Journal of Statistical Software
March 2023, Volume 106, Issue 8. *doi: 10.18637/jss.v106.i08*

Probabilistic Estimation and Projection of the Annual Total Fertility Rate Accounting for Past Uncertainty: A Major Update of the bayesTFR R Package

Peiran Liu
University of Washington

Hana Ševčíková
University of Washington

Adrian E. Raftery
University of Washington

Abstract

The **bayesTFR** package for R provides a set of functions to produce probabilistic projections of the total fertility rates for all countries, and is widely used, including as part of the basis for the United Nations official population projections for all countries. Liu and Raftery (2020) extended the theoretical model by adding a layer that accounts for the past total fertility rate estimation uncertainty. A major update of **bayesTFR** implements the new extension. Moreover, a new feature of producing annual total fertility rate estimation and projections extends the existing functionality of estimating and projecting for five-year time periods. An additional autoregressive component has been developed in order to account for the larger autocorrelation in the annual version of the model. This article summarizes the updated model, describes the basic steps to generate probabilistic estimation and projections under different settings, compares performance, and provides instructions on how to summarize, visualize and diagnose the model results.

Keywords: **bayesTFR**, autoregressive model, Bayesian hierarchical model, Markov chain Monte Carlo, R, United Nations, world population prospects, annual projections, past TFR uncertainty.

1. Introduction

In 2015 for the first time, the United Nations (UN) adopted the Bayesian method described by Alkema *et al.* (2011) for their official population projections for all countries, the world population prospects (WPP) 2015 (United Nations 2015). This method is probabilistic and based on a principled statistical footing, replacing the previous deterministic method. One of the major components is the projection of the total fertility rate (TFR) which is implemented in the **bayesTFR** R package (Ševčíková, Alkema, and Raftery 2011). This package is widely used in research on fertility rates and population projections (Abel, Barakat, Samir, and Lutz 2016; Gerland, Biddlecom, and Kantorová 2017; Ševčíková and Raftery 2016; Ševčíková, Raftery, and Gerland 2018).

While the projection of TFR is probabilistic, the method does not take uncertainty about the past into account. Liu and Raftery (2020) addressed this issue by developing a Bayesian model that takes past TFR observations from the World Fertility Data database (United Nations 2019a) as raw data, and combines the uncertainty from the data with the uncertainty from the model. Out-of-sample validation showed improved performance of the overall projection model, while providing users with information about the uncertainty of estimates of past fertility rates. A major overhaul of **bayesTFR** was required to incorporate the Liu and Raftery (2020) methodology into the package.

The original framework implemented in **bayesTFR** was designed to work with five-year estimates and produced projections on a five-year time interval basis. This has the disadvantage of missing TFR fluctuations and pattern changes within the five-year periods. There is a growing interest by the UN to publish population estimates and projections on an annual basis, and in response we have extended **bayesTFR** to work with annual data. The update revealed that an additional autoregressive component is needed to account for the larger autocorrelation and thus, to model the uncertainty in the fertility transition well.

The new version of the package, version 7.3-2 (Ševčíková, Alkema, Liu, Raftery, Fosdick, and Gerland 2023), now produces uncertainty information about the past which is propagated into the projections and is able to estimate and project on an annual basis. This article describes the methodological changes, and also provides instructions on how to generate probabilistic estimations and projections under different settings. These include with and without accounting for past TFR estimation, with annual or five-year data, and with and without the autoregressive component in the fertility transition phase of the model. Other updates to the package are also introduced and elaborated.

The package **bayesTFR** (Ševčíková *et al.* 2023) is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=bayesTFR`.

The article is organized as follows. Section 2 summarizes the theoretical models developed by Alkema *et al.* (2011); Raftery, Alkema, and Gerland (2014); Liu and Raftery (2020), and the autoregressive model in the fertility transition phase. Section 4 describes how to use the package, using a step-by-step approach with different model settings. Section 5 presents experiments on the performance of the models and the selection of the various settings. The article concludes with a discussion in Section 6.

2. Annual TFR model with uncertainty about the past

Here, we first summarize the original TFR model developed for five-year time periods (Alkema *et al.* 2011). We then review the new methodology for probabilistic estimation and projection of TFR for all countries of the world accounting for uncertainty about the past, as proposed by Liu and Raftery (2020). Finally, we describe the changes in the methodology to work for annual estimation and projections.

TFR can be defined as the number of children a woman would have if she were subject to the prevailing fertility rates at all ages from a single given year, and survived throughout her childbearing years. Alkema *et al.* (2011) defined a three-phase model for the evolution of TFR over time in a country:

- Phase I: Pre-transition phase with fluctuations at high fertility level.
- Phase II: Transition from high to low fertility, where decrements are modeled by a random walk with drift given by a double logistic function.
- Phase III: Post-transition phase where fertility fluctuates around the replacement level (a level close to 2.1), modeled by an autoregressive AR(1) process.

We will use the same notation as Ševčíková *et al.* (2011). Specifically, $f_{c,t}$ denotes the TFR in country c and time period t, τ_c denotes the start period of phase II for country c, λ_c is the start period of phase III for country c, while $g(\boldsymbol{\theta}_c, f_{c,t})$ and $\boldsymbol{\theta}_c$ denote the parametric decline function and the corresponding country-specific parameters, respectively.

2.1. Existing model with five-year estimates

The pre-transition phase (phase I) is not modeled, as all countries have already entered phase II. Thus, for the purpose of projecting into the future it is not needed.

The fertility transition phase (phase II) is modeled by a random walk with drift. This is specified by

$$f_{c,t+1} = f_{c,t} - d_{c,t} \quad \text{for} \quad \tau_c \leq t < \lambda_c \,. \tag{1}$$

The decrement $d_{c,t}$ in Equation 1 is modeled as the sum of a function of the level of the TFR and the noise, as follows:

$$d_{c,t} = d(\boldsymbol{\theta}_c, \lambda_c, \tau_c, f_{c,t}) = g(\boldsymbol{\theta}_c, f_{c,t}) + \varepsilon_{c,t} \tag{2}$$

where $g(\boldsymbol{\theta}_c, f_{c,t})$ are the double logistic decrements, which are determined by the country-specific parameter vector $\boldsymbol{\theta}_c = (\Delta_{c1}, \Delta_{c2}, \Delta_{c3}, \Delta_{c4}, d_c)$ and given by

$$\frac{-d_c}{1 + \exp\left(-\frac{2\ln(p_1)}{\Delta_{c1}}(f_{c,t} - \sum_{i=1}^{4} \Delta_{ci} + 0.5\Delta_{c1})\right)} + \frac{d_c}{1 + \exp\left(-\frac{2\ln(p_2)}{\Delta_{c3}}(f_{c,t} - \Delta_{c4} - 0.5\Delta_{c3})\right)}. \tag{3}$$

The random distortions $\varepsilon_{c,t}$ in each period have normal distributions as follows:

$$\varepsilon_{c,t} \sim \begin{cases} N(m_t, s_t^2), & \text{for } t = \tau_c \,, \\ N(0, \sigma(f_{c,t})^2) & \text{otherwise} \,. \end{cases} \tag{4}$$

The quantity $\sigma(f_{c,t})$ is the standard deviation of the distortions during the later periods with

$$\sigma(f_{c,t}) = c_{1975}(t) \left(\sigma_0 + (f_{c,t} - S)(-aI_{[S,\infty)}(f_{c,t}) + bI_{[0,S]}(f_{c,t}))\right) \,. \tag{5}$$

The constant $c_{1975}(t)$ is added to model the higher error variance of the distortions before 1975. For further details about the model and its priors, see Ševčíková *et al.* (2011). For the purpose of this article, we only point to the definition of two parameters, namely the country-specific maximum decrement d_c, and the hyperparameter for the maximum standard deviation of the distortions σ_0. The d_c parameter is defined as

$$\begin{aligned} d_c^* &= \log\left(\frac{d_c - 0.25}{2.5 - d_c}\right), \\ d_c^* &\sim N(\chi, \psi^2). \end{aligned} \tag{6}$$

The prior distribution of σ_0 is $\sigma_0 \sim U[0.01, 0.6]$.

The TFR in the post-transition phase (phase III) is modeled by a first order autoregressive time series model (Raftery *et al.* 2014) as

$$f_{c,t+1} \sim N(\mu_c + \rho_c(f_{c,t} - \mu_c), \sigma^2) \text{ for } t \geq \lambda_c\,, \tag{7}$$

where μ_c is the country-specific long-term mean fertility rate, and ρ_c is the autoregressive parameter with $\rho_c \in (0, 1)$. In **bayesTFR** these parameters can be estimated via the Markov chain Monte Carlo (MCMC) method. Alternatively, country-independent values can be predefined or estimated by maximum likelihood.

The start period of phase II, τ_c, is defined as

$$\tau_c = \begin{cases} \max\{t : (M_c - L_{c,t}) < 0.5\}, & \text{if } L_{c,t} > 5.5; \\ \text{first estimation year}, & \text{otherwise}, \end{cases} \tag{8}$$

where M_c is the maximum observed TFR outcome in country c, and $L_{c,t}$ denote local maxima.

The start period of phase III for country c, λ_c, is defined as the period where two consecutive increases of TFR below 2 have been observed. More formally,

$$\lambda_c = \min\{t : f_{c,t} > f_{c,t-1}, f_{c,t+1} > f_{c,t} \text{ and } f_{c,p} < 2 \text{ for } p = t-1, t, t+1\}\,. \tag{9}$$

2.2. Probabilistic TFR estimation with uncertainty

The method described above uses observed TFR values as input to estimate the model parameters. In the UN context, these input values are taken from the latest revision of the WPP. Such TFR data are in fact estimates of the observed values, often derived from multiple data sources and involve varying amounts of uncertainty. The TFR model from the previous section however, treats these estimates as true values.

Liu and Raftery (2020) developed a method that assesses the uncertainty around past estimates of the observed TFR values and propagates it into the projections. The medians of the resulting posterior distributions can be used as point estimates of the past TFR, reducing the need for manual analysis and assessments by individual UN analysts. In addition, TFR projections resulting from the method of Liu and Raftery (2020) show better out-of-sample validation, especially better coverage of the prediction intervals, than the existing method.

