

DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 30, ARTICLE 35, PAGES 1011–1034

PUBLISHED 2 APRIL 2014

<http://www.demographic-research.org/Volumes/Vol30/35/>

DOI: 10.4054/DemRes.2014.30.35

Research Article

Regional probabilistic fertility forecasting by modeling between-country correlations

Bailey K. Fosdick

Adrian E. Raftery

© 2014 Bailey K. Fosdick & Adrian E. Raftery.

This open-access work is published under the terms of the Creative Commons Attribution NonCommercial License 2.0 Germany, which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/by-nc/2.0/de/>

Table of Contents

1	Introduction	1012
2	Methodology	1015
2.1	Current model	1015
2.2	Correlation model	1017
2.3	Exploratory analysis	1018
3	Parameter estimation	1021
4	Results	1024
4.1	Model validation	1024
4.2	Effect of taking account of correlation	1027
4.3	Comparison to previous results	1030
5	Discussion	1030
6	Acknowledgements	1031
	References	1032

Regional probabilistic fertility forecasting by modeling between-country correlations

Bailey K. Fosdick¹

Adrian E. Raftery²

Abstract

BACKGROUND

The United Nations (UN) Population Division constructs probabilistic projections for the total fertility rate (TFR) using the Bayesian hierarchical model of Alkema et al. (2011), which produces predictive distributions of the TFR for individual countries. The UN is interested in publishing probabilistic projections for aggregates of countries, such as regions and trading blocs. This requires joint probabilistic projections of future country-specific TFRs, taking account of the correlations between them.

OBJECTIVE

We propose an extension of the Bayesian hierarchical model that allows for probabilistic projection of aggregate TFR for any set of countries.

METHODS

We model the correlation between country forecast errors as a linear function of time invariant covariates, namely whether the countries are contiguous, whether they had a common colonizer after 1945, and whether they are in the same UN region. The resulting correlation model is incorporated into the Bayesian hierarchical model's error distribution.

RESULTS

We produce predictive distributions of TFR for 1990-2010 for each of the UN's primary regions. We find that the proportions of the observed values that fall within the prediction intervals from our method are closer to their nominal levels than those produced by the current model.

¹ Statistical and Applied Mathematical Sciences Institute (SAMSI) and Duke University, USA. E-Mail: bfosdick@samsi.info.

² University of Washington, USA. E-Mail: raftery@u.washington.edu.

CONCLUSIONS

Our results suggest that a substantial proportion of the correlation between forecast errors for TFR in different countries is due to the countries' geographic proximity to one another, and that if this correlation is accounted for, the quality of probabilistic projections of TFR for regions and other aggregates is improved.

1. Introduction

The United Nations (UN) Population Division produces population estimates and projections for all countries every two years and publishes them in the biennial *World Population Prospects* (WPP). These projections are used by UN agencies and governments for planning, monitoring development goals, and as inputs to climate change and other models. They are also widely used by social and health science researchers and the private sector. The UN produces these population forecasts by projecting countries' age- and sex-specific fertility, mortality, and migration rates, and combining them to obtain age- and sex-specific population sizes using the standard cohort component method.

In this paper we focus on the fertility component. Country fertility in a given time period is summarized by the period total fertility rate (TFR), which is the average number of children a woman would bear if she lived past the end of the reproductive age span and at each age experienced the age-specific fertility rate of the given country and time period. Projections of future TFR are decomposed using forecasted age schedules to obtain projections of age-specific fertility rates.

The WPP reports three projection variants (low, medium, and high) for the population and vital rates based on expert opinion and models of historical patterns. The low and high variants correspond to TFR half a child below and above the medium value, respectively. One drawback of these projections is that the range given by the low and high variants has no probabilistic interpretation and hence does not reflect meaningful uncertainty in the forecasts.

For the 2010 WPP (United Nations 2011), the UN used as its medium projection the predictive median of TFR from a Bayesian hierarchical model developed by Alkema et al. (2011). We refer to this model as the "current model". This model produces predictive probability distributions of each country's TFR, although the distributions were not used in the 2010 WPP. The model is based on the demographic transition, where countries move from high birth and death rates to low birth and death rates, and is composed of three phases: before, during, and after the fertility transition. Predictions from this model

are typically summarized by the median country TFR prediction and the 80% or 95% prediction interval.

In addition to producing population estimates at the country level, the UN also provides projections for country aggregates such as geographic regions and trading blocs. The country TFR projections from the current Bayesian hierarchical model of Alkema et al. (2011) can be combined to obtain regional probabilistic TFR projections, provided the current model takes account of the dependence between countries' fertility rates. However, if dependence exists between country TFRs that is not accounted for in the Bayesian hierarchical model, treating the country-specific projections as independent may underestimate the uncertainty about the future TFRs and populations of aggregates.

Figure 1 shows the UN's 22 primary regions of the world and Table 1 summarizes the coverage probability of the out-of-sample TFR prediction intervals for these regions based on the current model. The coverage probabilities of the region-specific predictive intervals are smaller than the nominal levels, even though the country-specific coverages have been found to be approximately correct (Alkema et al. 2011). This suggests that the assumption of independent forecast errors may not be appropriate.

Figure 1: Primary regions of the world as identified by the UN

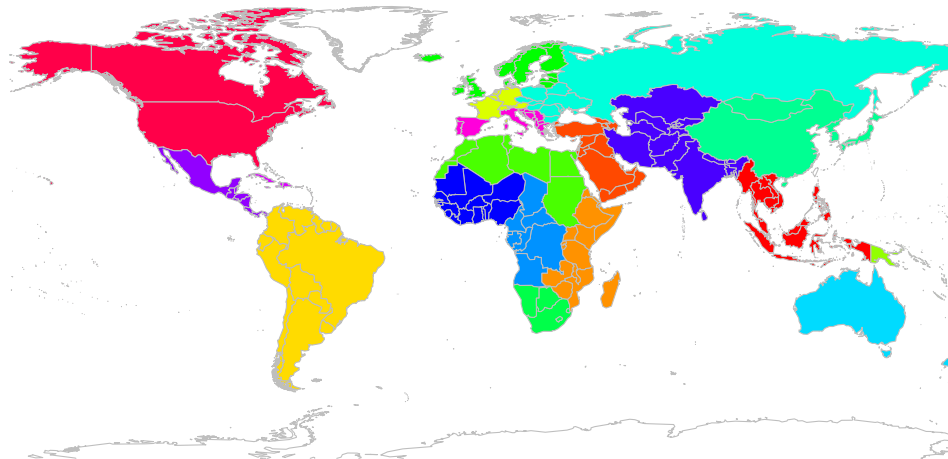


Table 1: Proportion of observed regional TFRs that fall within the specified out-of-sample prediction intervals (PI) obtained from the current Bayesian hierarchical model of Alkema et al. (2011)

