Why probabilistic population projections can hardly be evaluated

Christina Bohk and Roland Rau

September 23, 2011

Abstract

Population projections have a high societal impact. Valid evaluation tools are needed to analyze their methodology and outcome. Error measures have been proposed and successfully applied in the past. However, essential improvements in projection methodology—like capturing future's uncertainty in probabilistic approaches—motivate to ask whether error measures are still qualified for evaluation tasks. In this paper, we reconsider using error measures to evaluate *probabilistic* population projections. When evaluating a probabilistic projection, the projected result distribution should be compared with the *real* distribution of possible developments, though the *real* distribution is not observable. The unobservability of the *real* distribution of potential future developments leads to erroneous evaluation results when using conventional error measures. Hence, we propose to use *ordinal similarity* between projected and actually observed outcome as a criterion to evaluate probabilistic population projections. Ordinal similarity indicates whether fundamental development of actual fertility, mortality, and migration has been projected.

1 Introduction

Population projections are often used as a basis for political, and economic decisions. Due to their importance for public policy it is indispensable to evaluate them with appropriate indicators.

A large number of evaluation criteria have been proposed and successfully applied in the past, but their qualification needs to be reconsidered due to essential improvements in projection methodology. In this paper, we raise some concerns over using conventional error measures to evaluate probabilistic population projections.

2 Population projection process

Several steps are needed to conduct a population projection: After selecting an appropriate projection model, assumptions have to be made for the development of fertility, mortality, and migration, i. e. the three core demographic parameters that shape population size and structure in the future.

To evaluate a population projection comprehensively, each step of a population projection process should be investigated thoroughly [1]. Objectivity, validity, reliability, and usability are possible criteria to evaluate input, computational model, and output of a population projection [2, 3, 4].

3 Evaluating projections with error measures

The most popular evaluation criterion is accuracy of the projected outcome. Several error measures have been developed to quantify projection error, i. e. the deviation between actual observed and projected population counts [5, 6, 7]. Error measures can be defined in different ways; they evaluate accuracy or bias of projected outcome, and they can be sensitive or robust towards outliers. According to these features, error measures lend

themselves to different purposes. They are most commonly used to analyze the predictive power of population projection models. Besides, they are used to scrutinize the impact of single model parameters on outcome quantities, to calibrate a projection model, and to compare different projection methods [8, 9]. When comparing projection methods, it is important to choose an appropriate error measure as well as to look at the projection conditions; for instance, the (in)stability of demographic developments or the length of the projection horizon can have a great impact on evaluation results. Hence, if projection methods are compared, observed projection errors can not be unambiguously assigned to the methods themselves, but also to the projection conditions [10].

4 Applicability of error measures for projection evaluation

4.1 Deterministic approaches

Conventional error measures typically evaluate deterministic population projections. Deterministic population projections assume one trajectory for each parameter and, therefore, generate only one result.[1] If this result is meant to forecast the actual future development, error measures can be easily applied to quantify projection error(s). However, deterministic population projections fail to capture inherent uncertainty of population projections. If the model is appropriate, and computational as well as data errors can be excluded, sources of uncertainty are triggered by the future development of fertility, mortality, and migration.

4.2 Probabilistic approaches

Probabilistic population projections capture and quantify this uncertainty by considering multiple weighted assumption paths per model parameter in order to estimate a multitude of results. Typically, probabilistic population forecasts summarize all results in an empirically derived distribution. Thus, evaluating probabilistic population projections implies to evaluate a distribution instead of a point estimate (as generated by deterministic population projections).

With respect to evaluating the projection outcome, probabilistic approaches are a mixed blessing. On the one hand, probabilistic approaches capture inherent uncertainty of a projection, and reflect reality more closely than deterministic projections. On the other hand, probabilistic approaches can hardly be evaluated because the requirements of a valid evaluation can not be satisfied.

In reality, we can only observe the actual development ex post. This may lead to the simple idea that reality was deterministic, i. e. that there really was only one development that can occur. Anyway, we assume that reality is rather stochastic than deterministic, i. e. that the actually observed development is only one realization of multiple possible realizations. The problem is, that we neither get to know all *real* possible developments nor their *true* occurrence probabilities.

This has serious consequences for evaluating probabilistic population projections. A valid evaluation of a probabilistic projection requires a comparison of the projected result distribution with the *real* distribution of possible developments, though the *real* distribution is not observable.

