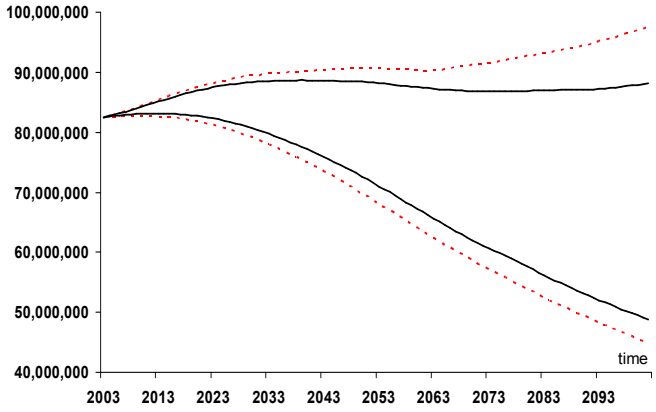
The aim of the following application of both types of the PPPM is to show how the prediction or confidence intervals vary between these two types due to their different structure. According to the evolutionary history of the theory and the method of the PPPM, we expected wider prediction intervals for the Open Type than for the Limited Type.

For this application of the PPPM, we generated several assumption matrices for all input parameters. Simplistically summarised, there are 4 Sets for the Set Type "fertility",

15 Sets for the Set Type "mortality", and 7 Sets for Set Type "migration". This implies in this example 18 fertility-, 30 mortality- and 14 migration-assumption matrices for the Open Type. All these assumption matrices differ more or less from the assumptions of the 10. coordinated population forecast of Germany, which is calculated by the Federal Statistical Bureau of Germany. After generating the assumptions for all input parameters, 5000 trials for each type of the PPPM were computed.

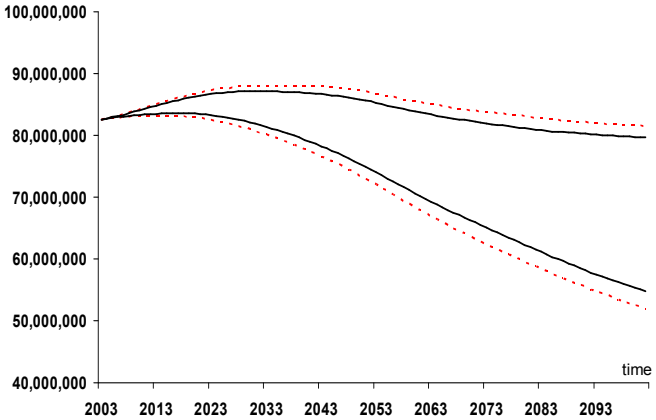
The 95 and the 80 per cent prediction or confidence interval of the total population at the end of a year are surprising (see figure 6 and 7).

Figure 6: 95% prediction interval of the total population at the end of a year



In this example, the solid line represents the upper and lower bound of the Open Type, and the dotted line represents the upper and lower bound of the Limited Type of the 95 per cent prediction interval after 5000 trials.

Figure 7: 80% prediction interval of the total population at the end of a year



In this example, the solid line represents the upper and lower bound of the Open Type, and the dotted line represents the upper and lower bound of the Limited Type of the 80 per cent prediction interval after 5000 trials.

Disproving our expectations, the outcome range of the Open Type is significantly narrower than that of the Limited Type of the PPPM. This finding can be explained with a closer look to the implausible combinations.

We thought that we are able to reduce the range of the outcomes by the elimination of the implausible combinations through the Set Types and Sets in the Limited Type. But, as an analysis of the results shows, these implausible combinations cause mainly middle or average result paths. This is because of the change between low and high assumption matrices from subpopulation to subpopulation in one trial.

For example, a combination of a TFR value of 1.1 for subpopulation A, 1.8 for subpopulation B and 1.3 for their descendents (subpopulation C) causes a middle result path. Roughly described, high assumptions are balanced by low assumptions.

Consequently, the frequency of middle result paths is higher in the Open Type and the range of the prediction or confidence interval becomes even narrower.

But, this does not mean that the creation of the Set Types and Sets does not reduce the uncertainty. They actually reduce uncertainty in the computation process; but in fact, there is much more uncertainty about a population's future evolution or development. This fact is expressed by the wider prediction intervals of the results of the Limited Type of the PPPM. In other words, the Open Type of the PPPM may cause its users to imagine an uncertain forecast to be certain.

Literature:

Alho, Juha M. (1990): "Stochastic methods in population forecasting", International Journal of Forecasting, 6, 521-530.

Bohk, Christina (2004): "Der Einbau probabilistischer Annahmen in ein Bevölkerungsprognosemodell" (Master Thesis, University of Rostock)

Carter, Lawrence R. and Ronald D. Lee (1992): "Modelling and forecasting US sex differentials in mortality", International Journal of Forecasting, 8, 393-411.

Cerone, Pietro (1987): "On Stable Population Theory with Immigration", Demography, 24, 431-438.

Dinkel, H. Reiner (2006): "Demographie. Bd. 3: Demographie der Migration", forthcoming

Edmonston, Barry and Jeffrey S. Passel (1992): "Immigration and immigrant generations in population projections", International Journal of Forecasting, 8, 459476.

Espenshade, T.J., L.F. Bouvier, and W.B. Arthur (1982): "Immigration and the Stable Population Model", Demography, 19, 125-133.

Keyfitz, Nathan (1981): "The limits of population forecasting" Population and Development Review, 7, 579-594.

Keyfitz, Nathan (1982): "Can knowledge improve forecasts?", Population and Development Review, 8(4), 729-751.

Keilman, Nico and Dinh Quang Pham (2000): "Predictive intervals for age specific fertility" European Journal of Population, 16, 41-66.

Lee, Ronald D. and Carter, Lawrence R. (1992): "Modelling and forecasting US mortality" Journal of the American Statistical Association, 47, No. 14, 659-671.

Lee, Ronald D. (1992): "Stochastic demographic forecasting", International Journal of Forecasting, 8, 315-327.

Lee, Ronald D. and Shripad Tuljapurkar (1994): "Stochastic Population Forecasts for the United States: Beyond High, Medium and Low", Journal of the American Statistical Association, 89, 1175-1189.

Lutz, Wolfgang and Scherbov, Sergei. (1998): "An expert-based framework for probabilistic national population projections: The example of Austria", European Journal of Population, 14, 1-14.

Lutz, Wolfgang, Sanderson ,Warren C. and Scherbov, Sergei. (1996): „Probabilistic population projections based on expert opinion", In: The future population of the world. What can we assume today ?, Lutz, Wolfgang (ed.), Laxenburg (Austria), International Institute for Applied Systems Analysis, 397-428.

Mitra, S. (1983): "Generalization of the Immigration and the Stable Population Model", Demography, 20, 111-115.

Pflaumer, Peter (1988): "Confidence Intervals for population projections based on Monte Carlo Methods", International Journal of Forecasting, 4, 135-142.

Pflaumer, Peter (1992): "Forecasting US population totals with the Box-Jenkins approach", International Journal of Forecasting, 8, 329-338.

Saboia, J. L. M. (1974): "Modelling and forecasting populations by time series: The Swedish Case." Demography, 2, 483-492.

Schmertman, Carl P. (1992): "Immigrants' Ages and the Structure of Stationary Populations with Below-Replacement Fertility", Demography, 29, 595-612.

Stoto, M. E. (1983): "The accuracy of population projections" Journal of the American Statistical Association, 78, 13-20.

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

Sykes, Z. M. (1969): "Some stochastic versions of the matrix model for population dynamics" Journal of the American Statistical Association, 44, 111-130.