Time Period	80% PI	90% PI	95% PI
1990-1995	0.73	0.86	0.95
1995-2000	0.68	0.73	0.86
2000-2005	0.59	0.73	0.82
2005-2010	0.73	0.82	0.91
All	0.68	0.78	0.89

The importance of modeling between-country correlations in stochastic projections is well recognized and others have proposed methods for addressing it. Lutz, Sanderson, and Scherbov (1997) combined stochastic projections, where the within and/or between region country correlation was zero or one, to incorporate forecast uncertainty due to between country correlation. This procedure was formalized by Alho and Spencer (2005) based on the use of random seeds and it was used to generate regional TFR forecasts in Statistics Netherlands (2005) (see also Alho et al. (2006)). Although this method is able to produce forecast errors with any desired marginal correlation, the individual forecasts come from a mixture distribution of two extreme scenarios, neither of which is realistic, and confidence intervals for aggregate values based on the empirical quantiles of the forecast simulations may not be valid. Lutz, Sanderson, and Scherbov (2001) generated correlated fertility forecast deviations assuming a between region correlation of 0.7, but provided no empirical justification for this value.

Keilman and Pham (2004) and Alho (2008) estimated correlations between TFR forecast errors for a set of European countries for which long and high quality time series data are available, and for which the TFRs have been low for a long time in most cases. Similarly, Wilson and Bell (2007) developed probabilistic population projections for Queensland and the rest of Australia using an empirical correlation between TFR errors. These studies represent best case scenarios, for which empirical estimates of the correlations are reasonably accurate and further modeling is probably unnecessary. This is not the case for our problem here: The number of empirical forecast errors is extremely limited for many countries, leading to undefined or unreliable empirical estimates of correlations.

In this article, we propose an extension to the Bayesian hierarchical model that produces TFR estimates for any aggregate of countries by modeling the residual correlation between country TFRs. Our extension adds a correlation structure to the error distribution in the hierarchical model, where the correlation between a pair of countries is modeled as

a linear function of time invariant covariates. Three covariates are chosen: whether two countries are contiguous, whether they had a common colonizer after 1945, and whether they are in the same UN region. This model provides estimates of the correlation between any pair of countries, even when empirical estimates are not available.

This paper is organized as follows. In the next section we review the current model and introduce the correlation model extension. We also describe the exploratory analyses that led to the choice of model extension. An estimation procedure based on a pairwise likelihood function is described, and model validation results are then presented for the prediction of the TFR in each of the UN's regions. We show theoretically which regional prediction intervals are most affected by the correlation model and compare the pairwise country correlation values from our model and those obtained in previous studies.

2. Methodology

2.1 Current model

The Bayesian hierarchical model of Alkema et al. (2011) divides the evolution of TFR in a country into three phases: before, during, and after the fertility transition. During the fertility transition, the TFR for country c in time period t , $f_{c,t}$, is modeled as following a systematic decline curve with normally distributed random errors. After the fertility transition is complete, the TFR is modeled as a first order autoregressive process that ultimately fluctuates around 2.1, which is considered replacement level fertility (see Section 5 for extensions that relax this assumption). If $\mathbf{f}_t = (f_{1,t}, \dots, f_{C,t})$ is the TFR for all countries at time t , the model can be written as follows:

$$\mathbf{f}_t = \mathbf{m}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(0, \Sigma_t = \text{diag}(\tilde{\sigma}_t^2)) \quad (1)$$

Fertility transition phase:

$$m_{c,t} = f_{c,t-1} - d(\boldsymbol{\theta}_c, f_{c,t-1}),$$

$$\tilde{\sigma}_{c,t} = \sigma_{c,t}(\boldsymbol{\theta}_c, f_{c,t-1}).$$

Post-transition phase:

$$m_{c,t} = 2.1 + 0.8859(f_{c,t-1} - 2.1),$$

$$\tilde{\sigma}_{c,t} = s = 0.1016,$$

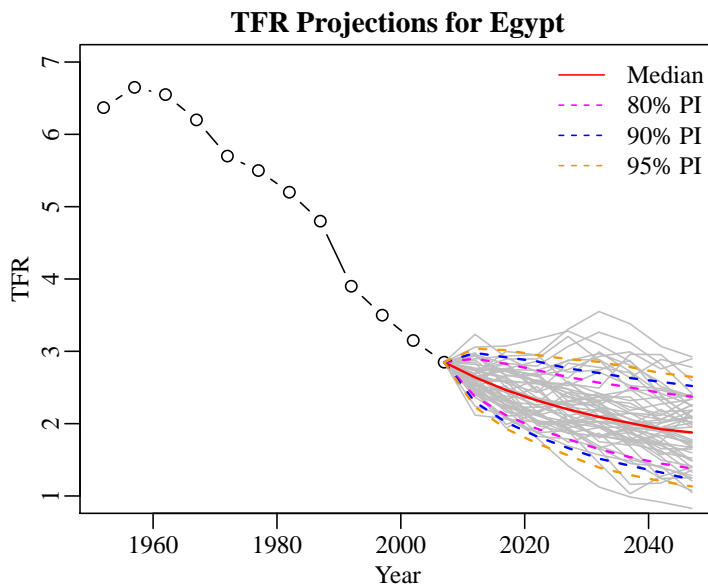
where $\text{diag}(\mathbf{x})$ is a diagonal matrix with the elements of the vector \mathbf{x} along the diagonal.

In (1), the quantities in bold font are vectors whose elements correspond to different countries, $d(\boldsymbol{\theta}_c, f_{c,t-1})$ is a double logistic function controlling the rate of the fertility decline, and $\boldsymbol{\theta}_c$ is a vector of country-specific parameters specifying country c 's unique decline pattern. The elements of $\boldsymbol{\theta}_c$ determine various aspects of the decline, such as the starting fertility level, the overall pace of the transition, and the timing and acceleration of the decline (see Alkema et al. (2011) for more details). The expected TFR in the next

time period, $m_{c,t}$, and the variances of the random errors, $\tilde{\sigma}_{c,t}^2$, differ in the transition and post-transition phases. Because a country's TFR is not modeled before the fertility decline, the vector \mathbf{f}_t for any time point t contains only those countries that have started or completed their fertility transition.

