Regional Probabilistic Fertility Forecasting by Modeling Between-Country Correlations

Bailey K. Fosdick and Adrian E. Raftery
University of Washington

October 14, 2012

Abstract

The United Nations Population Division releases country fertility estimates and projections every two years, currently using Alkema et al. (2011)'s model for total fertility rate (TFR). This Bayesian hierarchical model produces a predictive distribution of TFR for each country. We extend this model to allow probabilistic projection of the TFR for any set of countries, such as a region or trading bloc. We model the correlation between country TFRs that is not captured by the original model 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. This correlation structure is incorporated into the original model's error distribution and is shown to improve the calibration of predictive intervals for the future TFR of regions.

1 Introduction

The United Nations (UN) Population Division produces population estimates and projections every two years for all countries, and publishes them in the biennial *World Population Prospects* (WPP). For the 2010 report, the UN used a Bayesian hierarchical model for total fertility rate (TFR) developed by Alkema et al. (2011). This model produces probabilistic country fertility prediction intervals based on the posterior distribution of each country's TFR.

In addition to producing population estimates at the country level, the UN also provides projections for numerous regions of interest, such as geographical regions and trading blocs. Treating the country-specific projections as independent when obtaining probabilistic projections of regional aggregates can severely underestimate uncertainty about future regional TFR if there is excess dependence between country TFRs that is not accounted for in the Bayesian hierarchical model. To address this, we propose an extension to the Bayesian hierarchical model that produces regional TFR estimates for any specified region by modeling the residual correlation.

Alkema et al. (2011)'s Bayesian hierarchical model is based on the demographic transition, where countries move from high birth and death rates to low birth and death rates. The model is composed of two phases: 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 $f_{c,t} = f_{c,t-1} - d_{c,t}(\theta_c, f_{c,t-1}) + \epsilon_{c,t}$,

where $d_{c,t}(\theta_c, f_{c,t-1})$ is a drift term of systematic decline, $\epsilon_{c,t}$ is a normally distributed random distortion, and θ_c is a vector of country specific parameters. After the fertility transition is complete, the TFR is modeled as an AR(1) process centered about 2.1, which is considered replacement level fertility. This model produces a predictive distribution of TFR in future time periods. These predictions are typically summarized by the median TFR estimate and the 80% or 95% prediction intervals.

Even though the original model is hierarchical and shares information across countries, correlation was found to exist between country residuals. The optimal method for accounting for such dependence may be to incorporate it into the Bayesian hierarchical model. However, this would vastly increase the complexity of the original model so here we develop a simpler way of modeling it. Since the country TFR posterior predictive distributions given by the original model have been validated, a requirement of any extension to this model is that country marginal distributions remain unchanged.

2 Methods

Exploratory analysis of the one-time-period-ahead forecast errors (i.e. the differences between the a predicted values and the observed TFRs in the next time period) suggests that excess correlation mostly exists in Phase II when TFR values are low. As countries progress in their demographic transition, eventually all countries will be in Phase II of the model so estimating correlation between countries at low TFR is most important for future projections. However, projections in the near future can be improved by modeling the correlation at all TFR levels. Thus, we propose a piecewise correlation model that distinguishes between time periods when both countries have low TFR and when at least one has a larger TFR.

The Bayesian hierarchical model with our correlation structure extension can be written in terms of the two fertility transition phases, with additional structure on the error components. If $\boldsymbol{f_t} = (f_{c_1,t}, ..., f_{c_n,t})$ is the TFR for all countries at time t, the model can be written in the following way:

Model:
$\boldsymbol{f_t} = \boldsymbol{m_t} + \boldsymbol{\epsilon_t} \qquad \boldsymbol{\epsilon_t} \sim N(0, \Sigma_t = \widetilde{\boldsymbol{\sigma}}_{\boldsymbol{t}}^{\boldsymbol{T}} R_t \widetilde{\boldsymbol{\sigma}}_{\boldsymbol{t}})$

During fertility transition:
$m_{c,t} = f_{c,t-1} - d_t(\theta_c, f_{c,t-1})$
$\widetilde{\sigma}_{c,t} = \sigma_{c,t}$

