MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2024-016 I July 2024
Revised September 2023
https://doi.org/10.4054/MPIDR-WP-2024-016

Calibrating Probabilistic Forecast Paths on Past Forecast Errors: An Application to the Finnish Total Fertility Rate

Ricarda Duerst I duerst@demogr.mpg.de
Jonas Schöley
Julia Hellstrand
Mikko Myrskylä

Calibrating Probabilistic Forecast Paths on Past Forecast Errors: An Application to the Finnish Total Fertility Rate

Ricarda Duerst[a,b], Jonas Schöley[a], Julia Hellstrand[b,a], Mikko Myrskylä[a,b,c]

[a] *Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1, Rostock, 18057, Germany*
[b] *Helsinki Institute for Demography and Population Health, University of Helsinki, Unioninkatu 33, Helsinki, 00170, Finland*
[c] *Max Planck - University of Helsinki Center for Social Inequalities in Population Health, Konrad-Zuse-Straße 1, Rostock, 18057, Germany & Unioninkatu 33, Helsinki, 00170, Finland*

Abstract

In probabilistic forecasting, the key to reasonable results is good calibration of the forecast uncertainty. Analyzing forecasting errors offers an empirical solution for the calibration of such forecasts. We propose a novel quantile-mapping approach whereby we map non-calibrated forecasted outcome trajectories from a forecast model to a target distribution derived from historical out-of-sample forecasting errors. We present probabilistic forecasts of the Finnish Total Fertility Rate (TFR) from 2024 to 2070 calibrated on the historically observed distribution of forecasting errors. The forecasts come from two scenario-based models. The postponement time series model (PPS) assumes that fertility postponement will gradually decline and eventually stop. The second model is a naive freeze-rates approach to forecasting fertility. The validation shows that our TFR forecasts calibrated on historical data outperform the non-calibrated TFR forecasts in coverage and in-

terval score metrics. The results demonstrate the efficacy of empirical error quantification and quantile-mapping in calibrating probabilistic demographic forecasts.

Keywords: conformal prediction, quantile-mapping, empirical prediction intervals, fertility forecasting, Finland

---

1. Introduction

Demographic forecasting involves predicting future population trends based on factors such as birth rates, death rates, migration patterns, and aging. These forecasts are essential for governments, businesses, and policymakers to plan for future resource needs, including infrastructure, healthcare, and social services. However, demographic forecasting is subject to uncertainties due to unpredictable changes in behavior, policy, and environmental factors. Forecast uncertainty arises from variability in data inputs, model assumptions, and external events such as pandemics or economic crises. Managing this uncertainty and expressing it in the forecasts requires is essential to providing a comprehensive view of future demographic outcomes.

In the 1960s and early 1970s, Finland, and other Nordic countries, have experienced a sharp decline in fertility. After a period of recovery in the 1980s, followed by relatively stable period fertility in the 1990s and 2000s, Finland's Total Fertility Rate (TFR) fell again in the 2010s and reached a historical low of 1.26 in 2023 (Statistics Finland (2024)). Starting in the 1970s, fertility postponement, the delay of childbearing to older ages, has depressed fertility levels. However, the strong decrease in TFR in the recent decade is not only due to further acceleration of fertility postponement. Instead,

the decrease in Finnish period fertility likely is a quantum effect (Hellstrand et al. (2020)). This was not foreseen by demographic theories, making the future of Finnish fertility particularly interesting yet highly challenging to predict.

In light of yet another fluctuation in the trend of Finnish fertility, it is important to capture and convey the uncertainty that comes with forecasting Finland's fertility. Probabilistic forecasts, in contrast to deterministic point-forecasts, offer a solution to this problem. These type of forecasts assign a probability to each possible outcome and therefore allow one to distinguish between likely and extreme scenarios and to plan accordingly (for an introduction to probabilistic forecasting see e.g. Lee (1998) and Keilman (2018)). However, probabilistic forecasts are only of use, if they are well calibrated. Good calibration means that the forecasted probabilities of an outcome are consistent with the observed relative frequencies of that outcome, or put differently, forecasted rare events occur rarely and forecasted typical events occur regularly. Further, because of the highly fluctuating nature of the Finnish period fertility in the past, forecast models that extrapolate observed trends of the past into the future are unlikely to provide reasonable results, as they are unable to predict trend changes. Therefore, we propose the use of scenario-based approaches that can take the changing nature of the fertility development into account.

We present probabilistic forecasts of Finland's Total Fertility Rate (TFR) from 2024 to 2070 from two scenario-based forecast models. The first model, introduced in Nisén et al. (2020), is the Postponement Time Series Model (PPS). The PPS model operates under the assumption that the trend of de-

laying childbirth to later ages, known as fertility postponement, will continue but at a gradually slower pace, until it eventually stops. Hence, the increase in the average age of childbearing slows down and stabilizes. As childbirth is postponed, fertility temporarily decreases because children are born later during the life course. Therefore, the TFR is lower, compared to a scenario where childbirth is not delayed. When fertility postponement slows down, fertility increases. To mitigate the delay's impact on period fertility measures, we use the tempo-adjusted TFR. This adjusted rate is an estimate of what the TFR would be if the timing of childbearing did not change (Bongaarts and Feeney (1998)). The PPS model assumes that the TFR and the tempo-adjusted TFR will converge due to the assumed slowing and stopping of fertility postponement by 2050.

We use a second model as a naïve baseline to compare the results from the PPS model to. The freeze-rates model operates under the assumption that the recent decrease in Total Fertility Rate (TFR) isn't a result of delayed childbirth, meaning that fertility has declined without postponement of births by the population. While this hypothesis is theoretically feasible, considering demographic data, it seems improbable. The second assumption is that this fertility decline will come to a halt, and age-specific fertility rates will stay at the levels observed in the last recorded year, 2023.

To ensure a good calibration of the probabilistic forecasts, we use an empirical approach to calibration which has its roots in the analysis of errors of demographic forecasts (Williams and Goodman (1971); Stoto (1983); Cohen (1986); Smith and Sincich (1988); Alho and Spencer (2005); Alho et al. (2008)). Notably, Keilman and Pham (2004) constructed empirical predic-

tion intervals around forecasts of Nordic fertility, deriving the intervals from the standard deviation of errors of historical forecasts published by statistical agencies. After comparing the width of the prediction intervals with those from time-series models, the authors concluded that empirical forecast errors provide useful information when constructing prediction intervals for TFR forecasts. However, they noted that their empirical errors might not be normally distributed and "one has to be cautious" when using them.

In the field of machine learning, the methodology is known under the term conformal prediction (CP, Shafer and Vovk (2008)) and is used to construct probabilistic forecasts calibrated on out-of-sample errors. Since its introduction in Gammerman et al. (1998) until today, there have been strong methodological advances in this field, resulting in a variety of CP methods for different applications including, in more recent publications, time series forecasting (Fontana et al. (2023); Angelopoulos et al. (2024)). In climate modeling, calibration of predictions using historical data is widely done under the term quantile-mapping as a form of bias-correction of forecast distributions (Cannon (2018); Qian and Chang (2021)).

