Do Top Methods for Cohort Fertility Completion also Perform Best when Forecasting Period Fertility?

Christina Bohk-Ewald *[1], Peng Li[1], and Mikko Myrskylä[1]

[1]Max Planck Institute for Demographic Research, Rostock, Germany

Submission to the PAA Annual Meeting in 2019

Abstract

Forecasters could easily be overwhelmed by the plethora of methods for predicting cohort and period fertility. To shed light on which method to choose when predicting fertility we validate their forecast accuracy based on data of the Human Fertility Database and the UN World Population Prospects. We found for cohort fertility completion that forecast accuracy does not necessarily improve with method complexity. That is, the baseline Freeze rates (that holds latest fertility rates constant) belongs to the top methods in all world regions, and it is only outperformed by two extrapolation methods and two Bayesian approaches. As a follow-up we investigate if this finding for cohort fertility completion also holds true for period fertility forecasting, adopting the same study design. We introduce the *forecast performance spectrum* as a visualization tool that supports forecasters in their decision for the most appropriate method(s) in various fertility settings by joining all evaluation results.

*BohkEwald@demogr.mpg.de

1 Introduction

How many children will be born in the future?—is an urgent question for decision makers around the globe for planning e.g. local infrastructure and national welfare systems, particularly in the face of ongoing demographic change that is driven by small numbers of births and increasing shares of the elderly. Despite a plethora of methods that exist to predict fertility it is unclear which method(s) perform best in various forecast situations, which can be specified by different levels, patterns over age, and trends over time of cohort and period fertility. We therefore systematically validate and rank all available forecast methods with fertility data of the Human Fertility Database (2016) and the UN World Population Prospects (2017) in order to infer how robustly they produce accurate forecasts of cohort and period fertility across many different fertility settings.

Since 1940 more than 20 methods and hundreds of variants for them were developed to forecast fertility. Some of these methods are broadly applicable to forecast cohort and period fertility, whereas other methods are designed to forecast either cohort or period fertility.

Cohort fertility quantifies real lifetime reproduction of women from an actual birth cohort, whereas period fertility quantifies artificial lifetime reproduction of women from a synthetic birth cohort in a calendar year. For example, completed fertility measures the average number of children per woman of the same birth cohort over their entire reproductive lifespan, whereas total fertility rate measures average number of children born per woman of different birth cohorts but in the same calendar year.

Indicators of cohort and period fertility both have their advantages and disadvantages. While completed fertility is unaffected by timing effects (e.g. of a delay in childbearing to older maternal ages) it takes long waiting time (about 50 years) until it provides an estimate of real lifetime reproduction for a birth cohort. On the contrary, although period fertility is affected by timing effects it does provide immediate estimates of artificial lifetime reproduction for each calendar year. Together, completed fertility and total fertility rate give a comprehensive picture of lifetime reproduction.

Forecasts of cohort and period fertility both have their own right and value as they provide a valuable foundation for decision makers for planning local infrastructure (e.g. demand for kindergarten, primary, secondary, and tertiary education) and national welfare systems (e.g. pension entitlement).

Given more than 20 methods to forecast fertility—and hundreds of variants for them due to different parametrization—a forecaster could be easily challenged to choose the best method from this huge basket of optional methods. We therefore validate the overall forecast performance of all fertility forecast methods with as many data as possible in order to infer what are the top methods for cohort fertility completion and forecasting total fertility rate. More specifically, we systematically validate and compare the forecast performance of all methods—in terms of forecast accuracy and uncertainty estimates (where applicable)—based on all available fertility data of the Human Fertility Database (2016) and the UN World Population Prospects (2017).

Part one of this validation study focuses on methods of cohort fertility completion (Bohk-Ewald et al., 2018), part two of this validation study focuses on methods that forecast period fertility. For cohort fertility completion we find that forecast accuracy does not necessarily improve with method complexity. That is, the baseline method Freeze rates (that holds latest fertility rates constant) belongs to the top methods in all world regions across the globe, and it is only outperformed by two simple extrapolation methods and two hierarchical Bayesian approaches. In part two of this validation study we show if this finding for cohort fertility completion also holds true for period fertility forecasting adopting the same study design.

Together, the two parts of our comprehensive validation study—cohort fertility completion and

forecasting total fertility rate—constitute / represent the dimensions of a *forecast performance spectrum* that we introduce to visually compare the methods in different forecast settings at a glance. This *forecast performance spectrum* provides a multidimensional comparison of the forecast methods across various settings (and measurements) and is as such highly useful to derive recommendations for choosing the most appropriate method(s) from the large basket when completing cohort fertility and forecasting total fertility rate.

In section 2 we describe the data used, forecast methods applied, and the design of the validation study adopted, in section 3 we present tentative results for forecasts of period fertility, in section 4 we summarize and discuss the main findings, and in section 5 we finally draw conclusions.

2 Data and Methods

2.1 Fertility data

We base our validation study on fertility rates by single years of age and calendar year in 29 countries of the Human Fertility Database (HFD; 2016) and 201 countries of the UN World Population Prospects (UNWPP; 2017). In the HFD-based analysis, we include fertility data of the constituents Eastern Germany, Western Germany, Northern Ireland, Scotland, England and Wales and, consequently, excluded Germany and the UK. In the UNWPP-based analysis, we interpolate the five year data estimates provided by the UNWPP using the R function `spline.smooth` and the quadratic optimization method (Michalski et al., 2018). Tables 4 through 7 in Appendix A provide details on the data coverage by calendar year for the HFD and the UNWPP, respectively.

2.2 Forecast methods

We use five method types to classify the 20 methods to forecast fertility. That is, we have (1) the baseline method *Freeze rates* that simply holds constant fertility rates at their latest observed level, (2) Parametric curve fitting methods (PARs), (3) Extrapolation methods (EMs), (4) Bayesian approaches (BAs), and (5) Fertility context specific methods (CONs).