The data used to estimate the country parameters θ_c in the current model are the five-year time period TFR estimates from 1950 to 2010 in the 2010 WPP. A posterior distribution of the parameters is produced which indicates the probable values of the parameters given the data. In addition, a predictive distribution of TFR values for each country can be obtained by forecasting future values using the relations in (1). Figure 2 shows the predictive distribution of TFR for Egypt from 2010 to 2050. This distribution is summarized by the median prediction and the 80%, 90%, and 95% prediction intervals.

Figure 2: TFR projections (grey) and probabilistic prediction intervals for Egypt from 2010 to 2050 from the Bayesian hierarchical model



2.2 Correlation model

As discussed above, the regional TFR prediction intervals from the current model tend to be too narrow (see Table 1). This suggests there is excess correlation between countries' TFRs that is not accounted for in the current model. To capture this excess correlation, we propose modifying the error structure in (1) to allow for correlation between countries as follows:

$$\varepsilon_t \sim N(0, \Sigma_t = \text{diag}(\tilde{\sigma}_t) \cdot R_t \cdot \text{diag}(\tilde{\sigma}_t)). \quad (2)$$

The (i, j) element of the matrix R_t is the correlation between the TFR forecast errors (i.e. the deviations from the mean predicted values $m_{c,t}$) for country i and country j in time period t .

Our exploratory analyses, described below, indicated that the correlations had a different pattern when both countries had low fertility than otherwise, and our model allows for this. We sought to model the correlations using temporally stable characteristics of the country pairs, so that they could reasonably be used for projection. Thus, the elements of the correlation matrix are modeled as follows:

$$R_t[i, j] = \begin{cases} 1 & \text{if } i = j, \\ \rho_{i,j}^{(1)} & \text{if } f_{i,t-1} < \kappa \text{ and } f_{j,t-1} < \kappa, \\ \rho_{i,j}^{(2)} & \text{if } f_{i,t-1} \geq \kappa \text{ or } f_{j,t-1} \geq \kappa, \end{cases} \quad (3)$$

$$\rho_{i,j}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} \text{contig}_{i,j} + \beta_2^{(k)} \text{comcol}_{i,j} + \beta_3^{(k)} \text{sameRegion}_{i,j} \text{ for } k \in \{1, 2\}, i \neq j,$$

where $\text{contig}_{i,j} = 1$ if countries i and j are contiguous and 0 if not, $\text{comcol}_{i,j} = 1$ if they had a common colonizer after 1945, and $\text{sameRegion}_{i,j} = 1$ if they are in the same UN region. As mentioned above, a key feature of this model is that all covariates are time invariant, eliminating the need for a model that forecasts the covariates. If, for example, we modeled the correlation between two countries as a function of the ratio of their GDPs in the previous year, a model that predicted countries' future GDP would be necessary.

The correlation model in (3) states that when countries i and j both have TFR below κ , the correlation of their errors in the next time period is $\rho_{i,j}^{(1)}$, and when at least one of them has a TFR greater than κ , the correlation is $\rho_{i,j}^{(2)}$. In both cases, the correlation between two countries is modeled as a linear combination of the three pairwise country covariates. The parameters to be estimated therefore include the threshold κ , $\{\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)}\}$ for the correlation when both countries have TFR less than κ , and $\{\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)}\}$ for the correlation when at least one of the two TFRs is greater than κ .

Because the diagonal elements of R_t are equal to one, the joint predictive distribution of all country TFRs will have the same country marginal predictive distributions as those

from the current model. Thus expanding the model to allow for correlation will not change the marginal country-specific predictive distributions, which is desirable given the good performance of the current model for individual countries.

2.3 Exploratory analysis

Exploratory analysis of one-time-period-ahead forecast errors from the model of Alkema et al. (2011) and WPP data from 1950 to 2010 guided specification of the correlation model structure. For each time period and country, the forecast error is the difference between the observed TFR and the average predicted value given TFR in the previous time period. Estimating the correlations between these forecast errors is difficult because the estimates are based on a small number of data points (at most 11 five-year periods), and because the country-specific predictive means and variances are given by the Bayesian hierarchical model. To obtain empirical estimates of the correlation between the forecast errors for two countries, conditional on their predictive variances, we used the posterior mean with an arc-sine prior. This estimator was proposed by Fosdick and Raftery (2012), who showed it to have good small sample performance compared to other frequentist and Bayesian estimators.

Table 2 shows the number of overlapping five-year time periods from 1955 to 2010 for each country pair after both had started their fertility decline. These counts represent the number of forecast errors available to compute each correlation estimate. As a number of countries have only recently started their fertility decline, many pairwise correlation estimates were based on only a few observations or, in the case of only two overlapping time periods, were not computed at all. We therefore chose to model the correlation structure rather than directly use the noisy empirical estimates from the raw data.

Table 2: Numbers of unique country pairs with overlapping five-year time periods since the start of each country’s fertility decline

No. time periods	2	3	4	5	6	7	8	9	10	11
No. country pairs	389	385	1,131	3,009	1,635	2,550	2,508	2,750	1,978	2,775

The representation of TFR in the Bayesian hierarchical model differs when it is high (during the transition) compared to when it is low (post-transition), creating concern that country correlations might vary with TFR level as well. To investigate this, separate posterior mean correlation estimates were computed based on the forecast errors when both countries had TFR above three and when both countries had TFR below three. Table 3 shows the results from significance tests on these correlation estimates which examined

whether each estimate is non-zero (see Fosdick and Raftery (2012) for more details on these tests). Results are shown for only the correlation estimates that are based on at least five overlapping time periods. Based on 0.05-level tests, over 17% of the correlation estimates for country pairs with low TFRs were deemed significantly non-zero. This is much larger than the 5% we would expect if there was no correlation in forecast errors. The percentage significant when countries have high TFR was 3.3, close to that expected in the absence of correlation. The stark difference in the percentage of non-zero correlations when both countries have low TFR compared to when they both have high TFR led us to specify one model for the correlation when the TFRs of both countries are below a threshold κ , and a different model when at least one country has TFR above κ , where κ is to be estimated from the data.