The existing methodology provides probabilistic forecasts calibrated on historical data in the form of prediction intervals, meaning lower and upper bounds in which the forecast outcome will lie with a given probability, e.g. 95%. However, so far, research has not provided solutions for forecast outcomes in the form of calibrated time series trajectories that correspond to these bounds. Simulated trajectories of demographic outcomes are needed as input for down-stream modeling, e.g. for the modeling of determinants of social security systems. Therefore, we propose a flexible methodology to obtain

forecasts of demographic measures in the form of simulated trajectories that have been calibrated on historical data. We provide two sets of forecasts of the Finnish TFR from 2024 to 2070 from the PPS model and the naïve model that are designed to be transparent, probabilistic, well calibrated, and easy to integrate into further probabilistic modeling downstream. By "forecasting" TFR for the Nordic countries (Finland, Sweden, and Norway) for the years 1973 to 2023 we are able to learn the distribution of forecasting error of the two models and calibrate our future TFR forecasts accordingly. We build on the prevailing research, by proposing to combine existing approaches of empirical forecast calibration. First, we use the idea of learning a smoothed, time-varying distribution of historical forecasting error, as described in the "scorecaster" approach by Angelopoulos et al. (2024), which allows for bias in the forecasts. Second, we combine this approach with the technique of quantile-mapping (Cannon (2018)) in order to calibrate stochastic forecasting paths to a target distribution of empirical forecasting errors, as opposed to providing upper and lower bounds of prediction intervals around a point forecast.

In the remainder of this paper, we first describe the PPS and the naïve forecast model and our methodology of empirical forecast calibration in detail. After we present the results of Finland's probabilistic TFR forecasts, we validate our forecast models and summarize the quality of the probabilistic forecast with several calibration metrics.

2. Methods and data

We construct prediction intervals around time series forecasts of the Finnish Total Fertility Rate (TFR) which are calibrated on out-of-sample forecasting errors for Finland, Sweden, and Norway, and change dynamically over the forecasting horizon. The width of the prediction interval should reflect the distribution of forecasting errors as observed in the past for the same prediction model. We pool the forecasting errors from Finland, Sweden and Norway to achieve more robust estimates of the forecasting error distribution by decreasing autocorrelation (Alho et al. (2008)) and validate the results on data from Finland, Sweden, Norway, and Denmark. The empirical prediction intervals are then used to calibrate 5,000 forecast paths of future Finnish TFR. We source the time series of the fertility data from the Human Fertility Database (Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria)) for the years 1944 to 2023 for Finland and Sweden, years 1968 to 2023 for Norway, and years 1946 to 2023 for Denmark.

2.1. Forecast Models

We use two different forecast models to forecast the fertility:

1. *Postponement Time Series Model (PPS):* The scenario is based on the demographically meaningful assumption that fertility postponement would continue but gradually slow down and eventually stop. Due to biological factors, fertility postponement cannot continue forever, as fecundability declines with age, and especially fast for women in their mid to late 30s (Rothman et al. (2013)). According to Gold-

stein (2006), the mean age at first birth could plausibly rise to around 33 years before reaching biological and social limits. Further, Sobotka (2017) suggests that fertility postponement would continue another two to three decades. In line with these previous findings, in our scenario, fertility postponement stops in 2050 and the mean age at childbirth approaches 33, which is approximately the current highest value reported in the Human Fertility Database (32.9 years in the Republic of Korea in 2020, Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) (2024)). This assumption about the fertility postmponement is implemented in the forecast model by calculating the tempo-adjusted TFR for 2023 and forcing the TFR and the tempo-adjusted TFR to converge by 2050. After the convergence is completed, the TFR stayes fixed until the end of the forecast period. The model was introduced in Nisén et al. (2020). The time series of 5-year age group fertility rates are forecasted by a random-walk-with-drift model $\ln(\hat{y}_{x,t}) = \beta_{x,t} + \ln(y_{x,t-1}) + \epsilon_{x,t}, \epsilon_{x,t} \sim N(0, \sigma_x^2)$, where x is the age group, t is the calendar year, $\beta_{x,t}$ is the model drift and $\epsilon_{x,t}$ is the error term. The drift term $\beta_{x,t}$ is forecast under two calculation assumptions. First, the increase in the average age at birth slows down and the TFR approaches the tempo-adjusted TFR. Thereafter, there is no drift term, i.e. $\beta_{x,t} = 0$. Finally, the TFR is obtained by adding up the age-group-specific fertility rates and multiplying it with the width of the age-groups: $\text{TFR}_t = 5 * \sum_{x=15}^{45} \hat{y}_{x,t}$.

2. *Naïve freeze-rates model:* In this model, no drift-term is added to the random walk, so that the average 5-year age-group fertility rate of the

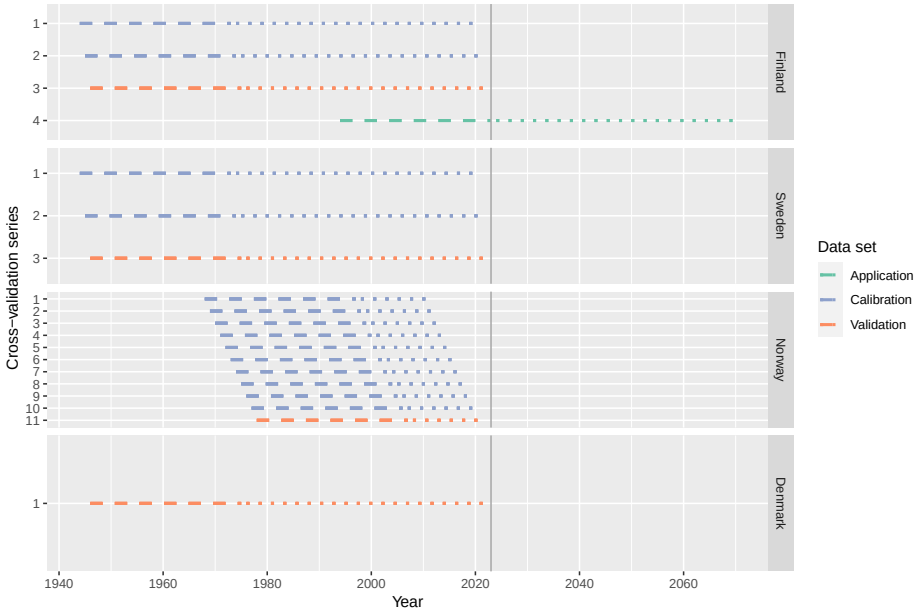
future is at the same level as the last observed value of 2023, resulting in the following model: $\ln(\hat{y}_{x,t}) = \ln(y_{x,t-1}) + \epsilon_{x,t}, \epsilon_{x,t} \sim N(0, \sigma_x^2)$. Similarly, the TFR is calculated as the sum of the forecast age-group-specific fertility rates multiplied with the width of the age-groups. The model serves as a naïve baseline model.