The method uses the World Fertility Data database (United Nations 2019a) for past raw TFR estimates from surveys, reports and vital registrations for most regions in the world. We denote these data points by $y_{c,t,s}$, i.e., the raw TFR estimate for country c, time t and source s. The source s may refer to a census, a survey, vital registration statistics or other sources. For each observation, there are features $\boldsymbol{x}_{c,s}$ that describe the sources, estimating methods, recall lags and other aspects of data collection and estimation, often measures of the quality of the data. The observed $y_{c,t,s}$ are modeled as:

$$\begin{aligned} y_{c,t,s} \mid f_{c,t} &\sim N(f_{c,t} + \delta_{c,s}, \rho_{c,s}^2), \\ \mathsf{E}[\delta_{c,s}] &= \boldsymbol{x}_{c,s}\boldsymbol{\beta}, \\ \mathsf{E}[\rho_{c,s}] &= \boldsymbol{x}_{c,s}\boldsymbol{\gamma}. \end{aligned} \tag{10}$$

The $\delta_{c,s}$ and $\rho_{c,s}$ are country-specific parameters which are estimated by maximum likelihood. In Liu and Raftery (2020), the features used are the sources of the data and the corresponding estimation methods, but the model allows for any user-specified features. This part is combined with the existing Bayesian hierarchical model implemented in **bayesTFR**. Here, the past TFRs are considered as unknown, and are part of the parameters to estimate. The complete model is described in the Appendix.

If we are using TFR for five-year intervals, as for example in the `tfr` dataset in the **wpp2019** package (United Nations Population Division 2020), the true TFR at any time stamp is considered to be the linear interpolation of two consecutive five-year TFRs, namely

$$f_{c,t} = \frac{1}{5}[(t_{\ell+5} - t)f_{c,t_\ell} + (t - t_\ell)f_{c,t_{\ell+5}}] \quad \text{for any } t \in [t_\ell, t_{\ell+5}]\,.$$

If we are estimating from annual TFR, for each observation of the raw data, we take the floor of t. For example, if an observation in the raw data is recorded at 1975.5, we consider this observation as an estimate of the calendar year 1975.

Since we are now also modeling the past, not just the future as in the extant method, we need to model the pre-transition phase (phase I), which is not necessary for projecting the future. The TFR in this phase will be modeled by a random walk, specified by

$$f_{c,t+1} = f_{c,t} + \varepsilon_{c,t} \quad \text{for } t < \tau_c,$$

where the distributions of the random distortions in each period are given by

$$\varepsilon_{c,t} \sim N(0, s_t^2)\,.$$

Here, we simplify the model by setting the variance to be the same as the variance in the first period of the TFR decrements. This is a reasonable assumption because the starting period of phase II is linked to phase I, and the expected decline of TFR at the starting period of phase II is small. Thus, the distortions of TFR share similar behavior.

The estimation of all country-specific parameters and hyperparameters conditional on the TFRs, other than the TFRs themselves, in the phase II model remains the same as described by Ševčíková *et al.* (2011). To estimate past TFR, the model for phase III is estimated together with the phase II model via an MCMC algorithm (Gelfand and Smith 1990). This algorithm is a combination of Gibbs sampling, Metropolis-Hastings (for Δ_{ci} in Equation 3 and TFR), and slice sampling steps (Neal 2003).

The estimation yields a set of TFR trajectories about the past. To project into the future, we apply the existing projection method as described in Ševčíková *et al.* (2011) starting with the last estimation period of each trajectory. This is in contrast with the previous version, where the projection for each country starts from a single data point, namely the last observed TFR.

2.3. Annual version of bayesTFR

The original model described in Section 2.1 was designed to work with five-year data. Several modifications needed to be made in order to estimate and project annual data well.

Most importantly, we found that unlike in the five-year version, the residuals of the phase II model are highly autocorrelated when using annual data. We found that the lag 1 autocorrelation coefficients are about 0.7 for residuals of phase II model for some major countries. Thus, we modified the phase II model defined in Equations 1 and 2 by adding an additional first-order autoregressive component. The random walk with drift model is then specified as

$$d_{c,t+1} - g(\boldsymbol{\theta}_c, f_{c,t+1}) = \phi(d_{c,t} - g(\boldsymbol{\theta}_c, f_{c,t})) + \varepsilon_{c,t}. \tag{11}$$

The prior distribution of ϕ is set to be Uniform$(0,1)$ and is not country-specific. For the random distortions $\varepsilon_{c,t}$, the distribution is considered to be the same as in Equations 4 and 5.

The same prior distributions as in the five-year version is used for most parameters. One exception is σ_0 where the lower bound was decreased by a factor of the square root of five, i.e., $\sigma_0 \sim U[0.0045, 0.6]$. The upper bound was kept the same to allow for the possibility of additional correlation.

The definition of the maximum decrement defined in Equation 6 was changed to be one-fifth of that for the five-year model:

$$d_c^* = \log\left(\frac{d_c - 0.05}{0.5 - d_c}\right).$$

No changes have been made to the model of the post-transition phase of TFR, phase III. It is modeled by a first-order autoregressive time series model as defined in Equation 7.

The rule for determining the start period of phase II, τ_c, as defined in Equation 8, was unchanged. However, since the local maxima are calculated using annual TFR data, the results can differ from those obtained from a five-year dataset.

To determine the start periods of phase III, λ_c, as defined in Equation 9, we first obtain five-year averages of TFR and apply the same rule as in the five-year version, namely that phase III starts when two consecutive increases of TFR below 2 are observed.

2.4. Changes in TFR projections

There are three main differences in the TFR projections between the new implementation and the one described by Ševčíková *et al.* (2011).

The first difference (which we alluded to at the end of Section 2.2), relates to the fact that by accounting for the past TFR uncertainty (Equation 10), instead of observed point estimates, the model results in a set of TFR trajectories about the past which changes the starting values of the forecast. To project $f_{c,T+1}$ where T is the last period of the estimation, the i-th sample from the MCMC output is given by $f^{(i)}_{c,T+1} = f^{(i)}_{c,T} - d^{(i)}_{c,T} + \varepsilon^{(i)}_{c,T}$. Thus, the uncertainty in the first forecast period will be wider than if we use a model without accounting for past uncertainty, in which case $f^{(i)}_{c,T} = f_{c,T}$ is the same for all trajectories.

The second difference relates to the annual model described in Section 2.3. When the additional autocorrelation of phase II is taken into account (Equation 11), the past noise is carried over to the next time period. Specifically, to project $f_{c,t+1}$ for a country c that is in Phase II at time t, the i-th sample is given by $f_{c,t+1}^{(i)} = f_{c,t}^{(i)} - d_{c,t}^{(i)} + \varepsilon_{c,t}^{(i)}$, where $d_{c,t}^{(i)} = g(f_{c,t}^{(i)}, \boldsymbol{\theta}_c^{(i)})$, and $\varepsilon_{c,t}^{(i)}$ is drawn from $N(\phi^{(i)}\varepsilon_{c,t-1}^{(i)}, \sigma^{(i)}(f_{c,t}^{(i)}))$. For the first forecast, i.e., at the time period $T+1$, the distortion of the last estimation period T is calculated before starting the projections.

Finally, the last difference regards the updated phase III model as described in Raftery *et al.* (2014), where country-specific long-term means μ_c and autocorrelation coefficients ρ_c were incorporated into the model (Equation 7) and estimated by MCMC. However, this change has been available in **bayesTFR** since version 3.0-0 was published in 2013. Using this approach, to project $f_{c,t+1}$ for a country c that is in phase III at time t, the i-th MCMC sample is drawn from a normal distribution $N\left(\mu_c^{(i)} + \rho_c^{(i)}(f_{c,t}^{(i)} - \mu_c^{(i)}), \sigma^{(i),2}\right)$.

3. Overview of the package updates

The updated package **bayesTFR** retains all the functionalities of the previous version, which implements the model of Alkema, Raftery, Gerland, Clark, and Pelletier (2012). Its new features allow the user to conduct estimation of past TFR while accounting for uncertainty as described in Liu and Raftery (2020), as well as to forecast TFR for both five-year and one-year time intervals, as requested by the UN.

These new functionalities are implemented in the form of either new functions or additional arguments to existing functions. For convenience, especially for users who are familiar with the previous version of the package, this section summarizes the various changes. For users who are new to the package we recommend skipping to Section 4 where the usage is explained in more detail.

The following are established **bayesTFR** functions that were updated.

- `run.tfr.mcmc`: This is the core function for MCMC estimation of fertility transition model parameters. The following optional arguments were added:
  - `annual`: Logical argument determining whether the model is trained based on annual TFR data (`TRUE`) or on the five-year data (`FALSE`). The default is `FALSE`.
  - `ar.phase2`: Logical argument. If `TRUE`, Model 11 will be used in the estimation, and the parameter ϕ will be estimated through the MCMC process. This is relevant only if `annual = TRUE`.
  - `uncertainty`: Logical argument determining whether the model described in Liu and Raftery (2020) is estimated (`TRUE`) or if the default behavior of treating observed data as true values is used (`FALSE`). If `TRUE`, the past TFR values for all countries and time periods are estimated as additional parameters. Furthermore, Phase III of the TFR transition model is estimated simultaneously and thus, there is no need to call `run.tfr3.mcmc` separately.
  - `my.tfr.raw.file`, `covariates`, `cont_covariates, iso.unbiased`: These are arguments relevant to estimating past TFR. They allow the user to pass a file with raw TFR estimates, to set categorical and continuous covariates for estimating bias and measurement error variance of raw data, as well as to determine for which countries the vital registration TFR estimates are considered accurate. The arguments are considered only if `uncertainty = TRUE`.

- `tfr.predict`: This is the core function for TFR prediction. There was one optional argument added:
  - `uncertainty`: Logical argument. If `TRUE` and the corresponding estimation was produced via `run.tfr.mcmc(..., uncertainty = TRUE)`, then each prediction trajectory starts from a trajectory representing the past. Otherwise all prediction trajectories start from the same point, namely the last observed TFR.
- `run.tfr.mcmc.extra`: Originally, this function has been implemented in order to estimate the TFR transition model for very small countries or countries with unusual historical patterns. These countries were excluded from `run.tfr.mcmc` in order not to bias the world parameters, and were estimated separately via this function. In this update, the function has been extended to recompute past TFR estimates on a country-specific basis. This allows users to analyze the impact of changes on the raw TFR of individual countries without needing to run a new simulation for the whole world. Added arguments to this function have the same meaning as for `run.tfr.mcmc`:
  - `uncertainty`, `my.tfr.raw.file`, `covariates` and `cont_covariates`, `iso.unbiased`
- `tfr.trajectories.table`, `tfr.trajectories.plot`: These functions give projection quantiles in tabular and graphical formats, respectively. They have been extended to include uncertainty information about the past, if such information exists.

