Evaluating probabilistic population forecasts*

Nico Keilman

nico.keilman@econ.uio.no

4 March 2020

Abstract

We demonstrate how a probabilistic population forecast can be evaluated, when observations for the predicted variables become available. Statisticians have developed various scoring rules for that purpose, but there are hardly any applications in population forecasting literature. A scoring rule measures the distance between the probability distribution of the predicted variable, and the actual outcome. We use scoring rules that reward accuracy (the outcome is close to the expected value of the prediction) and sharpness (the predictive distribution has low variance, which makes it difficult to hit the target).

We evaluate probabilistic population forecasts for France, the Netherlands, and Norway. For all three countries, we use results from the UPE-project ("Uncertain Population of Europe"). We inspect prediction intervals for population size in the period 2004-2019 and 3000 sample paths for population pyramids for the year 2010. For the Netherlands and for Norway, we compare the UPE-results with findings from the official probabilistic population forecast by Statistics Netherlands (2001-2019) and from a probabilistic forecast for Norway (1997-2019). All forecasts were computed using the cohort-component method and stochastically varying parameters for fertility, mortality and migration.

We show that the UPE-forecasts for the Netherlands and for Norway performed better than the other forecasts for these two countries. The error in the jump-off population caused a bad score for the French forecast.

We evaluate the 3000 UPE-simulations of the age and sex composition predicted for the year 2010. When normalized for population numbers in each age-sex category, the predictions for the Netherlands received the best scores, except for the oldest old. The age pattern for the Norwegian score reflects the under-prediction of immigration after the enlargement of the European Union in 2005.

Key words: probabilistic population forecast, scoring rule, cohort component model

* Note: this is the author's version of a paper accepted for publication in "Economie et Statistique / Economics and Statistics".

1. Introduction

Most statistical agencies in the world, who compute population forecasts, do so using a deterministic approach (NRC 2000). They analyse historical trends in fertility, mortality, and migration, and extrapolate those trends into the future, using expert opinion and statistical techniques. These extrapolations reflect their best guesses. In addition to computing a likely development of population size and structure, many agencies also compute a high and a low variant of future population growth, in order to tell forecast users that future demographic developments are uncertain. For example, the previous official population forecast for France indicates 76.5 million inhabitants in 2070, if current trends continue (Blanpain and Buisson 2016). However, population growth to 2070 might be weaker or stronger than what current trends suggest, leading to population sizes between 66.1 and 87.6 million persons. The forecasters assumed high and low trajectories for future fertility (1.8 or 2.1 children per woman on average after 2020), life expectancy of men (between 87.1 and 93.1 years in 2070) and women (between 90 and 96 years), and international migration (a migration surplus between 20 000 and 120 000 persons annually).

One important drawback of such a deterministic approach is that it fails to quantify uncertainty. We do not know if chances are 30, 60, or 90 per cent that France in 2070 will have between 66.1 and 87.6 inhabitants. Yet in many planning situations, it is important for the users to know how much confidence they should have in the predicted numbers. How robust should the pension system be with respect to fast or slow increases in life expectancies? Should we plan for extra capacity in primary schools, in case future births turn out to be much higher than expected? Indeed, as Keyfitz (1981) wrote almost 40 years ago: "Demographers can no more be held responsible for inaccuracy in forecasting population 20 years ahead, than geologists, meteorologists, or economists when they fail to announce earthquakes, cold winters, or depressions 20 years ahead. What we can be held responsible for is warning one another and our public what the error of our estimates is likely to be".

This is why the statistical agencies of some countries have started to publish their forecasts in the form of probability distributions, following common practice in, for example, meteorology and economics. Statistics Netherlands pioneered the field; see Alders and De Beer (1998). Statistics New Zealand (2011) and Statistics Italy (ISTAT, 2018) are the other two known examples. In this connection, one should also mention the Population Division of the United Nations, which is responsible for regular updates of population forecasts for all countries of the world. In 2014, the Population Division issued the first official probabilistic population forecasts for all countries, using the methodology developed by Raftery et al. (2012); see also http://esa.un.org/unpd/wpp/Graphs/Probabilistic/POP/TOT/. The aim of a probabilistic forecast is *not* to present estimates of future trends that are more accurate than a deterministic forecast, but rather to give the user a more complete picture of prediction uncertainty.

Demographers in these statistical agencies could build on work and methods developed by demographers and statisticians since the 1980s. Two developments are noteworthy. First, early contributions applied an analytical approach, assuming a stochastic cohort component model, in which the statistical distributions for fertility, mortality, and migration parameters were transformed into statistical distributions for the size of the population and its age-sex structure. One needed strong assumptions, or derived only approximate expressions for the second moments of the age-sex distributions. Nowadays, a simulation approach is common. It avoids the simplifying assumptions and the approximations of the analytical approach. The idea is to compute several hundreds or thousands of forecast variants ("sample paths") based on random draws for the input parameter values of fertility, mortality, and migration. The simulation results are stored in a database. Keilman (2009) gives an example for France. A second methodological change is that from a predominantly frequentist

approach to a Bayesian view of probability. In the frequentist tradition, the probability of an event is linked to its relative frequency of occurrence. In contrast, in the Bayesian approach a probability is interpreted as a person's subjective belief. It is particularly useful when models rely on expert opinions, and when one combines this kind of information with data. The change from a frequentist to a Bayesian approach in population forecasting was part of a more general trend towards "Bayesian demography", which started to gain popularity about 10 years ago (Bijak and Bryant 2016). The probabilistic UN forecasts mentioned earlier provide important examples of the Bayesian approach. Costemalle (this issue) applies this approach to the case of France.

Once a probabilistic forecast has been published, some 10-20 years later its accuracy can be evaluated, when *ex-post facto* observed data for population size and age structure have become available. However, accuracy assessment is difficult to carry out directly because it requires comparing a forecaster's predicted probabilities with the actual but unknown probabilities of the events under study. For that reason, statisticians have developed "scoring rules", also called "scoring functions". These are empirical distance measures between the predicted distribution of the demographic variable in question, and the empirical value it actually turns out to have. Gneiting and Raftery (2007) and Gneiting and Katzfuss (2014) review the field. The score that one finds for a certain variable has no intrinsic meaning. Only in a comparative perspective, one can interpret the scores in a useful manner. This explains why scoring functions are frequently used in comparing two competing probabilistic forecasts.

Although the methodology around evaluation of probabilistic forecasts and scoring rules has been known for some time, there are very few applications of scoring rules to population forecasting. Shang et al. (2016) evaluated the accuracy of probabilistic cohort-component forecasts for the UK, and compared two forecasting methods. They used a scoring rule for prediction intervals. Shang (2015) and Shang and Hyndman (2017) evaluated interval forecasts for age-specific mortality rates of various countries, and used interval scores to select the best among several methods of mortality forecasting. Alexopoulos et al. (2018) employed interval scores to prediction intervals of age-specific mortality of England & Wales and New Zealand, and evaluated the predictive performance of five different mortality prediction models. All four papers use holdout samples to evaluate the probabilistic demographic forecasts. Genuine out-of-sample evaluation of probabilistic demographic forecasts has not been attempted before, to the best of our knowledge.

The aim of this paper is to show how methods for evaluating probabilistic forecasts developed elsewhere can be applied to probabilistic population forecasts. We present and apply scoring rules for prediction intervals, and for simulated samples of future population size and age structures. We illustrate the scoring rules using data for France, the Netherlands, and Norway, and compare probabilistic forecasts computed by different researchers. The comparisons serve three purposes. First, we investigate how fast the accuracy of a given probabilistic forecast changes with lead-time, i.e. when it looks further into the future. Second, we compare the accuracy of two ("competing") probabilistic forecasts for the same country. Finally, the relative performance of forecasts across countries is analysed.

Section 2 discusses the way the results of a probabilistic forecast are made available: as prediction intervals, or by means of a database. Section 3 presents a number of scoring rules and their characteristics. Empirical illustrations are given in Section 4. We evaluate various probabilistic predictions for total population size and the population pyramid of the three countries. Section 5 concludes the paper.