2.2. Split data

The time series of observed annual ASFR values y_t for each country is partitioned into three data sets: the calibration data $\mathcal{D}_{\text{cal}}$, validation data $\mathcal{D}_{\text{val}}$, and application data $\mathcal{D}_{\text{app}}$. See Figure 1 for a visual representation of the data splitting and resulting data sets. We use $\mathcal{D}_{\text{cal}}$ to estimate the time-dependent distribution of the forecasting error around y_t and to calibrate the empirical prediction intervals accordingly. The hold-out set $\mathcal{D}_{\text{val}}$ is used to validate the properties of the empirical prediction intervals and $\mathcal{D}_{\text{app}}$ holds the data we want to forecast. The calibration data of Finland and Sweden hold two cross-validation series and the validation data one series, each with 30 years of training data, y_{Train}, followed by 47 years of test data, y_{Test}. Due to data constrains, we split the Norwegian data into 11 smaller cross-validation series with 30 years of training data, each, and reduced the test data to 15 years. The last one of these series is the validation data. The Danish data holds one cross-validation series of validation data of the same length as the Swedish and Finnish data. This increases the amount of left-out data, i.e. data that was never used in calibration, to validate on. The validation series of the other countries are partially overlapping with series from the calibration data. Thus, excluding Denmark from the calibration data set avoids over-fitting. The application series holds a single Finnish

series y_{Train} of length 30. See Tables 1, 2 and 3 for the input years of each data set.

Figure 1: Visualization of data splitting into cross-validation series.



2.3. Generate forecasts

Given a prediction model $\hat{y}_{T+h} = f(h|y_{\text{Train}})$ we produce central TFR forecasts $\hat{y}_t$ over forecasting horizon $h \in 1, 2, \ldots, H$. Forecasts are produced for $\mathcal{D}_{\text{cal}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{app}}$ and for all cross-validation series c. The models considered are the PPS model and the naïve model described in Section 2.1.

2.4. Calculate forecasting errors

We define a scoring function, $S(y_t, \hat{y}_t)$, to measure the deviance between the observed TFR value, y_t, and the prediction, $\hat{y}_t$, as $S(y_t, \hat{y}_t) = \ln(y_t/\hat{y}_t) = s_t$. The log-ratio scoring function, being a measure of relative error, is suitable

for positive values which vary in scale, like the TFR (Alho et al. (2008)). We calculate $s_{T+h} = S(y_{T+h}, \hat{y}_{T+h})$ for each point prediction over the forecast horizon of $\mathcal{D}_{\text{cal}}$. We denote the inverse of the scoring-function as $S^{-1}(s_t, \hat{y}_t) = \exp(s_t) \cdot \hat{y}_t$.

2.5. Model time-dependent distribution of forecasting errors

Given the observed forecasting errors u_{T+h} we estimate the cumulative error distribution $\hat{F}_{U_{T+h}}(u) = \mathrm{P}(U_{T+h} \leq u)$ over the forecast horizon of $\mathcal{D}_{\text{cal}}$. We employ a *Time-varying Skew-normal model (SN)* for the distribution of the error: $U_{T+h} \sim \text{SkewNormal}(\mu, \sigma_h, \tau)$, where $\sigma_h = \exp(\beta_1 + \beta_2 \cdot h)$. The estimated distribution function, $\hat{F}_{U_{T+h}}$, and quantiles, $\hat{Q}_{U_{T+h}}$, are then analytically given by the Skew-Normal distribution. Due to data limitations in the training data resulting in fewer long-term forecasts, it is necessary to model the distribution of the forecasting errors rather than using their distribution as is. Figure 4, which is a scatter plot of the calculated forecasting errors with the modelled error distribution at the 0.025 and 0.975 quantiles, illustrates the need for modelling. Angelopoulos et al. (2024) take on the same approach and call the model of the forecasting error distribution a scorecaster. In our case, we have chosen a time-varying skew-normal model as our scorecaster which allows for bias in the forecasts, because we expect the bias in the forecasting error distribution to also be present in the future. We added the constraint that the scorecaster's width is not allowed to narrow with time, because we assume that as the forecast length increases, so does the forecast uncertainty.

2.6. Calibrate forecast paths with quantiles of forecasting error distribution

We calibrate the forecasted TFR paths such that their distribution at forecasting step h reflects the modelled historical distribution of out-of-sample forecasting errors at h. We do so for the central forecasts $\hat{y}_{T+h}$ around which we construct empirical prediction intervals, and for the forecasted fertility paths $\hat{y}_{T+h,i}$.

1. *Probability transform method:* Because the forecasting error U_t is a transformation of the random variable Y_t, $U_t = S(Y_t, \hat{y}_t)$, with central forecast $\hat{y}_t$ being treated as non-random and scoring function S being a smooth and monotonic function over the range of Y_t, we have $F_{Y_t}(y) = F_{U_t}(S(y))$. Thus, the p quantile of the distribution of predicted values Y_{T+h} can be derived from the corresponding quantile of the error distribution U_{T+h} via $Q_{Y_{T+h}}(p) = S^{-1}(\hat{F}^{-1}_{U_{T+h}}(p), \hat{y}_{T+h})$.
2. *Quantile-mapped paths:* We calibrate a set of $i \in 1, 2, \ldots, N$ forecast fertility paths, $\hat{y}_{T+h,i}$, such that the marginal distribution of the calibrated paths $F_{\hat{Y}^*_{T+h}}$ is equal to some target distribution $F^+_{Y_{T+h}}$, which reflects the historical out-of-sample forecasting error. This calibration is achieved by having calibrated paths $\hat{y}^*_{T+h,i} = Q^+_{Y_{T+h}}(F^{\text{ecdf}}_{\hat{Y}_{T+h}}(\hat{y}_{T+h,i}))$, where $F^{\text{ecdf}}_{\hat{Y}_{T+h}}$ is the empirical cumulative distribution function over the non-calibrated forecast paths and $Q^+_{Y_{T+h}}$ is the quantile function of the target distribution at $T + h$. To prevent the calibrated paths from taking infinite values, we adjust $F^{\text{ecdf}}_{\hat{Y}_{T+h}}$ for the maximum values of the forecast paths, $\hat{y}_{T+h,max}$. Instead of $F^{\text{ecdf}}_{\hat{Y}_{T+h}}(\hat{y}_{T+h,max})) = 1$, which results in $Q^+_{Y_{T+h}}(F^{\text{ecdf}}_{\hat{Y}_{T+h}}(\hat{y}_{T+h,max})) = \infty$, we set $F^{\text{ecdf}}_{\hat{Y}_{T+h}}(\hat{y}_{T+h,max})) = 1 - 0.5^{\frac{1}{N}}$. This adjustment places the empirical cumulative distribution function's

value halfway between 1 and the value for the second-highest $\hat{y}^*_{T+h,i}$, ensuring that the calibrated paths remain finite. The target distribution is constructed via the *Probability transform method* described above.