The following are new functions added to the package. They are described in Section 4.4 in more detail.

- `get.tfr.estimation`: Allows exploring the estimated trajectories as well as any quantiles of the past TFR estimates.
- `tfr.estimation.plot`: Function for plotting estimates of past TFR for individual countries.
- `tfr.bias.sd`: Allows exploring the bias and standard deviation estimated for the raw TFR estimates.

4. Using the updated bayesTFR

Previous versions of the **bayesTFR** package which implemented the model described in Section 2.1 have been used by UN analysts and others to train TFR projection models based on past five-year estimates. New UN requirements added the need to update the package so that analysts can conduct estimation of past TFR accounting for uncertainty, and make corresponding forecasts for both five-year and annual time periods.

To make probabilistic TFR projections accounting for past TFR uncertainty, the user will follow four steps in the following order:

1. Data assembly (optional):

   (a) Prepare a dataset of raw TFR values. By default, the World Fertility Data 2019 (United Nations 2019a) is used.
   (b) Assemble a dataset of reference (initial) TFR values for all countries and time periods. By default, the *world population prospects* (United Nations 2019b) in the **wpp2019** package is used.

2. Model estimation:
   (a) Train linear models to estimate systematic bias and standard deviation for each observation from the raw TFR dataset.
   (b) Given the reference TFR, calculate the start period of phase II and the start period of Phase III for each country (τ_c and λ_c).
   (c) Run the MCMC process to obtain posterior samples of the phase II and phase III model parameters, and posterior samples of the past TFR for all countries.
3. Generate future TFR trajectories as discussed in Section 2.4.
4. Analyze the outputs using a set of functions that summarize, plot, diagnose and export the results of the three steps above.

As described by Ševčíková *et al.* (2011), steps 2 and 3 are relatively time-consuming. Adding the estimation uncertainty feature as well as working with annual estimates adds even more run time. Even though we optimized the code wherever possible, it takes several hours to complete these steps in a production-like setting.

The following sections describe the steps above in more detail, especially the parts that are different from Ševčíková *et al.* (2011). We will elaborate on how to use the new features, as well as how to use the package in the original way. We will demonstrate the package on a realistic example with a relatively large number of MCMC iterations, which might take several hours to process. Therefore, users who wish to explore the functionality quickly should reduce the number of iterations to the order of magnitude of 10. However note that since the Metropolis-Hastings step for the TFR updates will have an acceptance rate of around 30%, a small number of iterations will result in estimation plots that are less smooth than what we will present in this article.

4.1. Data assembly and estimation settings

The datasets assembled in this step will be passed to the main estimation function, `run.tfr.mcmc`, which now has additional arguments for this purpose. It can be specified what raw TFR data to use, whether to estimate and predict annually (logical argument `annual`), and whether to use the AR(1) model in phase II as defined in Equation 11 (logical argument `ar.phase2`).

The argument `uncertainty = TRUE` specifies that uncertainty about the past is incorporated into the estimation. In this case, a raw TFR dataset can be provided. By default, the World Fertility Data 2019 (United Nations 2019a) is used. This dataset contains $12,079$ records for 201 countries, each of which includes the corresponding estimation method and data source. These are then used by the model as data quality covariates in Equation 10.

The default raw TFR dataset can be viewed via

```
R> data("rawTFR", package = "bayesTFR")
R> head(rawTFR)

  country_code year  tfr   method source
1            4 1965 7.97 Indirect Census
2            4 1966 8.21 Indirect Census
3            4 1967 8.32 Indirect Census
4            4 1968 8.23 Indirect Census
5            4 1969 8.07 Indirect Census
6            4 1970 7.98 Indirect Census
```

The default covariates are `c("source", "method")`. Users can provide their own file and covariates of their choice. Required columns are `country_code`, `year` and `tfr`. The name of this file is passed to the argument `my.tfr.raw.file`, names of categorical variables to the argument `covariates`, and names of continuous variables to the argument `cont_covariates`.

An additional option allows an analyst to specify that vital registration data from selected countries are unbiased, if there is a belief that these data are not systematically biased in a particular direction. (Note that this is not the same as saying that these data are perfect.) The UN country codes of those countries can be passed into the argument `iso.unbiased`. In this case, the bias and standard deviation of the records of those countries for which the `source` column specifies "VR" (as vital registration) are forced to be equal to 0 and to be near 0, respectively (in practice the standard deviation is set to 0.0161). This option targets fine-tuning of the estimation of developed countries, especially because the annual TFR estimates are often not openly accessible.

The second dataset to assemble is a dataset on a reference, or initial, TFR. Its file name is passed into the argument `my.tfr.file`. If `uncertainty = FALSE`, this dataset is considered in the estimation as the true observed TFR. Otherwise, it is used as the starting points of TFR in the MCMC process. By default, if `my.tfr.file` is not given, the `tfr` dataset from the **wpp2019** package is used for this purpose, which is a five-year dataset. Thus, if `annual = TRUE`, a linear interpolation of the default dataset is computed.

4.2. Fitting the TFR model

Most arguments of `run.tfr.mcmc` remain the same as described in Ševčíková *et al.* (2011). Importantly, `start.year` and `present.year` set the first and the last year of the time series included in the computation, respectively. The arguments `nr.chains`, `iters` and `output.dir` determine the number of chains, the number of iterations and the directory to store the MCMC simulated values, respectively.

In the prior version of **bayesTFR**, the function `run.tfr.mcmc` was designed to obtain a posterior sample of phase II model parameters. The estimation of phase III parameters (as defined in Equation 7) is implemented in the function `run.tfr3.mcmc`. When building a full probabilistic model as described in Liu and Raftery (2020), the MCMC steps for updating TFR will affect both phases. Thus, if `uncertainty = TRUE`, the new `run.tfr.mcmc` function combines the estimation of both phases, and there is no need to invoke the `run.tfr3.mcmc` function explicitly. We call this method a "one-step estimation". However, this is not the case if uncertainty about the past is not taken into account. In this case, the workflow of estimating phase II and phase III separately remains and is referred to as a a "two-step estimation".

annual	uncertainty	
	TRUE	FALSE
TRUE	**A** One-step estimation; Phase II-AR(1) allowed	**B** Two-step estimation; Phase II-AR(1) allowed
FALSE	**C** One-step estimation	**D** Two-step estimation

Table 1: Possible combinations in fitting TFR projection model.

The various combinations of the possible settings of the arguments `annual` and `uncertainty` are summarized in Table 1. We have marked each cell with a letter which will be referred to in the remainder of this section.

As described in Section 2.3, when using the annual model (cells A and B), adding the autoregressive component in phase II as defined in Equation 11 should be considered. The option is controlled via the logical argument `ar.phase2` which should be passed to the `run.tfr.mcmc` function. If `ar.phase2` is set to `TRUE` the MCMC process performs an extra slice sampling step for estimating ϕ, an extra country-independent parameter in the model. The argument is ignored if `annual` is `FALSE`.

Starting a new simulation with two-step estimation

The two-step estimation should be performed if uncertainty about the past is not taken into account (cells B and D in Table 1). The main differences between cells B and D are the setting of prior distributions as described in Section 2.3, and whether the autoregressive component can be included in the model. Here we give an example of a simulation with annual data (cell B) without the autoregressive component. However, we will not use this example further in the text, as the main focus of the article is on cell A which will be discussed in the next section.

Our example simulation consists of three MCMC chains, each of which is $5,000$ iterations long where thinning is disabled. (As noted earlier, the user is advised to decrease the number of iterations to the order of ten for faster processing.) We will save the simulation results to a directory called “annual”.

```
R> annual <- TRUE
R> nr.chains <- 3
R> total.iter <- 5000
R> thin <- 1
R> simu.dir <- file.path(getwd(), "annual")
```

The first step is to launch an estimation of phase II:

```
R> m2 <- run.tfr.mcmc(output.dir = simu.dir, nr.chains = nr.chains,
+    iter = total.iter, thin = thin, annual = annual)
```

The second step is to start an estimation of phase III:

```
R> m3 <- run.tfr3.mcmc(sim.dir = simu.dir, nr.chains = nr.chains,
+    iter = total.iter, thin = thin)
```

Here, we are using the same number of chains and iterations for phase II and phase III. However, this is not a requirement, but rather depends on the MCMC convergence. Even the $3 \times 5,000$ iterations might be not enough to reach convergence, but will usually give realistic outputs. Setting `total.iter = 62000` or `total.iter = "auto"` will most likely result in full convergence.

Starting a new simulation with one-step estimation

We now show an example of a simulation with uncertainty which is performed with one step only (cells A and C in Table 1). In particular, here we set `annual` to `TRUE` (cell A), but the same function would be used if `annual` is `FALSE` (cell C). We will also include the phase II-AR(1) model (`ar.phase2`) which would not have any effect in cell C. The results will be saved in the directory `"annual_unc"`. We will use this simulation throughout the article.

```
R> annual <- TRUE
R> ar.phase2 <- TRUE
R> nr.chains <- 3
R> total.iter <- 5000
R> thin <- 1
R> simu.dir.unc <- file.path(getwd(), "annual_unc")
```

As in the previous case, this setting may not be enough to yield fully converged MCMC simulations, but will still give realistic outputs. The processing time is within a range of a couple of hours. For faster processing, set `total.iter = 50` for a toy simulation. In addition, the `parallel` argument can be set to `TRUE`, in which case the three chains will be processed in parallel. In Section 5.2, we will give recommendations for settings that yield fully converged MCMC simulations. When appropriate, we will use results from such converged simulations to present outputs.

As mentioned in Section 4.1, additional arguments of `run.tfr.mcmc` allow one to pass user-specific raw TFR data (`my.tfr.raw.file`), names of categorical covariates (`covariates`), names of continuous covariates (`cont_covariates`), or to specify countries with unbiased vital registration data (`iso.unbiased`). If the `iso.unbiased` option is used, the `covariates` argument should include the variable `source`, or more specifically, the variable defined by the argument `source.col.name` which is `source` be default. In such a case, the function reduces the bias and standard deviation of records where the `source` column specifies “VR”. In our example we will specify that the VR data of Canada and the USA (codes 124 and 840) are unbiased.

