Forecast accuracy hardly improves with method complexity when completing cohort fertility

Christina Bohk-Ewald[a,1], Peng Li[a], and Mikko Myrskylä[a,b,c]

[a]Max Planck Institute for Demographic Research, 18057 Rostock, Germany; [b]Department of Social Policy, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and [c]Department of Social Research, University of Helsinki, 00014, Helsinki, Finland

Edited by Adrian E. Raftery, University of Washington, Seattle, WA, and approved August 3, 2018 (received for review December 22, 2017)

Forecasts of completed fertility predict how many children will be born on average by women over their entire reproductive lifetime. These forecasts are important in informing public policy and influencing additional research in the social sciences. However, nothing is known about how to choose a forecasting method from a large basket of variants. We identified 20 major methods, with 162 variants altogether. The approaches range from naive freezing of current age-specific fertility rates to methods that use statistically sophisticated techniques or are grounded in demographic theory. We assess each method by evaluating the overall accuracy and if provided, uncertainty estimates using fertility data of all available birth cohorts and countries of the Human Fertility Database, which covers 1,096 birth cohorts from 29 countries. Across multiple measures of forecast accuracy, we find only four methods that consistently outperform the naive freeze rates method, and only two methods produce uncertainty estimates that are not severely downward biased. Among the top four, there are two simple extrapolation methods and two Bayesian methods. The latter are demanding in terms of input data, statistical techniques, and computational power but do not consistently complete cohort fertility more accurately at all truncation ages than simple extrapolation. This broad picture is unchanged if we base the validation on 201 United Nations countries and six world regions, including Africa, Asia, Europe, Latin America and the Caribbean, northern America, and Oceania.

fertility forecast methods | validation | forecast errors

Forecasts of completed fertility (CF) predict how many children will be born on average by women over their entire reproductive lifetime. They are a key element in research on population dynamics and forecasting (1), and they are used by decision makers throughout the society: from social security planning to marketing. Cohort fertility complements the reproduction picture given by period fertility: the former is informative about the experience of real cohorts of women and immune to timing effects of fertility, and the latter can vary substantially as a result of fertility timing but is useful for summarizing fertility within a period. This paper provides a systematic evaluation of the accuracy of existing methods for completing cohort fertility.

Over the last century, demographers have introduced dozens of methods to forecast cohort fertility—with continuously rising requirements for input data, advanced statistical techniques, and computing power—to pursue two primary objectives: (*i*) to increase forecast accuracy and recently, (*ii*) to provide reliable uncertainty estimates. However, new methodological innovations are often introduced without careful comparison of how the new approach performs with respect to existing alternatives. In fact, nothing is known about how to choose a forecasting method from a large basket of variants.

In the presence of this enormous methodological variety, a key question that forecasters face today is as follows: which method should they use to produce accurate cohort fertility forecasts with reliable uncertainty information? Furthermore, if forecast performance increased over time, what were the main methodological breakthroughs? Finally. what are the unresolved issues that can show us in what directions we should put our efforts to further improve the forecast methodology?

In this paper, we provide an assessment of all cohort fertility forecast methods. Our comprehensive survey identified 20 major methods, with 162 variants arising from different parameterizations. Each of these methods aims at completing lifetime fertility of women who have not yet reached their last reproductive age at the time that a forecast is made.

The existing approaches for cohort fertility forecasting range from naive freezing of current age-specific fertility rates to methods that use statistically sophisticated techniques or are grounded in demographic theory. We distinguish the baseline method freeze rates (also referred to as constant rates method), which holds fertility rates constant at their latest observed level, and four broader types of method, which are (*i*) parametric curve fitting methods (PARs) (*SI Appendix*, Table S46 provides a list of abbreviations), (*ii*) extrapolation methods (EMs), (*iii*) Bayesian approaches (BAs), and (*iv*) fertility context-specific methods (CONs). Some methods combine elements across many of these typologies (hybrid models), and therefore, the classification is unavoidably subjective.

The objective of PARs is to detect a universal pattern of fertility by age—which often resembles a bell shape, with fertility

Significance

Information on cohort fertility is critical for the understanding of population dynamics, but only in historical settings can it be calculated without forecasting. Several forecasting methods exist, but their strengths and weaknesses have not been evaluated. Relying on the Human Fertility Database, the largest high-quality fertility dataset to date, and the globally representative United Nations World Population Prospects, we present an assessment of all major methods that complete cohort fertility. This analysis is crucial to advance the understanding of benefits and drawbacks of state-of-the-art methods. We analyze forecast accuracy and uncertainty quantification, identify methodological breakthroughs, and uncover unresolved issues. This study constitutes an evaluation benchmark for cohort fertility forecasting and may inspire establishment of similar evaluation benchmarks in related fields.

Author contributions: C.B.-E. and M.M. designed research; C.B.-E. and P.L. performed research; C.B.-E., P.L., and M.M. contributed new reagents/analytic tools; C.B.-E., P.L., and M.M. analyzed data; and C.B.-E. and M.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The R code used for this study has been deposited in GitHub, https://github.com/fertility-forecasting/validate-forecast-methods.

[1]To whom correspondence should be addressed. Email: BohkEwald@demogr.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1722364115/-/DCSupplemental.