5 Evaluating probabilistic approaches

5.1 Problems

The unobservability of the *real* distribution of potential future developments leads to erroneous evaluation results when using conventional error measures. To illustrate this, we compare the evaluation results of four cases that differ in their actually observed development, and its location in the projected and in the *real*

[1] In the typical scenario-based deterministic projections, these are obviously as many results as there are combinations of the three parameters.

distribution of possible developments.

We presume that (1) reality is probabilistic, and that (2) the actually observed development is usually assumed to be the most likely one to occur, though this—depending on the *real* distribution—may or may not be true.

Figure 1 illustrates the four cases; the solid and dashed lines represent the projected and *real* distribution of possible developments respectively, and the red circle represents the actually observed development:

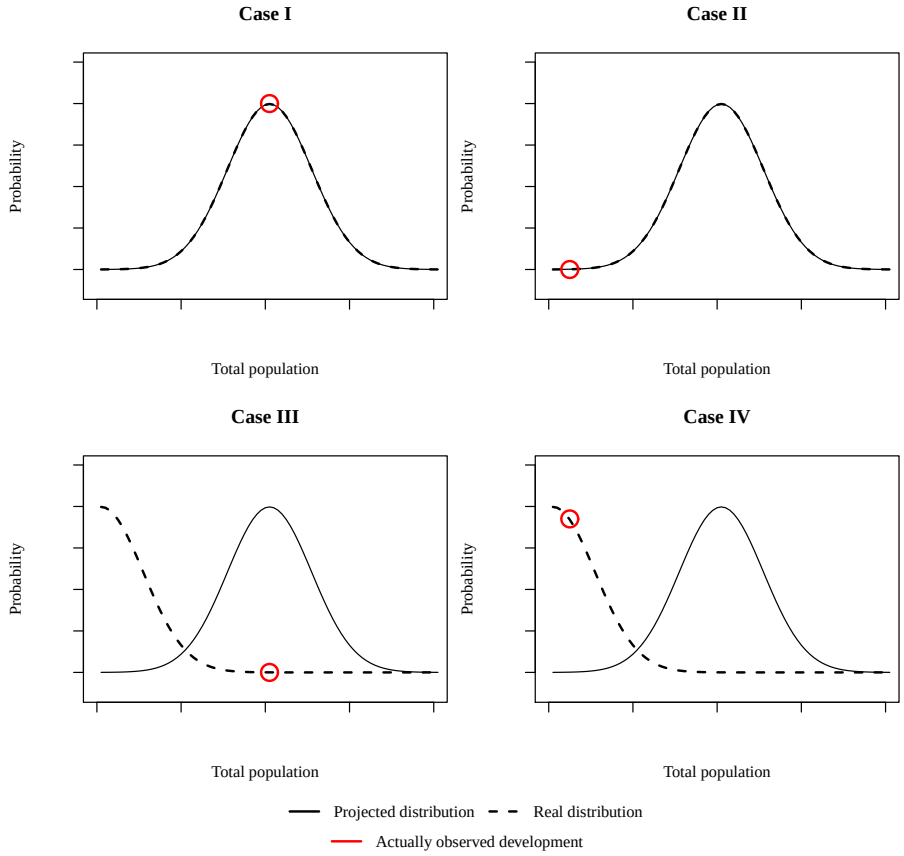

Figure 1: Four cases that illustrate wrong and correct evaluation results.

The projected and the *real* distribution of possible developments *coincide* in cases I and II (upper left and right plot in Figure 1), whereas they *differ* in cases III and IV (lower left and right plot in Figure 1). In each of these two cases, the actually observed development is projected to have a *high* occurrence probability (left plots in Figure 1), and to have a *low* occurrence probability (right plots in Figure 1).

Usual evaluations typically investigate if the actually observed development has been projected with a high occurrence probability. Presuming that we know the *real* distribution of possible developments, we compare these usual with the true evaluation results for each case:

Case I The projected and the *real* distribution of possible developments coincide, and the actually observed development is correctly projected to have a high occurrence probability. In this case, the projection would be usually evaluated to be correct. As there are no discrepancies to the real distribution, this usual evaluation result is truly correct.