To estimate both phase II and phase III, one could do

```
R> m <- run.tfr.mcmc(output.dir = simu.dir.unc, nr.chains = nr.chains,
+    iter = total.iter, annual = annual, thin = thin,
+    uncertainty = TRUE, ar.phase2 = ar.phase2, iso.unbiased = c(124, 840))
```

In comparison to the two-step model, the training process here has an extra Metropolis-Hastings step per iteration for generating posterior TFR samples.

Continuing an existing simulation

If an existing simulation needs to be extended by more iterations, one would proceed as in the previous version of **bayesTFR**:

- Use `continue.tfr.mcmc` if the MCMCs were originally created via `run.tfr.mcmc`, regardless of whether one is in the one-step or the two-step estimation mode.
- Use `continue.tfr3.mcmc` if the MCMCs were originally created via `run.tfr3.mcmc`.

Now suppose we want to extend the simulation in the previous section by 100 iterations. Then we could do

```
R> m <- continue.tfr.mcmc(output.dir = simu.dir.unc, iter = 100)
```

(Set the `iter` argument to 10 if working with a toy simulation.) This will continue running both TFR phases in a one-step estimation while inheriting all settings from the original simulation, including `uncertainty`, `annual`, `ar.phase2` or settings about the raw data. At the end of the processing, each chain will be $5,100$ iterations long.

4.3. Generating projections

The main function for generating projections is called `tfr.predict`. The new version of the package adds the argument `uncertainty`. If it is `TRUE` and the model was estimated taking uncertainty about the past into consideration, then that past uncertainty will be carried over to the projections. In this case, each future trajectory starts from a trajectory representing the past.

Suppose we want to generate projections represented by $1,000$ posterior trajectories until the year 2100 based on the simulation stored in the directory given by `simu.dir.unc`, with burn-in of the first $2,100$ iterations for each chain. This can be done using the following command:

```
R> pred <- tfr.predict(sim.dir = simu.dir.unc, end.year = 2100,
+    burnin = 2100, nr.traj = 1000, uncertainty = TRUE)
```

The function takes the existing $5,100$ iterations in each chain, removes the first $2,100$ values and generates $1,000$ TFR trajectories based on $1,000$ equally spaced parameter values and past TFR, out of the remaining $3,000 \times 3 = 9,000$ iterations. For a toy simulation, use `burnin = 20` and `nr.traj = 50`.

If `uncertainty` is set to `FALSE`, all future trajectories start from the last observed data point. If the estimation process accounted for uncertainty, but the projection does not, the starting value of the projections is the initial TFR value for the last observed time period. This is however not recommended but may serve the purpose of apples-to-apples comparisons.

4.4. Analyzing results

If results are to be explored at a later time point, one can load the estimation object from disk using the command

```
R> m <- get.tfr.mcmc(simu.dir.unc)
```

For one-step estimation, this object contains information about both phases. For a two-step simulation, or if the phase III object is to be extracted explicitly, use

```
R> m3 <- get.tfr3.mcmc(simu.dir.unc)
```

Similarly, to load the prediction object from disk, do

```
R> pred <- get.tfr.prediction(simu.dir.unc)
```

Summary functions

For the summary statistics of the estimation object in this section, we will use the following thinning and burn in settings:

```
R> thin <- 3
R> burnin <- 2100
```

Use `thin <- 1` and `burnin <- 20` if you're working with the toy simulation.

To view a summary of country-independent parameters, one can use

```
R> summary(m, thin = thin, burnin = burnin)
```

Since the object `m` was got by one-step estimation, the output includes estimation summaries for both phases. In comparison to the previous version of the package, phase II contains one additional parameter, namely `"rho_phase2"` which represents ϕ in Model 11. As with any other parameter, the name, or multiple parameter names, can be passed to the function to view summary statistics for those selected parameters.

```
R> summary(m, par.names = c("rho_phase2", "sigma0"), thin = thin,
+    burnin = burnin)
```

```
MCMCs of phase II
=================
Number of countries: 201
Hyperparameters estimated using 201 countries.
WPP: 2019
Input data: TFR for period 1950-2020
Time interval: annual

Iterations = 2103:5100
Thinning interval = 3
Number of chains = 3
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
```

```
              Mean      SD  Naive SE Time-series SE
rho_phase2 0.6753 0.09237 0.0016864       0.008547
sigma0     0.0606 0.01477 0.0002696       0.001473

2. Quantiles for each variable:

              2.5%     25%     50%     75%   97.5%
rho_phase2 0.47033 0.67203 0.69641 0.71269 0.73374
sigma0     0.05062 0.05485 0.05793 0.06168 0.08991
```

The full list of parameter names for phase II can be obtained via

```
R> tfr.parameter.names(meta = m$meta)
```

```
 [1] "alpha"        "alphat"        "delta"        "Triangle4"     "delta4"
 [6] "psi"          "chi"           "a_sd"         "b_sd"          "const_sd"
[11] "S_sd"         "sigma0"        "mean_eps_tau" "sd_eps_tau"    "rho_phase2"
```

Passing the `meta` argument is needed to identify that the simulation contains a phase II-AR(1) estimation, and thus it contains the `"rho_phase2"` parameter. Phase III parameter names are not dependent on the simulation, thus no `meta` argument is needed:

```
R> tfr3.parameter.names()
```

```
[1] "mu"        "rho"        "sigma.mu"  "sigma.rho" "sigma.eps"
```

Specifying a country in the `summary` function will show results for the country-specific parameters of that country. This is done via the `country` argument which accepts either the name or the numerical code of the country, as well as an ISO-2 or ISO-3 character code. This is the case for any function in the package that accepts the `country` argument, as is shown throughout the paper.

As for the parameters to summarize, functions `tfr.parameter.names.cs()` and `tfr3.parameter.names.cs()` list the allowed parameter names for phase II and phase III, respectively. For a simulation that took into account uncertainty about the past, there is an additional country-specific parameter, called `"tfr"`, capturing that uncertainty. It is not listed explicitly via the above functions, but it can be explored like any other parameter. For the `summary` function it means passing it to the `par.names.cs` argument. For example, to view summary statistics of TFR estimation for Nigeria, we can do

```
R> summary(m, country = "Nigeria", par.names.cs = "tfr", thin = thin,
+    burnin = burnin)
```

The tabular sections of the output contain one row per past observed period each (by default 71, i.e., from 1950 to 2020). To select a subset we can specify which time periods we are interested in as `tfr_<time>`. For example, to view results for time periods 1, 30 and 70 (corresponding to 1950, 1979 and 2019) we do

```
R> summary(m, country = "Nigeria", par.names.cs = c("tfr_1", "tfr_30",
+    "tfr_70"), thin = thin, burnin = burnin)
```

```
...
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

             Mean      SD  Naive SE Time-series SE
tfr_1_c566  6.371 0.19266 0.003518       0.025403
tfr_30_c566 6.740 0.07461 0.001362       0.009197
tfr_70_c566 5.200 0.26172 0.004778       0.099382

2. Quantiles for each variable:

             2.5%   25%   50%   75% 97.5%
tfr_1_c566  5.931 6.262 6.378 6.493 6.745
tfr_30_c566 6.587 6.689 6.741 6.796 6.876
tfr_70_c566 4.824 4.968 5.189 5.374 5.779
```

Exploring TFR estimation

In addition to summary statistics, one can explore the estimated trajectories as well as any quantiles of the past TFR estimates. For example,

```
R> nigeria_obj <- get.tfr.estimation(country = "NG", sim.dir = simu.dir.unc,
+    burnin = burnin, thin = thin, probs = c(0.025, 0.1, 0.5, 0.9, 0.975))
```

returns a list where trajectories are contained in the element `tfr_table`. The number of rows corresponds to the number of trajectories (here $3000 = 3[\text{chains}] \cdot (5100 - 2100)/3[\text{thin}]$, or $120 = 3(60-20)$ for the toy simulation), while columns correspond to time periods (here 71).

```
R> dim(nigeria_obj$tfr_table)
```

```
[1] 3000   71
```

The quantiles are contained in the element `tfr_quantile`:

```
R> head(nigeria_obj$tfr_quantile)
```

```
       2.5%      10%      50%      90%    97.5% year
1: 5.931205 6.134300 6.377760 6.614433 6.744657 1950
2: 5.977130 6.148832 6.383231 6.580278 6.701673 1951
3: 6.021235 6.168258 6.387442 6.579906 6.665112 1952
4: 6.039646 6.181161 6.376453 6.577242 6.684508 1953
5: 6.094877 6.163529 6.367027 6.599665 6.699566 1954
6: 6.052390 6.171088 6.359326 6.598861 6.691794 1955
```

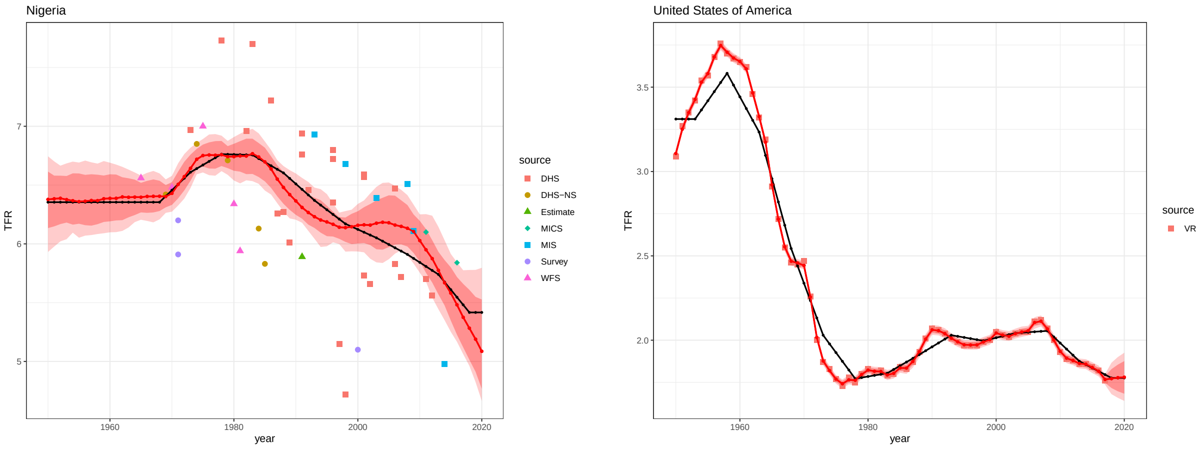


Figure 1: Annual TFR estimation for Nigeria (left panel) and the United States (right panel), resulting from a converged simulation. The red line shows the posterior median, while the red shaded area shows the pointwise 80% intervals, and the pink shaded areas shows the corresponding 95% intervals. The UN's 2019 WPP (interpolated) five-year estimates are shown by the black line.