Published online August 27, 2018.

levels being high at middle reproductive ages and low at very young and old reproductive ages—and to express it with a mathematical function. The number and the interpretation of parameters are the distinctive features of models within this group. Simple PARs model one hump (2–4) and complex PARs model a second hump for young mothers (4, 5) in the age distribution of fertility. Relational PARs (6–8) forecast fertility using deviations between current and standard fertility age schedules; 10 of our 20 methods are classified as PARs.

EMs are based on the assumption that past trends are helpful in predicting the future. EMs fit a model to observed fertility trends and extrapolate the trajectory based on the model. Complex EMs (9, 10) consider fertility dynamics in all three dimensions (age, calendar year, and birth cohort), whereas simple EMs (11–16) reduce complexity and reflect fertility trends in less detail. EMs often use time series models (17) to extrapolate trajectories (9–16) and to provide uncertainty estimates (10–15). We classify 6 of 20 methods as EMs.

The characteristic feature of hierarchical BAs (18–20) is to augment information about fertility in a country of interest with information about typical levels and trends of fertility in other countries. Borrowing strength from a large data pool can be advantageous for countries that have data of poor quality, exhibit unsteady fertility developments, or have incomplete fertility age schedules. The forecast performance of hierarchical BAs depends on the spatiotemporal composition of the data pool; 2 of 20 methods are hierarchical BAs.

CONs (21, 22) are designed to model particular fertility developments, such as a decline in CF and a delay of childbearing to higher maternal ages (23–25), which reduces their overall applicability. We identified only 1 CON method in 20 methods.

We further identify three hybrid models (8, 26, 27) that combine elements of PARs and EMs or BAs. They extrapolate key parameters of the fertility age schedule (26, 27) or make use of information from other countries (8). To keep the classification of methods manageable, we have assigned each of them to one of four method types: PARs, EMs, BAs, or CONs.

We describe the major methods and their variants in more detail in *SI Appendix*, section 1. We excluded four additional methods (11, 18, 21, 28) and their 14 variants from the main analysis, because their data requirements are so demanding that they can be implemented in only a handful of countries. *SI Appendix*, Table S31 provides information about which methods eventually entered the main analysis.

Our main comparison of the 20 methods is based on testing their forecasting performance in the largest high-quality dataset of fertility: the Human Fertility Database (HFD) (29). The version of the data that we use covers 29 countries, years 1891–2013, and 1,096 birth cohorts altogether. This massive database includes a large variety of levels, shapes, and trends of fertility in mostly developed countries (see *SI Appendix*, section 2 for more details on the HFD).

We complement the HFD-based validation with a comparison based on fertility data from the United Nations (UN) (30), which include six world regions: Africa, Asia, Europe, Latin America and the Caribbean, northern America, and Oceania. The UN dataset covers 201 countries, years 1950–2015, and 8,241 (i.e., 201×41) birth cohorts altogether. Combining the high-quality HFD data with the broad geographic and contextual variation available in the UN data increases the external validity and robustness of our findings. Results of the UN-based analysis are in *SI Appendix*, section 4. We put special emphasis on Africa in *SI Appendix*, section 5, which exhibits, on average, the highest fertility levels worldwide (31).

In our validation, we forecast CF for each birth cohort using data only up to a certain truncation age, which ranges from 20 to 39 yr old, forecast fertility up to age 40 yr old, and calculate the error that is the difference between the forecasted and true CF at age 40 yr old. With this procedure, we obtain 8,652 errors in the HFD-based comparison for 29 countries and 62,310 errors in the UN-based comparison for 201 countries. The errors are indexed by truncation age, country, and birth cohort for each of the 20 major methods. Details of how we obtain forecast errors are given in *SI Appendix*, section 2.

We include in the final comparison only the best-performing variant of each method; *SI Appendix*, section 2, provides details on this selection procedure. For methods that include uncertainty bounds, we test the coverage of these bounds using a similar protocol.

Results

Accuracy of CF Forecasts.

Stochastic dominance. Fig. 1 ranks the methods based on absolute percentage errors (APEs) and Kolmogorov–Smirnov (KS) test of stochastic dominance. For a total of 380 pairwise method comparisons (excluding the comparison with self), the test determines if the cumulative density of the APEs of method A (row) is greater than that of method B (column) ($P = 0.05$). The methods are ordered top down with a decreasing number of significant results of stochastic dominance.

In our full set of errors based on the HFD, the simple EM of Myrskylä et al. (16) performs best and has, on average, strictly smaller APEs than 17 other methods. This method is closely followed by the BA of Schmertmann et al. (20) and the simple EM of de Beer (12), which both produce CF forecasts that are, on average, strictly more accurate than those of 16 other methods. We cannot infer which of these two methods performs best in this setting because of a cross-over of the cumulative densities of their APEs. The BA of Ševcíková et al. (19) is on rank 4. Freeze rates is the fifth most accurate method, stochastically dominating 11 of the other 19 methods in this setting.