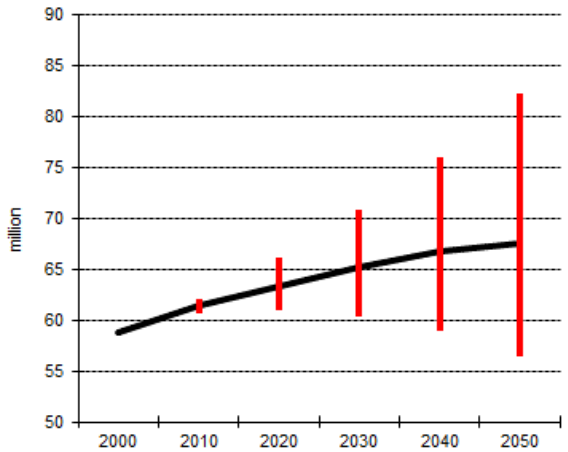
2. Publishing a probabilistic population forecast

The methods one uses to evaluate a probabilistic forecast depend strongly on the way the forecast results are made available. There are two main methods. One is to publish prediction intervals for population variables. Alternatively, one may give the users access to a database with sample paths.

Costemalle (this issue) presents prediction intervals for the population of France, computed by a Bayesian approach. For instance, his Figure 16 shows that there is an 80 per cent probability that total population size in 2070 will be between 68.1 million and 75.0 million persons. The graph also shows 95 per cent prediction intervals. These are much wider, because they cover more extreme situations. Other scholars (see Section 4 for examples) present their probabilistic forecasts in terms of 67 per cent prediction intervals.

Figure 1 plots 80 per cent prediction intervals for the population of France taken from the so-called UPE-project, to be discussed below. The jump-off year of this probabilistic forecast was 2003. Thus in 2050, 47 years into the future, the 80 per cent prediction interval is 82.2 – 56.5 = 25.7 million persons wide. This is much wider than Costemalles 80 per cent interval of 75.0 – 68.1 = 6.9 million persons (after 46 years). Different perceptions of prediction uncertainty for future fertility, mortality, and international migration lead to sharper (optimistic) or wider (pessimistic) prediction intervals.

Figure 1. Median values (black line) and 80 per cent prediction intervals (red lines) for total population of Metropolitan France. Chances are 50 per cent that population size in 2050 will be less than 67.7 million; a population larger than 67.7 million is equally likely. There is an 80 per cent probability that total population in 2050 will be between 56.5 and 82.2 million. Source: Keilman (2009).



These examples illustrate a more common finding, namely that different authors use different coverage probabilities for their prediction intervals. Selecting a coverage probability of 67 or 80 per cent covers the majority of forecasts but excludes the more volatile tail of the error distribution. Those who use a coverage probability of 95 per cent do so, probably, because they have in mind a tradition in social science that implies constructing 95 per cent *confidence* intervals or performing hypothesis tests with a low probability (5 per cent) for type I errors (i.e. that one rejects a null hypothesis whereas in

fact it is true). On the other hand, a prediction interval with coverage probability of 67 or 80 per cent gives the user of the forecast an idea of how things might deviate from the point forecast. This is very different from constructing confidence intervals and from hypothesis testing. We will use both 67 and 80 per cent prediction intervals in the empirical examples of Section 4.

Prediction intervals present only a summary of the complete probability distribution for the variable in question. Sometimes one can assume that the underlying distribution is approximately normal. In such cases, one can infer the parameters of the distribution from the upper and lower bounds of the interval. However, some population variables are restricted to a certain part of the real line, such as the share of the elderly in the population (between zero and one), and a normal distribution is not appropriate. In such cases one loses much information by publishing prediction intervals only, and not the underlying distributions.

The most of information becomes available when all simulated trajectories are stored in a database, to which the user has access (Alho and Spencer 2005). A prominent example is the set of probabilistic population forecasts for 18 European countries, commonly known as the UPE-forecasts ("Uncertain Population of Europe"). The cohort-component model was applied 3000 times for each country, with a deterministic jump-off population (as of 1 January 2003) and probabilistically varying values for age-specific fertility, mortality, and net migration. The forecast horizon was 2050. The UPE-forecasts have two attractive features. First, an explicit aim was to quantify uncertainty in such a way that it would reflect historical volatility in fertility, mortality, and international migration. Second, the project provided the first comprehensive look at empirical correlatedness of forecast errors in fertility, mortality, and migration across countries. More information, including a number of published and unpublished papers, is available at the UPE website at http://www.stat.fi/tup/euupe/index_en.html. The website contains a databank with simulation results (N=3000) for men and women in five-year age groups in each country at ten-year (2010(10)2050) time intervals. This means that the user can build his or her histogram(s) for the variable(s) of interest. In Section 4, we will use the forecasts of the population pyramids for 2010 for France, the Netherlands, and Norway to illustrate the scoring rules discussed in Section 3.

3. Evaluation

Write the variable for which one computes a forecast as X, with cumulative distribution function (CDF) defined as $F(x) = \mathrm{P}(X \leq x)$. The probability density function (PDF) of X is $f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$. We assume throughout the existence of the integrals and various moments of the probability distributions. More fundamental treatments based on probability-theoretic considerations can be found in, for instance, Gneiting and Katzfuss (2014), and Gneiting and Raftery (2007). Write y for the observed value of X. A scoring function $S(F(x),y)$ assigns a numerical value (a "score") to the forecast $F(x)$, given the observation y. $S(F(x),y)$ takes values in the real line $\mathbb{R}$ (including, possibly, plus and minus infinity).

A natural starting point for defining a scoring function is the following: *a forecast that predicts the actual outcome with high probability should receive a good score*. This works well for categorical forecasts, when X is a discrete random variable. However, we are dealing with forecasts for the number of persons (by age, sex, and forecast year), and X is closer to a continuous than a discrete random variable (unless the forecast is for a very small population). Henceforth we shall assume that

the forecast and the scoring function are based on a continuous random variable. Many of the scoring functions start from the following two principles. First, an observation close to the median or the expected value of the predictive distribution gives a good score – the closer the better. In that case, the scoring rule is sensitive to distance (Staël von Holstein 1970, Murphy 1970). Second, given an observation, a narrow ("sharp") predictive distribution gives a good score – the narrower the better. For example, an 80 per cent prediction interval that covers a certain observation represents a better forecast than an equally wide 67 per cent interval that covers the same observation, because it is relatively difficult to hit the target when the PDF has low variance. However, the two principles are not equally important. One may argue that when the observation is "too far" from the median or expected value, one should no longer reward a narrow PDF. In other words, if the forecaster is "takes a chance" (i.e. predicts a narrow PDF), the forecast should have a good score when the forecast is close to the median or expected value, but not when it is too far away. What one means by "too far away" is unclear, and it differs between scoring rules. The example above puts it as "the observation falls outside the prediction interval". This choice may be criticized: it rests on an extremely sharp dichotomy. In a very small interval around the upper or the lower bound of the prediction interval, the forecast changes very abruptly from having a good score to being punished for having an observation just outside the interval. To put it differently, given the predictive distribution and the observation, a prediction interval with the lower bound slightly lower than the observation gives a good score, whereas a bad score arises when the lower bound is slightly higher than the observed value. Coverage probabilities are arbitrary (80 % is often used, but 81% or 79% work equally well). Therefore, one should be careful when defining the notion of "too far away".

Some of the scoring rules that we will discuss below indeed follow the idea that closeness is more important than sharpness. However, as we shall see, what we mean by "too far" is different for different scoring rules. Other scoring rules treat the two principles as independent.

We say that a scoring function is *negatively* oriented when a lower score implies a better forecast, and the other way round for a *positively* oriented scoring function. Hence, a negatively oriented scoring function may be interpreted as a cost function, whereas a positively oriented scoring function reflects rewards.

Many different scoring rules have been proposed, depending on the nature of the forecast. Gneiting and Raftery (2007) and Jordan et al. (2019) give extensive overviews of the field. We will restrict ourselves to scoring rules for continuous random variables. One class of scoring rules applies to density forecasts based on closed-form expressions of the CDF or the PDF. An example is the logarithmic score $\mathrm{LogS}(F(x),y) = -\log(f(y))$. A different class of scoring rules, more appropriate for the subject of this paper, evaluates simulated samples – in that case, the predictive distribution is not available analytically. A second distinction is that between univariate forecasts and multivariate forecasts. In the latter case, both the predicted variable X and the observation y consist of a vector. Jordan et al. (2019) developed the **scoringRules** package for R, which covers a wide range of situations in applied work.

Below we will introduce three types of scoring rules: those based on the first two moments of the predictive distribution only (Section 3.1), those stemming from the simulated complete predictive distribution, available as a sample (Section 3.2), and finally those for which one only has prediction intervals (Section 3.3).

3.1 A variance-based scoring function

Assume a unimodal PDF of the forecast. When the actual outcome is close to the centre of the predicted density (as characterized by the mean, the mode, or the median), this forecast is better than one for which the outcome is far away from the centre. Stated differently, when there is little variation in *X* around *y*, the forecast scores better than when there is much variation. This leads to a variance-based scoring function, written as *VS* henceforth, and defined as follows.