Case II The projected and the *real* distribution of possible developments coincide, and the actually observed development is correctly projected to have a low occurrence probability. In this case, the projection would be usually evaluated to be less correct, because the actually observed development is barely in the projected distribution, and it has been projected to have a low probability to occur. However, the projection is truly correct, because the actually observed development is projected to be less likely and it actually had only a small probability to occur. Hence, usual evaluations would make a Type II error, i. e. a false negative evaluation, in case II.

Case III The projected and the *real* distribution of possible developments do not coincide, and the actually observed development is wrongly projected to be most likely although it actually had a low probability to occur. In this case, the projection would be usually evaluated to be correct, because the actually observed development is in the center of the projected distribution, and it has been projected to have a high occurrence probability. However, the projection is truly less correct, because the actually observed development really had only a small probability to occur. Hence, usual evaluations would make a Type I error, i. e. a false positive evaluation, in case III.

Case IV The projected and the *real* distribution of possible developments do not coincide, and the actually observed development is wrongly projected to be less likely although it actually had a high probability to occur. In this case, the projection would be usually evaluated to be less correct, because the actually observed development is barely in the projected distribution, and it has been projected to have a low probability to occur. However, the projection is truly less correct, because the actually observed development really had a high probability to occur. Hence, the usual evaluation would be correct, but only due to wrong reasons.

Table 1 summarizes these findings:

Usually evaluated as:	Correct	False
Truly:		
Correct	Case I	Case II
False	Case III	Case IV

Table 1: Possible evaluation results

Using conventional error measures to compare probabilistically projected with actually observed development may lead—in two of four cases—to wrong evaluation results. Cases II and III represent wrong evaluation results, and cases I and IV represent correct evaluation results, though case IV is a correct evaluation, but only due to wrong reasons.

Erroneous evaluation results may have a bad impact on the development of projection methodology. Improving a projection model on the basis of erroneous evaluation results may cause wrong calibration, and comparing different projection methods may evaluate actually accurate projection methods to be less accurate, and actually less accurate projection methods to be more accurate. This can have broad ramifications as population projections are often used as a basis for political, social, and economic decisions.

5.2 Proposed solutions

Evaluation possibilities for probabilistic projections depend to some extent on the purpose of the projection.

The purpose of a probabilistic projection can be the revelation of potential future developments, or the projection of the actual future development. If a probabilistic projection is meant to reveal potential future developments, its purpose is to project the *real* distribution of possible developments. A probabilistic projection with such a purpose can hardly be evaluated, because even though the actual development is known ex post, we still do not know which other developments would have been possible, and what their *true* occurrence probabilities would have been (see section 5.1). This is a fundamental problem that can not be solved easily. However, if—as pointed out by Keilman [11]—a probabilistic projection is meant to forecast the actual future development, and the uncertainty range represents expectable projection errors,[2] some possibilities remain to evaluate the success of such a probabilistic projection.

To examine remaining evaluation possibilities for probabilistic projections (of the latter category), we re-project the western German population for the ages 0 to 90+ from 1961 to 2000 with the Probabilistic Population Projection Model (PPPM) [12]. We assume actually observed mortality, fertility, and migration as mean estimates, and add random, normally distributed errors around them to represent projection uncertainty.[3] We conduct a total of 1000 trials.

In a first step, we can analyze whether the actually observed development has been projected at all, i. e. if it is in the result distribution of the probabilistic projection. Figure 2 therefore shows the projected as well as the actual total population for western Germany from 1961 to 2000.

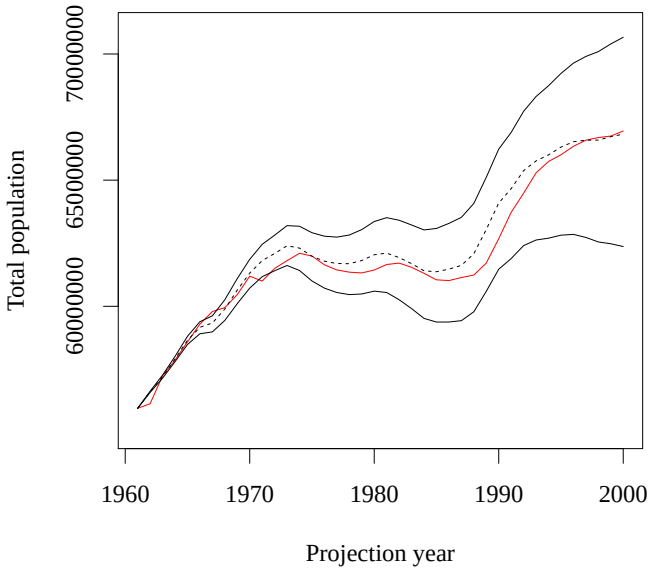

Figure 2: Projected (black) and actually observed (red) total population for western Germany from 1961 to 2000. Projected uncertainty is represented by quantiles 0, 0.5 (dashed line), and 1.