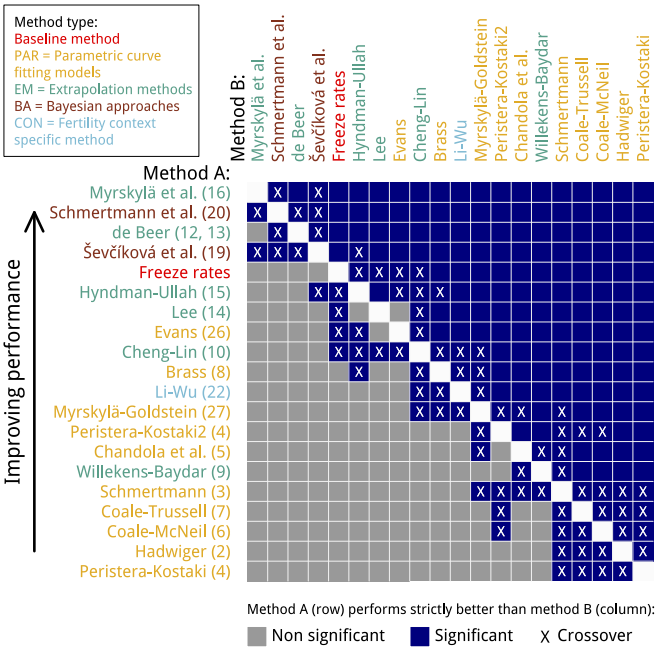

Fig. 1. Two-sample, one-sided KS test statistic for stochastic dominance: KS (method A, method B). Blue indicates that the cumulative density of the APEs of method A (row) is significantly greater than that of method B (column), and gray indicates nonsignificant results for the same test. Crosses indicate inconclusive test results caused by cross-overs of cumulative densities. Testing data are from the HFD (details are in *SI Appendix*, section 2).

Regarding the method types, it is striking that the forecasts of simple EMs (12, 14–16) and BAs (19, 20) are, on average, more accurate than those of more complex EMs (9, 10), CONs (22), and most PARs (2–7). In addition, the forecasts of hybrid methods that combine properties of PARs and EMs (26, 27) or BAs (8) are, on average, more accurate than the pure PARs. Forecasts of simple PARs (2–4, 6, 7) are, on average, less accurate than those of more complex PARs (4, 5).

Thresholds for APE. Fig. 2 shows thresholds, below which 50, 80, 90, and 95% of the APEs are for each method. The methods are ordered first by method type and then by publication date. This order reveals that forecast accuracy did not consistently improve over time, although increasingly more high-quality data and statistical methods have become available for developing approaches. Specifically, the hybrid methods of Evans (26) and Brass (8) and the EM of de Beer (12) disrupt this continuous line of improvement with their outstanding performance.

The threshold analysis is in line with the findings of the tests of stochastic dominance: the thresholds below which a specific percentage of errors falls are relatively small for the simple EMs (12, 14–16), BAs (19, 20), and freeze rates in comparison with those of more complex EMs (9, 10), hybrid methods (8, 26, 27), and PARs (2–7).

Ranking. Table 1 ranks the 20 methods according to seven metrics: KS test; 50, 80, 90, and 95% APE thresholds; mean APE; and root mean square error (RMSE). The overall ranking is based on the KS test; ties are solved with mean APE. For example, the BA of Schmertmann et al. (20) and the simple EM of de Beer (12) have the same rank (rank 2) according to the KS test statistic; adding information of the overall mean breaks the tie in favor of the method of Schmertmann et al. (20).

The overall ranking has the simple EMs of Myrskylä et al. (16) and de Beer (12) in the first and third positions, respectively; the BAs of Schmertmann et al. (20) and Ševcíková et al. (19) in the second and fourth positions, respectively; and freeze rates in the fifth position. The simple EM of Lee (14) is in the sixth position.

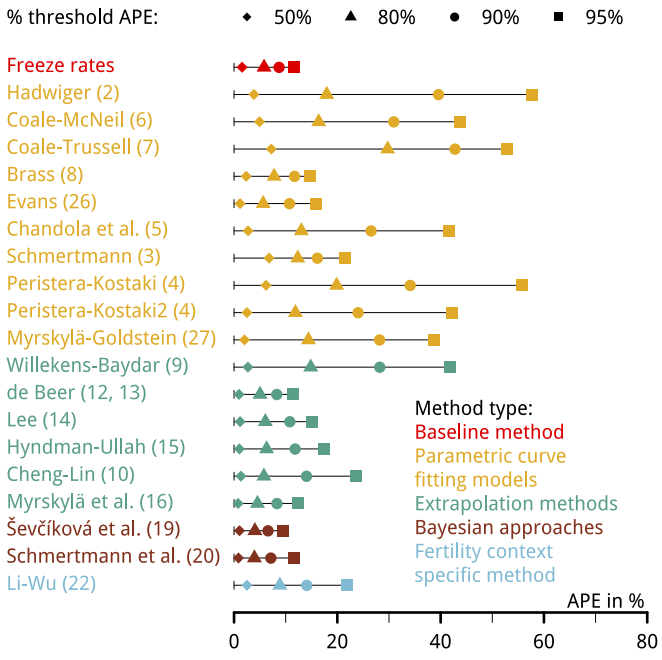

Fig. 2. Thresholds below which 50% (♦), 80% (▲), 90% (●), and 95% (■) of APEs are for each method. Testing data are from the HFD (details are in *SI Appendix*, section 2).

Table 1 also provides useful information on the magnitude of the errors. The threshold below which one-half of the errors are is 1.6% or less for the five best-performing methods. For the 90% threshold, the corresponding line is 8.7%. Whether these errors are considered large or small depends on the application.