Let *VS* be the variance of *X* around the observed value *y*, or

$$VS = \int (x - y)^2 f(x) dx \, . \tag{1}$$

For *y* equal to the expectation of *X* (written as μ), *VS* reduces to the variance of *X*, written as σ^2. Expression (1) leads to

$$VS = \sigma^2 + (\mu - y)^2. \tag{2}$$

This defines a simple variance-based scoring function, which one could use to assess the quality of a unimodal predictive PDF. Gneiting and Raftery (2007) list it as a scoring function that corresponds to the so-called predictive model choice criterion or PMCC. One may apply it for analytical density functions as well as simulated samples. In the latter case, one uses estimates of σ^2 and μ from the sample. This scoring function is negatively oriented: a lower score indicates a better forecast. It rewards both accuracy - when *y* coincides with μ, the forecast is of optimal quality - and sharpness - a small variance gives a good score, irrespective of how far off the forecast was.

For a deterministic (point) forecast, σ^2 is zero and the forecast is μ. In that case, *VS* reduces to the squared error of the forecast. Errors of this kind form the basis of the Mean Squared Error frequently used in evaluations of deterministic population forecasts (Alho and Spencer 2005; Smith et al. 2001; Keilman 1990).

An alternative scoring function, also based on the first two moments of the predictive distribution, is the Dawid-Sebastiani score (e.g. Gneiting and Katzfuss 2014)

$$DSS = \ln(\sigma^2) + (\mu - y)^2/\sigma^2. \tag{3}$$

This scoring function is similar to the variance-based score *VS* in expression (2), but it gives different weight to the forecast variance σ^2. A low variance leads to a good (low) score as long as $\frac{dDS}{d\sigma^2} = \frac{1}{\sigma^2} - \frac{(\mu - y)^2}{\sigma^4} > 0$, or $\sigma > |\mu - y|$. Whereas *VS* always rewards predictive distributions with low variance, *DSS* does so if the observation *y* is less than one standard deviation away from the expectation of the predictive distribution.

Imagine a forecaster, who knows that her probabilistic forecast in due time will be evaluated by the scoring rules (2) or (3). Assume that at a certain stage of the production process of the forecast, the issue is to calibrate the forecast model. Use of scoring rule (2) or (3) implies that this calibration should focus on selecting an appropriate value for the mean μ of the predictive distribution – not the median or any other parameter of location. Indeed, there is a close relationship between model calibration and forecast evaluation. The situation is clear when there is only one user. However, things become more complicated when there are many users with different scoring rules (or with unknown scoring rules).

3.2 The continuous ranked probability score *CRPS*

The continuous ranked probability score might serve as a standard score in evaluating probabilistic forecasts of real-valued variables (Gneiting and Raftery 2007). It is defined in terms of the predictive CDF $F(x)$ as

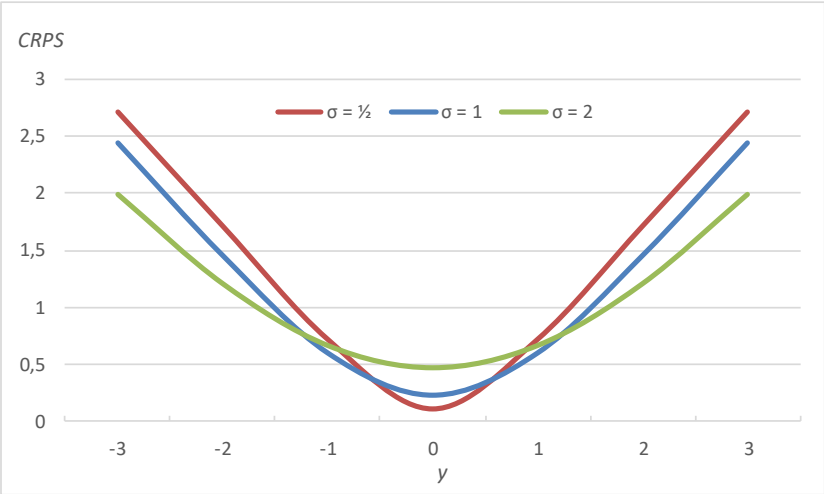
$$CRPS(F, y) = \int (F(z) - \mathbb{I}\{y \leq z\})^2 dz, \qquad (4)$$

where $\mathbb{I}\{y \leq z\}$ denotes the indicator function which is one if $y \leq z$ and zero otherwise. The particular form of the *CRPS* originates from the Brier score (Brier 1950). The Brier score or probability score (PS) is a mean squared error of a categorical probability forecast. Murphy (1970) adapted it to the case of ordered categories for X, which led to the Ranked Probability Score or RPS. Matheson and Winkler (1976) proposed a RPS for the case of a continuous random variable, the CRPS.

Readily computable solutions to the integral above are few. Jordan et al. (2019) list the known cases. For instance, when $F(z)$ is the standard normal distribution $\Phi(.)$ with density $\varphi(.)$, $CRPS(\Phi,y)$ equals $y(2\Phi(y) - 1) + 2\varphi(y) - 1/\sqrt{\pi}$. The normal distribution with general expectation μ and standard deviation σ gives $\sigma CRPS(\Phi,(y - \mu)/\sigma)$.

It is worth to analyse a few concrete cases of the *CRPS*. We take the example of a normal distribution and assume, without loss of generality, that μ equals zero. Figure 2 plots this *CRPS* as a function of y, in other words, its sensitivity to distance. We show three cases, namely standard deviations of ½, 1, and 2. By construction ($\mu = 0$), the curves are symmetric around zero. As we might expect, the best score arises when y equals zero. The score becomes worse when y increases in absolute value, in other words, when y is far from μ. Sharpness of the predictive PDF (a low standard deviation) is only rewarded within a certain y-interval around zero. For instance, a perfect forecast (y equal to zero) scores better for $\sigma = ½$ (*CRPS* = 0.1168) than for $\sigma = 2$ (*CRPS* = 0.4674). However, the PDF with $\sigma = 2$ scores better than the one with $\sigma = ½$ for observations y larger than approximately 0.9 in absolute value. For low σ-values, the interval where sharpness is rewarded is shorter than for high values.

Figure 2. Continuous ranked probability scores (CRPS) for a normal distribution with expected value μ equal to zero and observations y ranging from -3 to +3. Standard deviations σ of ½, 1, and 2.



Probabilistic population forecasts are commonly computed as simulated distributions, and one cannot compute the integral in (4). In that case, a useful starting point is the fact that (4) can be written as

$$CRPS(F, y) = \mathrm{E}_F|X_1 - y| - \tfrac{1}{2}\mathrm{E}_{F,F}|X_1 - X_2|, \quad (5)$$

where X_1 and X_2 are independent random variables with distribution F (Gneiting and Raftery 2007). The *CRPS* measures how close the observation y one can expect to be to the predicted variable X, corrected for the expected distance between all possible pairs of values of X. The latter expected distance is small when the standard deviation of F is small. Other things being the same, an increase in the standard deviation leads to a better score. However, when the standard deviation changes also the first expectation $\mathrm{E}_F|X_1 - y|$ changes. Whether this score rule always rewards sharpness, or only on a certain interval, remains an empirical issue.

The *CRPS* reduces to the absolute error if F is a deterministic forecast.

Assume that we have a forecast available in terms of a simulated distribution. Then the CDF is $\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{X_i \leq x\}$, where m is the size of the sample, and (5) becomes

$$CRPS(\hat{F}_m, y) = \frac{1}{m}\sum_{i=1}^{m} |X_i - y| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} |X_i - X_j|.$$

Implementation of this expression is inefficient, because it is of computational order o(m^2). A more efficient and algebraically equivalent representation is (Jordan et al. 2019, p. 6)

$$CRPS(\hat{F}_m, y) = \frac{2}{m^2}\sum_{i=1}^{m}(X_{(i)} - y)\left(m\mathbb{I}\{y < X_{(i)}\} - i + \tfrac{1}{2}\right), \quad (6)$$

where $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, …, $X_{(m)}$, is the sorted simulated sample. The *CRPS* as defined in (6) is always positive, because each term in the sum is positive.