After fertility transition:
$m_{c,t} = 2.1 + 0.9(f_{c,t-1} - 2.1)$
$\widetilde{\sigma}_{c,t} = s$

Correlation matrix (R_t):
$R_t[i,i] = 1$
$R_t[i,j] = \mathrm{I}\left[(f_{c_i,t-1} < \kappa) \cap (f_{c_j,t-1} < \kappa)\right] \cdot \rho_{ij}^{(1)} + \mathrm{I}\left[(f_{c_i,t-1} \geq \kappa) \cup (f_{c_j,t-1} \geq \kappa)\right] \cdot \rho_{ij}^{(2)} \quad \text{for } i \neq j$

where I[$\cdot$] represents an indicator function that is one if the condition inside the brackets is true and zero otherwise. The bold terms represent vectors and R_t contains the correlation between country deviations from the original model predicted values. When both countries i

and j have TFR below κ, their correlation is $\rho_{ij}^{(1)}$, and the correlation is $\rho_{ij}^{(2)}$ in time periods when at least one of them has a TFR greater than κ. By restricting the diagonals of R_t to be one, the joint posterior distribution with have the same country marginal distributions as those in the original model.

Since some countries began their fertility decline in the last few decades and many have completed their decline, sample correlation estimates of $\rho_{ij}^{(1)}$ and $\rho_{ij}^{(2)}$ are not available for all country pairs, regardless of the value of κ. Therefore, a model is needed to provide estimates of these correlations. The Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) is a French research center that provides information on country pairwise characteristics, such as whether two countries are contiguous or share an official language (Mayer and Zignago (2006)). The posterior mean correlation assuming an arc-sine prior was computed for countries that have at least three time periods of observed TFR values after the start of their fertility decline (Fosdick and Raftery (2012)). Based on regression analyses involving the estimated correlations and covariates from CEPII, we found that the variables that are most highly predictive of country correlation are whether two countries are continuous (contig), had a common colonizer after 1945 (comcol), and are within the same UN region (sameRegion). This suggested the following linear model for country correlation:

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

The parameters in the correlation model include κ as the cutoff between the two pieces of the model, $\{\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)}\}$ 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 TFR is greater than κ.

To estimate the parameters of the correlation model, the posterior mean estimate of the standardized forecast residual $\overline{e}_{c,t}$ was computed for each country c and time period t. Ideally, we would maximize the multivariate normal likelihood of the correlation parameters in R_t given these residuals. However, this is difficult as many values of the parameters result in correlation matrices that are negative definite. To circumvent this problem, we define a pseudo-likelihood function, which acts as a surrogate for the multivariate normal likelihood, and use it to obtain parameter estimates.

Define the *Aggregated Partial Likelihood (APL)* to be

$$\begin{aligned}\mathrm{L}_{\mathrm{APL}}(\kappa, \boldsymbol{\rho^{(1)}}, \boldsymbol{\rho^{(2)}}|\overline{\boldsymbol{e}}) = \prod_{t=1}^{T}\prod_{i<j}\Big[&\mathrm{L}_1(\rho_{ij}^{(1)}|\overline{e}_{i,t}, \overline{e}_{j,t}) \cdot \mathrm{I}\left[(f_{i,t-1} < \kappa) \cap (f_{j,t-1} < \kappa)\right] \\ &+ \mathrm{L}_2(\rho_{ij}^{(2)}|\overline{e}_{i,t}, \overline{e}_{j,t}) \cdot \mathrm{I}\left[(f_{i,t-1} \geq \kappa) \cup (f_{j,t-1} \geq \kappa)\right]\Big]\end{aligned}$$

where T is the number of observed time periods, and L_1 and L_2 are bivariate normal likelihoods with correlations $\rho_{ij}^{(1)}$ and $\rho_{ij}^{(2)}$, respectively. This can be maximized in terms of $\{\kappa, \beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)}, \beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)}\}$ using the Nelder-Mead optimization procedure in `R` for fixed values of κ. Table 1 shows the parameter values corresponding to the largest APL likelihood value obtained over κ values $\{0.5, 0.6, 0.7, ..., 8.9, 9.0\}$.