Alternative rankings than the one shown in Table 1 would be possible. For example, the method by Ševčíková et al. (19) ranks on top in terms of RMSE, suggesting that it has fewer very large errors than other best-performing methods. In fact, three of the top five methods rank as number one in at least one of seven metrics. However, none of the methods outside the top five rank as number one on any of the metrics considered.

Truncation age. Fig. 3 displays the accuracy of the best five methods compared with freeze rates by truncation age based on absolute errors (AEs) on the relative and absolute scales. We spot three patterns. First, forecasts of simple EMs (12, 14, 16) are up to 40% more accurate than those of freeze rates for truncation ages above 30 yr old. For truncation ages below 25 yr old, EMs have mostly no advantage and often have a disadvantage compared with freeze rates. Second, the BA of Ševčíková et al. (19) shows an opposite pattern: its forecasts are particularly accurate for truncation ages below 25 yr old, but for truncation ages above 35 yr old, the advantage over freeze rates is small or turns to a disadvantage. Third, forecasts of the BA of Schmertmann et al. (20) have a pattern over truncation age that is between those of simple EMs and the BA of Ševčíková et al. (19). This method performs well for truncation ages 25–35 yr old; for younger and older truncation ages, the advantage over freeze rates is small. Furthermore, only the forecasts of the BA of Schmertmann et al. (20) and of the simple EM of de Beer (12) are, on average, almost consistently more accurate than those of freeze rates.

Uncertainty Quantification. Only 8 of 20 methods (8, 14–16, 19, 20, 26, 27) provide prediction intervals to quantify forecast uncertainty. To assess the calibration of the prediction intervals, we analyze how well their empirical coverage matches with the nominal coverage of the prediction intervals. Fig. 4 shows the empirical coverage of the nominal 80% (▲) and 95% (■) prediction intervals. The empirical coverage varies strongly across methods. Except for the EM of Hyndman and Ullah (15) and the BA of Schmertmann et al. (20), the methods substantially underestimate uncertainty, and of these two, the method of Hyndman and Ullah (15) substantially overestimates the uncertainty.

Discussion

Only a few studies have dealt extensively with the verification of forecast methods (reviews of, for example, refs. 32–34). Although they focus on different quantities and are from other fields, their conclusions are consistent with our broad findings: more complex methods do not necessarily outperform simpler methods.

We assess, compare, and rank the overall performance of 20 major methods that complete cohort fertility on exactly the same testing dataset of the high-quality HFD, and therefore, any differences in terms of forecast accuracy and uncertainty estimates can be traced back to methodological differences. We find that the baseline freeze rates method is consistently outperformed by only four methods: the simple EMs of Myrskylä et al. (16) and de Beer (12) and the BAs of Schmertmann et al. (20) and Ševcíková et al. (19). Comparing their performances among each other, we find that any one of the top four methods could be in the lead depending on the applied metric. While some of the best-performing methods produce large gains with respect to freeze rates for some truncation ages, only two of them consistently outperform freeze rates across (almost) all truncation ages. Moreover, all but two probabilistic methods [i.e., the methods of Hyndman and Ullah (15) and of Schmertmann et al. (20)] underestimate forecast uncertainty.

Table 1. Rank methods based on different measures of forecast accuracy

		Threshold APE, %					
Overall ranking	KS	50	80	90	95	Mean	RMSE
1. Myrskylä et al. (16)	17 (1)	0.8 (1)	4.5 (3)	8.3 (4)	12.5 (5)	2.9 (4)	5.9 (5)
2. Schmertmann et al. (20)	16 (2)	0.9 (2)	3.9 (1)	7.1 (2)	11.5 (3)	2.7 (2)	5.4 (3)
3. de Beer (12, 13)	16 (2)	1.0 (3)	5.0 (4)	8.3 (3)	11.4 (2)	2.9 (3)	5.3 (2)
4. Ševcíková et al. (19)	15 (4)	1.1 (5)	4.0 (2)	6.6 (1)	9.5 (1)	2.4 (1)	4.2 (1)
5. Freeze rates	11 (5)	1.6 (9)	5.8 (7)	8.7 (5)	11.5 (4)	3.3 (5)	5.5 (4)
6. Lee (14)	11 (5)	1.2 (7)	6.1 (8)	10.8 (7)	15.1 (7)	3.6 (6)	6.9 (6)
7. Evans (26)	11 (5)	1.2 (6)	5.6 (5)	10.7 (6)	16.0 (8)	3.8 (7)	7.7 (8)
8. Hyndman and Ullah (15)	11 (5)	1.1 (4)	6.3 (9)	11.8 (9)	17.5 (9)	3.9 (8)	7.8 (9)
9. Brass (8)	8 (9)	2.4 (11)	7.8 (10)	11.7 (8)	14.7 (6)	4.5 (9)	7.1 (7)
10. Cheng and Lin (10)	8 (9)	1.3 (8)	5.8 (6)	14.0 (10)	23.5 (12)	4.9 (10)	11.2 (12)
11. Li and Wu (22)	8 (9)	2.5 (12)	8.9 (11)	14.1 (11)	21.8 (11)	5.7 (11)	10.7 (11)
12. Myrskylä and Goldstein (27)	5 (12)	2.0 (10)	14.4 (15)	28.2 (15)	38.8 (13)	8.5 (13)	16.2 (13)
13. Willekens and Baydar (9)	4 (13)	2.7 (14)	14.9 (16)	28.2 (16)	41.9 (15)	9.5 (14)	19.6 (15)
14. Chandola et al. (5)	4 (13)	2.7 (15)	13.0 (14)	26.5 (14)	41.6 (14)	9.6 (15)	24.4 (18)
15. Peristera and Kostaki (4)	4 (13)	2.5 (13)	11.9 (12)	24.0 (13)	42.3 (16)	10.3 (16)	29.8 (19)
16. Schmertmann (3)	0 (16)	6.8 (19)	12.3 (13)	16.2 (12)	21.6 (10)	8.2 (12)	10.6 (10)
17. Coale and McNeil (6)	0 (16)	4.9 (17)	16.4 (17)	30.9 (17)	43.7 (17)	10.8 (17)	18.4 (14)
18. Hadwiger (2)	0 (16)	3.9 (16)	18.0 (18)	39.6 (19)	57.6 (20)	12.0 (18)	21.9 (16)
19. Peristera and Kostaki (4)	0 (16)	6.2 (18)	19.9 (19)	34.1 (18)	55.8 (19)	15.0 (19)	33.0 (20)
20. Coale and Trussell (7)	0 (16)	7.2 (20)	29.7 (20)	42.8 (20)	52.9 (18)	15.1 (20)	23.1 (17)
No. of inversions	0	17	17	17	21	9	24