2.7. Validate prediction intervals

The validation of the empirical prediction intervals is done on the validation data set $\mathcal{D}_{\text{val}}$ over the years of the training data, y_{Train}. Three calibration scores are evaluated and compared for the model-based prediction intervals of the PPS and the naïve model and the empirical prediction intervals stemming from the quantile-mapped paths of the PPS model. We evaluate the calibration scores over the full range of the forecast years (47 years) and over sub-sections of the forecast years (1 to 5, 6 to 15, 16 to 25, and 26 to 47 years).

1. *Coverage:* We compare the nominal 80%, 90% and 95% coverage with the actual coverage of the prediction intervals across the cross-validation series c. The actual coverage is the fraction of observations n that are inside the bounds of the declared prediction interval of all observations N: $Cov = \frac{n\{l \leq y_t \leq u\}}{N}$ where l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles.
2. *Mean Interval Width:* We define the mean interval width as the mean difference between the upper and lower bounds of the prediction interval over the forecast years: $\overline{W} = \frac{\sum^t u_t - l_t}{N}$, where l_t and u_t are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles at time t.
3. *Mean Interval Score (MIS):* The interval score S takes both the coverage and the width of the prediction intervals into account. Given the same actual coverage, a prediction interval that is on average wider

is penalized by the interval score (Gneiting and Raftery (2007)). The score is defined as follows:

$$\mathrm{S}^{int}_{\alpha,t}(l_t, u_t; y_t) = \begin{cases} (u_t - l_t) + \frac{2}{\alpha}(l_t - y_t) & \text{for} \quad y_t < l_t \\ (u_t - l_t) + \frac{2}{\alpha}(y_t - u_t) & \text{for} \quad y_t > u_t \end{cases} \tag{1}$$

where l_t and u_t are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles at time t, and $\frac{2}{\alpha}(l_t - y_t)$ and $\frac{2}{\alpha}(y_t - u_t)$ are the penalty terms for observations that fall below or above the bounds, respectively. The penalty is proportional to the $1-\alpha$ level. We aggregate the interval score for every observation y over time t using the mean to be able to compare the different forecast models (Bracher et al. (2021)): $MIS^{int}_{\alpha} = \frac{\sum^{t} S^{int}_{\alpha,t}}{N}$. In terms of interpretation, the smaller the MIS, the better the prediction interval.

3. Results

In the following, first, we present the forecast results of the PPS and the naïve model, followed by the calibration of the PPS forecasts with the historical error data. Second, we present the results from the validation analysis.

3.1. Forecast Results

Figure 2 and 3 show an intermediate outcome of the two forecast models in form of the forecast paths of the age-specific fertility rates (ASFR) of Finland, together with the observed ASFR of 5-year age-groups. The observed ASFR values, represented by black dots, cover the period from 1944 to 2023, followed by 250 of the 5,000 forecast paths up to the year 2070. As described in Section 2, the ASFR paths are then used to calculate the forecast TFR

		Training Period		Calibration Period	
Country	**Series Nr.**	**Start**	**End**	**Start**	**End**
Finland	1	1944	1973	1974	2020
	2	1945	1974	1975	2021
Sweden	1	1944	1973	1974	2020
	2	1945	1974	1975	2021
Norway	1	1968	1997	1998	2012
	2	1969	1998	1999	2013
	3	1970	1999	2000	2014
	4	1971	2000	2001	2015
	5	1972	2001	2002	2016
	6	1973	2002	2003	2017
	7	1974	2003	2004	2018
	8	1975	2004	2005	2019
	9	1976	2005	2006	2020
	10	1977	2006	2007	2021

Table 1: Calibration data D_{cal}

paths, shown in Figure 6 for the naïve model, and in Figure 5 (panel a) for the PPS model.

After calculating the forecasting errors of the PPS model for the cross-validation series of the calibration data, we model the error distribution as described in 2. The empirical distribution of the forecasting error, see Figure 4, shows that the PPS model tends to overpredict the Total Fertility Rate (TFR) in Finland. The asymmetry of the forecasting error is already visible

		Validation Period	
Country	**Nr. of Series**	**Start**	**End**
Finland	3	1976	2022
Sweden	3	1976	2022
Norway	11	1976	2022
Denmark	1	2008	2022

Table 2: Validation data D_{val}

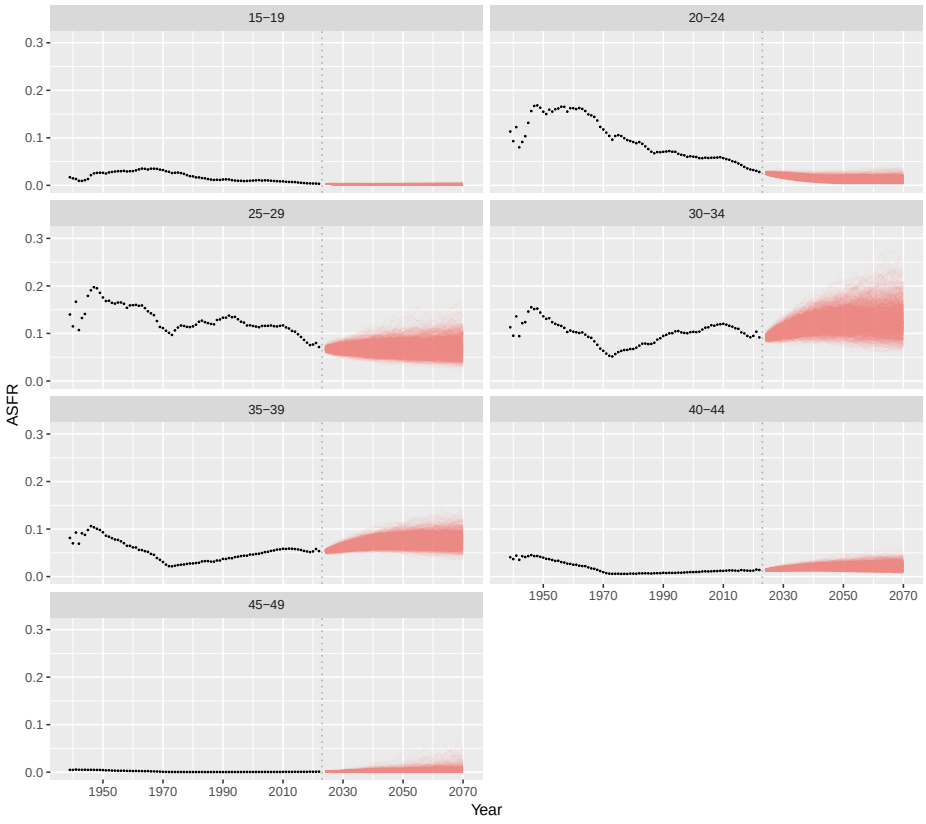
Country	**Training Period**		**Forecast Period**	
	Start	**End**	**Start**	**End**
Finland	1994	2023	2024	2070

Table 3: Application data D_{app}

in the first three forecast years and increases with the forecast length, justifying the use of the Time-varying Skew-normal model to model the forecasting error distribution. Another reason for modeling the forecasting error distribution rather than using raw quantiles is the need for interpolation, because the data scarcity results in fewer data points with increasing forecast length.