Table 3: Results from level-0.05 significance tests on the correlation estimates testing whether each estimate is non-zero

	Number of country pairs	Number significant	Percent significant
Both country TFRs above 3	1857	60	3.3
Both country TFRs below 3	1356	235	17.3

Notes: Only the correlation estimates that were based on at least five overlapping time periods were considered here.

The average estimated correlation between countries in the same UN-defined region when both have TFR below 3 was 0.37, and for countries in different regions was 0.09, using only correlation estimates based on at least eight time periods. This suggests that the correlation between forecast errors at low TFR levels may be related to geographical proximity, and motivates modeling the correlation as a function of geographical predictors.

As our aim is to make long-term projections, we consider only predictors that are essentially time invariant. A database of country pairwise covariates is available from the Centre d'Études Prospectives et d'Informations Internationales (CEPII) (Mayer and Zignago 2006). In order to select which covariates to include in the correlation model, we used a Bayesian model averaging approach as implemented in the BMA package in the statistical program R (Raftery 1995; Raftery, Painter, and Volinsky 2005; Raftery et al. 2013). This method considers all regression models corresponding to subsets of the covariates and based on the evidence in the data in favor of each model, it computes the posterior probability of each covariate being in the model (i.e. the probability the variable's regression coefficient β is non-zero).

Table 4 shows the results for all variables in the CEPII database, as well as two additional indicator variables that signify whether pairs of countries are in the same UN area and same UN region. Note here we are only considering correlation estimates based on errors when both countries had TFR less than 3 and share at least five overlapping time periods. We chose to focus on estimates when TFR is low as these were determined to contain the most excess correlation and this correlation will have the most impact on long-term forecasts. The posterior inclusion probabilities for the covariates related to whether countries are contiguous (contig), share a common colonizer after 1945 (comcol), and are in the same UN region (sameRegion) are equal to one. This suggests there is strong evidence in the data for each of these covariates as predictors of country correlation. Barbieri and Berger (2004) showed that the single regression model with the predictors whose posterior inclusion probabilities are above 50% is predictively optimal, and so following them we choose the model with these three covariates.

Table 4: Output from the BMA package in R

	$P(\beta \neq 0)$	Estimate (β)	Std Deviation
intercept	100.0	.062	.001
contiguous (contig)	100.0	.211	.034
common colonizer after 1945 (comcol)	100.0	.270	.039
same UN region indicator (sameRegion)	100.0	.117	.020
colonial relationship after 1945	41.8	.056	.072
colonial link	19.5	.018	.039
were/are the same country	14.5	-.014	.039
distance weighted by city pop: harmonic mean	9.5	0	0
distance weighted by city pop: arithmetic mean	9.1	0	0
share a language spoken by at least 9%	2.5	0	.001
common official language	2.2	.001	.001
currently in a colonial relationship	0.0	0	0
geodesic distance by most important cities	0.0	0	0
geodesic distance by capital cities	0.0	0	0
same UN area indicator	0.0	0	0

Notes: The second column contains the posterior probability of each variable being in the regression model, the third and fourth columns contain the regression coefficient posterior mean estimates and posterior standard deviations, weighted and averaged over all possible models containing subsets of the covariates.

3. Parameter estimation

Our method for estimating the parameters of the correlation model in (3) relies on the one-time-period-ahead standardized forecast errors. The Bayesian hierarchical model of Alkema et al. (2011) was fit to the 2010 WPP TFR estimates from 1950 to 2010, and posterior distributions of θ_c given the data were obtained for each country. Using these parameter estimates and the TFR in a given time period, a predictive distribution of the expected TFR $m_{c,t}$ for the next time period was computed. The value of $m_{c,t}$ for a parameter vector θ_c is

$$\hat{m}_{c,t}|f_{c,t-1}^{WPP}, \theta_c = \begin{cases} f_{c,t-1}^{WPP} - d(\theta_c, f_{c,t-1}^{WPP}), & \text{during the fertility transition,} \\ 2.1 + \rho(f_{c,t-1}^{WPP} - 2.1), & \text{after the fertility transition,} \end{cases}$$

where $f_{c,t-1}^{WPP}$ is the 2010 WPP TFR estimate for country c at time period $t - 1$. For each sample $\theta_c^{(k)}$ from the posterior distribution, there is a corresponding expected TFR value at time t .

We define the parameter-specific standardized forecast error $e_{c,t}(\theta_c)$ for country c , time period t , and parameter vector θ_c as

$$e_{c,t}(\theta_c) = \frac{f_{c,t} - E[f_{c,t}|f_{c,t-1}, \theta_c]}{SD[f_{c,t}|f_{c,t-1}, \theta_c]} = \frac{f_{c,t}^{WPP} - \hat{m}_{c,t}|f_{c,t-1}^{WPP}, \theta_c}{\tilde{\sigma}_{c,t}}.$$

We define the standardized one-time-period-ahead forecast error $\bar{e}_{c,t}$ as the average over the posterior samples of the parameter-specific forecast errors, namely

$$\bar{e}_{c,t} = \frac{1}{K} \sum_{k=1}^K e_{c,t}(\theta_c^{(k)}),$$

where K is the number of parameter samples from the posterior distribution.

The standardized errors can be viewed as samples from a multivariate normal model with correlation matrix R_t , $\bar{e}_t(\theta) \sim N(\mathbf{0}, R_t)$ for $t = 1955, \dots, 2010$. Ideally we would estimate the correlation model parameters $\{\kappa, \beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)}, \beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)}\}$ via maximum likelihood estimation based on the multivariate normal model. However, this is made challenging by the fact that for any time period t , the vector \bar{e}_t contains standardized errors for only those countries that have started their fertility decline by time t , and that for many parameter values the estimated correlation matrix is not positive definite, making the likelihood undefined.