The overall ranking of methods in column 1 is based on the KS test statistic of stochastic dominance (column 2) and the overall mean (column 7). Testing data are from the HFD (details are in *SI Appendix*, section 2).

We examine how consistent these findings are for (*i*) older truncation ages 30–39 yr old and all years, (*ii*) truncation ages 20–39 yr old and years 1990 and later, and (*iii*) older truncation ages 30–39 yr old and years 1990 and later in *SI Appendix*, section 3. We find that four of the top five methods that did well in the full error set are also at the top in the three error subsets, indicating robust performance. Freeze rates moves slightly down to rank eight, and the simple EM of Hyndman and Ullah (15) moves to the top five.

Our analysis relied on the high-quality HFD that covers mostly developed countries with comparatively low-fertility settings and delay of childbearing since 1990. Such stable trends may explain the high performance of freeze rates when completing cohort fertility, which is immune to fertility timing. To analyze whether our key findings are strongly context specific, perhaps varying by region or fertility level, we replicated our validation exercise using data from the UN (30). The data quality may be less consistent than in the HFD, but the reconstructed UN data cover effectively the whole globe. We analyzed the performance of the 20 methods in 201 UN countries and six world regions: Africa, Asia, Europe, Latin America and the Caribbean, Oceania, and northern America; data availability allows us to forecast only from year 1990.

Fig. 5 shows that the broad picture of the HFD-based validation holds across the globe—method complexity does not necessarily increase forecast accuracy. The group of the top-performing methods (red in Fig. 5) stays rather stable across regions as does the group of the lower-ranked methods (blue in Fig. 5). Although the UN-based rankings differ slightly across regions, we find that the simple EMs of de Beer (12) and Myrskylä et al. (16) and the BAs of Schmertmann et al. (20) and Ševcíková et al. (19) are at the top complemented by the simple EMs of Hyndman and Ullah (15) and Lee (14). There is no evidence in any region that the complex parametric, complex extrapolation, or Bayesian methods would consistently perform better than simple extrapolation. A detailed analysis of the 201 UN countries and Africa is in *SI Appendix*, sections 4 and 5.

We also show in *SI Appendix*, sections 3–5, that bias in CF is overall small for top methods and that it varies by truncation age and region. Absolute bias peaks at young truncation ages and does not exceed 0.10 in the HFD data. In UN and Africa, the bias is, unlike in HFD countries, consistently positive at all truncation ages and also larger, ranging from 0.1 to 0.3 at age 20 yr old across top methods. This positive and higher bias may be due to overall levels of fertility being higher in Africa and the

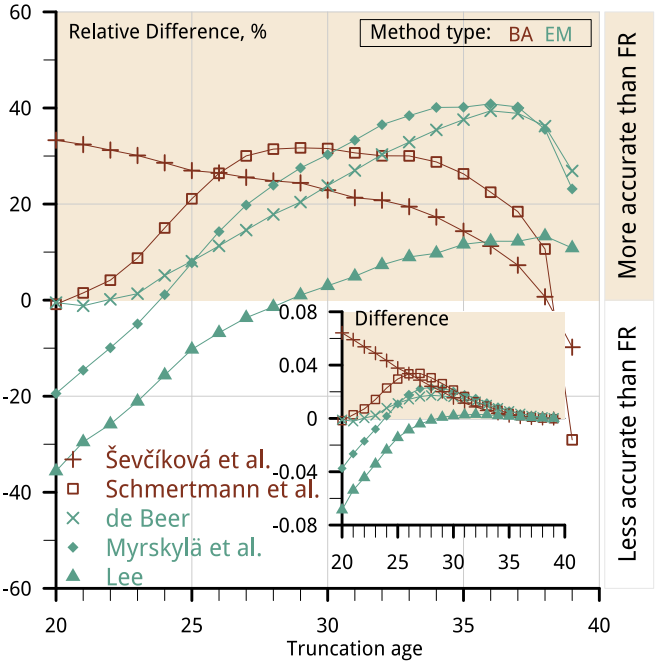

Fig. 3. Mean AE of the top five methods compared with freeze rates (FRs) by truncation age. Relative difference in percentage. (*Inset*) Difference on an absolute scale. Testing data are from the HFD (*SI Appendix*, section 2).

