The unexpected results produced by a specific design of a Probabilistic Population Projection Model

Thomas Salzmann and Christina Bohk

Population projections have a high relevance for several societal aspects, for instance for political, economical and social decisions. The progress from deterministic to probabilistic projection models was an important attempt to capture the forecast uncertainty of the future evolution of the vital rates. Previous probabilistic approaches (e.g., Sykes 1969 and Saboia 1974) were followed by extended and improved ones (see: Keyfitz 1981, Stoto 1983, Alho 1990, Pflaumer 1992, Lee-Carter 1992, Lee and Tuljapurkar 1994, Lutz and Scherbov 1998). But most of these probabilistic or stochastic approaches still capture the forecast uncertainty with some restrictions.

In our presentation we discuss our findings using different types of one probabilistic population projection model. But in spite of following some different well known stochastic projection approaches [abovementioned], we are going to introduce a novel probabilistic population projection model (PPPM). The PPPM does not generate probability by generating assumption sequences with stochastic processes that are implemented into cohort-component matrices. Rather, our PPPM focuses on the method of using probability in a complex projection model. Therefore, a different understanding of probability is applied.

The meaning of probability in the PPPM

The main characteristic of the PPPM is the understanding of probability in the projection model. Each existing probabilistic projection approach uses a certain method to calculate the probability distribution for the future population size: Keyfitz (1981) and Stoto (1983) elaborate past population projections and use their observed errors to construct confidence intervals for the computed future population sizes of present projections. Furthermore, other approaches use complex time series models with random elements (e.g. ARIMA) to generate future vital rates (see: Alho 1990, Lee-Carter 1992, Lee 1992, Lee and Tuljapurkar 1994), while for the same purpose Pflaumer (1988) uses a Monte Carlo Simulation. The method of the last two mentioned approaches can be roughly described as an iteration process of the assumption generation with time series models for k times with k random sequences for specific vital rates for a period t to $t + n$. By inserting every single sequence into the cohort-component matrices, one obtains k outcomes for k population sizes. To denote the probability of occurrence of the k population sizes, several confidence intervals can be computed.

In our PPPM, the probability for the incidence of a specific output parameter[1] has another origin. The procedure of the PPPM consists of the use of various exogenous assumption sequences for each input parameter[2]. The exogenous assumption sequences are generated with no predetermined method, therefore they can be generated by, e.g., simple or complex extrapolation methods, expert judgement (see: Keyfitz 1982, Lutz, Sanderson and Scherbov 1996), stochastic processes (see: Pflaumer 1988, Alho 1990, Lee and Carter 1992, Lee 1992, Pflaumer 1992, Lee and Tuljapurkar 1994, Keilman and Pham 2000) or a mixture of different methods.

The only sufficient condition for the PPPM is the existence of assumptions sequences for the specific input parameters for all subpopulations[3] for the 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 a probability of occurrence 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 probabilities of occurrence each assumption sequence gets a special weight. The higher the probability of occurrence of an assumption sequence, the more often it will be chosen by the random number generator 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 confidence intervals.

The structure of the Probabilistic Population Projection Model (PPPM)

Our 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 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

[1] e.g. population size, age dependency ratios etc.
[2] see section: "The input parameters of the PPPM"
[3] see section: "The structure of the Probabilistic Population Projection Model (PPPM)"

B – the children, grandchildren and great grandchildren etc. of the migrants. (Subjected to a common fertility distribution of industrialized countries up to seven generations could be expected to be computed with a projection horizon of 100 years.) Because of its self-producing generations (see: Dinkel 2006), the assembled subpopulation C becomes even larger with continuing time with a Net Reproduction Rate greater or equal 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 the whole 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 in form of fertility rates, death rates and the net migration in an annual age-sex composition. We use this method as a basic framework and enlarge it 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 regarded more sophisticated: 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. First of all, we introduce the term called Set Type. A Set

Type comprises k variables[4]. Each of these k variables is a combination of a specific input parameter and of different i subpopulations. The number (i) of subpopulations containing this specific input parameter corresponds to the number of variables included in a Set Type ($k = i$). Each variable is assigned with the assumption sequences of this input parameter.

With these predeterminations, it is possible to explain the difference between the Open Type and the Limited Type of the PPPM. The first one is the extent of a given combination level for a) the assumption sequences of the variables and b) the variables themselves. This distinction implicates the level of the extent of implausible combinations in the PPPM. Implausible combinations are randomly chosen assumption sequences, for instance, for fertility over all subpopulations in one trial which are in general not reasonable.[5] In general, the occurrence probability for such a case is not very high but its existence is not negligible. These implausible combinations are a part of the so called *Open Type* of the PPPM. The second difference concerns the allocation of occurrence probabilities to each assumption sequence. In the Open Type, each assumption sequence for an input parameter of a subpopulation has an occurrence probability. The accumulation of these occurrence probabilities has to be one.

The Limited Type of the PPPM is based on another idea. It tries to eliminate the implausible combinations through a presetting of combinations of a) the variables and b) the assumption sequences of these variables. The combination presetting of variables is performed by the implementation of *Set Types*, while the combination presetting of the assumption sequences of these variables is performed by the implementation of *Sets*.

The extension of the PPPM by Set Types and Sets eliminates the cause and the problem of implausible combinations. To illustrate the function of Set Types and Sets, an example is given. The variables of one input parameter of different subpopulations are assembled to a certain Set Type – for example, the fertility rates of all subpopulations can be combined in 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 the assumption sequences of the fertility rates. One of these Sets could contain lower fertility rates as assumption sequences; while another Set could comprise higher fertility rates. (It is important to notice that the assumption sequences of the fertility rates in a Set for the Set Type

[4] A variable is a combination of one specific input parameter and a subpopulation (e.g. 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), …])

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

"Fertility" are given separately for each subpopulation.) Additionally, a Set can contain more than one assumption sequence of the fertility rates for a subpopulation. There is no upper limit for the number of the assumption sequences for a subpopulation in a Set.
In the Limited Type, occurrence probabilities are assigned to a) all Sets of a Set Type (sum of all these allocated occurrence probabilities has to be one) and to b) all assumption sequences of a specific input parameter of a specific subpopulation in a Set (sum of all these allocated occurrence probabilities has to be one as well).

One characteristic of the Open Type of the PPPM is the freedom in the combination of all assumption sequences. Therefore, the results are not influenced by predetermined combinations but by the probabilities of occurrence. The distribution of the occurrence probabilities for each assumption sequence is – clearly defined - able to regulate that the unrealistic, the realistic and the more realistic combinations have a low, high and higher occurrence probability, respectively. By the use of this procedure and this knowledge, the implausible combinations can be reduced to a minimum in the Open Type of the PPPM, but they are still existing.

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 sequences which could lead to a restriction of the probability in the PPPM.

Findings of the application of both types of the PPPM

The findings concern the difference of the variance – e.g., the total population size at the end of the projection horizon - between both types of the PPPM. According to our preliminary ideas, we have expected that the range of the results of the Open Type would be wider than the range of the Limited Type of the PPPM. By reason of implausible combinations, we expected a larger variance in the outcomes of the Open Type compared to the Limited Type. But, disproving our expectations, the range of the outcomes 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.

Literature:

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

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

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.

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