Instead, we took a pseudo-likelihood approach that approximates the multivariate normal likelihood by a product of bivariate normal likelihoods (Besag (1975), Lindsay

(1988)). This pseudo-likelihood is a composite marginal likelihood, and is more specifically referred to as a pairwise likelihood (PL) as it is the product of the marginal distributions of all country pairs (Cox and Reid (2004); see Varin, Reid, and Firth (2011) for a review). It is defined by

$$L_{\text{PL}}(\kappa, \boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)} | \bar{\mathbf{e}}) = \prod_{t=1}^T \prod_{i < j} \left[L_1(\rho_{i,j}^{(1)} | \bar{e}_{i,t}, \bar{e}_{j,t}) \cdot \mathbf{I}[(f_{i,t-1} < \kappa) \text{ and } (f_{j,t-1} < \kappa)] \right. \\ \left. + L_2(\rho_{i,j}^{(2)} | \bar{e}_{i,t}, \bar{e}_{j,t}) \cdot \mathbf{I}[(f_{i,t-1} \geq \kappa) \text{ or } (f_{j,t-1} \geq \kappa)] \right], \quad (4)$$

where T is the number of observed time periods and $\mathbf{I}(A)$ is an indicator function that equals 1 if the condition A is true and 0 otherwise. L_1 and L_2 are bivariate normal likelihoods with zero means, variances equal to one, and correlations $\rho_{i,j}^{(1)}$ and $\rho_{i,j}^{(2)}$, respectively.

The PL can be maximized separately over $\{\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)}\}$ and $\{\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)}\}$ for a fixed value of κ . For each value of the threshold κ from 0.5 to 9 children at intervals of 0.1, we estimated the model parameters by maximizing the PL in (4) numerically using a Nelder-Mead method subject to the constraint that the correlations $\rho_{i,j}^{(1)}$ and $\rho_{i,j}^{(2)}$, defined by the β s, lie in the interval $[-1, 1]$ for all i and j .

The PL was maximized at $\kappa = 5$ children, and the corresponding regression coefficients are shown in Table 5. These estimates mirror the exploratory analysis result that correlations are larger on average when both countries have lower TFR. At TFR values below κ , the correlation between two countries that are contiguous and in the same region but do not share a common colonizer is 0.46. The correlation between countries that are not in the same region but are contiguous is 0.37. The corresponding values when at least one TFR is above κ are 0.13 and 0.11, respectively. Country pairs with no colonial or geographic relationship have a correlation of 0.11 when both TFRs are below κ , and 0.05 otherwise. This illustrates that the correlation between two countries is associated with their geographic and colonial relationship.

Table 5: Parameter estimates for correlation model (3) with estimate of the threshold κ being 5

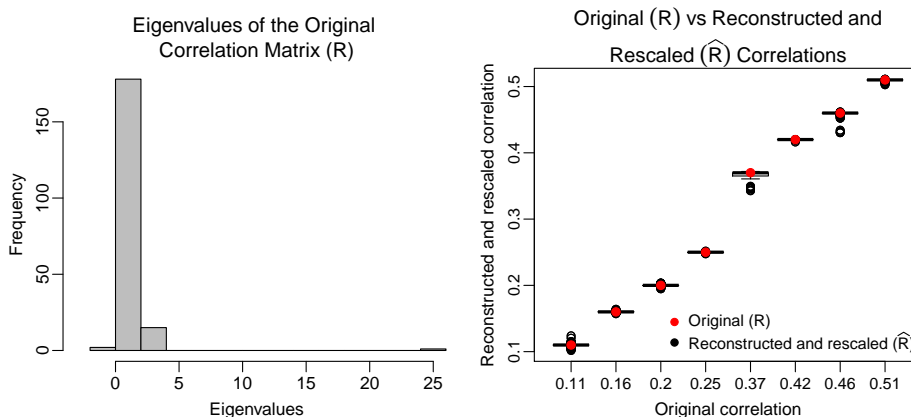
	intercept	contig	comcol	sameRegion
	β_0	β_1	β_2	β_3
Both country TFRs below κ	0.11	0.26	0.05	0.09
At least one country TFR greater than κ	0.05	0.06	0.00	0.02

For many time periods, the PL estimates of the parameters result in estimated correlation matrices R_t that are symmetric but not positive semidefinite. However, the correlation matrix must be positive semidefinite to use it for simulation of forecast errors. The symmetric positive semidefinite matrix closest in Frobenius norm to a given symmetric matrix is obtained by zeroing out all negative eigenvalues of the original matrix (Driessel 2007) and then reconstructing the matrix. Thus, at each time point t for which R_t is not positive semidefinite, we perform the following procedure:

1. Compute the eigenvalue decomposition of R_t to express it as $R_t = UDU^T$, where U is an orthogonal matrix of eigenvectors and D is a diagonal matrix of eigenvalues.
2. Replace all the negative eigenvalues by zero. The matrix D is thereby changed to \tilde{D} .
3. Compute the reconstructed matrix $\tilde{R}_t = U\tilde{D}U^T$.
4. The diagonal elements of \tilde{R}_t will not equal one unless the original matrix was positive semidefinite. Therefore, treat \tilde{R}_t as a covariance matrix and rescale it to obtain a reconstructed and rescaled correlation matrix \hat{R}_t to use in the projections.

Figure 3 illustrates the effect this reconstruction and rescaling has on the original correlations for the correlation matrix when all countries are in the post-transition phase. The histogram on the left shows that there were two small negative eigenvalues. The ratio of the absolute value of the sum of the negative eigenvalues to the sum of the positive eigenvalues is approximately 0.0008. This confirms what is seen in the plot on the right in Figure 3, that zeroing out the negative eigenvalues has very little effect on the correlations. The rescaling of the correlation matrix in step 4 ensures that the predictive distribution of TFR for any individual country remains the same as from the current Bayesian hierarchical model of Alkema et al. (2011). Thus, only joint predictive distributions of the TFRs in more than one country are affected. Note that the matrix approximation \hat{R}_t that results from this procedure is singular unless the original matrix R_t is positive definite, but the predictive distributions remain well defined.

Figure 3: Histogram of the eigenvalues of the correlation matrix R_t assuming all countries are in the post-transition phase (left plot) and boxplots of the correlation values from the reconstructed and rescaled correlation matrix \widehat{R}_t when all countries are in the post-transition phase (right plot). Original correlation values represented by red dots



4. Results

4.1 Model validation

We assessed the model by estimating the current hierarchical model parameters from the data for 1950 to 1990, projecting regional TFR for the UN's 22 primary regions from 1990 to 2010 using the error correlation structure, and comparing the probabilistic projections with the actual observations for the four held-out five-year periods. We approximated regional TFR by a weighted average of country-specific TFRs, with weights proportional to the current female populations of each country. A similar approximation was used for regional life expectancy by Raftery et al. (2013b).