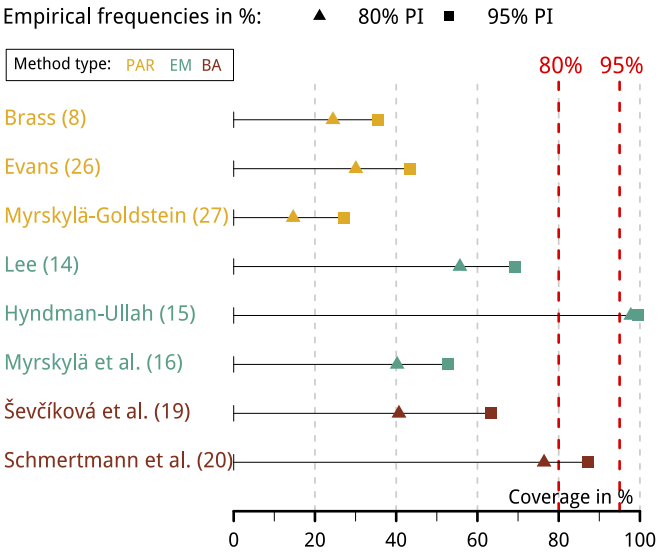

Fig. 4. Empirical coverage of nominal 80 and 95% prediction intervals (PIs) of probabilistic methods. Testing data are from the HFD (details are in *SI Appendix*, section 2).

inability of some methods, particularly freeze rates, to capture strong declines in fertility.

Although freeze rates performs less accurately in some UN world regions than in HFD countries compared with top methods, it is remarkable that it consistently outperforms most of the sophisticated and less sophisticated forecast methods, and was in top 5 of all 20 methods in the core evaluation using the high-quality HFD data. Moreover, throughout our analysis and across the globe, we found no evidence that the simple EMs would be consistently outperformed by more complex approaches in terms of forecast accuracy. This finding not only raises the question of how much the extra effort of more complex methods is worth, but it also suggests that it should be established as a standard procedure to assess whether proposed new methods outperform at least the naive method freeze rates and the simple EMs of de Beer (12) and Myrskylä et al. (16). To facilitate such benchmarking, we provide an implementation of the forecasting methods evaluated in this study at https://github.com/fertility-forecasting/validate-forecast-methods.

Since increasingly available resources in terms of high-quality data, statistical methods, and computing power did not necessarily lead to a continuous improvement of forecast performance over time (Fig. 2), we ask the following question: what are the key methodological breakthroughs of the top methods compared with the lower ranked methods? Superior performance seems to be linked to feeding forecasts with temporal fertility trends in a country of interest and experiences of many other countries. Pure parametric curve fitting models (2–7) are less accurate than extrapolation (9, 10, 12, 14–16) and BAs (19, 20), because they do not rely on temporal fertility trends and borrow strength from other countries. However, the hybrid parametric methods of Evans (26) and Brass (8) adopt features of simple extrapolation or BAs, respectively, and consequently, they have relatively high forecast accuracy. Moreover, complex extrapolation (9, 10) underperforms against simple extrapolation (12, 14–16), perhaps because it may overfit observed trends by age, period, or birth cohort.

Complementing the analysis of methodological breakthroughs with forecast errors by country (*SI Appendix*, section 6) uncovers what forecast situations are challenging, even for the top methods. Not surprisingly, major challenges across regions are fertility declines that deviate from continuous trends or the experience of other countries. The top methods produce relatively large forecast errors for eastern Germany when fertility fell to unprecedentedly low levels after German reunification (35) and for Algeria when fertility sharply declined from over eight to only two children per woman between the 1970s and the 2000s (36). No method has found an effective remedy against such unstable fertility developments so far.

Our analyses have consistently shown that more complex methods do not necessarily outperform simpler methods when completing cohort fertility. The scope of our findings does not extend to long-term forecasts of cohort or period fertility due to lack of sufficiently long data time series and restricted applicability of methods. However, they do show that a thorough evaluation of forecast performance is crucial to identify the best methods, to retroactively grade methodological breakthroughs, and to proactively uncover unresolved issues. Establishing extensive validation as a standard evaluation benchmark for new fertility forecast approaches would help to better focus efforts aimed at methodological development.

Materials and Methods

Readers have access to R code on GitHub: https://github.com/fertility-forecasting/validate-forecast-methods. *SI Appendix* describes forecast methods in section 1; validation procedure to select the best variant per method in section 2; results of HFD-based error subsets in section 3; results of UN- and Africa-based validation in sections 4 and 5, respectively; and errors by country in section 6.

Forecast CF. CF is the sum of fertility rates over ages 15–40 yr old for women of the same birth cohort; it gives the average number of children for women born in the same calendar year. CF forecasts complete lifetime fertility for women who have not yet reached the last reproductive age. For example, to complete fertility for women whose truncation age is 30 yr old, we forecast their remaining fertility for ages 31–40 yr old; their CF contains fertility that is observed until age 30 yr old and forecasted above age 30 yr old.