Parametric curve fitting methods aim to mathematically describe a universal pattern over age of fertility. Extrapolation methods rely on trends over time to forecast fertility—that is, they fit a model to observed fertility of previous years and extrapolate the parameter estimates into years to come. Hierarchical Bayesian approaches forecast fertility in a country of interest with information about levels, patterns over age, and trends over time of fertility in other (reference) countries. Fertility context specific methods are applicable only to forecast fertility in certain contexts like fertility postponement.

In our validation study of methods for cohort fertility completion (Bohk-Ewald et al., 2018), we consider one baseline method, ten Parametric curve fitting methods (Hadwiger, 1940; Coale and McNeil, 1972; Coale and Trussell, 1974; Brass, 1974; Evans, 1986; Chandola et al., 1999; Schmertmann, 2003; Peristera and Kostaki, 2007; Myrskylä and Goldstein, 2013), six Extrapolation methods (Willekens and Baydar, 1984; de Beer, 1985; Lee, 1993; Hyndman and Ullah, 2007; Cheng and Lin, 2010; Myrskylä et al., 2013), two hierarchical Bayesian approaches (Schmertmann et al., 2014; Ševčíková et al., 2016), and one Fertility context specific method (Li and Wu, 2003). A detailed description of these methods is given in the Supporting Information, section one, of Bohk-Ewald et al. (2018).

Wherever it is required and at all possible we slightly adapt the forecast procedure of some of these forecast methods so that they are also applicable to predict period fertility. Specifically, we fit the Parametric curve fitting methods of Hadwiger (1940); Coale and McNeil (1972); Coale and Trussell (1974); Evans (1986); Chandola et al. (1999); Schmertmann (2003); Peristera and Kostaki (2007) to age schedules of period (and not cohort) fertility in the observation window, and

extrapolate their parameter estimates in order to predict period fertility. Note that we cannot use the PAR of Brass (1974), the EM of Cheng and Lin (2010), the BA of Schmertmann et al. (2014), and the CON of Li and Wu (2003) to forecast period fertility.

The comparison of methods to forecast period fertility relies on 16 methods, whereas the comparison of methods for cohort fertility completion is based on 20 methods. Table 1 lists the forecast methods and specifies their applicability and method type classification when forecasting completed fertility (CF) and total fertility rate (TFR).

Table 1: Applicability and classification of forecast methods.

Method	Applicable to forecast CF	TFR	Method type when forecasting CF	TFR
Baseline method:				
Freeze rates	YES	YES	Baseline	Baseline
Designed as Parametric curve fitting method (PAR):				
Hadwiger (1940)	YES, as designed.	YES, modified.	PAR	PAR + EM
Coale and McNeil (1972)	YES, as designed.	YES, modified.	PAR	PAR + EM
Coale and Trussell (1974)	YES, as designed.	YES, modified.	PAR + EM	PAR + EM
Brass (1974)	YES, as designed.	NO.	PAR + EM + BA	--
Evans (1986)	YES, as designed.	YES, modified.	PAR + EM	PAR + EM
Chandola et al. (1999)	YES, as designed.	YES, modified.	PAR	PAR + EM
Schmertmann (2003)	YES, as designed.	YES, modified.	PAR + EM	PAR + EM
Peristera and Kostaki 1 (2007)	YES, as designed.	YES, modified.	PAR	PAR + EM
Peristera and Kostaki 2 (2007)	YES, as designed.	YES, modified.	PAR	PAR + EM
Myrskylä and Goldstein (2013)	YES, as designed.	YES, modified.	PAR + EM	PAR + EM
Designed as Extrapolation method (EM):				
Willekens and Baydar (1984)	YES, as designed.	YES, as designed.	EM	EM
de Beer (1985)	YES, as designed.	YES, modified.	EM	EM
Lee (1993)	YES, with modification.	YES, as designed.	EM	EM
Hyndman and Ullah (2007)	YES, modified.	YES, as designed.	EM	EM
Cheng and Lin (2010)	YES, as designed.	NO	EM	--
Myrskylä et al. (2013)	YES, as designed.	YES, modified.	EM	EM
Designed as Bayesian approach (BA):				
Schmertmann et al. (2014)	YES, as designed.	NO	BA	--
Ševcíková et al. (2016)	YES, modified.	YES, as designed.	BA	BA
Designed as Fertility context specific method (CON):				
Li and Wu (2003)	YES, as designed.	NO	CON	--

2.3 Validation methods

To assess the forecast performance of each method, we—depending on the specific application—complete cohort fertility or forecast total fertility rate for several calendar years, and compare these fertility predictions with their corresponding (true) observations; applying the typical procedure of a validating forecast. With each method, we conduct as many validating forecasts as possible using fertility data of all available countries, birth cohorts, and calendar years of the HFD (2016) and the UNWPP (2017). We then measure the forecast performance of each method with forecast errors. For example, we use the absolute percentage error (APE) to quantify forecast accuracy

$$APE = \frac{\mid F - O \mid}{O} \cdot 100 \tag{1}$$

The APE gives the percentage of the absolute deviation between forecasted (F) and observed (O) fertility from the true value (O). Based on these single APEs—characterized by country, age, birth cohort, and calendar year—we use different metrics to summarize the overall accuracy across all validating forecasts for each method. For example we use a statistical test of stochastic dominance (i.e. the Kolmogorov-Smirnov test statistic (e.g. Levy, 1992; Heathcote et al., 2010; Barrett and