Apparently, the actually observed total population has been projected in almost all projection years.[4] Moreover, the trajectory of the actual total population is very similar to the projected development of quantile

[2] The uncertainty range consists of projection errors that can be interpreted as expectable deviations between most likely projected and actual development.

[3] According to the terminology of the PPPM, we have one representative assumption path with stochastic variation per model parameter.

[4] The few exceptions are presumably due to some irregularities in the official statistical data.

0.5. With such an analysis, we can evaluate whether the actual total population has been projected at all, but we can not evaluate if the projected occurrence probability is correct, since the *real* distribution of possible future developments is unknown.

In a second step, we can evaluate how *similar* the projected and the actually observed developments are. Both, size and (age)structure of a population, can be tested for similarity. For instance, if there was a population increase in reality, we can analyze how many projected results predict this population increase, and to what extent they follow the actually observed trajectory. We can also analyze if the projected age structure is similar to the actually observed one. Figure 3 therefore compares projected and actually observed age structure for western Germany in the projection year 2000.

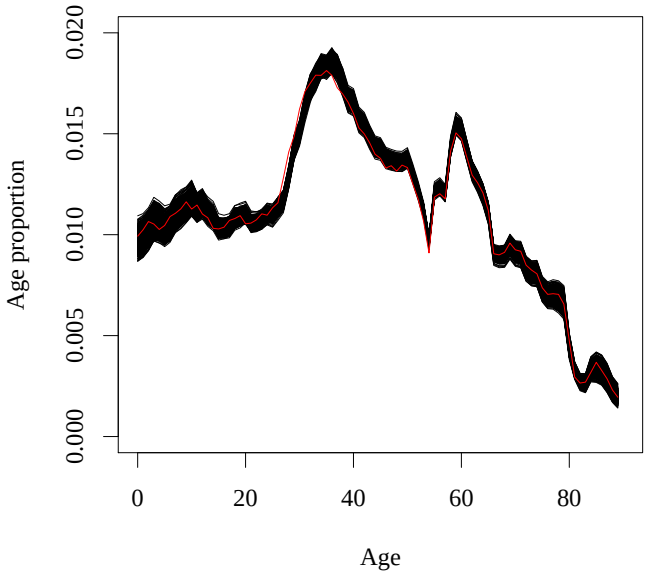

Figure 3: Projected (black) and actually observed (red) age structure for western Germany in 2000.

The more similar the projected and the actually observed outcome are, the better is the predictive power of a probabilistic projection. When analyzing similarity, we rather focus on *ordinal* than on cardinal similarity, to analyze whether the fundamental tendency of the actual development has been projected or not. Hence, ordinal similarity indicates whether fundamental developments have been predicted for the three core demographic parameters fertility, mortality, and migration, which influence future size and structure of a population.

A comparison of similarity patterns for multiple projection years can reveal, for instance,

1. if the projected age-structure is similar, too young, or too old
2. if dissimilarity starts to develop in the beginning, mid or end of a projection
3. if dissimilarity develops continuously
4. if there are ages with less, middle, and high dissimilarity in short-, medium-, and long-run projections
5. if there are cohort effects

6. if there are period effects

We made several attempts to test ordinal similarity between projected and actually observed outcome. As this is still work in progress, our focus lies rather on population structure than on population size for now. To compare the projected and actually observed population structure, we take *age proportions*, i. e. *relative* age structures.