3.3 Interval scores

Many probabilistic population forecasts are presented as interval forecasts, not as (simulated) probability distributions; see Section 2. Consider a central (1 - α) prediction interval, with lower and upper endpoints that are the predictive quantiles at levels $\alpha/2$ and (1 - $\alpha/2$), respectively.[1] Write l and u for the lower and upper quantiles. Gneiting and Raftery (2007) define the following score function

$$(u - l) + \frac{2}{\alpha}\left[(l - y)\mathbb{I}\{y < l\} + (y - u)\mathbb{I}\{y > u\}\right]. \quad (7)$$

Given α, the Gneiting-Raftery interval score (written as *GRIS* henceforth) rewards forecasts for narrow prediction intervals that capture the observation y: when two competing forecasts have different prediction intervals for a given α, the forecast with shortest prediction interval gets the best (lowest) score. However, a value of y outside the prediction interval gives a bad (higher) score. The penalty for missing the prediction interval is larger for small than for large α. *GRIS* can be readily applied to the

[1] Note that we assume that the two quantiles are known. In case we want to evaluate interval forecasts when the nominal coverage level is specified, but the quantiles on which intervals are based are not specified, one cannot employ the approach outlined here (Askanazi et al. 2018).

prediction interval of a variable for different lead times: 1 year ahead, 2 years ahead, 3 years ahead … etc.

There are situations in which *GRIS* does not reward sharpness, even when the interval captures the realization. Assume two competing forecasts with the same prediction interval $[l,u]$ that have different coverage probabilities. For instance, one forecast attaches 67 per cent probability to the prediction interval $[l,u]$, whereas the other one has a coverage probability of 80 per cent for the same interval. The second forecast is sharper and should receive a better score when the observation y falls inside $[l,u]$. However, this is not the case, as *GRIS* is independent of α in this situation. To solve the issue, one may use a slightly modified version of *GRIS*, namely

$$GRISmod = \alpha(u-l) + \beta[(l-y)\mathbb{I}\{y<l\} + (y-u)\mathbb{I}\{y>u\}], \qquad (8)$$

where $\beta > 0$ is a parameter that determines how fast the score deteriorates when the observation is further away from either the upper or the lower bound of the prediction interval. A high β-value incurs a larger penalty than a low value. *GRISmod* rewards sharpness both for fixed α and different prediction intervals, and for the situation where one has a fixed prediction interval but different values of α. When β equals two, *GRISmod* equals α.*GRIS*. In case one uses a β-value equal to the probability α, *GRISmod* reduces to $\alpha(u - y)$ for $y < l$ and to $\alpha(y - l)$ when $y > u$.

As an alternative to using scoring functions for prediction intervals, one could check how often actual data fall within the intervals. For instance, Raftery et al. (2012) validated their Bayesian method of forecasting populations for 159 countries by estimating the model based on data for the 40-year period 1950–1990 to generate a predictive distribution of the full age- and sex-structured population for the 20-year period 1990–2010. They compared the resulting 80 per cent and 95 per cent prediction interval distributions with the actual observations, and checked the proportion of the validation sample that fell within their intervals. These proportions were close to the nominal values of 80 and 95 per cent; therefore, the authors concluded that their approach was satisfactory. One important drawback of this method is that one compares data and prediction intervals for many variables, for instance the population size for all 56 countries in Africa at a certain date. However, regional correlations for fertility, mortality, and/or migration imply that the 56 population sizes are not independent. One has less data than originally thought and observed proportions cannot be compared directly with nominal values (Alho and Spencer 2005, 248).

3.4 Scoring functions used in the empirical applications

In Section 4, we use the *CRPS* in expression (6) to evaluate forecasts for which detailed simulation results are available. In case we only have prediction intervals, we use the variance-based score *VS* of expression (2), the Dawid-Sebastiani score (*DSS*) of expression (3), and the interval scores (*GRIS* and *GRISmod*) of expressions (7) and (8). For *GRISmod* we assume a value for the parameter β equal to the probability α that was used to define the interval. *VS* and *DSS* use the expectation and the standard deviation of the predictive distribution. Since only upper and lower interval bounds are available, we assume normality and take the expectation as the mean of the two bounds, while we estimate the standard deviation as half the width of the interval for 67 per cent intervals, and as the interval width divided by 2.564 for 80 per cent intervals.

Note that the scores depend on the scale of the variable *X* for which we have a predictive distribution (which is the same scale as that of the observation *y*). Hence, when we compare the scores of two population forecasts for countries with very different population sizes, the smaller population will receive the best score, irrespective of its accuracy. For a fair comparison, we need to account for population size. We have normalized *VS*, *DSS*, *CRPS*, *GRIS*, and *GRISmod* as follows:

- we divided *VS* by μ^2, i.e. the square of the expected value of the predictive distribution;
- we normalized *DSS* by subtracting $2\ln(\mu)$ from the original *DSS*-value;[2]
- we divided *CRPS*, *GRIS*, and *GRISmod* by μ.

4. Findings

Below we illustrate the scoring rules mentioned in Section 3.4 by evaluating probabilistic population forecasts for three countries: France, the Netherlands, and Norway. We focus on total population size (Section 4.1) and on the population pyramid (Section 4.2) of the three countries. The data stem from various sources.

1. At the UPE-website (see Section 2), samples (N = 3000) for the forecasts of the population pyramid for the three countries are available for the years 2010, 2020, …, 2050. We use results for 2010.

2. Alho and Nikander (2004) report 80 per cent prediction intervals and medians for total population size, amongst others, for each year in the period 2004-2050 for all UPE-countries. We use results for 2004 – 2019.

3. For the Netherlands, we have information about the official probabilistic population forecast with jump-off year 2000; see Statistics Netherlands (2001). The tables give 67 per cent prediction intervals and expected values for total population for each year during the period 2000 – 2050, and for five-year age groups of men and women at five-year intervals.

3. For Norway, we use results of the so-called StocProj ("Stochastic Projections") project (Keilman et al. 2002). The purpose was to compute a probabilistic population forecast with jump-off year 1996. Detailed simulation results are no longer available, but we use instead 80 per cent prediction intervals for total population size for the years 1997-2019.

4.1 Population size

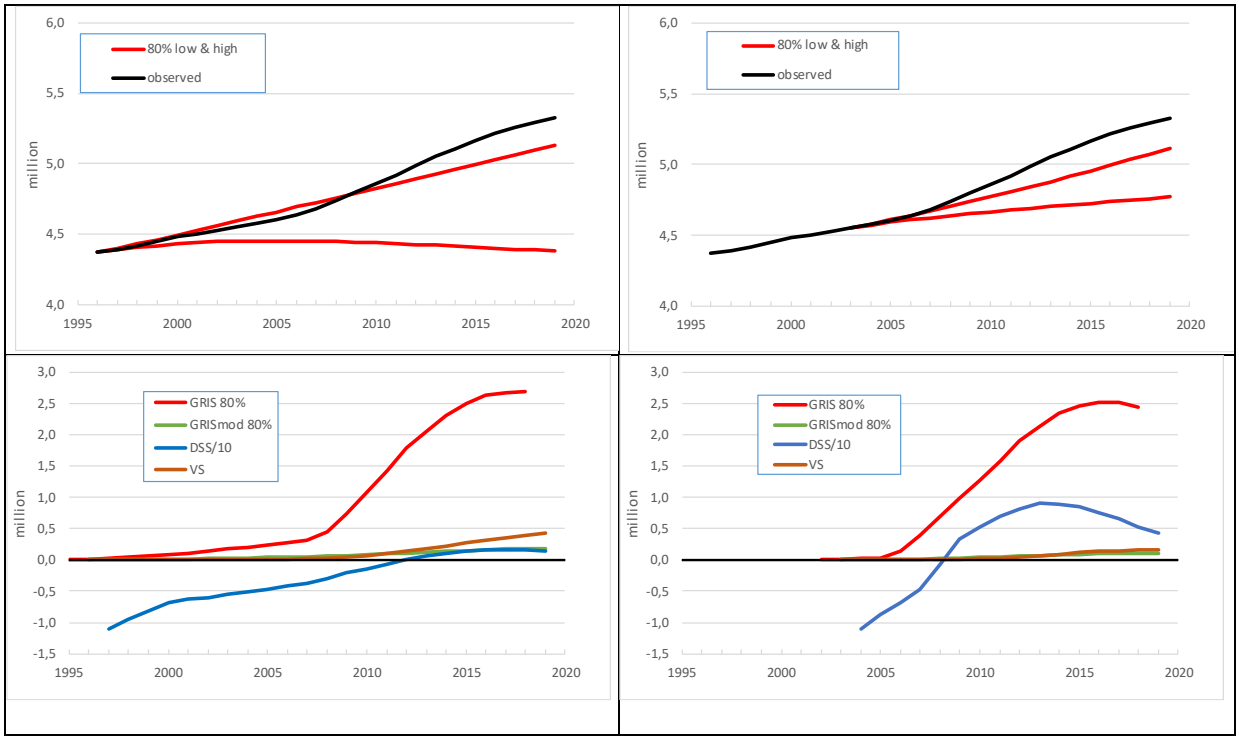
Figure 3 gives our findings for Norway. There are four graphs, two for the StocProj forecast (left), and two for the UPE-forecast (right). The upper two graphs plot 80 per cent prediction intervals and observed values for total population sizes, while the lower two graphs show the scores of the two forecasts.