As mentioned above, the full country correlation matrix at a given time period from the correlation model is often negative semidefinite, creating an invalid covariance matrix for prediction. The symmetric positive semidefinite matrix closest in Frobenius norm to a given

symmetric negative definite matrix is obtained by zeroing out all negative eigenvalues of the original matrix (Driessel (2007)). This mathematical approximation is used during prediction for each time period in which a correlation matrix that is not positive definite is obtained. The resulting matrix is treated as a covariance matrix and rescaled to be a correlation matrix with diagonal elements equal to one to ensure the marginal distribution of country predictions are unchanged.

Table 1: Parameter values that maximize the APL.

	$\beta^{(1)}$: Both country TFRs below κ				$\beta^{(2)}$: At least one country TFR greater than κ			
κ	intercept	contig	comcol	sameRegion	intercept	contig	comcol	sameRegion
5	0.11	0.26	0.05	0.09	0.05	0.06	0.00	0.02

3 Results

The primary motivation for modeling the excess correlation between country TFR forecast errors is to construct regional TFR estimates. This requires individual country TFR values, as well as the age-specific population of childbearing women in each country. We evaluate the correlation model extension using the prediction of a weighted average of TFR values for countries within each of the UN's 22 primary regions in the last twenty years. Country weights are equal to the proportion of the region female population that currently resides in each country. A similar approximation was used for regional life expectancy in Raftery et al. (2012). Posterior distributions of the Bayesian hierarchical model parameters were obtained based on data from 1950-1990 and projections were made for the four 5-year time periods from 1990-2010 under the original model assuming independent errors and the extended model which estimates country correlations. The posterior distributions of the weighted average TFRs for each of the 22 regions was compared to the observed weighted average values.

Table 2 shows the proportion of observed weighted averages that fell within the 80%, 90%, and 95% posterior prediction intervals. A greater proportion of observed regional TFR averages are contained in the extended model prediction intervals, suggesting that this additional correlation structure is more accurately representing the variability in regional TFRs. Figure 1 shows box plots of the posterior distribution of regional average TFR for four regions with the observed regional average shown in red. The box associated with a given period and projection method represents the 80% prediction interval and the ends of the whiskers correspond to the 95% interval. Since the correlation parameter estimates, $\beta_j^{(k)}$, are larger in the portion of the model when both countries have low TFR values, bigger differences between the original model and this extended model are seen for regions like Northern and Southern Europe, for whom the majority of the countries have completed most of the fertility decline. Regions that have very few countries with TFR less than 5, such as Eastern and Western Africa, showed little change in the predictive intervals from the original model and model with the correlation extension.

Table 2: Proportion of observed regional weighted average TFRs that fall within the specified prediction intervals.

Time Period	Model	80% CI	90% CI	95% CI
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

4 Conclusion

We have developed a correlation model extension to the Alkema et al. (2011) TFR model that accounts for excess country dependence and allows for inference about TFR at the regional level. The new model is shown to improve upon the original model by providing more accurate prediction interval coverage. As time passes and more data becomes available, the correlation model parameters can be reestimated to further improve estimation and prediction accuracy.

References

Alkema, L., A. E. Raftery, P. Gerland, S. J. Clark, F. Pelletier, T. Buettner, and G. K. Heilig (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography 48*, 815–839.

Driessel, K. R. D. (2007). Computing the Best Positive Semi-Definite Approximation of a Symmetric Matrix Using a Flow. *Institute for Mathematics and its Applications: Applications in Biology, Dynamics, and Statistics March*.

Fosdick, B. K. and A. E. Raftery (2012). Estimating the Correlation in Bivariate Normal Data with Known Variances and Small Sample Sizes. *The American Statistician 66*(1), 34–41.

Mayer, T. and S. Zignago (2006). Notes on CEPIIs distances measures. *CEPII*.

Raftery, A., J. L. Chunn, P. Gerland, and H. Ševčíková (2012). Bayesian probabilistic projections of life expectancy for all countries. *Demography 49*, to appear.