This element is missing if the `probs` argument is not given.

For example, to plot the estimation with user-defined intervals, do:

```
R> plot <- tfr.estimation.plot(country = 566, sim.dir = simu.dir.unc,
+     burnin = burnin, thin = thin, pis = c(80, 95), plot.raw = TRUE)
R> print(plot)
```

The function uses the **ggplot2** package (Wickham 2016) to visualize estimation uncertainty. Figure 1 shows results of the function call for Nigeria (as above) and the USA (`country = 840` or `country = "USA"`) using a converged simulation.

Several arguments in this function need to be clarified:

- `sim.dir`: Users can specify the location of the simulation set, or use the `mcmc.list` argument to pass the `m` object directly. For example `tfr.estimation.plot(mcmc.list = m, ...)`.

- `pis`: Specifies which probability intervals will be plotted. It is a vector of at most two elements.

- `plot.raw`: Logical argument which determines whether the raw data used for estimating past uncertainty are plotted. If `TRUE` and the estimation process was not based on the default data, it is recommended to provide the argument `grouping`, which should be one of the categorical covariates in the raw data set. This covariate defines the color and shape of the points, as seen in Figure 1 where the default grouping is the `source` column of the `rawTFR` dataset.

- `save.image`: (not used in the call above) If `TRUE`, the plot will be saved as a PDF file in the directory specified in the argument `plot.dir`, named `"tfr_country`<*code*>`.pdf"`.

	Estimate	Std. Error	t value	P(> |t|)
`(Intercept)`	−0.43	0.10	−4.37	0.00
`covariate_1DHS-NS`	−0.31	0.17	−1.76	0.09
`covariate_1Estimate`	−0.14	0.37	−0.39	0.70
`covariate_1MICS`	0.72	0.27	2.68	0.01
`covariate_1MIS`	0.29	0.16	1.79	0.08
`covariate_1Survey`	−0.21	0.23	−0.94	0.35
`covariate_1WFS`	−0.18	0.17	−1.06	0.30
`covariate_2Indirect`	0.80	0.11	7.05	0.00

Table 2: Linear model for bias obtained from `summary(bias_sd$model_bias)` for Nigeria.

Method	Source	Bias	Std
Indirect	WFS	0.18	0.13
Indirect	DHS-NS	0.06	0.09
Direct	Survey	−0.64	0.64
Indirect	DHS	0.37	0.28
Direct	WFS	−0.61	0.61
Direct	DHS-NS	−0.74	0.74

Table 3: Bias and standard deviation of each observation obtained from `bias_sd$table` for Nigeria.

Exploring bias and standard deviation of observations

Information about the bias and standard deviation of observations will give users an indication of the quality of the observations and whether these quantities were poorly estimated.

Now suppose we are interested in the bias and standard deviation estimates of the observations for Nigeria. Then we could use

```
R> bias_sd <- tfr.bias.sd(sim.dir = simu.dir.unc, country = 566)
```

The function will return a list with elements `model_bias`, `model_sd` and `table`. The `model_bias` and `model_sd` objects are of class '`lm`' and contain the linear models used to estimate the bias and standard deviation, respectively, while the '`table`' object includes the observed data points, data quality covariates, and the actual estimates for the specified country, here for Nigeria.

```
R> summary(bias_sd$model_bias)
R> head(bias_sd$table)
```

The results are shown in Tables 2 and 3.

To generate the estimates in the '`table`' object, the `predict` S3 method is applied to the '`model_*`' objects. Then the following steps are performed:

1. For some countries, the number of data points is very small for several groups. This could lead to a large bias, but a very small variance. As a result, the estimation will be unreasonably concentrated on the bias-adjusted data points. In this case, the

standard deviation estimates were adjusted as max(0.1, |bias|/2). The reason for such an adjustment is that it is unlikely that one observation is very biased but with a very small relative standard deviation. It is also unlikely that there is a source of data that is very precise (with standard deviation less than 0.1), but is only collected once.

2. For countries included in `iso.unbiased`, the model estimates are overwritten with zero or close to zero values as explained in Section 4.1.

3. Duplicates are dropped so that the combinations of data quality covariates are unique.

The output can help to detect problematic estimates on certain data points so that adjustments can be made by the analyst if necessary. In the example above, the estimated bias and standard deviation for the Indirect method and nationwide DHS surveys were 0.06 and 0.09, respectively. These estimates were derived based on only three data points in this category, and all of them were very close to the UN estimates (three of the brown dots in Figure 1 in year 1969, 1974 and 1979). Since the number of data points from nationwide DHS estimates is small (3 data points), the estimated bias (0.06) and standard deviation (0.09) may be too small.

Exploring TFR prediction

Plotting the posterior sample of projected TFR trajectories is done via the `tfr.trajectories.plot` function. The updated version of the package incorporates uncertainty about the past, if taken into account in the estimation and projection. For example, to plot the prediction of TFR for Burkina Faso contained in the `pred` object created in Section 4.3 or at the beginning of Section 4.4, use

```
R> tfr.trajectories.plot(pred, country = "Burkina Faso", nr.traj = 20,
+    pi = c(80, 95), uncertainty = TRUE)
R> tfr.trajectories.plot(pred, country = "Burkina Faso", nr.traj = 20,
+    pi = c(80, 95), uncertainty = FALSE)
```

Here, the parameter `uncertainty` is used to specify whether the uncertainty about the past TFR should be plotted together with the prediction. If `uncertainty` is `TRUE`, the optional parameters `thin, burnin, col_unc` can be used to define the burn-in, thinning and the color for the past uncertainty plot.

The code above applied to a converged simulation results in the plots shown in Figure 2.

If the user selects `uncertainty = FALSE` for a simulation where past uncertainty was taken into account (similarly to the right panel of Figure 2), the past TFR used for the initialization of the model is shown as the observed TFR. In this case, there could be a discontinuity between the last observed and the first projected time period.

These new arguments are also accepted by the `tfr.trajectories.plot.all` function which generates projection plots for all countries at once, as described by Ševčíková *et al.* (2011).

TFR predictions in a tabular format can be explored using the `tfr.trajectories.table` and `summary` functions which work the same way as in the previous versions of the package, except that in the former case, the output now contains uncertainty information about the past.

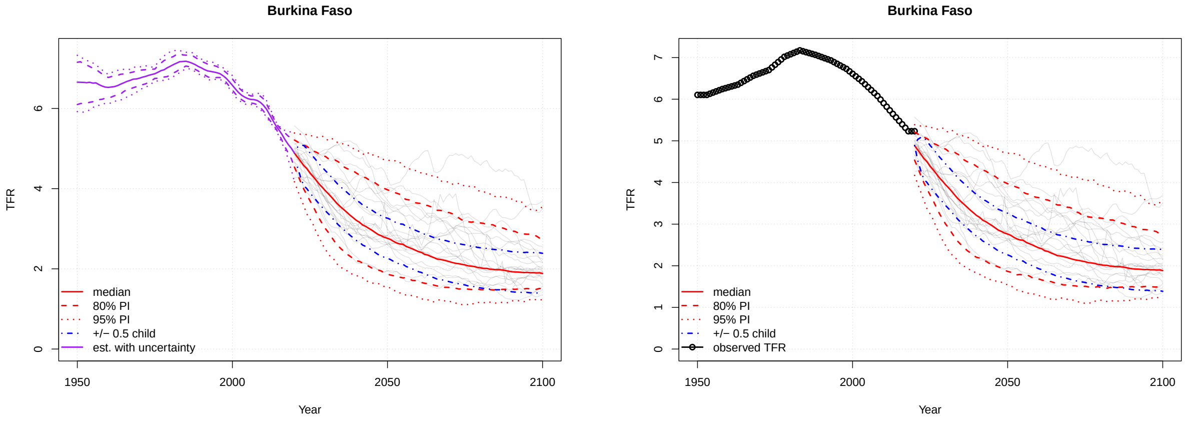


Figure 2: TFR prediction (from a converged simulation) for Burkina Faso with uncertainty about the past TFR (left panel) and without it (right panel). The black dots in the right panel represent the TFR used for initializing the simulation. In both panels, the red curves (solid, dashed and dotted) show the probabilistic prediction (median, 80% and 95% probability intervals), while the blue lines show the traditional UN scenarios of adding and removing a half a child to/from the main projection, here the median TFR.

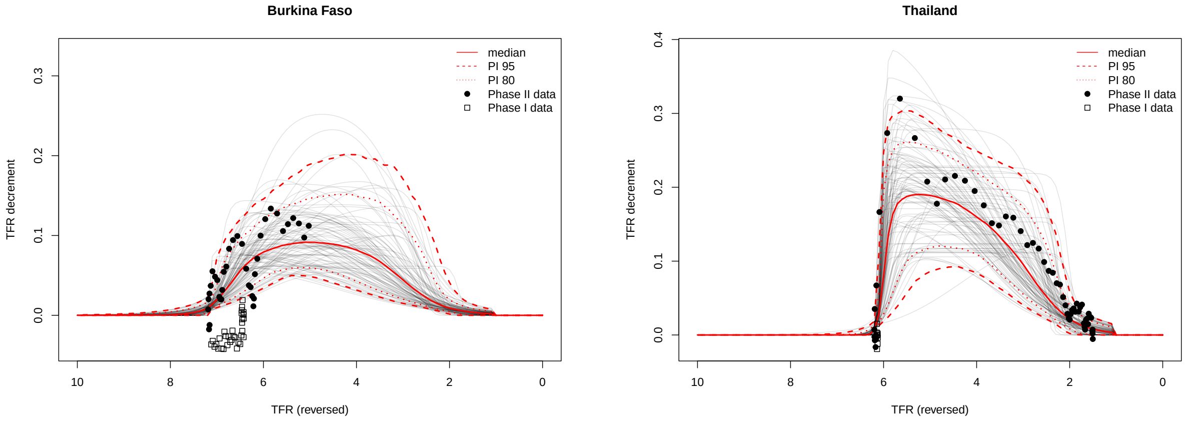
```
R> tfr.trajectories.table(pred, country = "Burkina Faso")
```

```
       median    0.025      0.1      0.9    0.975 -0.5child +0.5child
1950 6.655622 5.920312 6.096360 7.150919 7.325010        NA        NA
1951 6.652163 5.898786 6.121541 7.164751 7.290802        NA        NA
1952 6.649168 5.910383 6.125930 7.111552 7.235108        NA        NA
1953 6.642280 5.929401 6.130181 7.080910 7.200169        NA        NA
1954 6.652158 5.961541 6.152881 7.044679 7.167199        NA        NA
...
2095 1.912930 1.196287 1.508627 2.863504 3.516937  1.412930  2.412930
2096 1.902512 1.235173 1.499763 2.864060 3.482927  1.402512  2.402512
2097 1.904649 1.216530 1.492337 2.839775 3.470921  1.404649  2.404649
2098 1.897675 1.239371 1.488701 2.818639 3.458704  1.397675  2.397675
2099 1.899654 1.229692 1.510012 2.767684 3.514113  1.399654  2.399654
2100 1.887052 1.226206 1.500331 2.749430 3.537936  1.387052  2.387052
```

```
R> summary(pred, country = "Burkina Faso")
```

```
Projections: 80 ( 2021 - 2100 )
Trajectories: 1000
Phase II burnin: 2100
Phase II thin: 9
Phase III burnin: 2100
Phase III thin: 9

Country: Burkina Faso
```