Forecast Error. The AE measures accuracy as the absolute difference between forecasted and observed CF, and the APE relates this AE to observed CF. If observed CF is 2, an APE of 10% indicates that the absolute deviation between forecasted and observed CF (i.e., AE) was 0.2 and that the forecasted CF was either 1.8 or 2.2.

KS Test Statistic of Stochastic Dominance. The KS test statistic is widely applied to test for equality and stochastic dominance of two distributions (37–39); it is a nonparametric and distribution-free test statistic that requires large but not equal sample sizes (40–42). We use the two-sample, one-sided KS test statistic (42) to determine for each pair of methods, A and B, if

Method type: Baseline PAR EM BA CON Truncation ages 20 to 39 and JOYs 1990+

	HFD29	UN201	Africa	Asia	Europe	LAC	Oceania	NA
Completed fertility:	1.87	3.67	5.32	3.67	1.91	3.17	3.82	1.91
Myrskylä et al. (16)	1	2	2	1	2	3	1	1
de Beer (12, 13)	2	1	1	2	1	2	3	4
Ševčíková et al. (19)	3	6	5	5	6	5	8	6
Hyndman-Ullah (15)	4	4	4	6	3	4	4	5
Schmertmann et al. (20)	5	5	6	4	5	6	7	2
Evans (26)	6	12	11	10	8	7	6	9
Lee (14)	7	3	3	3	4	1	2	3
Freeze rates	8	7	8	7	7	8	5	8
Li-Wu (22)	9	11	12	11	9	10	9	11
Cheng-Lin (10)	10	10	10	12	10	11	12	7
Brass (8)	11	8	7	8	11	12	10	10
Myrskylä-Goldstein (27)	12	19	20	19	12	18	17	14
Peristera-Kostaki2 (4)	13	16	17	16	16	13	14	12
Willekens-Baydar (9)	14	9	9	9	14	9	11	15
Chandola et al. (5)	15	14	18	13	13	19	16	16
Hadwiger (2)	16	20	19	17	17	20	19	17
Peristera-Kostaki (4)	17	18	16	20	19	16	15	13
Schmertmann (3)	18	15	15	18	18	15	18	19
Coale-McNeil (6)	19	13	14	14	15	14	13	18
Coale-Trussell (7)	20	17	13	15	20	17	20	20

Fig. 5. Ranking of the methods based on 29 HFD countries, 201 UN countries, and world regions: Africa, Asia, Europe, Latin America and the Caribbean (LAC), Oceania, and northern America (NA). Truncation ages are 20–39 yr old, and years are 1990+. Testing data are from the HFD and UN data (details are in *SI Appendix*, sections 2, 4 and 5).

method A stochastically dominates method B: KS(A, B) with a significance level of 0.05. To identify a possible cross-over, we conduct the statistical test from both sides: KS(A, B) and KS(B, A).

Number of Inversions. To quantify the dissimilarity between rankings, we compare their number of inversions. Each pair of ranks $i < j$ that is out of sort $r_i > r_j$ is an inversion (43). The overall method ranking is in ascending order $r_1, r_2, \ldots, r_{20}$ and has zero inversions. The greater the number of inversions is for any other ranking, the greater is its dissimilarity with the overall ranking.

Relative Performance Based on AEs. We quantify how much more accurate forecasts of a method are compared with freeze rates with the simple difference of AEs between freeze rates and another method on the absolute scale and with the relative difference of AEs, which relates the simple difference of AEs to the AEs of freeze rates. Positive values indicate that a method is more accurate than freeze rates and vice versa.

ACKNOWLEDGMENTS. We are grateful for comments from the anonymous referees as well as from the audience of the 2nd HFD Symposium in Berlin in 2016, the Social Statistics Seminar at the University of Helsinki in 2017, and the Annual Meeting of the Population Association of America in 2017. We are also grateful for helpful comments from members of the Max Planck Institute for Demographic Research, in particular Dmitri A. Jdanov, Pavel Grigoriev, Sebastian Klüsener, and Héctor Pifarré i Arolas. M.M. was supported by the European Research Council Grant 336475 (COSTPOST).