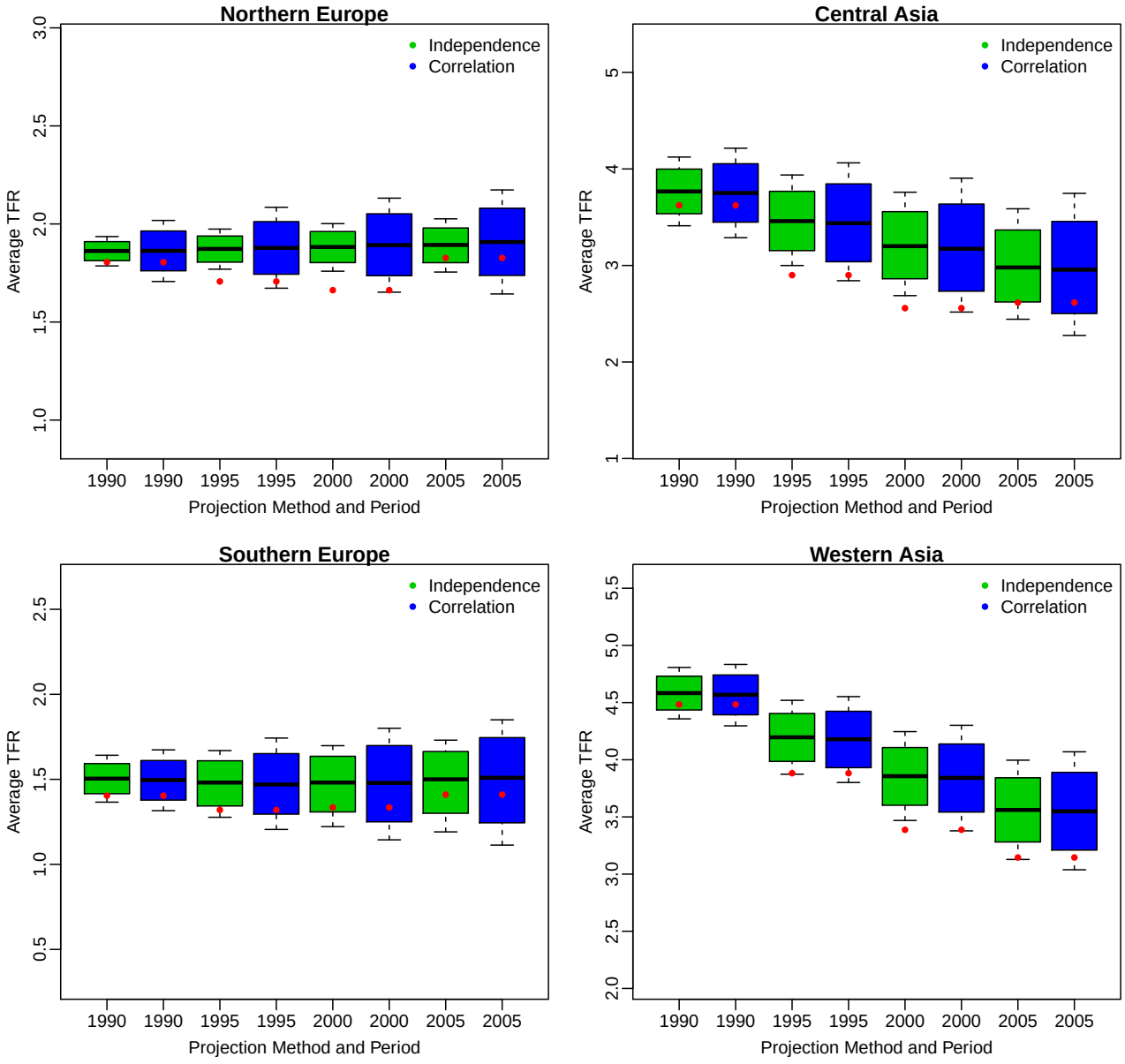

Figure 1: Boxplots show the 80% and 95% prediction intervals for the regional weighted average TFR for the original model and the new model. The box of each box plot represents the 80% prediction interval and the ends of the whiskers mark the endpoints of the 95% prediction interval. The corresponding observed TFR average is shown in red.