Both forecasts underpredicted total population from around 2005 onwards. The most important explanation is that after the enlargement of the European Union, labour immigration to Norway from

[2] The interest is in the *DSS*-value for a scaled random variable X/N and scaled value y/N of y (N non-random and positive), written as $DSS(y/N)$. Then $DSS(y/N) = 2\ln(\sigma/N) + [(\mu/N - y/N)/(\sigma/N)]^2 = DSS(y) - 2\ln(N)$. For N we select expected population size μ.

Baltic and East-European countries was much higher than expected. Note that at each forecast lead-time, the prediction intervals for StocProj are *wider* than the UPE-intervals. The modified interval score *GRISmod* rewards sharpness, and hence it is lower and thus better for UPE than for StocProj, although the difference is small; cf. the green lines. The modified interval score *GRISmod* and the variance-based score *VS* show the same trend: both forecasts become gradually worse for longer lead times. The blue curves show the Dawid-Sebastiani score *DSS* divided by ten, so that we could plot it in the same graph as the other three scores. *DSS* starts at negative values in both cases, because the standard deviations σ of both forecasts are small (measured in millions) in the first few years. For instance, for StocProj in 1997, σ = 0,0039, which gives $\ln(\sigma^2) = -11{,}1162$. Since $((\mu\text{-}y)/\sigma)^2 = 0{,}0309$, *DSS* equals -11,0853, plotted as -1,1085 in Figure 3. *DSS* increases steeply for UPE, because it does not reward sharpness anymore as soon as the standard deviation of the predictive distribution is smaller than the absolute error $|\mu - y|$; cf. Section 3.1. This occurs in all years for which we have UPE-data, i.e. from 2004 onwards. On the other hand, for StocProj the situation with too small standard deviation to reward sharpness does not occur until 2008, 12 years into the future. On the other hand, score functions *GRISmod* and *VS* do not punish "over-optimistic" forecasts (i.e. forecasts for which the variance of the predictive distribution is too small). Note that for StocProj, *DSS* stabilizes from around 2016, 20 years into the future.

Figure 3. Total population size, Norway. Prediction intervals and observed values in the upper panels, interval scores (*GRIS* and *GRISmod*), Dawid-Sebastiani (*DSS*), and variance-based (*VS*) scores in the lower panels. StocProj forecasts 1997-2019 (left) and UPE forecasts 2004-2019 (right). Prediction intervals, observed values, *GRIS*, *GRISmod*, and *VS* are in millions. Dawid-Sebastiani score is divided by ten.



Similar to the case of Norway, the UPE prediction intervals (80 per cent) for the Netherlands reflect a sharper forecast than the intervals of Statistics Netherlands' forecast (67 per cent); see Figure 4. In

both cases, the prediction intervals capture observed population size after 2011, which means that the modified interval score for the UPE forecast is much better than that of Statistics Netherlands' forecast. Interval scores miss the fact that observed values come closer to the centre of the intervals, because these scores do not include information about the mean, the median, or the mode of the predictive distribution. Judged by the Dawid-Sebastiani scores, the two forecasts are of equal quality. In both cases, *DSS* stabilizes from 2010 onwards. The reason is that the forecast error $|\mu - y|$ diminishes slowly over time, because observed population size approaches expected population size; this compensates the increase in the standard deviations of predicted population size in the two forecasts; cf. expression (3).

GRIS shows the same, rather irregular, time pattern as *DSS*, qualitatively speaking. This is very clear in Figure 4 for the Netherlands, but it is also visible in Figure 3 for Norway. In addition, *GRISmod* and *VS* develop very smoothly for the Netherlands, as we saw already for Norway.

Figure 4. Total population size, Netherlands. Prediction intervals and observed values in the upper panels, interval scores (*GRIS* and *GRISmod*), Dawid-Sebastiani (*DSS*), and variance-based (*VS*) scores in the lower panels. Statistics Netherlands forecasts 2000-2019 (left) and UPE forecasts 2004-2019 (right). Prediction intervals, observed values, *GRIS*, *GRISmod*, and *VS* are in millions. Dawid-Sebastiani score is divided by ten.

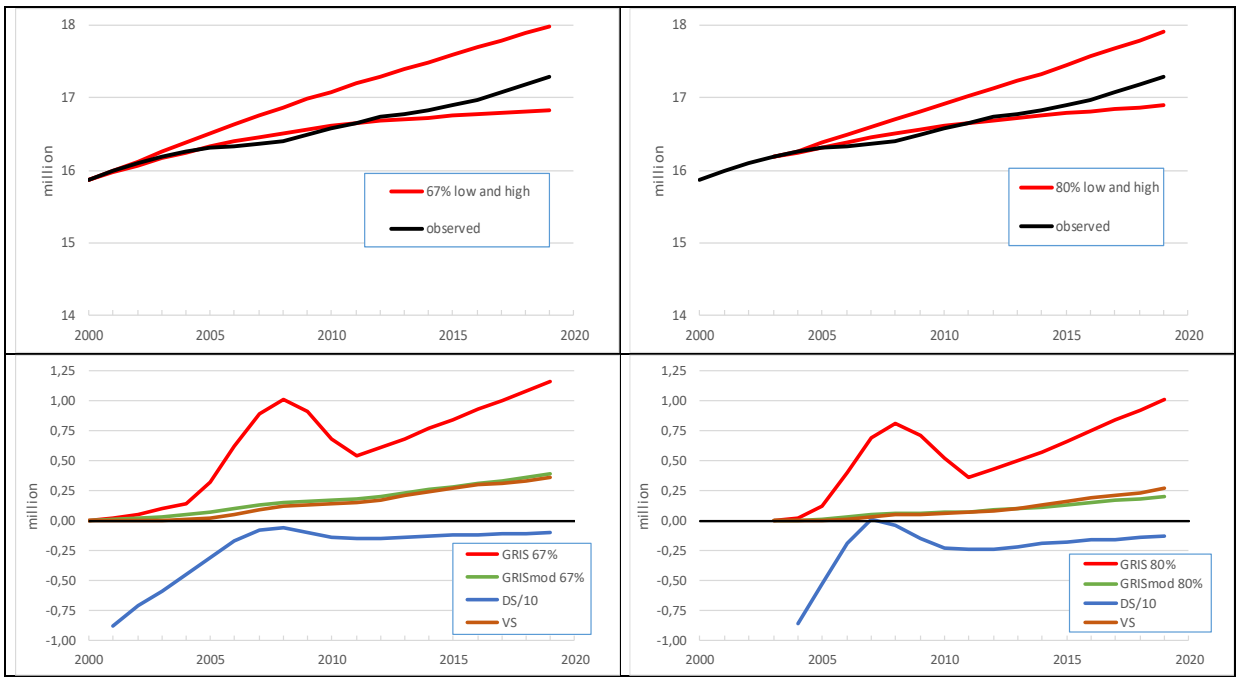
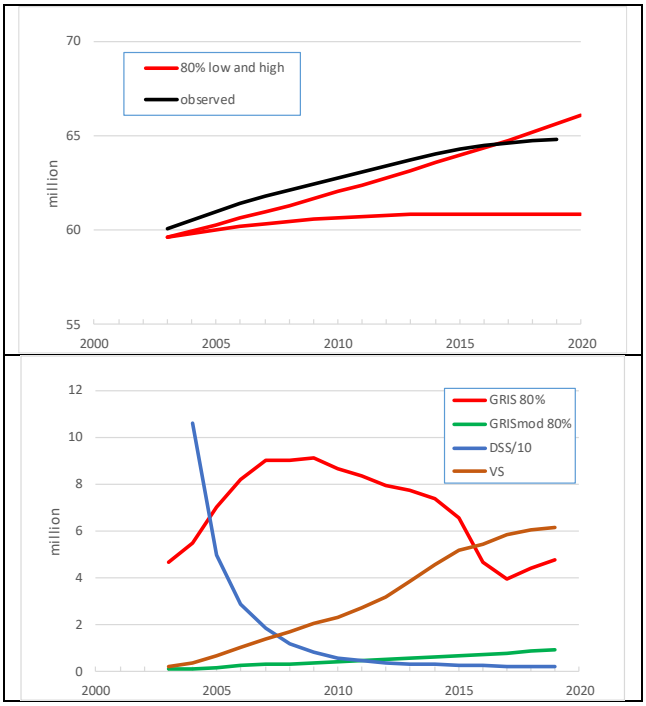


Figure 5 gives UPE scores for total population size of Metropolitan France. A striking feature is that the forecast jump-off population in 2003 is almost 500 000 persons lower than the current estimate for population size that year. Data from Eurostat, available in 2004, provided the basis for the UPE-simulations. Observed numbers in Figure 5 are from INSEE (2019). Obviously, the 2003 population number as reported by Eurostat in 2004 has been revised in later years.