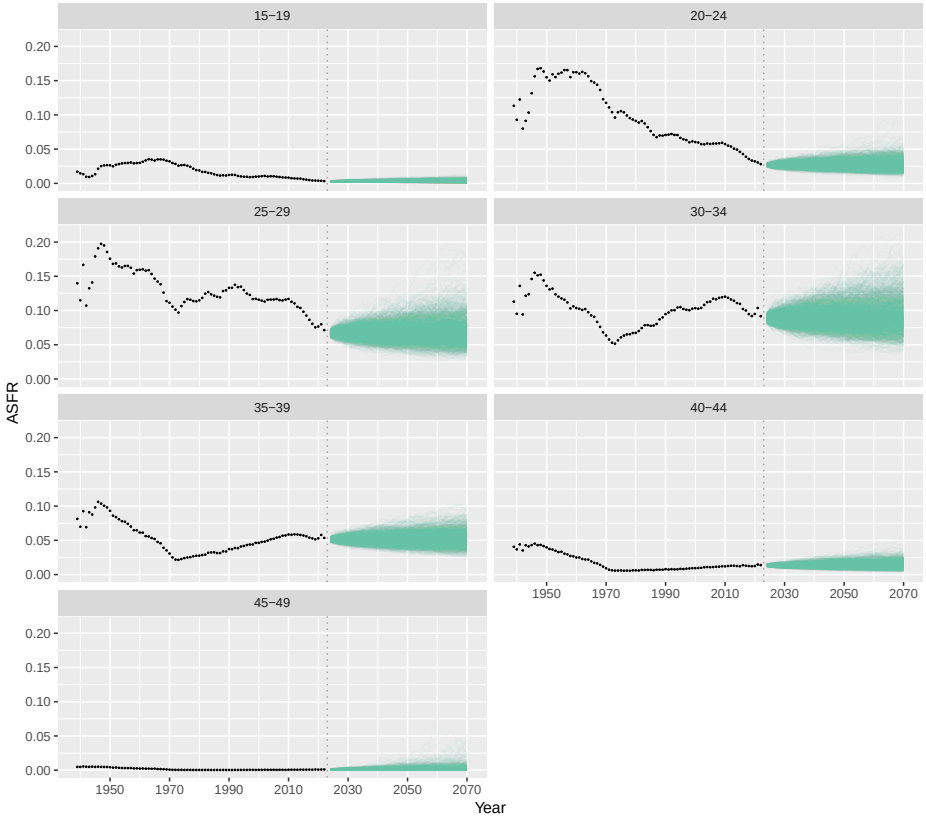
The asymmetry of the empirical forecasting errors towards over-prediction and their magnitude in the early forecast years leads to the differences in the probabilistic forecast for Finland that can be observed in Figure 5. Panel a shows the TFR paths from the PPS model without calibration to the historical error data. Due to the assumption of the PPS model that the fertility postponement will continue but slow down and eventually stop, the median of the forecast paths rises in the first 15 forecast years and then levels off to reach a TFR of 1.56 in 2070. The 95% prediction interval starts narrow

Figure 2: Observed ASFR of Finland and 250 forecast paths from PPS model by 5-year age-groups.



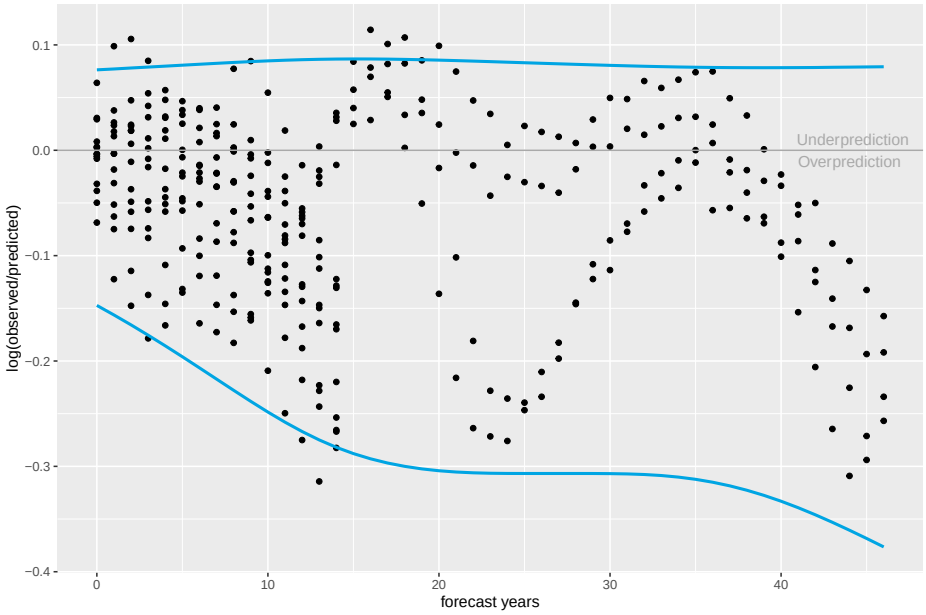
around the first forecast and increases with increasing forecast length, ranging from a TFR of 1.20 to 2.03. Panel b shows the TFR trajectories derived from the PPS model and calibrated to the historical forecast error distribution. In contrast to the non-calibrated forecasts, the 95% prediction interval is notably narrower, ranging from 1.07 to 1.69. As a result of the shape of the empirical foresting error distribution, the prediction interval is negatively skewed and starts notably wider in 2024 than the non-calibrated prediction interval. Further, the median forecast of 1.44 is lower. The changed shape

Figure 3: Observed ASFR of Finland and 250 forecast paths from naive model by 5-year age-groups.



and median of this forecast distribution is the result of calibrating the forecast paths to represent the historical forecasting error distribution that starts wider, widens more slowly, and is skewed towards overestimation. We illustrated the different distributions of forecast paths by adding the density plots for the years 2024 and 2070 (in grey color). These highlight the difference in width and skewness of the forecast distributions before and after the calibration to the historical error data. Overall, the calibrated PPS forecast predicts that Finland's TFR is likely to be higher in 2070 than last observed.