Posterior distributions of the Bayesian hierarchical model parameters were obtained based on data from 1950 to 1990. The parameter values in Table 5 were used in the correlation model and not re-estimated based on the restricted data set, since these estimates are already based on very limited data. Projections of TFR were obtained for the four five-year time periods from 1990 to 2010 under the current model assuming independent

errors and using our proposed error correlation structure. The predictive distributions of the weighted average TFR for each of the 22 regions were compared to the observed weighted average values.

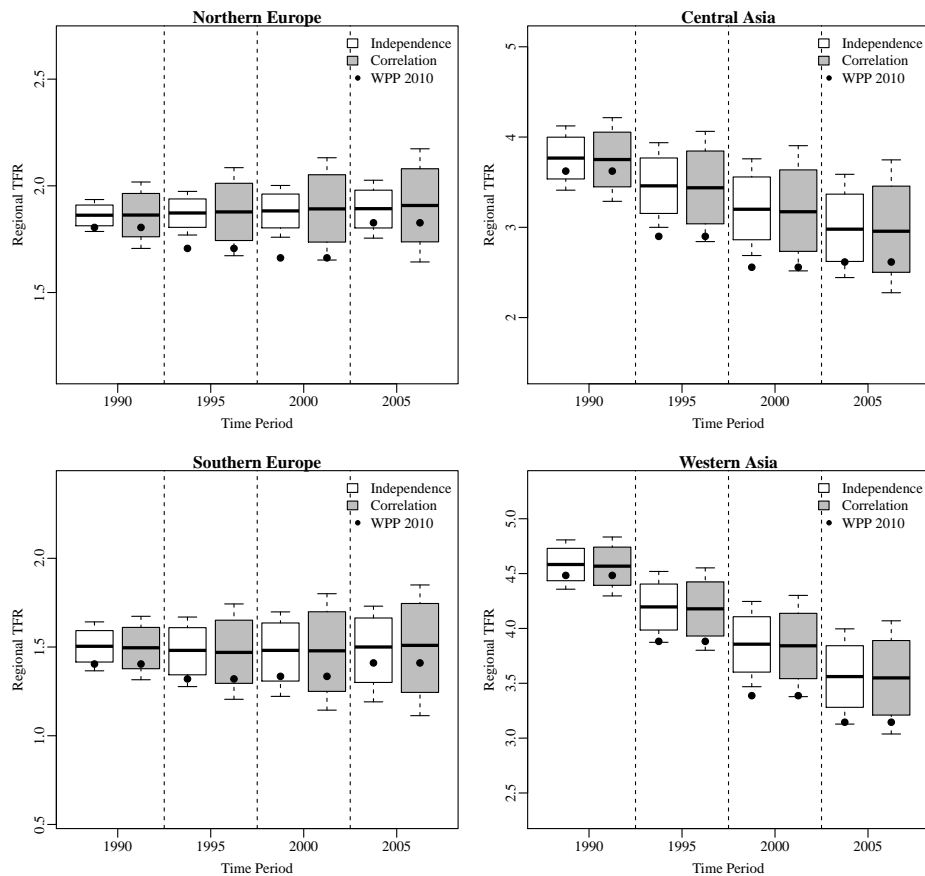
Table 6 shows the proportion of observed weighted averages that fell within the 80%, 90%, and 95% prediction intervals from both approaches. In each case the observed proportion was closer or as close to the theoretical value under the correlation model than under the independence model.

Table 6: Proportion of observed regional weighted average TFRs that fall within the specified prediction intervals (PI)

Time Period	Model	80% PI	90% PI	95% PI
1990-1995	Independence	0.73	0.86	0.95
	Correlation	0.86	0.91	0.95
1995-2000	Independence	0.68	0.73	0.86
	Correlation	0.73	0.86	0.95
2000-2005	Independence	0.59	0.73	0.82
	Correlation	0.64	0.73	0.95
2005-2010	Independence	0.73	0.82	0.91
	Correlation	0.77	0.86	0.91
All	Independence	0.68	0.78	0.89
	Correlation	0.75	0.84	0.94

Figure 4 shows the posterior distribution of regional TFR for four regions, along with the observed regional value. The box associated with a given period and projection method represents the 80% interval and the ends of the whiskers correspond to the 95% interval. For Northern Europe the current model prediction intervals do not cover the observed value in 1995 and 2000, but the correlation model prediction intervals do. Similar patterns are seen in the other regions: The prediction intervals based on the correlation model are wider, reflecting greater uncertainty, and contain more of the observed regional TFR values than the current model assuming independent errors.

Figure 4: Boxplots showing the 80% and 95% prediction intervals for the regional weighted average TFR for the current model assuming independent errors and that with the correlation error structure. Boxes represent the 80% prediction interval, ends of the whiskers mark the 95% prediction interval. The corresponding observed average TFR based on the 2010 WPP is also shown.



As the estimated correlations are larger when both countries have low TFR values, bigger differences between the current model and the correlation model prediction intervals are seen for regions such as Northern Europe and Central Asia, for which the majority of the countries have completed most of the fertility decline. Regions that have few countries with TFR less than 5, such as Eastern and Western Africa, showed little change between the prediction intervals from the current model and the correlation model, as expected.

4.2 Effect of taking account of correlation

For a given parameter vector θ and time period t , the effect of taking account of correlation on the variance of the regional TFR one-time-period-ahead can be quantified analytically. We denote by p_i the proportion of the region's female population that resides in country i , by f_i the TFR in country i in time period t , and by N the number of countries in the region. The regional weighted average TFR is then $\sum_i p_i f_i$ where the sum is over all countries in the region.

If the forecast errors are assumed to be independent as in the current model, the predictive variance of the regional TFR in time period t given all TFRs in time period $t - 1$ is $\text{Var}[\sum_i p_i f_i] = \sum_i p_i^2 \text{Var}[f_i]$, where $\text{Var}[f_i] = \tilde{\sigma}_{i,t}^2$. As we project TFR into the future, eventually all countries will be in the last phase of the model, having completed their fertility transition, where $\text{Var}[f_i] = s^2$. When all countries in the region are in the post-transition phase,

$$\text{Var} \left[\sum_i p_i f_i \right] = s^2 \sum_i p_i^2 \quad (5)$$

under the current model.

We will refer to $\sum_i p_i^2$ as the independence factor (IF) because it represents the ratio of the regional variance to the country-specific variance assuming independence in the post-transition phase. It indicates the effect of the distribution of the population across countries in the region and shows that the more evenly the regional female population is spread amongst the countries within the region, the greater the variability in the regional estimate. This comes from the fact that the IF is maximized when each $p_i = \frac{1}{N}$ for all countries in the region.