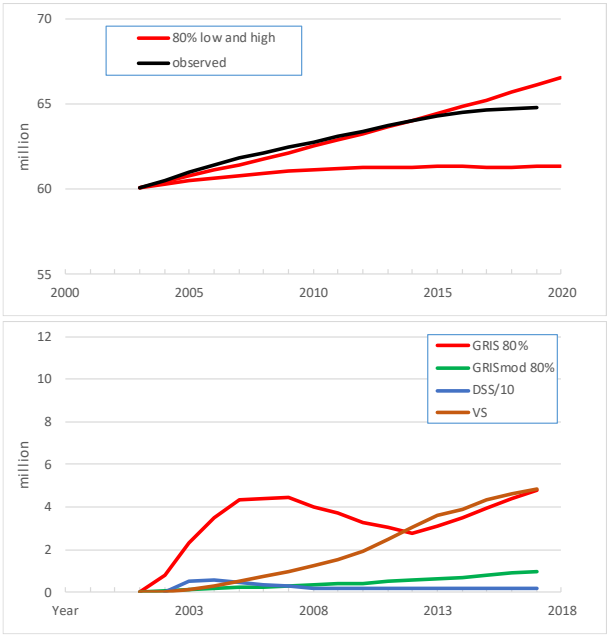
Figure 5. Total population size, Metropolitan France. Prediction intervals and observed values in the upper panel, interval scores (*GRIS* and *GRISmod*), Dawid-Sebastiani (*DSS*), and variance-based (*VS*) scores in the lower panel. UPE forecasts 2004-2019. Prediction intervals, observed values, *GRIS, GRISmod*, and *VS* are in millions. Dawid-Sebastiani score is divided by ten.



The jump-off error results in extremely bad values for the Gneiting-Raftery and the Dawid-Sebastiani score functions. What would these scores have been, in case the UPE-forecast would have started from the current estimate of total population size in 2003 (60.102 mln) rather than the number that was actually used (59.635 mln)? We can give an approximate[3] answer by lifting the 80 per cent prediction interval up by 467 000 persons. Figure 6 shows the results, with the same vertical scales as in Figure 5. *DSS* improves dramatically, to 5.2 and 5.6 in 2005 and 2006, respectively (instead of 49.6 and 28.6 for these years), while it stabilizes at a level around 1.6 – 1.7 after 2015 (rather than falling slowly to 2.0 in 2019). The interval scores and the variance-based score become slightly lower. These findings illustrate the importance of starting from the right jump-off population. At the same time, revision of population numbers occurs frequently, in particular in countries without a population register. In such cases, one should treat the jump-off population as stochastic, in addition to parameters for fertility, mortality and migration. Alho and Spencer (2005) give an example of random jump-off values for a probabilistic population forecast for Lithuania.

[3] Approximate, because we ignore the consequences for fertility and mortality of a higher jump-off population.

Figure 6. Total population size, Metropolitan France. Prediction intervals and observed values in the upper panel, interval scores (*GRIS* and *GRISmod*), Dawid-Sebastiani (*DSS*), and variance-based (*VS*) scores in the lower panel. Prediction intervals from UPE forecasts 2004-2019 are adjusted for jump-off error. Prediction intervals, observed values, *GRIS*, *GRISmod*, and *VS* are in millions. Dawid-Sebastiani score is divided by ten.



A common finding so far is that when we look further into the future, *GRISmod* and *VS* get worse over time, because prediction intervals become wider, and variances of predictive distributions increase. This, of course, reflects the fact that population forecasting is more difficult in the long-term than in the short-term. In contrast to *GRISmod* and *VS*, *DSS* stabilizes when forecast lead-times increase. The explanation lies in the definition of this particular scoring function. It is the sum of two terms; one term increases while the other one decreases when prediction variance goes up; see expression (3). Thus, one view is that *DSS* is not an appropriate measure for analysing how fast forecast quality deteriorates with increasing lead-time. However, a different view is that, exactly because *DSS* hardly changes over time, it controls for forecast lead-time. Still another possibility is to inspect the slopes in *GRISmod* and *VS*, since these two score functions increase quite smoothly with time. Further research into this issue, drawing upon data from many other forecasts (and controlling for different population sizes; see below) is clearly needed.

As mentioned earlier, one explanation for the relatively bad scores for France is the fact that the score functions depend of population size. For a comparison across countries, normalized scores are useful. We normalized the scores the way explained in Section 3.4. Table 1 gives results for the five forecasts in 2018.

After normalization, the scores for the French forecast and the two Dutch forecasts in the year 2018 become very similar; see the upper panel of Table 1. In many cases, the scores for these two countries are one order of magnitude better than those for Norway. For many years, observed population size in France and the Netherlands fell within the prediction intervals (cf. upper panels of Figures 4 and 6; the French intervals corrected for jump-off error). This contributes to the good scores for the two countries.

Table 1. Normalized interval, variance-based, and Dawid-Sebastiani scores for the year 2018 (upper panel) and for a lead-time of 15 years (lower panel).

	Norway		Netherlands		France[1]
	StocProj	UPE	StatNeth	UPE	UPE
	year 2018				
GRIS/μ	0,564	0,513	0,062	0,053	0,069
GRISmod/μ	0,038	0,022	0,021	0,011	0,014
VS/μ^2 (x 1000)	17,552	6,569	1,108	0,781	1,154
DSS – 2ln(μ)	-1,525	2,073	-6,797	-7,149	-6,639
	15 years ahead				
GRIS/μ	0,231	0,513	0,049	0,053	0,069
GRISmod/μ	0,021	0,022	0,016	0,011	0,014
VS/μ^2 (x 1000)	4,870	6,569	0,906	0,781	1,154
DSS – 2ln(μ)	-3,752	2,073	-6,903	-7,149	-6,639

Note 1. Adjusted for error in jump-off population.

The two forecasts for Norway still receive bad scores because of the under-prediction of net immigration mentioned above. For the high StocProj scores in 2018 there is an additional reason: the jump-off year of this forecast is 1996, and hence the forecast lead time in 2018 is 22 years – much longer than the UPE lead time in 2018 (15 years). The lower panel of Table 1 shows the normalized scores for StocProj after a forecast duration of 15 years (in 2011). Compared to the scores for the other two countries after 15 years, the situation has improved quite much, but StocProj-scores are still much higher than those for StatNeth and for UPE in France and the Netherlands.

The final evaluation of total population size forecasts is by means of the Continuous Ranked Probability Score (*CRPS*). We computed this score function using 3000 UPE-simulations for 2010. The *CRPS* depends of population size; see expression (6). To enhance comparison between the three countries, Table 2 gives normalized scores, defined as the *CRPS* divided by the mean of the 3000 simulations. The results confirm the good quality of the UPE-forecast for the Netherlands that we found earlier.

Table 2. Normalized *CRPS*-scores for total population size, UPE forecasts for 2010.