Figure 4: Forecasting errors from calibration data series with 95% quantiles of modeled error distribution (blue).

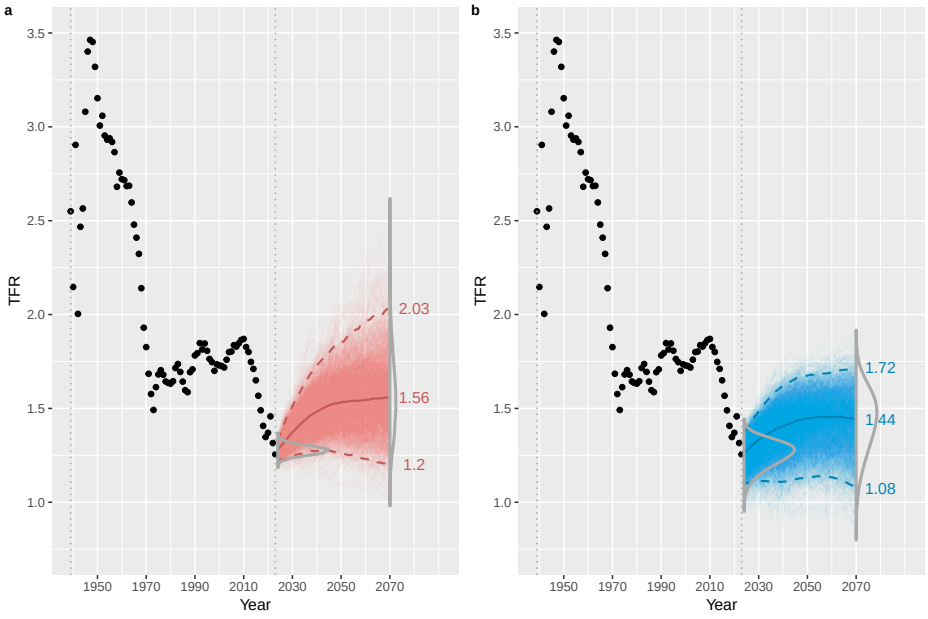


The probability that the TFR will fall below the level of 2023 (1.26) is 14%.

3.2. Forecast Validation

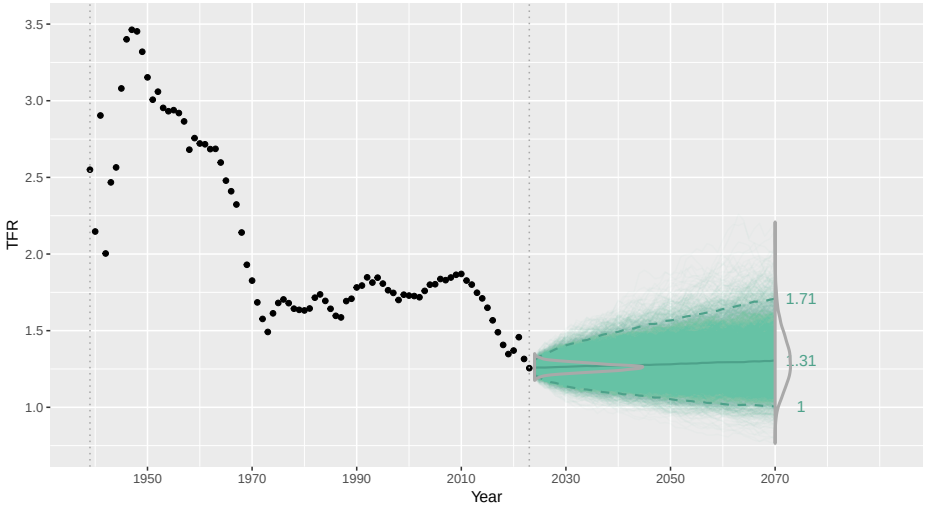
The validation analysis supports the visual differences between the PPS forecast paths that have been calibrated to historical error data and those that were not, and shows a consistently higher performance of the calibrated forecast paths for the validation data, compared to the non-calibrated PPS and naïve forecasts. Table 4 summarizes the calibration scores for three different nominal coverage levels (95%, 90% and 80%) for the full forecast horizon of 47 years. For all nominal coverage values, the calibrated PPS paths have the closest actual coverage, while also being better calibrated in terms of width, as indicated by smaller MIS values. Looking at the validation

Figure 5: Observed TFR of Finland and 250 forecast paths from PPS model (a) and after calibration on historical error data (b), with median and 95% quantiles, and density in 2024 and 2070.



results for short, medium and long term forecast years (see Table 5), we can see why the calibrated forecasts perform better. The non-calibrated prediction intervals from the naive and the PPS model have the problem of being too narrow in the beginning (indicated by small average width $\overline{W}$) while simultaneously having strong under-coverage, resulting in higher Mean Interval Scores (MIS). In contrast, the calibrated PPS forecasts, which start wider and widen more slowly, have better coverage and smaller values for the MIS.

Figure 6: Observed TFR of Finland and 250 forecast paths from naive model, with median and 95% quantiles, and density in 2024 and 2070.



Nominal Coverage	**Model**	**Cov.**	**W**	**MIS**
95%	PPS (cal.)	*91%*	*0.66*	*0.98*
	PPS (non-cal.)	73%	0.77	2.38
	Naive	82%	0.73	1.59
90%	PPS (cal.)	*84%*	*0.56*	*0.87*
	PPS (non-cal.)	65%	0.64	1.71
	Naive	77%	0.61	1.17
80%	PPS (cal.)	*75%*	*0.44*	*0.74*
	PPS (non-cal.)	47%	0.50	1.26
	Naive	70%	0.48	0.87

Table 4: Evaluation of model calibration for full forecast length of 47 years using the Coverage (Cov.), the Mean Interval Score (MIS) and the average width (W) of the prediction intervals.

Forecast Period	Model	Cov.	W	MIS
1-5 Years	PPS (cal.)	*87%*	0.43	*0.72*
	PPS (non-cal.)	33%	*0.25*	3.46
	Naive	40%	0.26	2.68
6-15 Years	PPS (cal.)	*77%*	0.58	*1.85*
	PPS (non-cal.)	53%	0.50	4.25
	Naive	60%	*0.49*	3.22
16-25 Years	PPS (cal.)	*100%*	*0.69*	*0.69*
	PPS (non-cal.)	77%	0.73	2.05
	Naive	87%	0.71	0.78
26-47 Years	PPS (cal.)	*95%*	*0.73*	*0.78*
	PPS (non-cal.)	89%	1.03	1.44
	Naive	100%	0.96	0.96

Table 5: Evaluation of model calibration for different forecast lengths with a nominal coverage of 95% using the Coverage (Cov.), the Mean Interval Score (MIS) and the average width (W) of the prediction intervals.