The variance of a region's TFR one-time-period-ahead under the correlation model is

$$\text{Var} \left[\sum_i p_i f_i \right] = \left(\sum_i p_i^2 \text{Var}[f_i] + 2 \sum_{i < j} p_i p_j \sqrt{\text{Var}[f_i] \text{Var}[f_j]} R_t[i, j] \right).$$

When all countries have completed the fertility transition, this becomes

$$\text{Var} \left[\sum_i p_i f_i \right] = s^2 \left(\sum_i p_i^2 + 2 \sum_{i < j} p_i p_j R_t[i, j] \right). \quad (6)$$

We will refer to $\left(\sum_i p_i^2 + 2 \sum_{i < j} p_i p_j R_t[i, j] \right)$ as the dependence factor (DF) as it is the multiplicative factor in the variance under the correlation model. Equation (6) shows that the larger the country correlations, especially between those countries with a relatively high proportion of the regional female population, the larger the one-time-period-ahead variance of the regional TFR.

The ratios of the dependence factor to the independence factor for the 22 UN regions are shown in Table 7. The regions with the largest ratios are those whose predictive distributions are most impacted by between-country correlations. For example, Western Asia and Eastern Africa's ratios are both greater than 2.5, indicating that the variance of the regional TFR predictive distributions is more than 2.5 times greater for the model with the correlation structure than that from the current model. Those regions with ratios close to 1, such as Northern America and Australia/New Zealand, have similar predictive distributions from the two models. These same phenomena are reflected in the 95% one-time-period-ahead prediction interval widths, which show greater differences between the model with and without correlation for those regions with high ratios and small differences for those regions with small ratios.

Both the between-country correlations and the proportion of the regional female population within each country influence the effect of the correlations. The number of countries in the region and the proportion of the regional female population that live in the largest country are also shown in Table 7. If a high proportion of the female population lives in a single country, correlations will not have a large effect on the prediction intervals for the region. Examples of this include Northern America and Eastern Asia which have low DF/IF ratios and high proportions of the female population in a single country. Overall, the regions whose predictive intervals will be most highly affected by between-country correlations in the future include Eastern and Northern Africa, Western and Central Asia, Eastern and Western Europe, South America, and the Caribbean.

Table 7: Effect of correlation model on variance of regional weighted average TFRs

Region	DF/IF	95% Pred. Int. Width		Max Proportion	N
		No Corr.	With Corr.		
Eastern Africa	3.02	0.14	0.25	0.26	18
Western Asia	2.57	0.16	0.26	0.33	18
Central Asia	2.10	0.22	0.32	0.45	5
South America	1.95	0.21	0.30	0.50	13
Caribbean	1.94	0.18	0.25	0.27	16
Northern Africa	1.92	0.20	0.27	0.39	7
Western Europe	1.91	0.22	0.31	0.43	7
Eastern Europe	1.87	0.21	0.29	0.49	10
South-Eastern Asia	1.76	0.19	0.26	0.40	10
Western Africa	1.67	0.22	0.28	0.52	16
Southern Europe	1.65	0.20	0.26	0.39	12
Middle Africa	1.65	0.23	0.29	0.52	9
Southern Asia	1.36	0.29	0.34	0.71	9
Northern Europe	1.34	0.26	0.30	0.63	11
Polynesia	1.33	0.24	0.28	0.48	3
Central America	1.25	0.29	0.33	0.73	8
Micronesia	1.22	0.29	0.32	0.62	2
Southern Africa	1.14	0.35	0.37	0.87	5
Melanesia	1.12	0.32	0.33	0.78	5
Northern America	1.10	0.36	0.38	0.90	2
Eastern Asia	1.09	0.34	0.35	0.85	8
Australia/New Zealand	1.08	0.34	0.35	0.83	2

Notes: The ratio of the dependence factor to the independence factor (DF/IF) indicates the multiplicative increase in the variance of the regional TFR when using the correlation model compared to the current model where forecast errors are assumed independent. Similarly, the 95% prediction interval widths using the no correlation (independence) model and the model with correlation demonstrate the absolute change in uncertainty between the two models when predicting TFR one-time-period-ahead. The "max proportion" column shows the largest proportion of the current region female population that is attributed to a single country. N is the number of countries in the region.

4.3 Comparison to previous results

Others have investigated the correlation between country forecast errors and obtained similar results to those reflected by the correlation model here. Keilman and Pham (2004) modeled TFR in 18 countries in the European Economic Area (EEA) with an autoregressive conditional heteroscedastic model and calculated the average correlation between country TFR errors to be 0.33. Although not all countries in the EEA are within the same UN region, in our model two countries with low TFR that are in the same region and not contiguous have a correlation of 0.20 and those not in the same region but contiguous have a value of 0.37.

When Alho (2008) further studied the correlation matrix obtained by Keilman and Pham (2004) (see also Statistics Netherlands (2005)), he found a stark contrast between the correlations between the Mediterranean countries (Portugal, Spain, Italy, and Greece) and all others. His estimate of the average correlation between forecast errors in Mediterranean and non-Mediterranean countries was 0.12 and the correlation within each of these groups was 0.3. Recall that in our model, the correlation between countries that have low TFR and have no geographic or colonial relation is 0.11.

Wilson and Bell (2007) modeled TFR using a random walk with drift and found the correlation between errors for Queensland and the rest of Australia to be 0.40. According to our correlation model, when TFR is less than 5, as it has been in Australia for many decades, the correlation between Australia and a hypothetical country contiguous to it would be 0.46. This is comparable to the result of Wilson and Bell (2007).

5. Discussion

When producing probabilistic population projections for country aggregates, it is critical to take account of between-country correlations in forecast errors of vital rates (Lutz 1996; Lee 1998; Bongaarts and Bulatao 2000). In this paper we have proposed a method for estimating between-country correlations in forecast errors of the TFR for all countries and using them to produce probabilistic TFR forecasts for aggregates of countries such as regions. For many country pairs there are few relevant data available, and so we estimate the correlations by modeling them as a function of three time-invariant predictors.