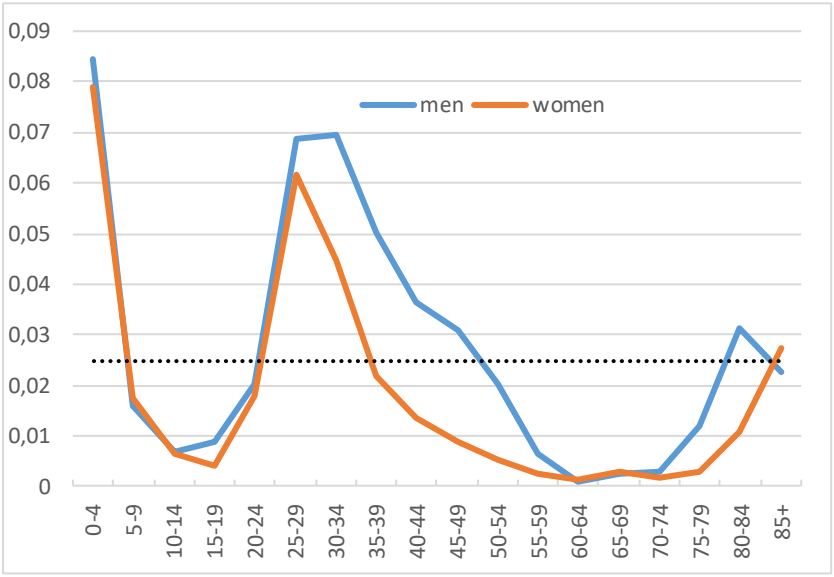
Norway	Netherlands	France
0,0249	0,0075	0,0492

4.2 Age and sex structures

Figures 7-9 plot normalized *CRPS*-scores for simulated populations broken down by sex and five-year age group on 1 January 2010 according to the UPE forecasts. The horizontal dotted lines represent *CRPS*-values for total population sizes from Table 2. The age patterns of the scores differ strongly between the three countries. The findings for Norway in Figure 7 are easy to interpret. High scores, i.e. low-quality forecasts, apply to young children, young adults, and the elderly. Scores are much better for ages 10-19 and 55-74. This age pattern reflects the under-prediction of immigration after 2005, already noted in Section 4.1. However, prediction errors for births and deaths may have contributed,

too. Indeed, the age pattern of the scores is qualitatively similar to the pattern found for absolute errors in point forecasts of age and sex structures (e.g. Keilman 2009). This reflects the fact that births, migration flows, and deaths are difficult to predict. The lead-time of the UPE-forecasts is only seven years. At such a short horizon, fertility has no impact on the age group 10-19. International migration and mortality influence these age groups only very little. The same holds for age group 55-74. Clearly, had the evaluation taken place after a lead-time of twenty years or more, the normalized *CRPS*-values for age groups 10-19 and 55-74 would have been much worse. Finally, note that the scores for men in ages 19-54 and 75+ are somewhat higher than those of women in these age groups. The reason is that men are more prone to migrate (19-54) or to die (75+) than women.

Figure 7. Normalized *CRPS*-scores for population by age and sex, Norway, UPE forecast 2010.



Whereas the Norwegian score agrees with what one might expect, the scores for the other two countries are more difficult to interpret. Normalized scores indicate that the Dutch forecast is of better quality than the other two, except for old ages. The French scores tend to decline with age. The pattern suggests that fertility was more difficult to predict accurately, than international migration or mortality. One may also think of several other explanations. First, the revision of the population numbers discussed above may have been stronger in some age groups than in others. We found (numbers not shown here) that revised numbers for men and women by five-year age group are approximately one per cent higher than those used in UPE. However, there are a few exceptions. Revisions were less than half a per cent in age groups 0-4 and 80+, while for men aged 20-24 the revised number was one per cent *lower* than the number used in UPE. This pattern caused by revisions between 2003 and 2010 is not reflected in Figure 9. A second explanation is that under- or over-prediction of net migration flows to France during the years 2003-2009 may also differ across age groups. Finally, our empirical data on age-sex distributions as of 2010 include the effects of so-called administrative corrections. Such corrections are necessary in case registration of births and deaths is incomplete. For register countries Norway and the Netherlands, errors in registered immigration and emigration are included as well in the administrative corrections. For Norway, the effect of these corrections is likely small, but the situation is worse for the Netherlands and France. For instance, data from Statistics Netherlands and INSEE show that total net-migration for the years 2003-2009 *without*

administrative corrections amounts to 214 000 and 601 000 persons respectively. Eurostat provides net migration data *including* such corrections. Using those data we find that the totals for net-migration during 2003-2009 are very different, namely 17 000 (the Netherlands) and 884 000 (France).[4] Because of the lack of reliable data on net migration and administrative corrections broken down by age for the Netherlands and France, we have not analysed this issue further. Note also that the UPE-forecasts do not include a separate variable that deals with administrative corrections (as is common practice for population forecasting).

Figure 8. Normalized *CRPS*-scores for population by age and sex, Netherlands, UPE forecast 2010.

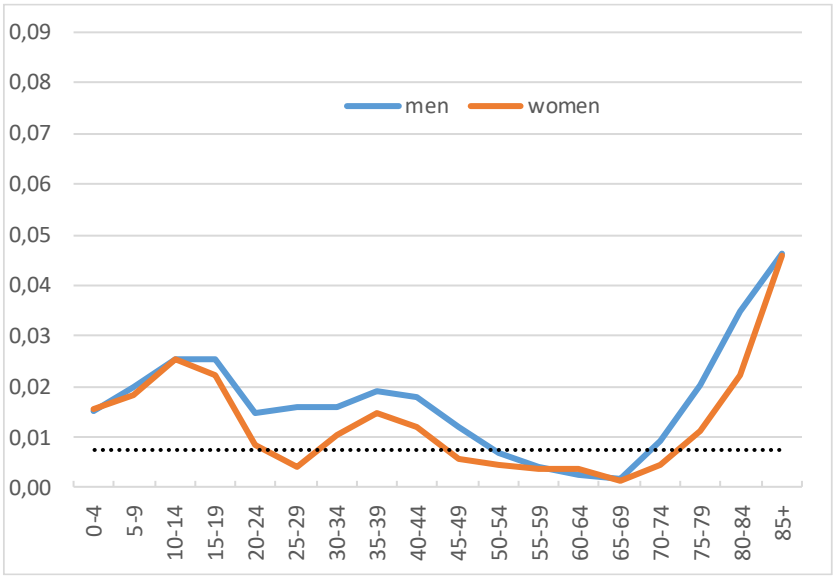
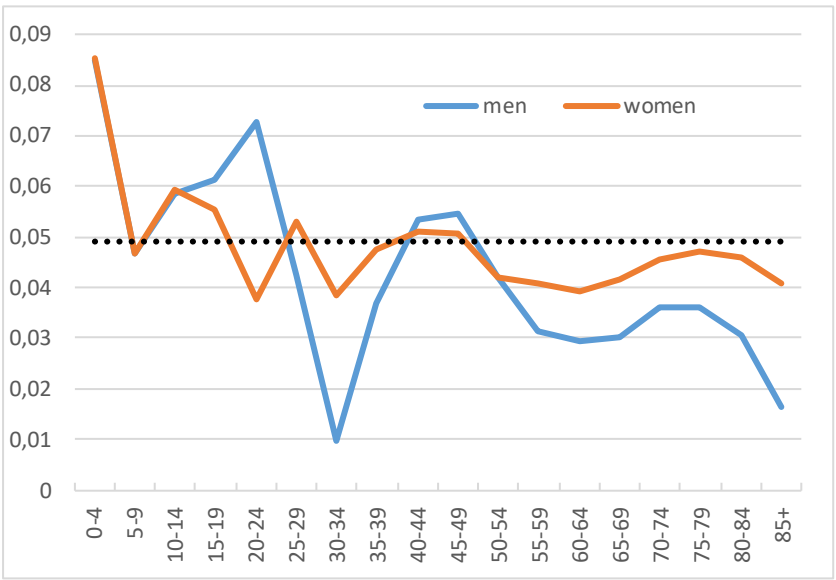


Figure 9. Normalized *CRPS*-scores for population by age and sex, Metropolitan France, UPE forecast 2010.



[4] For Norway, the numbers are 188 300 (without administrative corrections) and 187 800 (with corrections).

The general conclusion from this evaluation is that the UPE-forecast of the Dutch population pyramid for 2010, as measured by the normalized *CRPS* score, is better than the UPE-forecasts of Norway and France, except for the oldest-old. The age pattern for the Norwegian *CRPS* score is similar to that of absolute errors in point forecasts. It is difficult to indicate why the age patterns differ strongly between the three countries, due to data problems for international migration in particular.

5. Summary and conclusions

The purpose of this paper is to demonstrate how a probabilistic population forecast can be evaluated, when observations for the predicted variables become available. Statisticians have developed various scoring rules for that purpose, but there are hardly any applications in population forecasting literature. A scoring rule measures the distance between the probability distribution of the predicted variable, and the actual outcome. A score as such has no intrinsic meaning – we can only interpret it by comparing it to the score of another forecast. We have used scoring rules that reward accuracy (the outcome is close to the expected value of the prediction) and sharpness (the predictive distribution has low variance, which makes it difficult to hit the target). One may argue that accuracy is more important than sharpness: sharpness ought to be rewarded only when the outcome is not too far away from the central tendency of the predictive distribution. We discussed the notion of "too far away".