One possibility to test ordinal similarity between relative age structures is to investigate if they have a similar and continuously developing error structure over time. A continuous development of projection errors by means of slightly increasing errors indicates that fundamental developments in fertility, mortality, and migration are properly captured in a projection. Therefore, we compute the absolute error AE for each age proportion x, trial n, and projection year t:

$$AE_{x,n,t} = |F_{x,n,t} - A_{x,t}| \tag{1}$$

The absolute error $AE_{x,n,t}$ is the absolute difference between the projected (F) and the actually observed (A) age proportion x of trial n in the projection year t. Having conducted 1000 trials, we have 1000 absolute errors for each age proportion and year. To compare these 1000 absolute errors for each age proportion and year over the whole projection horizon, we compute the mean of these absolute errors over all trials:

$$MAE_{x,t} = \frac{\sum_{n=1}^{1000} AE_{x,n,t}}{1000} \tag{2}$$

Figure 4 shows these mean absolute errors between projected and actually observed age proportions (over all trials) for each projection year.

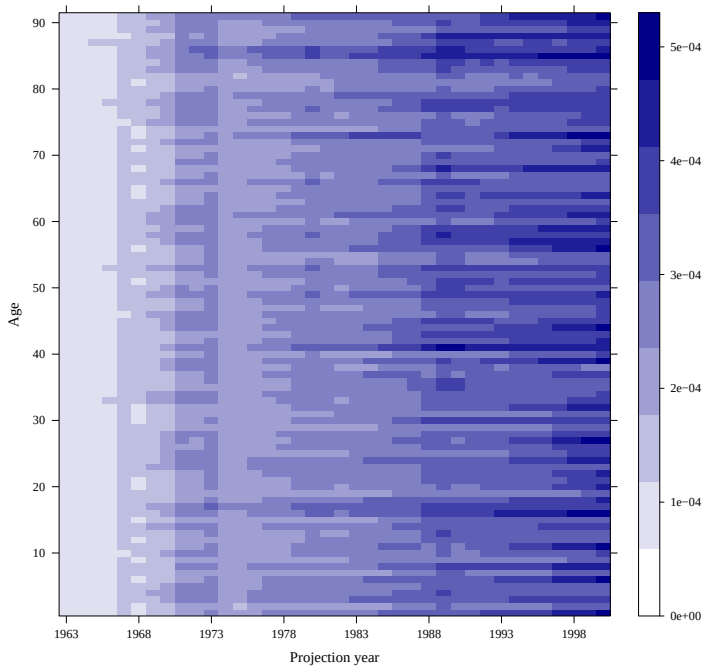

Figure 4: Mean absolute errors between projected and actually observed age proportions (over all trials) for each projection year. The higher a mean absolute error is, the darker is its color in the levelplot.

The comparison of the mean absolute errors for the age proportions over time reveals a continuous development, i. e. the errors slightly increase with time for all ages. This is no surprise, as we assumed actually

observed fertility, mortality, and migration in this projection. Hence, no apparent period or cohort effects can be observed. Nevertheless, mean absolute errors slightly differ among ages, especially in the later projection years.

To further analyze the error structure, we do not only look at the mean errors for single ages, but at the mean errors for the whole age structure. Therefore, we compute the mean of the absolute errors $AE_{x,n,t}$ over all ages:

$$MAE_{n,t} = \frac{\sum_{x=1}^{90+} AE_{x,n,t}}{91} \tag{3}$$

Since we have 1000 trials, we have 1000 of these mean absolute errors per projection year. To compare the error structure over time, we compute a cumulative density function for each projection year with the corresponding 1000 mean absolute errors.

Figure 5 shows these 1000 cumulative density functions of the mean absolute errors between projected and observed age structures for western Germany, from 1962 to 2000:

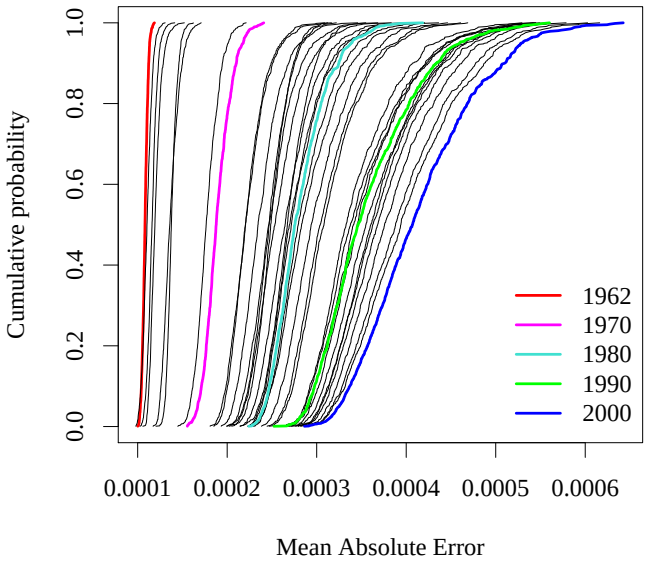

Figure 5: Cumulative density functions for the mean absolute errors between projected and actually observed age structures for western Germany, from 1962 to 2000. Single years are highlighted to show the right shift of the cumulative density functions with time. For instance, the blue line represents the cumulative density function of the 1000 mean absolute errors in the projection year 2000.