Burkina Faso

Thailand

Figure 3: Estimated double logistic curves (from a converged simulation) for Burkina Faso (left panel) and Thailand (right panel). The data points (black dots and squares) are the estimated median decrements per year.

```
Projected TFR:
     mean    SD 2.5%   5%  10%  25%  50%  75%  90%  95% 97.5%
2020 4.87 0.277 4.17 4.33 4.54 4.71 4.89 5.05 5.21 5.32  5.39
2021 4.77 0.334 3.96 4.12 4.36 4.57 4.78 4.98 5.17 5.28  5.38
2022 4.66 0.384 3.77 3.98 4.17 4.43 4.68 4.91 5.12 5.27  5.35
2023 4.56 0.434 3.57 3.82 3.99 4.29 4.58 4.85 5.09 5.24  5.36
...
```

Exploring double logistic function

The double logistic function defined in Equation 3 can be viewed using

```
R> DLcurve.plot(country = "BFA", mcmc.list = m, burnin = burnin,
+    pi = c(95, 80), nr.curves = 100)
```

Results can be seen in the left panel of Figure 3, while the right panel shows the result of the same call with `country = "Thailand"`.

If a simulation contains information about past uncertainty, then the phase II and I data (black dots and squares) represent decrements of the estimated TFR median. In case of an annual simulation, these are annual decrements, otherwise they would correspond to five-year decrements. If the projections were produced without taking past uncertainty into account, then the data points represent the observed decrements.

This also applies to the `DLcurve.plot.all` function which plots the double logistic curves for all countries at once.

MCMC traces, density and diagnosis

To explore traces of the MCMC parameters, the existing functions `tfr.partraces.plot` (for country-independent parameters) and `tfr.partraces.cs.plot` (for country-specific param-

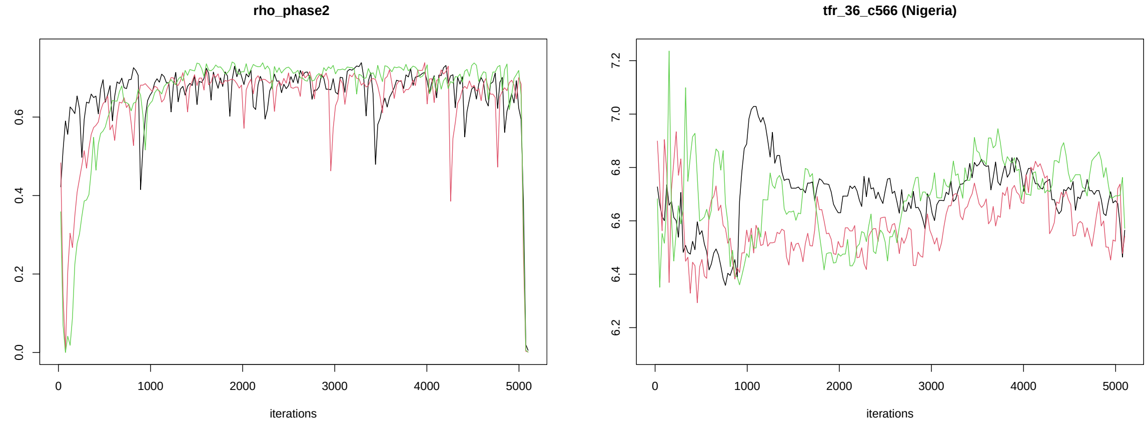


Figure 4: Trace plots for ϕ (left panel) and TFR of Nigeria in 1985 (right panel).

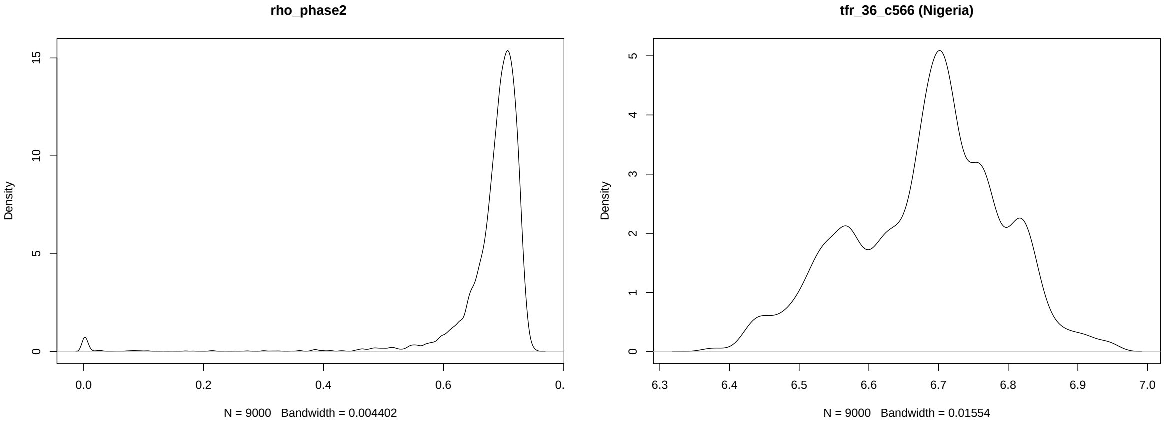


Figure 5: Density plots for ϕ (left panel) and TFR of Nigeria in 1985 (right panel).

eters) can be used. Similarly, for density plots, `tfr.pardensity.plot` and `tfr.pardensity.cs.plot` are available.

As mentioned previously, there are two additional parameters in this version of the package, namely `"rho_phase2"`, which is country-independent and defined in Model 11, and `"tfr"` which is a country-specific parameter. These two parameters can be used within the aforementioned functions, like any other parameters.

For example, the trace plots and the density plots of ϕ and Nigeria's TFR estimate in year 1985 (as shown in Figures 4 and 5) can be visualized via

```
R> tfr.partraces.plot(m, par.names = "rho_phase2", nr.points = 200)
R> tfr.partraces.cs.plot(m, country = "Nigeria", par.names = "tfr_36",
+    nr.points = 200)
R> tfr.pardensity.plot(m, par.names = "rho_phase2", burnin = burnin)
R> tfr.pardensity.cs.plot(m, country = "Nigeria", par.names = "tfr_36",
+    burnin = burnin)
```

To check if the MCMC algorithm has converged and adequately explored the parameter space, the `tfr.diagnose` function can be used; see Ševčíková *et al.* (2011) for more details. In the case of one-step estimation, the function checks parameters from phase II as well as Phase III. In the case of two-step estimation, one would use `tfr.diagnose` for assessing the convergence of phase II parameters, and `tfr3.diagnose` for assessing the convergence of phase III parameters. Both functions accept a logical argument `express` which can disable or reduce the checking of country-specific parameters in order to speed up the process.

If the estimation includes uncertainty about the past, the assessment of country-specific parameters include the `"tfr"` parameter for each country and time period, in our case more than $14,200$ `"tfr"` parameters. In practice, it is often impossible to achieve convergence for all of them. Thus, we introduced the rule of accepting the `"tfr"` parameters as having converged if 95% of them have converged.

To apply the convergence diagnostics to our simulation, one could do

```
R> tfr.diagnose(simu.dir.unc, thin = thin, burnin = burnin, express = TRUE)
```

As mentioned earlier, in our illustrative code examples the MCMC algorithm has not been run for long enough to achieve full convergence. See Section 5.2 for alternative settings. Note that the toy simulation we proposed earlier cannot be checked for convergence, as there is a requirement of a minimum number of iterations per chain, which the toy simulation does not satisfy.

4.5. Estimating a small set of countries

The Bayesian framework we have shown so far is designed to estimate all countries of the world at once, where the historical experience of an individual country influences the distribution of its own parameters as well as of the world parameters, while using the same settings for all countries. However, this is not always practical for several reasons:

1. Analysts might want to experiment with settings for individual countries without waiting several hours for a simulation of the whole world to finish.

2. Different sets of covariates might be needed to estimate different countries.

3. Countries with unusual historical patterns or very small countries might be excluded from the simulation in order not to bias the world parameters.

It was the last reason, as well as the need to include aggregations in the estimation, that motivated us to implement the `run.tfr.mcmc.extra` function in the original version of the package. The idea is that, while `run.tfr.mcmc` updates all parameters, the `run.tfr.mcmc.extra` function updates only the country-specific parameters of the specified countries, while re-using the existing distribution of the global parameters.

Since the function was designed for special cases of countries or aggregations, the original implementation allowed the user to process only the locations that had not been included in the world simulation. With the two additional use cases above, we have now relaxed that restriction and made it possible to rerun and overwrite existing estimations of country-specific parameters and past TFR estimates for individual countries, while allowing the user to change various estimation settings. However, several global settings are not subject to change, such as switching between annual and five-year estimation, or changing the `ar.phase2` argument.