A forecaster can make the probabilistic forecast available to the user in three different ways. The first is by publishing a prediction interval for the variable of interest. Coverage probabilities of 67 and 80 per cent are rather common. Some population forecasters present 95 per cent prediction intervals. We do not recommend this practice, because 95 per cent intervals are very wide as they stretch to quantiles where extreme events start to happen. The second method is to give the user access to a database that contains sample paths for the stochastically simulated development in population size and other forecast results. Sometimes, only the first moment (expectation) and the second moment (variance) of the prediction interval are available. We presented scoring rules that one may use for either type of forecast results. The scoring rules are negatively oriented: a lower score implies a better forecast.

We have evaluated probabilistic population forecasts for France, the Netherlands, and Norway. For all three countries, we have used results from the UPE-project. Since many scoring rules apply the same scale as population size, we proposed using normalized scoring rules when the interest is in comparing forecasts for different countries. We inspected prediction intervals for population size in the period 2004-2019 and 3000 sample paths for population pyramids for the year 2010. For the Netherlands and for Norway, we compared the UPE-results with findings from the official probabilistic population forecast by Statistics Netherlands (2001-2019) and from a probabilistic forecast for Norway (1997-2019). All forecasts were computed using the cohort-component method and stochastically varying parameters for fertility, mortality and migration.

Our evaluations show that the UPE-forecasts for the Netherlands and for Norway performed better than the other forecasts for these two countries, because the UPE-predictions were relatively sharp, with narrow prediction intervals. The UPE-forecast for France started from a jump-off population in 2003 that was estimated at 60.1 million persons at the time the forecast was computed. This number is almost 500 000 persons higher than the current estimate of the population in 2003 (59.6 million). The error in the jump-off population caused a bad score for the French forecast. To revise population statistics for inter-census years when data from a new population census become available, is common practice. In case one cannot be certain about the size and structure of a population during an inter-censal period, the correct approach is to treat the jump-off population of the forecast as stochastic.

We evaluated the 3000 UPE-simulations of the age and sex composition predicted for the year 2010. When normalized for population numbers in each age-sex category, the predictions for the Netherlands received the best scores, except for the oldest old. The age pattern for the Norwegian score reflects the under-prediction of immigration after the enlargement of the European Union in 2005. However, prediction errors for fertility and mortality may have played a role as well. The age-specific scores for France are difficult to interpret. They do not reflect the age pattern of the revision of the population data for 2003 mentioned above. Over- or under-prediction of fertility, mortality and migration may have played a role. In the cohort-component model, the age- and sex-composition of the population of 2010 is a complicated non-linear function of model parameters for mortality, fertility, and migration prior to 2010. Therefore, one cannot identify the contribution of these three components of change to the scores.

In addition to the issue of data revision, we were also confronted with the problem of "administrative corrections". This is a notion that statistical agencies sometimes use as a distinct component of change of population size and structure. When there are errors in the registration of births, deaths, and migrations, administrative corrections are necessary to obtain a correct set of bookkeeping statistics for population. Empirical population numbers for the Netherlands and France are strongly influenced by administrative corrections.

There is a rich literature that evaluates probability forecasts and that discusses a large number of scoring rules. Many apply to predictive distributions of a discrete random variable, and are of little interest for evaluating demographic forecasts. In case we limit ourselves to scoring rules for continuous random variables, the literature still proposes many scoring rules, of which we selected just a few. As we have shown in Sections 3 and 4, these scoring rules are very different, giving different weight to distance or to sharpness. Some rules give a bad score as soon as observed numbers fall outside the prediction interval. Others develop more smoothly when the observation is further and further away from the central tendency and from the interval bounds. Further work applied to scoring rules for probabilistic demographic forecasts is necessary, hopefully leading to guidelines for the selection of such rules in various situations.

Scoring rules are useful in *ex-post facto* evaluations of two or more probabilistic forecasts. Once we have concluded that, judged by a number of score functions, one forecast was better than another one, we have to ask ourselves *why* this was the case. To answer that question, one needs to analyse very carefully the many steps in the production process of the two probabilistic forecasts. This poses a new challenge, in particular when different scholars or different agencies computed the two forecasts.

Acknowledgement

Excellent comments from Laurent Toulemon and from three reviewers anonymous are gratefully acknowledged.

References

Alders, M. and De Beer, J. (1998) Kansverdeling van de bevolkingsprognose ("Probability distribution of the population forecast"). Maandstatistiek van de Bevolking 46, 8-11.

Alexopoulos, A., Dellaportas, P., Forster, J.J. (2018) Bayesian forecasting of mortality rates by using latent Gaussian models. Journal of the Royal Statistical Society: Series A (Statistics in Society).

Alho, J. and Nikander, T. (2004) Uncertain population of Europe: Summary results from a stochastic forecast. Available at http://www.stat.fi/tup/euupe/del12.pdf (accessed on 21 March 2019).

Alho, J. and Spencer, B. (2005) Statistical Demography and Forecasting. New York: Springer.

Askanazi, R., Diebold F.X., Schorfheide F., Shin M. (2018) On the comparison of interval forecasts. Journal of Time Series Analysis 39(6), 953 – 965.

Bijak, J., and Bryant, J. (2016). Bayesian demography 250 years after Bayes. Population Studies, 70(1), 1–19.

Blanpain, N. and Buisson, G. (2016) Projections de population à l'horizon 2070: Deux fois plus de personnes de 75 ans ou plus qu'en 2013. Insee Première no 1619, Novembre 2016.

Brier, G. (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78(1) 1-3.

Costemalle, V. (this issue) Projections probabilistes bayésiennes de population pour la France.

Gneiting, T. and Raftery, A. (2007) Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association 102(477) 359-378.

Gneiting, T. and Katzfuss, M. (2014) Probabilistic forecasting. Annual Review of Statistics and Its Applications 1: 125-151.

ISTAT (2018) Il futuro demografico del paese: Previsioni regionali della popolazione residente al 2065 (base 1.1.2017). Report Statistische 3 maggio 2018. Roma: ISTAT.

Jordan A., Krüger F., Lerch S. (2019) Evaluating Probabilistic Forecasts with **scoringRules**. Journal of Statistical Software.

Keilman, N. (1990) Uncertainty in national population forecasting: Issues, backgrounds, analyses, recommendations. Amsterdam and Rockland, MA: Swets and Zeitlinger Publishers.

Keilman, N. (2009) Erroneous population forecasts. In P. Festy and J.-P. Sardon (eds.) Profession démographe - Hommage à Gérard Calot, pp. 237-254. Paris: INED.

Keilman, N., Pham, D.Q, Hetland, A. (2002) Why population forecasts should be probabilistic - illustrated by the case of Norway. Demographic Research 6-15 May 2002, 409-454.

Keyfitz, N. (1981) The limits of population forecasting. Population and Development Review 8 (44), 579-593.

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. Management Science 22: 1087-1096.

Murphy, A. (1970) The ranked probability score and the probability score: A comparison. Monthly Weather Review December 98, 917-924.

NRC - National Research Council (2000) Beyond six billion: Forecasting the world's population. Panel on Population Projections. John Bongaarts and Rudolfo Bulatao ed. Washington DC: National Academy Press.

Raftery, A., N. Li, H. Ševčíková, P. Gerland, G. Heilig (2012) Bayesian probabilistic population projections for all countries. PNAS 109 (35), 13915-13921.

Shang, H.L. (2015) Statistically tested comparisons of the accuracy of forecasting methods for age-specific and sex-specific mortality and life expectancy. Population Studies 69(3), 317-335.

Shang, H.L., Smith, P., Bijak, J. and Wisniowski, A. (2016) A multilevel functional data method for forecasting population, with an application to the United Kingdom. International Journal of Forecasting 32, 629-649.

Shang, H.L. and Hyndman, R. (2017) Grouped functional time series forecasting: An application to age-specific mortality rates. Journal of Computational and Graphical Statistics 26(2), 330-343.

Smith, S., Tayman, J. and Swanson, D. (2001) State and Local Population Projections: Methodology and Analysis. New York: Kluwer Academic/Plenum Publishers.

Staël von Holstein, C.-A. (1970) A family of strictly proper scoring rules which are sensitive to distance. Journal of Applied Meteorology 9, 360-364.

Statistics New Zealand (2011) National Population Projections: 2011(base)–2061. Bulletin published 19 July 2012, ISSN 1178-0584. Available at http://archive.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalPopulationProjections_HOTP2011.aspx (accessed on 21 March 2019).