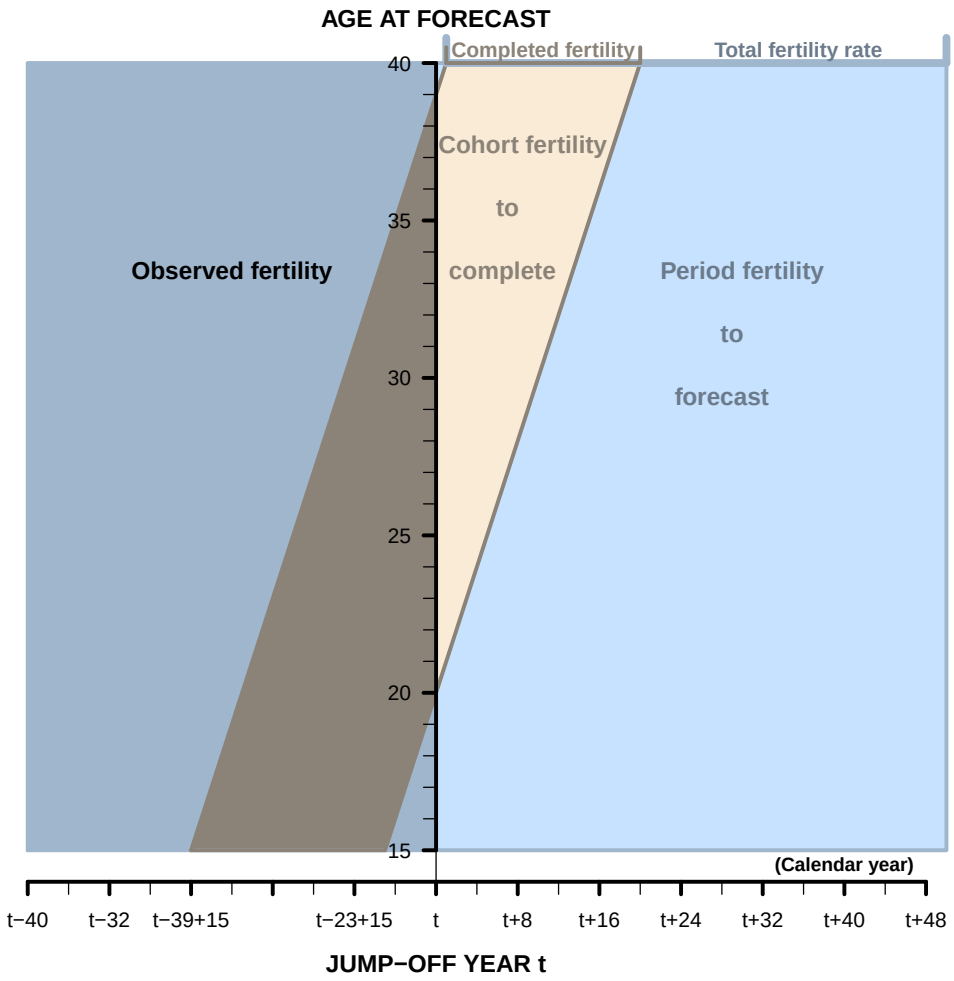

Figure 1: Procedure for cohort fertility completion (beige) and period fertility forecasting (blue).

Donald, 2003; Marsaglia et al., 2003; Massey, 1951; Sheskin, 2011)) and descriptive statistics like the mean, the median, and thresholds below which certain percentages of forecast errors fall. A detailed description of the validation procedure and error metrics can be found in Bohk-Ewald et al. (2018).

To ensure the validity of our evaluation study we make the comparison across methods as fair as possible—that is, we assess forecast accuracy of only the best variant (or parameterization) of each method with error datasets that are equal for all methods in terms of included countries, ages, birth cohorts, and calendar years. To identify the best variant for each method we adopt a similar validation strategy as for the main method comparison; Table 8 in Appendix B lists the best variant for each method to forecast period fertility. Since we assess the forecast performance of the methods across many different kinds of fertility levels, patterns over age, and trends over time that were experienced in countries worldwide, the external validity of our results is relatively high. Finally, to make this validation study replicable we have deposited the R code on GitHub at `https://github.com/fertility-forecasting/validate-forecast-methods`.

3 Tentative results for forecasts of period fertility

We compare forecast accuracy across methods for predicting total fertility rate based on seven (APE-)metrics: (1) the two-sample one-sided Kolmogorov-Smirnov test of stochastic dominance ($p = 0.05$), (2–5) the thresholds below which 50, 80, 90, and 95% of APEs fall, (6) the mean of APEs, and (7) the root mean square error (RMSE) of APEs. Tables 2 and 3 show the ranking of methods according to these seven metrics for forecasts up to 25 years ahead based on 29 countries of the HFD (2016) and 201 countries of the UNWPP (2017), respectively, from years 1990 and later.

Table 2: 29 HFD countries. Based on forecasts up to 25 years ahead for years 1990 and later.

		Threshold APE:					
Method	**KS**	**50%**	**80%**	**90%**	**95%**	**Mean**	**RMSE**
1. Freeze rates	11 (1)	5.87 (1)	13.19 (1)	19.8 (4)	27.05 (4)	8.89 (2)	13.36 (4)
2. Peristera-Kostaki2	11 (1)	5.92 (2)	13.28 (2)	19.87 (5)	26.96 (3)	8.92 (4)	13.42 (5)
3. Chandola et al.	9 (3)	5.97 (3)	13.42 (4)	19.91 (6)	27.17 (6)	9 (5)	13.54 (7)
4. Hadwiger	8 (4)	6.13 (5)	13.31 (3)	19.59 (3)	26.49 (2)	8.89 (1)	13.04 (2)
5. Ševcíková et al.	8 (4)	6.11 (4)	13.98 (7)	19.53 (2)	26.28 (1)	8.9 (3)	12.88 (1)
6. Peristera-Kostaki	8 (4)	6.27 (7)	13.43 (5)	20.07 (7)	27.73 (8)	9.06 (6)	13.31 (3)
7. Coale-McNeil	8 (4)	6.36 (8)	13.52 (6)	19.37 (1)	27.39 (7)	9.19 (7)	13.65 (8)
8. de Beer	8 (4)	6.16 (6)	14.18 (8)	20.56 (8)	27.09 (5)	9.2 (8)	13.53 (6)
9. Lee	6 (9)	7.21 (9)	17.73 (9)	26.14 (10)	35.51 (10)	11.75 (10)	24.09 (13)
10. Myrskylä et al.	5 (10)	7.5 (10)	17.74 (10)	27.56 (11)	38.61 (11)	11.34 (9)	16.42 (10)
11. Hyndman-Ullah	3 (11)	9.49 (11)	23.35 (12)	35.23 (12)	52.21 (12)	15.22 (12)	22.97 (11)
12. Evans	2 (12)	9.58 (12)	18.86 (11)	24.9 (9)	31.69 (9)	12.11 (11)	15.78 (9)
13. Schmertmann	2 (12)	9.93 (13)	25.62 (13)	40.27 (13)	54.75 (13)	16.06 (13)	23.42 (12)
14. Myrskylä-Goldstein	1 (14)	10.17 (14)	26.83 (14)	46.79 (14)	62.55 (15)	17.98 (14)	28.82 (14)
15. Coale-Trussell	0 (15)	33.03 (16)	51.36 (15)	55.39 (15)	57.2 (14)	33.32 (15)	37.25 (15)
16. Willekens-Baydar	0 (15)	18.25 (15)	58.45 (16)	110.55 (16)	234.84 (16)	49.54 (16)	102.15 (16)
Inversions	0	4	4	16	16	7	18