Suppose that after running the simulation with the default data from the World Fertility Data, the user wishes to experiment with their own data that exclude Nigeria's questionable data points, such as the Indirect DHS-NS data points identified in Table 3 as having unreasonably low standard deviations and biases. Unlike in the main simulation, the experiment will not force the vital registration (VR) data of the United States to have zero bias and variance. For that purpose, we will extract data for Nigeria (code 566) and the USA (code 840) from the default raw dataset discussed in Section 4.1, remove the Indirect DHS-NS points for Nigeria and store them into a file called `"raw_tfr_user.csv"`:

```
R> countries <- c(566, 840)
R> myrawTFR <- subset(rawTFR, country_code %in% countries)
R> myrawTFR <- subset(myrawTFR, !(country_code == 566
+    & method == "Indirect" & source == "DHS-NS"))
R> write.csv(myrawTFR, file = "raw_tfr_user.csv", row.names = FALSE)
```

For experimentation with the `run.tfr.mcmc.extra` function, we recommend copying the main simulation into a different directory and applying the function to the copy. This is because the processing overwrites the existing estimation results, and thus there is no way back to the original results in case the experiments do not yield satisfactory outputs. Here we will append `"_extra"` to the directory name stored in `simu.dir.unc` and copy the content from `simu.dir.unc` into it. This step is equivalent to the command `"cp -r annual_unc annual_unc_extra"` on unix-based systems:

```
R> simu.dir.extra <- paste0(simu.dir.unc, "_extra")
R> dir.create(simu.dir.extra)
R> file.copy(list.files(simu.dir.unc, full.names = TRUE), simu.dir.extra,
+    recursive = TRUE)
```

To run the new estimation for the two selected countries, we can do

```
R> run.tfr.mcmc.extra(sim.dir = simu.dir.extra, countries = countries,
+    iter = total.iter, burnin = burnin, uncertainty = TRUE,
+    my.tfr.raw.file = "raw_tfr_user.csv",
+    covariates = c("source", "method"))
```

We recommend using the same values of `total.iter` and `burnin` as in the main simulation.

To compare the new estimation results to those shown in Figure 1 we again use the `tfr.estimation.plot` function, now passing `simu.dir.extra` into the `sim.dir` argument. It can be seen in the left panel of Figure 6 that excluding the Indirect DHS-NS data points for Nigeria changed the estimates, especially for 1979. The uncertainty increased for the United States (right panel of Figure 6), since it was removed from the `iso.unbiased` set.

Finally, the option `uncertainty = TRUE` can be used even in two-step estimation where uncertainty about the past was not taken into account. This is possible because we do not expect the global parameters to be significantly different in the two situations (i.e., with and without uncertainty).

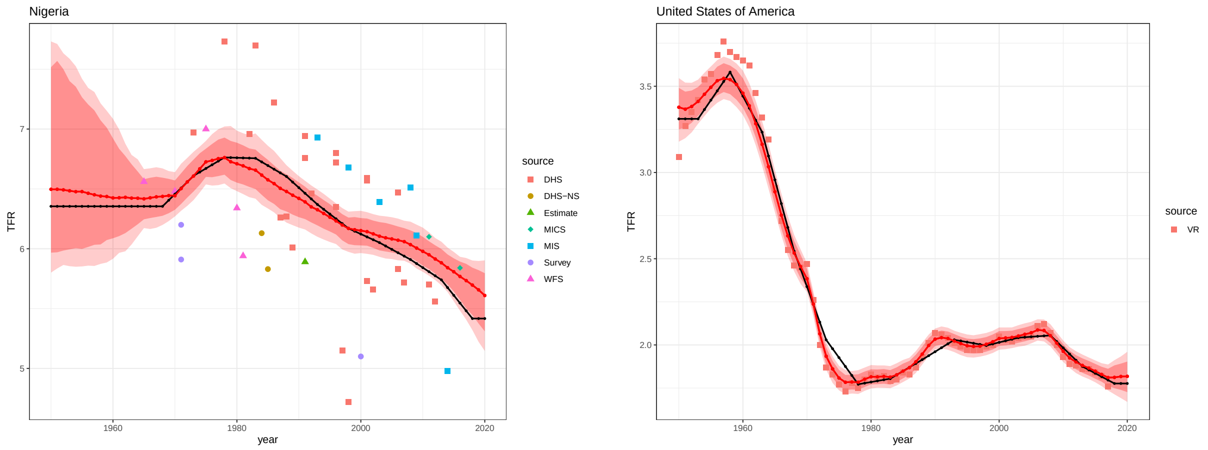


Figure 6: TFR estimation for Nigeria (left panel) and the United States (right panel), resulting from a non-converged simulation with modified data set.

4.6. Structure of the output directory

Having a look at the simulation directory, here `annual_unc`, one should see a structure similar to the following:

```
annual_unc
├── bayesTFR.mcmc.meta.rda
├── diagnostics
├── mc1
├── mc2
├── mc3
├── phaseIII
│   ├── bayesTFR.mcmc.meta.rda
│   ├── mc1
│   ├── mc2
│   └── mc3
├── predictions
└── thinned_mcmc_9_2100
    ├── bayesTFR.mcmc.meta.rda
    └── mc1
```

The directories `mc1`, `mc2` and `mc3` on the first level are generated by the `run.tfr.mcmc` function and contain results from the three chains of the Phase II estimation. Each of the directories contains one text file per parameter. The names of the hyperparameters and their corresponding notation are the same as described in Table 1 in Ševčíková *et al.* (2011). In addition, the parameter `rho_phase2` representing ϕ from Equation 11 is also stored as a hyperparameter if the phase II-AR(1) is considered. The names of the files storing country-independent parameters consist of the parameter name and the suffix `".txt"`, while in the case of the files storing country-specific parameters the parameter name is followed by the suffix `"_country`<*code*>`.txt"`.

If uncertainty is taken into account, the MCMC algorithm also generates estimates for the

$\bar{\mu}$	$\bar{\rho}$	σ_{μ}	σ_{ρ}	σ_{ε}
`mu`	`rho`	`sigma.mu`	`sigma.rho`	`sigma.eps`

Table 4: Country-independent parameters for phase III in Model 7, with their corresponding names in the code. They can be obtained using `tfr3.parameter.names()`.

μ_c	ρ_c
`mu.c`	`rho.c`

Table 5: Country-specific parameters for phase III in Model 7, with their corresponding names in the code. They can be obtained using `tfr3.parameter.names.cs()`.

past TFR data. These samples are considered as country-specific parameters, called `"tfr"`, and thus stored in files `"tfr_country`<*code*>`.txt"`. They contain matrices of size the number of (thinned) iteration times the number of time periods. In the example above, the default starting year is 1950, and the present year is 2020, i.e., 71 years. Therefore, each file contains TFR estimates in 5,100 rows and 71 columns.

The file `"bayesTFR.mcmc.meta.rda"` on the first level stores meta information about the phase II estimation, which is contained in the `m$meta` object. If uncertainty is taken into account, the raw data used to obtain the estimates of TFR are stored as an additional element, called `raw_data.original`. A logical element `ar.phase2` indicates whether the autoregressive component of phase II is considered in the estimation. In order to allow users to work with different subsets of countries with the same base of global estimates, information indicating whether the countries were processed separately has been also stored in the `meta` object. It is accessible via the `extra` element, created only if the `run.tfr.mcmc.extra` function has been invoked and if `uncertainty` is `TRUE`. Here, `extra_iter` and `extra_thin` are used to retrieve the settings for specific countries. The raw data in this case are stored in a list called `raw_data_extra`. It is overwritten every time `run.tfr.mcmc.extra` is called for the same country.

The results of phase III are stored in the directory `"phaseIII"`. It has the same structure as described above. It is generated either by the `run.tfr.mcmc` function in case of a one-step estimation, or by the `run.tfr3.mcmc` function, in case of a two-step estimation. The meta file contains meta information related to the phase III estimation. In the `"mc`x`"` directories, the names of the hyperparameters and their notations for phase III are listed in Table 4. Similarly, the country-specific parameters and their notations are listed in Table 5. All files in this case contain one value per (thinned) iteration. Note that the country-specific parameters for phase III are only estimated for countries which are already in phase III, which in our case is 41 countries.

The `"predictions"` directory is created by the `pop.predict` function and it holds binary files, one per country, each containing the predicted TFR trajectories for that country.

Other convenience directories might have been created for speeding up processing. For example, the `"thinned_mcmc_9_2100"` directory was created by `pop.predict` to hold the final chain for each parameter derived by applying the burnin, thinning and collapsing the three chains into one, in order to generate the predictions. Since we asked to generate 1,000 posterior TFR trajectories with burnin of 2,100 iterations, a thinning of 9 was applied to retrieve those trajectories: $3 \cdot (5,100 - 2,100)/9 = 1,000$. Thus, the parameter files in the `"mc1"` subdirectory here all contain $1,000$ rows. Note that these values will differ when working with a toy simulation.

If functions for convergence diagnostics have been used, the simulation directory contains a folder `"diagnostics"` which holds results from these runs, one file per unique combination of thin and burnin.

5. Experiments

We have shown how the updated **bayesTFR** package can handle different versions of the TFR projection model. In this section, we will present results of experiments under different settings and discuss the implications of these settings. Based on those experiments we will give recommendations for a reasonable configuration of the model. Finally, we will discuss future directions in the development of the package.

5.1. Experiments with settings

The new version of **bayesTFR** allows the user to handle different types of modeling needs, summarized in Table 1. An analyst can choose between a five-year and an annual model, as well as between accounting for past uncertainty or not. Flexibility is added by allowing the user to treat VR records for selected countries as unbiased, as well as using the autoregressive component in phase II.

However, a question of consistency of results between the various settings may arise. For example, a forecast should not change dramatically when switching from five-year to annual data. Currently, there are no annual observations collected for all countries, and only a few countries (such as New Zealand) have good annual VR data, the only available annual observations. Thus, if past uncertainty is not taken into account the model would be estimated on some version of interpolated data for most countries.

Countries in phase III with good records

The first major difference can be seen for countries in phase III, especially for countries with high quality VR records. We take Switzerland as an example. The left panel of Figure 7 shows TFR projections for a five-year model without accounting for past uncertainty (cell D in Table 1), while the right panel shows results from an annual model with uncertainty about the past (cell A in Table 1). It can be seen that the results on the right yield wider probability intervals. For countries like Switzerland, the bias and uncertainty of past estimation is very low. Since the estimating process takes the linear interpolated TFR as the reference, the process can add extra bias to these data. Even though this is not large, the uncertainty propagated from the beginning of the forecast period could lead to a large difference.