In general, these cumulative density functions of mean absolute errors over all ages can be used to evaluate the extent of ordinal similarity. If actual and projected age structures are equal, the cumulative density function is a vertical line at level *zero* for the mean absolute errors over all ages. If the difference between projected and actual age structure increases slightly with time, the cumulative density functions have a similar shape, and they shift successively to the right for subsequent projection years, without any crossings. Hence, a continuous right-shift of the cumulative density functions over time indicates that the fundamental tendency of actual fertility, mortality, and migration has been projected, so that we can observe *ordinal similarity* between actual and projected age structures. In contrast, *ordinal dissimilarity* can be observed if the difference between projected and actual age structure grows non-monotonically over time. Indicators for discontinuous differences

are, for example, cross-overs among the cumulative density functions, due to different shapes and locations on the error scale (see x-axis in Figure 5).[5]

In our projection for western Germany, the cumulative density functions have a similar shape, and they shift successively to the right without any cross-overs. All of these characteristics indicate ordinal similarity between actual and projected age structures. Exceptions are small differences in shape—from almost vertical lines in early projection years to cumulative normal distributions in later projection years—which indicate a slowly growing dissimilarity between projected and observed age structure with time. Moreover, the level of the mean absolute error over all ages that is reached by at least 50 per cent of all trials slightly increases with time, and therefore indicates ordinal similarity between actual and projected age structures as well.

6 Outlook

We also plan to investigate alternative approaches to evaluate ordinal similarity, such as:

1. a statistical hypothesis test for homogeneity between actual and projected age distributions
2. modeling differences between actual and projected age structures over time with a function

To our knowledge, ordinal similarity between actual and projected development has not been used as an evaluation criterion for probabilistic population projections so far. As our analysis provides some interesting additional insights compared to common evaluations of projection accuracy, we will continue our studies on ordinal similarity as an evaluation criterion for probabilistic population projections in future.

[5] Besides, a straight cumulative density function represents mean absolute errors that are uniformly distributed over all trials in a projection year. The less the slope of a straight cumulative density function, the larger is the error range.

References

[1] J. Scott Armstrong. *Principles of Forecasting*. Kluwer Academic Publishers, 2001.

[2] Stanley K. Smith, Jeff Tayman, and David A. Swanson. *State and Local Population Projections. Methodology and Analysis*. Kluwer Academic/Plenum Publishers, 2001.

[3] John F. Long. Complexity, Accuracy, and Utility of Official Population Projections. *Mathematical Population Studies*, 5(3):203–216, 1995.

[4] Dennis A. Ahlburg. Simple versus Complex Models: Evaluation, Accuracy, and Combining. *Mathematical Population Studies*, 5(3):281–290, 1995.

[5] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[6] J. Scott Armstrong. Evaluating forecasting methods. In J. Scott Armstrong, editor, *Principles of Forecasting*, pages 441–472. Kluwer Academic Publishers, 2001.

[7] J. Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8:69–80, 1992.

[8] Han Lin Shang, Heather Booth, and Rob J. Hyndman. Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research*, 25(5):173–214, 2011.

[9] Tom Wilson and Phil Rees. Recent Developments in Population Projection Methodology: A Review. *Population, Space and Place*, 11:337–360, 2005.

[10] Frans J. Willekens. Demographic forecasting: state of the art and research needs. In C. A. Hazeu and G. A. B. Frinking, editors, *Emerging Issues in Demographic Research*, pages 9–66. Elsevier Science Publishers B. V., 1990.

[11] Nico Keilman. UK national population projections in perspective: How successful compared to those in other European countries? *Population Trends*, 129:20–30, 2007.

[12] Christina Bohk, Roland Ewald, and Adelinde M. Uhrmacher. Probabilistic Population Projection with JAMES II. In M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, editors, *Proceedings of the Winter Simulation Conference*, 2009.