Table 3: 201 UN countries. Based on forecasts up to 25 years ahead for years 1990 and later.

		Threshold APE:					
Method	**KS**	**50%**	**80%**	**90%**	**95%**	**Mean**	**RMSE**
1. Myrskylä et al.	13 (1)	4.81 (3)	15.79 (2)	25.8 (2)	36.09 (2)	9.62 (1)	15.9 (1)
2. Ševcíková et al.	11 (2)	5.68 (5)	19.06 (5)	30.31 (5)	42.41 (4)	11.29 (3)	17.99 (2)
3. Hyndman-Ullah	10 (3)	5.6 (4)	17.79 (4)	28.82 (3)	41.48 (3)	11.75 (4)	25.99 (10)
4. Freeze rates	8 (4)	8.75 (6)	24.52 (7)	38.38 (8)	52.4 (7)	15.36 (6)	23.97 (3)
5. Peristera-Kostaki2	7 (5)	8.95 (7)	24.81 (9)	38.75 (11)	52.94 (11)	15.56 (7)	24.2 (5)
6. Chandola et al.	6 (6)	9.05 (8)	24.86 (10)	38.48 (9)	52.45 (8)	15.61 (8)	24.12 (4)
7. Lee	4 (7)	3.04 (1)	12.85 (1)	22.48 (1)	34.03 (1)	11.05 (2)	44.69 (15)
8. Peristera-Kostaki	4 (7)	9.39 (10)	25.15 (12)	38.98 (12)	52.84 (10)	15.91 (10)	24.39 (7)
9. de Beer	3 (9)	3.23 (2)	15.82 (3)	29.8 (4)	48.41 (5)	11.85 (5)	29.35 (11)
10. Hadwiger	3 (9)	9.24 (9)	24.91 (11)	38.56 (10)	52.49 (9)	15.76 (9)	24.22 (6)
11. Evans	2 (11)	10.03 (12)	24.76 (8)	38.33 (7)	53.24 (12)	16.2 (12)	24.68 (9)
12. Coale-McNeil	1 (12)	9.53 (11)	25.45 (13)	39.28 (13)	53.27 (13)	16.09 (11)	24.53 (8)
13. Myrskylä-Goldstein	0 (13)	10.49 (13)	20.82 (6)	33.08 (6)	50.1 (6)	16.88 (13)	32.95 (13)
14. Schmertmann	0 (13)	13.96 (14)	33.68 (14)	48.14 (14)	63.93 (15)	20.98 (14)	30.33 (12)
15. Coale-Trussell	0 (13)	33.14 (16)	45.17 (15)	52.74 (15)	58.55 (14)	34 (15)	36.77 (14)
16. Willekens-Baydar	0 (13)	16.88 (15)	88.93 (16)	288.52 (16)	501.99 (16)	84.75 (16)	188.49 (16)
Inversions	0	17	25	27	23	11	22

The overall ranking in column one is based primarily on the Kolmogorov-Smirnov test statistic, possible ties are broken with the mean APE. According to this overall ranking the top methods differ between the 29 HFD countries and the 201 UN countries. While only the baseline method Freeze rates and the Bayesian approach of Ševcíková et al. (2016) are among the top five methods in the overall ranking of both datasets, they are complemented by the complex parametric curve fitting methods of Peristera and Kostaki (2007); Chandola et al. (1999); Hadwiger (1940) in the HFD dataset and by the simple extrapolation methods of Myrskylä et al. (2013); Hyndman and Ullah (2007); Lee (1993) in the UNWPP dataset.

The two datasets also differ in the magnitude of errors that are on average larger for the 201 UN countries than for the 29 HFD countries. The rankings also appear to be heterogeneous across the seven metrics in both datasets. For example, in the UN-based evaluation, Freeze rates is on position four according to the overall ranking in column one and on positions six, seven, and eight according to the mean APE and other % threshold APEs. This heterogeneity in the rankings across metrics is even more pronounced for the extrapolation method of Lee (1993) that is on position seven according to the overall ranking and on top positions one and two according to the mean APE and other % threshold APEs. Such heterogeneous rankings indicate that the forecast errors of one method are not *consistently* smaller or larger than those of other methods. On the contrary, these heterogeneous rankings indicate that there are crossovers in the cumulative distributions of APEs between methods—that is, while smaller errors (lower quartile) might be smaller for one method than for other method(s), its larger errors (upper quartile) might also be larger than those of other method(s) and vice versa.

4 Summary and Discussion

A ranking of the many existing methods to predict fertility is essential for forecasters and decision makers alike in order to produce highly accurate fertility forecasts in many different fertility settings and to use them as a reliable foundation for planning e.g. local infrastructure and national welfare systems.

To provide sound recommendations for choosing the best method(s) to predict fertility (in any context) we systematically validated their overall forecast accuracy based on all available fertility data (by country, single years of age, birth cohort, and calendar year) of the Human Fertility Database (2016) and the UN World Population Prospects (2017).