The resulting method yields the same probabilistic projections of TFR for individual countries as the Bayesian hierarchical model of Alkema et al. (2011), which was used by the UN for its (deterministic) medium population projections for all countries in the 2010 WPP (United Nations 2011). In an out-of-sample validation experiment, our correlation extension yielded better coverage of predictive intervals than the current model of Alkema et al. (2011) that does not explicitly take account of between-country correlations. The posterior samples produced by our method can be incorporated into probabilistic popula-

tion projections in the same way as those produced by the current method of Raftery et al. (2012).

The assumption in (1) that in the long-term (post-transition phase) every country TFR will fluctuate around 2.1 can be relaxed. Raftery et al. (2013a) proposed an extension that allows each country to have its own long-term mean TFR. This extension has been incorporated into the `bayesTFR` package in R, along with the TFR correlation model presented here (Ševčíková, Alkema, and Raftery 2011). The 2010 WPP data used in this article are available in the `wpp2010` R package (Ševčíková and Gerland 2013).

While we have discussed procedures for aggregating country-specific TFR projections, the methods proposed here could be adapted to address excess correlation between any set of stochastic forecasts, such as mortality rates, migration rates, etc. The key task would be obtaining a set of time-invariant covariates that explain the between-country correlation. The existence of excess correlation is a function of the mean model specification and may not always be present. For example, Raftery et al. (2013b) found no evidence for cross country correlation in life expectancy forecasts. Nonetheless depending on the complexity and mean model fit, excess correlation may exist, and adapting the methods presented here may provide better uncertainty estimates for aggregate quantities.

6. Acknowledgements

This work was supported by NIH grants R01 HD054511 and R01 HD070936, and by Science Foundation Ireland under E.T.S. Walton visitor award 11/W.1/I2079. The authors thank Leontine Alkema, Samuel Clark, Patrick Gerland, Nial Friel, and the members of the University College Dublin Working Group on Statistical Learning for their helpful comments and discussion.

References

- Alho, J. (2008). Aggregation across countries in stochastic population forecasts. *International Journal of Forecasting* 24(3): 343–353. doi:10.1016/j.ijforecast.2008.05.003.
- Alho, J. and Spencer, B. (2005). *Statistical Demography and Forecasting*. New York: Springer.
- Alho, J., Alders, M., Cruijsen, H., Keilman, N., Nikander, T., and Pham, D.Q. (2006). New forecast: Population decline postponed in Europe. *Statistical Journal of the United Nations Economic Commission for Europe* 23(1): 1–10.
- Alkema, L., Raftery, A.E., Gerland, P., Clark, S.J., Pelletier, F., Buettner, T., and Heilig, G.K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography* 48(3): 815–839. doi:10.1007/s13524-011-0040-5.
- Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *Annals of Statistics* 32(3): 870–897. doi:10.1214/009053604000000238.
- Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society, Series D (The Statistician)* 24(3): 179–195. doi:10.2307/2987782.
- Bongaarts, J. and Bulatao, R.A.E. (2000). *Beyond Six Billion*. Washington, D.C.: National Research Council.
- Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91(3): 729–737. doi:10.1093/biomet/91.3.729.
- Driessel, K.R. (2007). Computing the best positive semi-definite approximation of a symmetric matrix using a flow. [electronic resource]. Institute for Mathematics and its Applications: Applications in Biology, Dynamics, and Statistics. <http://www.ima.umn.edu/AlgGeom/W3.5-9.07/activities/Driessel-Kenneth/poster.pdf>.
- Fosdick, B.K. and Raftery, A.E. (2012). Estimating the correlation in bivariate normal data with known variances and small sample sizes. *The American Statistician* 66(1): 34–41. doi:10.1080/00031305.2012.676329.
- Keilman, N. and Pham, D.Q. (2004). Empirical errors and predicted errors in fertility, mortality and migration forecasts in the European Economic Area. Research Department of Statistics Norway.
- Lee, R.D. (1998). Probabilistic approaches to population forecasting. *Population and Development Review* 24: 156–190. doi:10.2307/2808055.
- Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80: 221–39. doi:10.1090/conm/080/999014.

- Lutz, W., Sanderson, W., and Scherbov, S. (1997). Doubling of world population unlikely. *Nature* 387: 803–805. doi:10.1038/42935.
- Lutz, W., Sanderson, W., and Scherbov, S. (2001). The end of world population growth. *Nature* 412: 543–545. doi:10.1038/35087589.
- Lutz, W.E. (1996). *The Future Population of the World. What Can We Assume Today?* London: Earthscan.
- Mayer, T. and Zignago, S. (2006). Notes on CEPII's distance measures. [electronic resource]. CEPII. http://www.cepii.fr/welcome_en.asp.
- Raftery, A.E., Li, N., Ševčíková, H., Gerland, P., and Heilig, G.K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences* 109(35): 13915–13921. doi:10.1073/pnas.1211452109.
- Raftery, A.E., Painter, I., and Volinsky, C.T. (2005). BMA: An R package for Bayesian Model Averaging. *R News* 5(2): 2–8.
- Raftery, A., Hoeting, J.A., Volinsky, C., Painter, I., and Yeung, K.Y. (2013). BMA: Bayesian model averaging. [electronic resource]. R package version 3.16.1. <http://CRAN.R-project.org/package=BMA>.
- Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology* 25: 111–164. doi:10.2307/271063.
- Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92(437): 179–191. doi:10.1080/01621459.1997.10473615.
- Raftery, A., Alkema, L., and Gerland, P. (2013). Bayesian population projections for the United Nations. *Statistical Science* 28: in press.
- Raftery, A., Chunn, J.L., Gerland, P., and Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* 50(3): 777–801. doi:10.1007/s13524-012-0193-x.
- Ševčíková, H., Alkema, L., and Raftery, A.E. (2011). bayesTFR: An R package for probabilistic projections of the total fertility rate. *Journal of Statistical Software* 43(1): 1–29.
- Ševčíková, H. and Gerland, P. (2013). wpp2010: World Population Prospects 2010. [electronic resource]. <http://CRAN.R-project.org/package=wpp2010>.
- Statistics Netherlands (2005). Changing population of Europe: Uncertain future. Brussels: UPE Final Report.

United Nations (2011). World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* 21: 5–42.

Wilson, T. and Bell, M. (2007). Probabilistic regional population forecasts: The example of Queensland, Australia. *Geographical Analysis* 39(1): 1–25. doi:[10.1111/j.1538-4632.2006.00693.x](https://doi.org/10.1111/j.1538-4632.2006.00693.x).