1. Raftery AE, Li N, Ševčíková H, Gerland P, Heilig GK (2012) Bayesian probabilistic population projections for all countries. *Proc Natl Acad Sci USA* 109:13915–13921.
2. Hadwiger H (1940) Eine analytische reproduktionsfunktion für biologische Gesamtheiten. *Scand Actuarial J* 1940:101–113.
3. Schmertmann CP (2003) A system of model fertility schedules with graphically intuitive parameters. *Demographic Res* 9:81–110.
4. Peristera P, Kostaki A (2007) Modeling fertility in modern populations. *Demographic Res* 16:141–194.
5. Chandola T, Coleman DA, Hiorns RW (1999) Recent European fertility patterns: Fitting curves to "'distorted'" distributions. *Popul Stud* 53:317–329.
6. Coale A, McNeil D (1972) The distribution by age at first marriage in a female cohort. *J Am Stat Assoc* 67:743–749.
7. Coale AJ, Trussell TJ (1974) Model fertility schedules: Variations in the age structure of childbearing in human populations. *Popul Index* 40:185–258.
8. Brass W (1974) Perspectives in population prediction: Illustrated by the statistics of england and wales. *J R Stat Soc Ser A, General* 137:532–583.
9. Willekens F, Baydar N (1984) Age-period-cohort models for forecasting fertility (Netherlands Interuniversity Demographic Institute (NIDI), Voorburg, The Netherlands), working paper no. 45.
10. Cheng PR, Lin ES (2010) Completing incomplete cohort fertility schedules. *Demographic Res* 23:223–256.
11. Saboia JLM (1977) Autoregressive integrated moving average (ARIMA) models for birth forecasting. *J Am Stat Assoc* 72:264–270.
12. de Beer J (1985) A time series model for cohort data. *J Am Stat Assoc* 80:525–530.
13. de Beer J (1989) Projecting age-specific fertility rates by using time-series methods. *Eur J Popul* 5:315–346.
14. Lee RD (1993) Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level. *Int J Forecast* 9:187–202.
15. Hyndman RJ, Ullah MS (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Comput Stat Data Anal* 51:4942–4956.
16. Myrskylä M, Goldstein JR, Cheng YA (2013) New cohort fertility forecasts for the developed world: Rises, falls, and reversals. *Popul Dev Rev* 39:31–56.
17. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, Hoboken, NJ).
18. Alkema L, et al. (2011) Probabilistic projections of the total fertility rate for all countries. *Demography* 48:815–839.
19. Ševčíková H, Li N, Kantorová V, Gerland P, Raftery AE (2016) Age-specific mortality and fertility rates for probabilistic population projections. *Dynamic Demographic Analysis*, ed Schoen R (Springer International Publishing, Cham, Switzerland), pp 285–310.
20. Schmertmann C, Zagheni E, Goldstein JR, Myrskylä M (2014) Bayesian forecasting of cohort fertility. *J Am Stat Assoc* 109:500–513.
21. Sobotka T, Zeman K, Lesthaeghe R, Frejka T, Neels K (2011) Postponement and recuperation in cohort fertility: Austria, Germany and Switzerland in a European context. *Comp Popul Stud* 36:417–452.
22. Li N, Wu Z (2003) Forecasting cohort incomplete fertility: A method and an application. *Popul Stud* 57:303–320.
23. Lesthaeghe R (2014) The second demographic transition: A concise overview of its development. *Proc Natl Acad Sci USA* 111:18112–18115.
24. Frejka T, Sardon JP (2004) *Childbearing Trends and Prospects in Low-Fertility Countries: A Cohort Analysis* (Kluwer Academic Publishers, Dordrecht, The Netherlands).
25. Billari F, Kohler HP (2004) Patterns of low and lowest-low fertility in Europe. *Popul Stud* 58:161–176.
26. Evans MDR (1986) American fertility patterns: A comparison of white and nonwhite cohorts born 1903-56. *Popul Dev Rev* 12:267–293.
27. Myrskylä M, Goldstein JR (2013) Probabilistic forecasting using stochastic diffusion models, with applications to cohort processes of marriage and fertility. *Demography* 50:237–260.
28. Kohler HP, Ortega JA (2002) Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. *Demographic Res* 6:91–144.
29. Human Fertility Database (2016) Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org. Accessed April 7, 2016.
30. United Nations, Department of Economic and Social Affairs, Population Division (2017) World population prospects: The 2017 revision, DVD ed. 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. Accessed March 1, 2018.
31. United Nations, Department of Economic and Social Affairs, Population Division (2015) World fertility patterns 2015—data booklet (ST/ESA/SER.A/370). Available at http://www.un.org/en/development/desa/population/publications/pdf/fertility/world-fertility-patterns-2015.pdf. Accessed April 13, 2018.
32. Smith SK (1997) Further thoughts on simplicity and complexity in population projection models. *Int J Forecast* 13:557–565.
33. Makridakis S, Hibon M (2000) The M3-Competition: Results, conclusions and implications. *Int J Forecast* 16:451–476.
34. Keilman N, Pham DQ (2004) Empirical errors and predicted errors in fertility, mortality and migration forecasts in the European Economic area (Statistics, Kongsvinger, Norway), Technical Report, Discussion Paper 386.
35. Goldstein JR, Kreyenfeld M (2011) Has East Germany overtaken West Germany? Recent trends in order-specific fertility. *Popul Dev Rev* 37:453–472.
36. Ouadah-Bedidi Z, Vallin J (2013) Fertility and population policy in Algeria: Discrepancies between planning and outcomes. *Popul Dev Rev* 38:179–196.
37. Levy H (1992) Stochastic dominance and expected Utility: Survey and analysis. *Manag Sci* 38:555–593.
38. Heathcote A, Brown S, Wagenmakers E, Eidels A (2010) Distribution-free tests of stochastic dominance for small samples. *J Math Psychol* 54:454–463.
39. Barrett GF, Donald SG (2003) Consistent tests for stochastic dominance. *Econometrica* 71:71–104.
40. Marsaglia G, Tsang WW, Wang J (2003) Evaluating Kolmogorov's distribution. *J Stat Softw* 8:1–4.
41. Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 46:68–78.
42. Sheskin DJ (2011) *Handbook of Parametric and Nonparametric Statistical Procedures* (CRC, Boca Raton, FL).
43. Knuth DE (1998) *The Art of Computer Programming: Sorting and Searching* (Addison-Wesley, Redwood City, CA), 2nd Ed, Vol 3.