Specifically, we tested the forecast performance of the baseline method Freeze rates (that holds latest fertility rates constant), ten Parametric curve fitting methods (Hadwiger, 1940; Coale and McNeil, 1972; Coale and Trussell, 1974; Brass, 1974; Evans, 1986; Chandola et al., 1999; Schmertmann, 2003; Peristera and Kostaki, 2007; Myrskylä and Goldstein, 2013), six Extrapolation methods (Willekens and Baydar, 1984; de Beer, 1985; Lee, 1993; Hyndman and Ullah, 2007; Cheng and Lin, 2010; Myrskylä et al., 2013), two hierarchical Bayesian approaches (Schmertmann et al., 2014; Ševčíková et al., 2016), and one Fertility context specific method (Li and Wu, 2003). All of these 20 methods are applicable to complete cohort fertility, 16 of them are also applicable to forecast period fertility (although some of them needed to be modified to be applicable for this additional task).

Bohk-Ewald et al. (2018) found for cohort fertility completion that forecast accuracy does not necessarily improve with method complexity. That is, the baseline method Freeze rates belongs to the top methods in all world regions across the globe, and it is only outperformed by two simple extrapolation methods and two hierarchical Bayesian approaches. In addition, the methodologically more complex and computing-intensive BAs do not consistently complete cohort fertility more accurately than the two simple EMs. These findings are consistent for the 29 HFD (2016) countries, for the 201 UNWPP (2017) countries, and for six world regions.

In this paper we found for forecasting period fertility that only the baseline method Freeze rates and the Bayesian approach of Ševcíková et al. (2016) belong to the top five methods in both datasets, and that they are complemented by the complex parametric curve fitting methods of Peristera and Kostaki (2007); Chandola et al. (1999); Hadwiger (1940) in the HFD dataset and by the simple extrapolation methods of Myrskylä et al. (2013); Hyndman and Ullah (2007); Lee (1993) in the UNWPP dataset.

These differences in the ranking of methods in both datasets can perhaps be explained with the fertility settings they cover. While the fertility levels are mostly below replacement and relatively stable since the 1990s in many HFD countries, fertility levels range from high to low and are relatively unstable (including strong fertility declines) in the 201 UN countries. Given these mostly low and stable fertility settings in the HFD, it is not surprising that Freeze rates ranks number one and that the variants of the complex parametric curve fitting methods that freeze recent levels and trends (see Table 8) also yield top forecast results. On the contrary, given the diverse fertility levels and trends in the 201 UN countries, it is reasonable that simple extrapolation methods better capture changes in fertility levels and trends than Freeze rates; even though the latter method is still on rank five. The Bayesian approach of Ševcíková et al. (2016) confirms this finding from a different angle as it ranks in both datasets among the top five methods. That is, borrowing information from other countries seems to be the key for forecasting fertility throughout different settings, extrapolating past trends seems to be somewhat more useful in unstable settings (rank two in 201 UN countries) than in stable settings (rank five in 29 HFD countries).

The *forecast performance spectrum* displayed in Figure 2 provides even more insights into the performance of the methods across various fertility settings when completing cohort fertility (upper half) and when forecasting total fertility rate (lower half). Across all forecast methods, the *forecast performance spectrum* compares forecast accuracy (here: based on median APE) for the six world regions—Africa, Oceania, Asia, Latin America and the Caribbean (LAC), Northern America (NA), and Europe—ordered from highest fertility in Africa to lowest fertility in Europe. The median APEs