4. Discussion

We introduced a novel approach by combining the methods of empirical prediction intervals with the scorecaster method and with quantile-mapping to produce probabilistic demographic forecasts. By incorporating empirical forecasting errors into the forecast uncertainty we provide a data-driven estimate of the forecast uncertainty. Using a scorecaster, i.e. a model of the empirical forecasting error distribution, allows for greater generalizability through smoothing and interpolation. The chosen skew-normal model does

so using only four parameters. Moreover, the calibration of forecast paths with quantile-mapping, rather than providing uncertainty intervals, offers more flexibility for downstream modeling. Our results demonstrated strong performance in the validation analyses. These findings reflect the strength and versatility of the techniques and their combined use.

We presented probabilistic forecasts of the Finnish Total Fertility Rate (TFR) from 2024 to 2070. We use two different models to forecast the TFR of Finland. The first model is a scenario-based approach that assumes that the fertility postponement, i.e. the delay of childbearing to older ages, that started in Finland in the 1970s, will slow down and eventually stop. The second scenario-based model serves as a naïve baseline and assumes that the most recently observed age-specific fertility rates will remain constant. In a next step, we calibrate the results of the postponement scenario model (PPS) using a methodology based on empirical prediction intervals and quantile-mapping. More specifically, we construct prediction intervals around the time series forecasts that are based on a model of the out-of-sample forecast errors. We then calibrate 5,000 paths of future TFR values so that the distribution of the paths matches the distribution of the modeled historical out-of-sample errors. These TFR paths can be easily integrated into further analyses, such as population projection models or economic and health planning models.

The paths of the PPS model which have been calibrated on the historical error data predict the TFR to increase until 2050 and then level off. This is a result of the assumption about the trend in fertility postponement. Using this method, the median of the TFR paths reaches a value of 1.44 in 2070 (95% PI [1.08, 1.72]). The validation analysis shows that the uncertainty around

these calibrated forecasts has a better coverage and is better calibrated in terms of width, compared to the non-calibrated paths from the PPS model (median 1.56, 95% PI [1.20, 2.03] in 2070) and the naïve baseline model (median 1.31, 95% PI [1.00, 1.71] in 2070).

Similar to our results, the latest edition of the UN World Population Prospects (UNWPP 2024) projects Finland's TFR to slightly increase in their median variant and to level off at 1.51 children by 2100, reaching a value of 1.47 by 2070. The 95% prediction interval around this median ranges from 0.89 to 2.0 children in 2070, which is wider than the 95% prediction intervals of our results. The UN produces probabilistic projections with country-specific assumptions based on the country's past experience (see United Nations, Department of Economic and Social Affairs, Population Division (2024) for detailed methodology). The UN categorizes Finland as having entered a low-fertility post-transition phase. Finland's TFR is then projected using a time-series model, "assuming that the fertility level would approach and, in the long run, fluctuate around an ultimate country-specific level" (United Nations, Department of Economic and Social Affairs, Population Division, 2024, p. 30). In addition to the the median variant, other scenarios are published to give an idea of possible future fertility developments, including high and low fertility variants with +/- 0.5 children and a freeze-rate approach.

Calibrating time series forecasts using historical data is data intensive. This need for long available time series is the main limitation of the proposed methodology of empirical prediction intervals. In addition, the scenario nature of the PPS model restricted the applicability of the model to periods

where fertility change is strongly affected by tempo effects and the main assumption of continuing but slowing down fertility postponement holds. Therefore, we extended the training data set for Finland by including data from Sweden and Norway, which have similar childbearing patterns in terms of the timing and level of fertility. We also allowed the cross-validation series to overlap to increase the number of series. However, the overlapping of the cross-validation series carries the risk of over-fitting the model of the forecast error distribution. To mitigate this, we included data from Denmark that were only used in the validation analyses and not in the calibration of the forecasts.

However, there is another approach to obtaining a forecast error distribution to derive empirical prediction intervals that does not involve forecasting historical data using the same model as for the actual forecast. Keilman and Pham (2004) introduce the use of published historical forecasts from statistical agencies and other official sources to calculate forecast errors and thus derive empirical prediction intervals. In this way, the uncertainty of expert forecasts in the past informs the uncertainty of the forecasts today, regardless of the methodology used to produce the historical forecasts. This type of empirical prediction intervals can be a valuable tool for scenario-based forecast models, where the lack of applicability to historical data limits the data availability for the training data.

The underlying assumption that allows us to use empirical forecasting errors to derive measures of forecast uncertainty is that future forecasting errors will resemble past errors. One might ask, however, why this should be the case. Alho et al. (2008) argue that "[...] if one does not believe that they

will be, it is necessary to provide arguments as to why the future is expected to be different from the past". Demographers have for a long time been aware of the distorting impact of changes in fertility timing on period fertility (Hajnal (1947); Bongaarts and Feeney (1998). While not explicitly predicting the end of fertility postponement, they illustrated how fertility rises due to slowing down fertility postponement or catching up of postponed births. This has later been referred to as "the third phase of fertility recuperation" Sobotka (2017). Further, Finnish research of the past expected that the observed fertility postponement would not go on forever: "At present it seems reasonable to assume in long-range studies that fertility will stabilize at the level prevailing at the end of the 1980s" (Auvinen, 1989, p. 54). This believe is also reflected in the "high"-scenario of the 1984 population projection of Statistic's Finland (Hämäläinen and Honkanen (1984)). Therefore, we believe that applying our scenario-based PPS model to past periods is a valid choice. We see no reason why the historic forecasting errors of the PPS model should not be used to inform the uncertainty of the current PPS forecasts.

Although the validation analyses have shown that the prediction intervals calibrated to historical data perform notably better in terms of coverage and width than the non-calibrated ones, the calibration scores are not perfect. The main problem is the under-coverage of the prediction intervals for forecasts up to 15 years ahead due to over-prediction of the TFR. Although we have taken this forecast bias into account when modeling the forecasting error distribution, the problem persists to some extend as revealed in the validation. A possible reason for these results is the lack of data, which lead

to less robust validation results, as we only have four data series available for the validation analysis. In addition, the empirical error distribution shows that the overestimation of the TFR in the first forecast years is a problem inherent to the PPS model, because the slowing of the fertility postponement is assumed to start at the first forecast year. Looking at the TFR forecasts up to 2070 together with the observed data up to 2023 (Figure 5), the model would benefit from a smoother transition between the last observed value and the first forecast year that takes the short-term trend in the latest observed years into account.

In contrast to modelling the forecasting error distribution using a score-caster, empirical prediction intervals could also be derived by taking the raw quantiles of the error distribution. However, we chose to model it using a time-varying skew-normal distribution. The resulting prediction intervals are still informed by empirical forecasting errors. However, the modeling helps to increase their generalizability in the presence of data scarcity by smoothing and extrapolation.