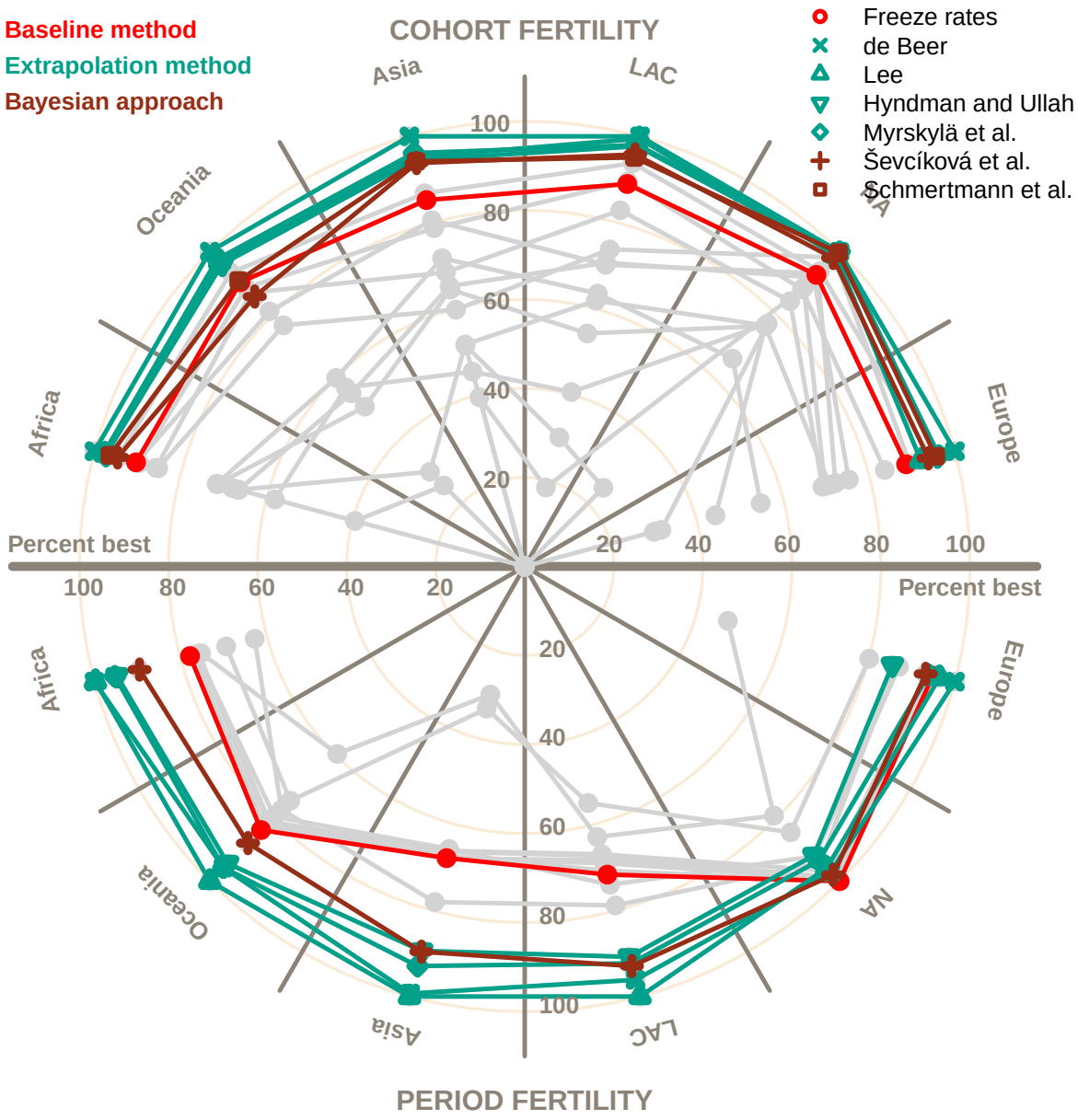

Figure 2: *Forecast performance spectrum* of methods, based on median APE, to compare forecast accuracy across methods when completing cohort fertility (upper half) and when predicting total fertility rate (lower half) for different fertility settings (represented by six world regions). The 50% threshold APEs are rescaled for each world region so that they range between minimum performance (0%, center) and maximum performance (100%, furthest circle) achieved of all methods per world region. Testing data are the UN world regions, years 1990 and later.

are rescaled for each world region so that they range between minimum performance (0%, innermost circle) and maximum performance (100%, furthest circle) achieved of all methods per world region. Other rescaling techniques are possible but we chose this one for the sake of comparability of forecast performance across methods, fertility settings (here: represented by world regions), and later also error metrics.

Splitting the rankings for the 201 UN countries into six rankings by world region, based on the median APE in the *forecast performance spectrum* displayed in Figure 2, shows that Extrapolation methods (green) have a strong advantage over Bayesian approaches (brown) and Freeze rates (red) in high fertility settings such as in Africa, Oceania, Asia, and Latin America and the Caribbean (LAC). This strong advantage of Extrapolation methods (green) diminishs in lower fertility settings such as in Northern America (NA) and Europe, where Extrapolation methods, Bayesian approaches, and Freeze rates seem to forecast fertility almost equally accurately. Although this general finding applies to both contexts, cohort fertility completion and total fertility rate prediction, it appears to be more pronounced in *period* fertility settings—that is, the median APE of Freeze rates is comparatively large when forecasting the total fertility rate in Africa, Oceania, Asia, and LAC. This finding indicates that the small errors of Freeze rates (below median APE) are comparatively large in settings with strong fertility declines from high to low levels (Africa, Oceania, Asia, and Latin America and the Caribbean), and comparatively low in settings with stable fertility at low levels (Northern America and Europe). It also indicates that the development of period fertility might be less stable than that of cohort fertility, which is reasonable as the total fertility rate is affected by timing effects in contrast to completed fertility.

In the full paper we will prepare the *forecast performance spectrum* for other error metrics and we will use forecast accuracy as well as empirical coverage of prediction intervals to comprehensively evaluate and compare forecast performance of methods when completing cohort fertility and forecasting period fertility.

5 Conclusion

Based on our comprehensive evaluation study we can start to answer the question posed in the title—that is, do the top methods for cohort fertility completion also perform best when forecasting period fertility? The simple answer is *yes*, the more detailed answer complements this simple *yes* with a more differentiating *but it depends*. That is, Freeze rates is among the top five methods when completing cohort fertility and when forecasting period fertility, but it appears to forecast total fertility rate substantially less accurate than it completes cohort fertility in high and unstable fertility settings such as in Africa, Oceania, Asia, and Latin America and the Caribbean. In addition, we introduced the *forecast performance spectrum* as a supportive visualization tool that joins all evaluation results in order to enable a forecaster to effectively select the most appropriate method(s) for cohort fertility completion and total fertility rate prediction at a glance.

References

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica 71*(1), 71–104.

Bohk-Ewald, C., P. Li, and M. Myrskylä (2018). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences 115*(37), 9187–9192.

Brass, W. (1974). Perspectives in Population Prediction: Illustrated by the Statistics of England and Wales. *Journal of the Royal Statistical Society. Series A (General) 137*(4), 532–583.

Chandola, T., D. A. Coleman, and R. W. Hiorns (1999). Recent European fertility patterns: Fitting curves to "'distorted"' distributions. *Population Studies 53*(3), 317–329.

Cheng, P. R. and E. S. Lin (2010). Completing incomplete cohort fertility schedules. *Demographic Research 23*(9), 223–256.

Coale, A. and D. McNeil (1972). The distribution by age at first marriage in a female cohort. *Journal of the American Statistical Association 67*(340), 743–749.

Coale, A. J. and T. J. Trussell (1974). Model Fertility Schedules: Variations in The Age Structure of Childbearing in Human Populations. *Population Index 40*(2), 185–258.

de Beer, J. (1985). A Time Series Model for Cohort Data. *Journal of the American Statistical Association 80*(391), 525–530.

Evans, M. D. R. (1986). American Fertility Patterns: A Comparison of White and Nonwhite Cohorts Born 1903-56. *Population and Development Review 12*(2), 267–293.

Hadwiger, H. (1940). Eine analytische Reproduktionsfunktion für biologische Gesamtheiten. *Scandinavian Actuarial Journal 1940*(3–4), 101–113.

Heathcote, A., S. Brown, E. Wagenmakers, and A. Eidels (2010). Distribution-free tests of stochastic dominance for small samples. *Journal of Mathematical Psychology 54*(5), 454–463.

Human Fertility Database (2016). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at `www.humanfertility.org`. Data downloaded on April 07, 2016.

Hyndman, R. J. and M. S. Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis 51*(10), 4942–4956.

Lee, R. D. (1993). Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level. *International Journal of Forecasting 9*, 187–202.

Levy, H. (1992). Stochastic Dominance and Expected Utility: Survey and Analysis. *Management Science 38*(4), 555–593.

Li, N. and Z. Wu (2003). Forecasting cohort incomplete fertility: A method and an application. *Population Studies 57*(3), 303–320.

Marsaglia, G., W. W. Tsang, and J. Wang (2003). Evaluating Kolmogorov's Distribution. *Journal of Statistical Software 8*(1), 1–4.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association 46*(253), 68–78.

Michalski, A. I., P. Grigoriev, and V. P. Gorlishchev (2018). R programs for splitting abridged fertility data into a fine grid of ages using the quadratic optimization method. MPIDR Technical Report TR 2018-002, Max Planck Institute for Demographic Research, Germany. Available at https://www.demogr.mpg.de/papers/technicalreports/tr-2018-002.pdf.

Myrskylä, M. and J. R. Goldstein (2013). Probabilistic Forecasting Using Stochastic Diffusion Models, With Applications to Cohort Processes of Marriage and Fertility. *Demography 50*, 237–260.

Myrskylä, M., J. R. Goldstein, and Y. A. Cheng (2013). New Cohort Fertility Forecasts for the Developed World: Rises, Falls, and Reversals. *Population and Development Review 39*(1), 31–56.

Peristera, P. and A. Kostaki (2007). Modeling fertility in modern populations. *Demographic Research 16*(6), 141–194.

Schmertmann, C., E. Zagheni, J. R. Goldstein, and M. Myrskylä (2014). Bayesian Forecasting of Cohort Fertility. *Journal of the American Statistical Association 109*(506), 500–513.

Schmertmann, C. P. (2003). A system of model fertility schedules with graphically intuitive parameters. *Demographic Research 9*(5), 81–110.

Sheskin, D. J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press.

United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Prospects: The 2017 Revision, DVD Edition. Available at https://esa.un.org/unpd/wpp/DVD/Files/1_Indicators%20(Standard)/EXCEL_FILES/2_Fertility/WPP2017_FERT_F07_AGE_SPECIFIC_FERTILITY.xlsx. Data downloaded on March 01, 2018.

Ševčíková, H., N. Li, V. Kantorová, P. Gerland, and A. E. Raftery (2016). Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections. In R. Schoen (Ed.), *Dynamic Demographic Analysis*, pp. 285–310. Springer International Publishing.

Willekens, F. and N. Baydar (1984). Age-period-cohort models for forecasting fertility, Working paper no.45, Voorburg, July 1984.

A Data coverage in HFD and UNWPP

Table 4: Data coverage in the HFD and UNWPP. Part I

Country	Calendar years
Human Fertility Database:	
Austria	1951-2014
Belarus	1964-2014
Bulgaria	1947-2009
Canada	1921-2011
Czech Republic	1950-2014
Estonia	1959-2013
Finland	1939-2012
France	1946-2013
Germany	1956-2012
Eastern Germany	1956-2012
Western Germany	1956-2012
Hungary	1950-2009
Iceland	1963-2010
Italy	1954-2012
Japan	1947-2012
Lithuania	1959-2013
Netherlands	1950-2012
Norway	1967-2014
Portugal	1940-2012
Russia	1959-2010
Slovakia	1950-2009
Slovenia	1983-2014
Sweden	1891-2014
Switzerland	1932-2011
Taiwan	1976-2010
Ukraine	1959-2013
United Kingdom	1974-2013
England & Wales	1938-2013
Scotland	1945-2013
Northern Ireland	1974-2013
USA	1933-2013
UNWPP, northern America:	
Canada	1950-2015
United States of America	1950-2015
UNWPP, Africa:	
Algeria	1950-2015
Angola	1950-2015
Benin	1950-2015
Botswana	1950-2015
Burkina Faso	1950-2015
Burundi	1950-2015
Cabo Verde	1950-2015
Cameroon	1950-2015
Central African Republic	1950-2015
Chad	1950-2015
Comoros	1950-2015
Congo	1950-2015
Democratic Republic of the Congo	1950-2015
Congo	1950-2015
Egypt	1950-2015
Equatorial Guinea	1950-2015
Eritrea	1950-2015
Ethiopia	1950-2015
Gabon	1950-2015
Gambia	1950-2015
Ghana	1950-2015
Guinea	1950-2015
Guinea-Bissau	1950-2015
Ivory Coast	1950-2015
Kenya	1950-2015
Lesotho	1950-2015
Liberia	1950-2015
Libya	1950-2015
Madagascar	1950-2015
Malawi	1950-2015
Mali	1950-2015
Mauritania	1950-2015
Mauritius	1950-2015
Mayotte	1950-2015
Morocco	1950-2015
Mozambique	1950-2015