The results of this study show how empirical prediction intervals and quantile-mapping serve to improve the quality of probabilistic demographic forecasts. We would like to emphasize that this is a flexible methodology that can be applied to all kinds of demographic (or non-demographic) measures, regardless of the type of forecast model or outcome. In this study we have applied it to a scenario-based model for forecasting Finland's Total Fertility Rate. However, any type of forecast model can be calibrated using the presented methodology, e.g. models based on expert opinions, or simple extrapolation models. The critical component for successful application is a

long time series of historical data to which the chosen forecast model can be applied. We encourage researchers to use probabilistic forecast methods, to be transparent about their assumptions, and to calibrate and validate their results using historical data. Time has shown that demographic forecasts made by researchers in the past to the best of their knowledge have turned out to be wrong. We see no reason why current forecasts should be any different. It is intuitive to us, therefore, to use the knowledge of past forecast errors to our advantage and let it inform our forecasts of today.

Author contributions

R.D.: Conceptualization, Methodology, Formal analysis, Software, Visualization, Writing - Original Draft. J.S.: Conceptualization, Methodology, Software, Supervision, Writing - Review & Editing. J.H.: Methodology, Software, Writing - Review & Editing. M.M.: Conceptualization, Methodology, Supervision, Writing - Review & Editing, Funding acquisition.

Acknowledgements

The authors would like to thank Nico Keilman, Rob Hyndman and Marie-Pier Bergeron-Boucher for discussions on empirical uncertainty quantification. R.D. gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS).

Funding sources

R.D. was supported by the Finnish Centre for Pensions (ETK2023031).

J.H. was supported by the Strategic Research Council (SRC) of the Academy of Finland, FLUX consortium (Family Formation in Flux—Causes, Consequences, and Possible Futures), decision numbers 345130 and 345131, and by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101019329).

M.M. was supported by the Strategic Research Council (SRC), FLUX consortium, decision numbers 345130 and 345131; by the National Institute on Aging (R01AG075208); by grants to the Max Planck – University of Helsinki Center from the Max Planck Society (Decision number 5714240218), Jane and Aatos Erkko Foundation, Faculty of Social Sciences at the University of Helsinki, and Cities of Helsinki, Vantaa and Espoo; and the European Union (ERC Synergy, BIOSFER, 101071773). Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Alho, J.M., Cruijsen, H., Keilman, N., 2008. Uncertain demographics and fiscal sustainability. Cambridge University Press, Cambridge. chapter Empirically based sprecification of forecast uncertainty. pp. 34–54.

Alho, J.M., Spencer, B.D., 2005. Statistical demography and forecasting.

Springer. volume 2016. chapter Uncertainty in Demographic Forecasts: Concepts, Issues, and Evidence. pp. 226–268.

Angelopoulos, A., Candes, E., Tibshirani, R.J., 2024. Conformal pid control for time series prediction. Advances in neural information processing systems 36.

Auvinen, R., 1989. Finland's low fertility and the desired recovery. Finnish Yearbook of Population Research , 53–59.

Bongaarts, J., Feeney, G., 1998. On the quantum and tempo of fertility. Population and development review , 271–291.

Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., 2021. Evaluating epidemic forecasts in an interval format. PLoS computational biology 17, e1008618.

Cannon, A.J., 2018. Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables. Climate dynamics 50, 31–49.

Cohen, J.E., 1986. Population forecasts and confidence intervals for Sweden: a comparison of model-based and empirical approaches. Demography 23, 105–126. doi:10.2307/2061412. publisher: Duke University Press.

Fontana, M., Zeni, G., Vantini, S., 2023. Conformal prediction: a unified review of theory and new challenges. Bernoulli 29, 1–23.

Gammerman, A., Vovk, V., Vapnik, V., 1998. Learning by transduction. arXiv:1301.7375.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378. doi:10.1198/016214506000001437. publisher: Informa UK Limited.

Goldstein, J.R., 2006. How late can first births be postponed? some illustrative population-level calculations. Vienna Yearbook of Population Research , 153–165.

Hajnal, J., 1947. The analysis of birth statistics in the light of the recent international recovery of the birth-rate. Population Studies 1, 137–164.

Hellstrand, J., Nisén, J., Myrskylä, M., 2020. All-time low period fertility in finland: Demographic drivers, tempo effects, and cohort implications. Population Studies 74, 315–329.

Hämäläinen, H., Honkanen, O., 1984. VÄESTÖENNUSTEET LÄÄNEITTÄIN KUNTAMUODON MUKAAN 1980 - 2000. Technical Report. Statistics Finland.

Keilman, N., 2018. Probabilistic demographic forecasts. Vienna Yearbook of Population Research 16, 25–36. URL: https://www.jstor.org/stable/26670702.

Keilman, N., Pham, D.Q., 2004. Time series based errors and empirical errors in fertility forecasts in the nordic countries. International Statistical Review 72, 5–18.

Lee, R.D., 1998. Probabilistic approaches to population forecasting. Population and Development Review 24, 156–190. URL: http://www.jstor.org/stable/2808055.

Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria), 2024. Human fertility database. URL: https://humanfertility.org/.

Nisén, J., Hellstrand, J.I.S., Martikainen, P., Myrskylä, M., 2020. Hedelmällisyys ja siihen vaikuttavat tekijät suomessa lähivuosikymmeninä. Yhteiskuntapolitiikka 85, 358–369.

Qian, W., Chang, H.H., 2021. Projecting health impacts of future temperature: a comparison of quantile-mapping bias-correction methods. International journal of environmental research and public health 18, 1992.

Rothman, K.J., Wise, L.A., Sørensen, H.T., Riis, A.H., Mikkelsen, E.M., Hatch, E.E., 2013. Volitional determinants and age-related decline in fecundability: a general population prospective cohort study in denmark. Fertility and sterility 99, 1958–1964.

Shafer, G., Vovk, V., 2008. A tutorial on conformal prediction. Journal of Machine Learning Research 9.

Smith, S.K., Sincich, T., 1988. Stability over time in the distribution of population forecast errors. Demography 25, 461–474. doi:10.2307/2061544. publisher: Duke University Press.

Sobotka, T., 2017. Post-transitional fertility: childbearing postponement and the shift to low and unstable fertility levels. Technical Report. Vienna Institute of Demography Working Papers.

Statistics Finland, 2024. Official statistics of finland (osf): Births [online publication]. URL: https://stat.fi/en/publication/cln3ad0zibipb0cutxy95o145.

Stoto, M.A., 1983. The accuracy of population projections. Journal of the American Statistical Association 78, 13–20. doi:10.1080/01621459.1983.10477916. publisher: Informa UK Limited.

United Nations, Department of Economic and Social Affairs, Population Division, 2024. World popoulation prospects 2024: Methodology of the united nations population estimates and projections.

Williams, W.H., Goodman, M.L., 1971. A simple method for the construction of empirical confidence limits for economic forecasts. Journal of the American Statistical Association 66, 752–754. doi:10.2307/2284223. publisher: Informa UK Limited.