to be cont. on next page

Table 5: Data coverage in the HFD and UNWPP. Part II

Country	Calendar years
UNWPP, Africa, cont.:	
Namibia	1950-2015
Niger	1950-2015
Nigeria	1950-2015
Réunion	1950-2015
Rwanda	1950-2015
Sao Tome and Principe	1950-2015
Senegal	1950-2015
Seychelles	1950-2015
Sierra Leone	1950-2015
Somalia	1950-2015
South Africa	1950-2015
South Sudan	1950-2015
Sudan	1950-2015
Swaziland	1950-2015
Togo	1950-2015
Tunisia	1950-2015
Uganda	1950-2015
United Republic of Tanzania	1950-2015
Western Sahara	1950-2015
Zambia	1950-2015
Zimbabwe	1950-2015
UNWPP, Oceania:	
Australia	1950-2015
New Zealand	1950-2015
Fiji	1950-2015
New Caledonia	1950-2015
Papua New Guinea	1950-2015
Solomon Islands	1950-2015
Vanuatu	1950-2015
Guam	1950-2015
Kiribati	1950-2015
Micronesia (Fed. States of)	1950-2015
French Polynesia	1950-2015
Samoa	1950-2015
Tonga	1950-2015
UNWPP, Asia:	
Afghanistan	1950-2015
Armenia	1950-2015
Azerbaijan	1950-2015
Bahrain	1950-2015
Bangladesh	1950-2015
Bhutan	1950-2015
Brunei Darussalam	1950-2015
Cambodia	1950-2015
China	1950-2015
China, Hong Kong SAR	1950-2015
China, Macao SAR	1950-2015
China, Taiwan Province of China	1950-2015
Cyprus	1950-2015
Dem. People's Republic of Korea	1950-2015
Georgia	1950-2015
India	1950-2015
Indonesia	1950-2015
Iran (Islamic Republic of)	1950-2015
Iraq	1950-2015
Israel	1950-2015
Japan	1950-2015
Jordan	1950-2015
Kazakhstan	1950-2015
Kuwait	1950-2015
Kyrgyzstan	1950-2015
Lao People's Democratic Republic	1950-2015
Lebanon	1950-2015
Malaysia	1950-2015
Maldives	1950-2015
Mongolia	1950-2015
Myanmar	1950-2015
Nepal	1950-2015
Oman	1950-2015
Pakistan	1950-2015
Philippines	1950-2015
Qatar	1950-2015
Republic of Korea	1950-2015

to be cont. on next page

Table 6: Data coverage in the HFD and UNWPP. Part III

Country	Calendar years
UNWPP, Asia, cont.:	
Saudi Arabia	1950-2015
Singapore	1950-2015
Sri Lanka	1950-2015
State of Palestine	1950-2015
Syrian Arab Republic	1950-2015
Tajikistan	1950-2015
Thailand	1950-2015
Timor-Leste	1950-2015
Turkey	1950-2015
Turkmenistan	1950-2015
United Arab Emirates	1950-2015
Uzbekistan	1950-2015
Viet Nam	1950-2015
Yemen	1950-2015
UNWPP, Latin America and the Caribbean:	
Antigua and Barbuda	1950-2015
Aruba	1950-2015
Bahamas	1950-2015
Barbados	1950-2015
Cuba	1950-2015
Curaçao	1950-2015
Dominican Republic	1950-2015
Grenada	1950-2015
Guadeloupe	1950-2015
Haiti	1950-2015
Jamaica	1950-2015
Martinique	1950-2015
Puerto Rico	1950-2015
Saint Lucia	1950-2015
Saint Vincent and the Grenadines	1950-2015
Trinidad and Tobago	1950-2015
United States Virgin Islands	1950-2015
Belize	1950-2015
Costa Rica	1950-2015
El Salvador	1950-2015
Guatemala	1950-2015
Honduras	1950-2015
Mexico	1950-2015
Nicaragua	1950-2015
Panama	1950-2015

Country	Calendar years
UNWPP, Latin America & the Caribbean, cont.:	
Argentina	1950-2015
Bolivia (Plurinational State of)	1950-2015
Brazil	1950-2015
Chile	1950-2015
Colombia	1950-2015
Ecuador	1950-2015
French Guiana	1950-2015
Guyana	1950-2015
Paraguay	1950-2015
Peru	1950-2015
Suriname	1950-2015
Uruguay	1950-2015
Venezuela (Bolivarian Republic of)	1950-2015
UNWPP, Europe:	
Albania	1950-2015
Austria	1950-2015
Belarus	1950-2015
Belgium	1950-2015
Bosnia and Herzegovina	1950-2015
Bulgaria	1950-2015
Channel Islands	1950-2015
Croatia	1950-2015
Czechia	1950-2015
Denmark	1950-2015
Estonia	1950-2015
Finland	1950-2015
France	1950-2015
Germany	1950-2015
Greece	1950-2015
Hungary	1950-2015
Iceland	1950-2015
Ireland	1950-2015
Italy	1950-2015
Latvia	1950-2015
Lithuania	1950-2015
Luxembourg	1950-2015
Malta	1950-2015
Montenegro	1950-2015
Netherlands	1950-2015
Norway	1950-2015

Table 7: Data coverage in the HFD and UNWPP. Part IV

Country	Calendar years
UNWPP, Europe, cont.:	
Poland	1950-2015
Portugal	1950-2015
Republic of Moldova	1950-2015
Romania	1950-2015
Russian Federation	1950-2015
Serbia	1950-2015
Slovakia	1950-2015
Slovenia	1950-2015
Spain	1950-2015
Sweden	1950-2015
Switzerland	1950-2015
TFYR Macedonia	1950-2015
Ukraine	1950-2015
United Kingdom	1950-2015

B Best variant of major methods for forecasting period fertility

Table 8 lists the best variant for each of the 16 methods according; based on test results for 25-year-ahead-forecasts of fertility for ages 15 through 44.

Table 8: Best variant of 16 methods to forecast period fertility.

Method	Best variant
Baseline method:	
Freeze rates	test not required
Hybrid methods: PAR + EM:	
Hadwiger (1940)	Freeze, 30/35/40
Coale and McNeil (1972)	Freeze, 30/35/40
Coale and Trussell (1974)	Freeze, 30/35/40
Evans (1986)	Freeze, 30
Chandola et al. (1999)	Freeze, 30/35/40
Schmertmann (2003)	Freeze, 35
Peristera and Kostaki 1 (2007)	Freeze, 30/35/40
Peristera and Kostaki 2 (2007)	Freeze, 30/35/40
Myrskylä and Goldstein (2013)	ARIMA, IFC, 40
Extrapolation methods:	
Willekens and Baydar (1984)	ARIMA, 30
de Beer (1985)	de Beer 1 (CARIMA(1,1,0)(1,0,0)), 35
Lee (1993)	Bt, ARIMA, Actual, 40
Hyndman and Ullah (2007)	30
Myrskylä et al. (2013)	test not required
Bayesian approach:	
Ševcíková et al. (2016)	SPPW, 30