Now we consider the VR records for a set of selected countries (OECD and some developed countries as unbiased; the list can be found in the Appendix). The corresponding TFR projections for Switzerland are shown in the right panel of Figure 8.

It can be seen that when compared to results from a five-year model (left panel), the differences between the two sets of projections are negligible. It is important especially for countries with nearly perfect historical data, such as Switzerland, that similar results be obtained whether annual or five-year data are used.

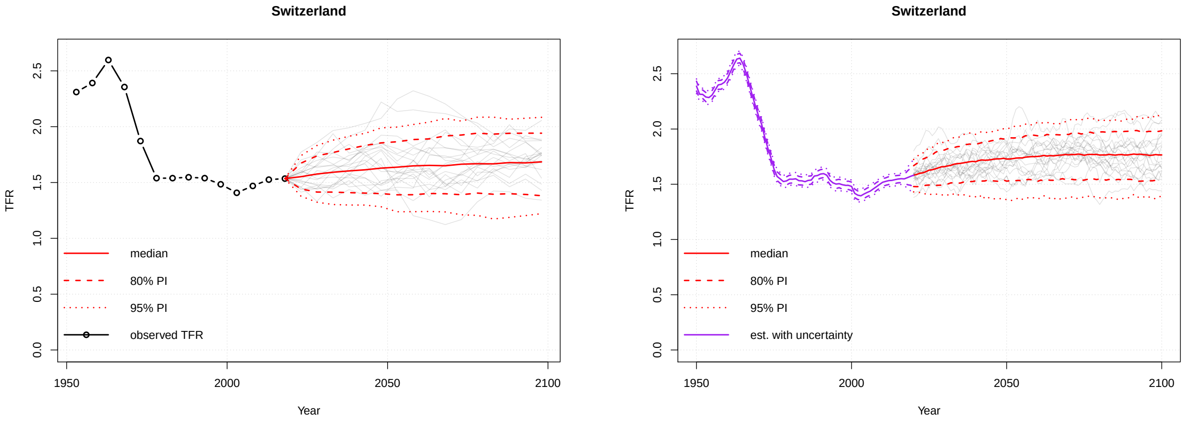


Figure 7: TFR predictions for Switzerland. Left panel: Original five-year model without accounting for past uncertainty. Right panel: Annual model with past uncertainty.

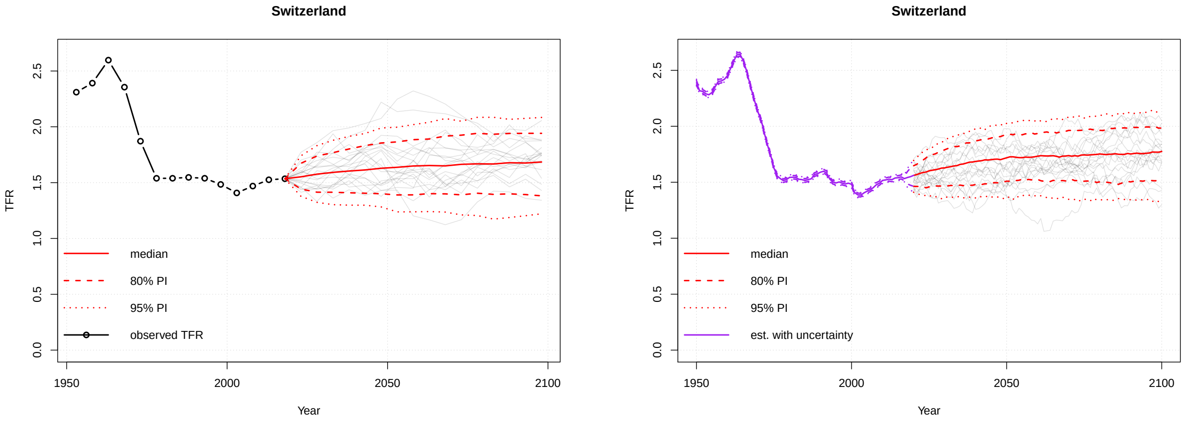


Figure 8: TFR prediction of Switzerland. Left panel: Original five-year model without accounting for past uncertainty. Right panel: Annual model with past uncertainty, with assuming VR records of selected countries (including Switzerland) as unbiased.

Countries in phase II

The second major difference relates to countries in phase II, such as Nigeria. Figure 9 shows the difference between a projection resulting from a five-year model without accounting for past uncertainty (left panel) and from an annual model with uncertainty about the past without applying the phase II-AR(1) component.

It can be seen that if we account for uncertainty and use annual data, the prediction shows a faster decline. Without performing an out-of-sample validation, it is impossible to say which of these projections is better. Nevertheless, a more detailed analysis revealed that the posterior median of the residuals $\varepsilon_{c,t}$ for all countries in Model 2 is highly autocorrelated. Figure 10 summarizes the estimates.

This suggests including the autocorrelation process in the modeling as defined in Equation 11. Figure 11 summarizes the differences. The decline has become slower, which is more in line

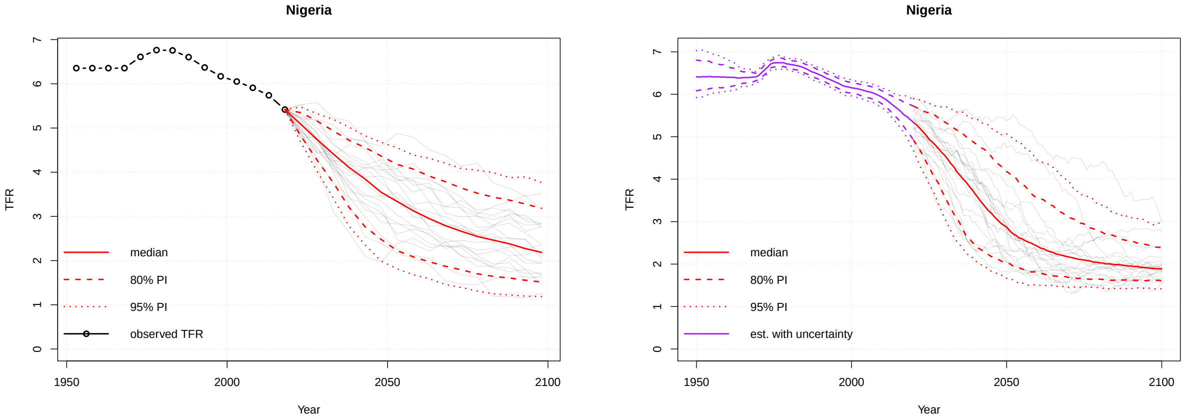


Figure 9: TFR prediction of Nigeria. Left panel: Original five-year model without accounting for past uncertainty. Right panel: Annual model with past uncertainty without phase II-AR(1).

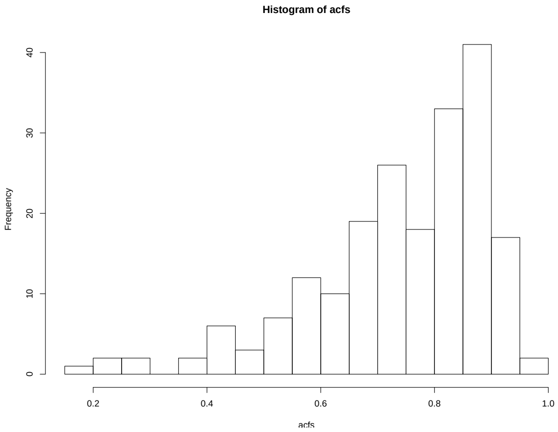


Figure 10: Histogram of autocorrelation for median phase II residuals of all countries.

with the five-year projections.

It can be seen however, that the starting point of the projections (year 2020) is now lower, and in fact it is significantly lower than the current UN estimates. The standard deviation of ε in Model 11 is less than 0.02, if the autoregressive component is included. This could be problematic, given that for developed countries with low TFR and relatively stable societies, the standard deviation of annual TFR changes is about twice as much as 0.02. This is likely a result of a possible smoothing of the data. To remedy that, we introduce a new lower bound on the σ_0 parameter (argument `sigma0.min` in `run.tfr.mcmc`) of 0.04, which becomes the new default. Figure 12 shows the relevant differences.

If the lower bound on σ_0 is applied, the prediction yields wider probability intervals as well as a higher median (top right panel), which better matches the five-year forecast. The estimation in this case (bottom right panel) also shows a better match with the raw data as well as with the UN estimates, which is another argument for using the new default for `sigma0.min`.

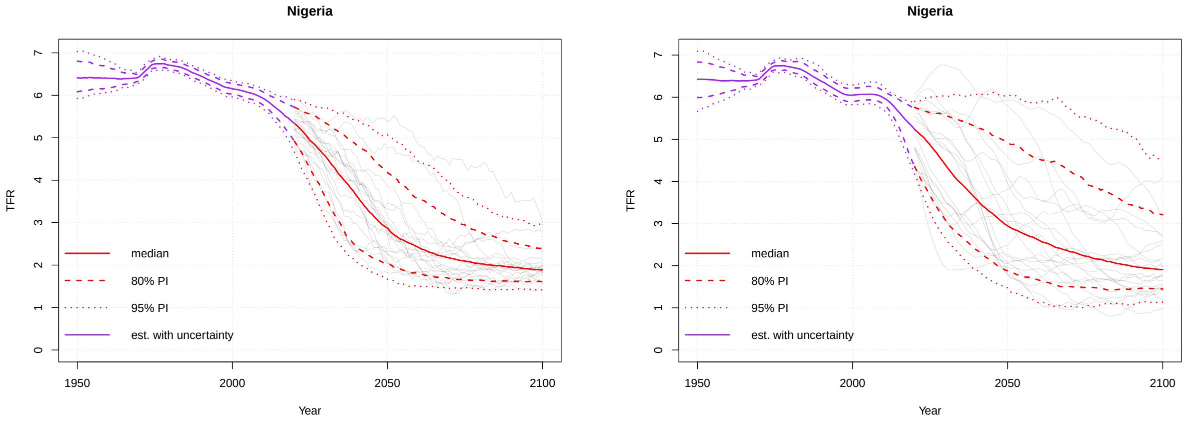


Figure 11: TFR prediction for Nigeria resulting from an annual model with past uncertainty without phase II-AR(1) (left panel) and with phase II-AR(1) (right panel).

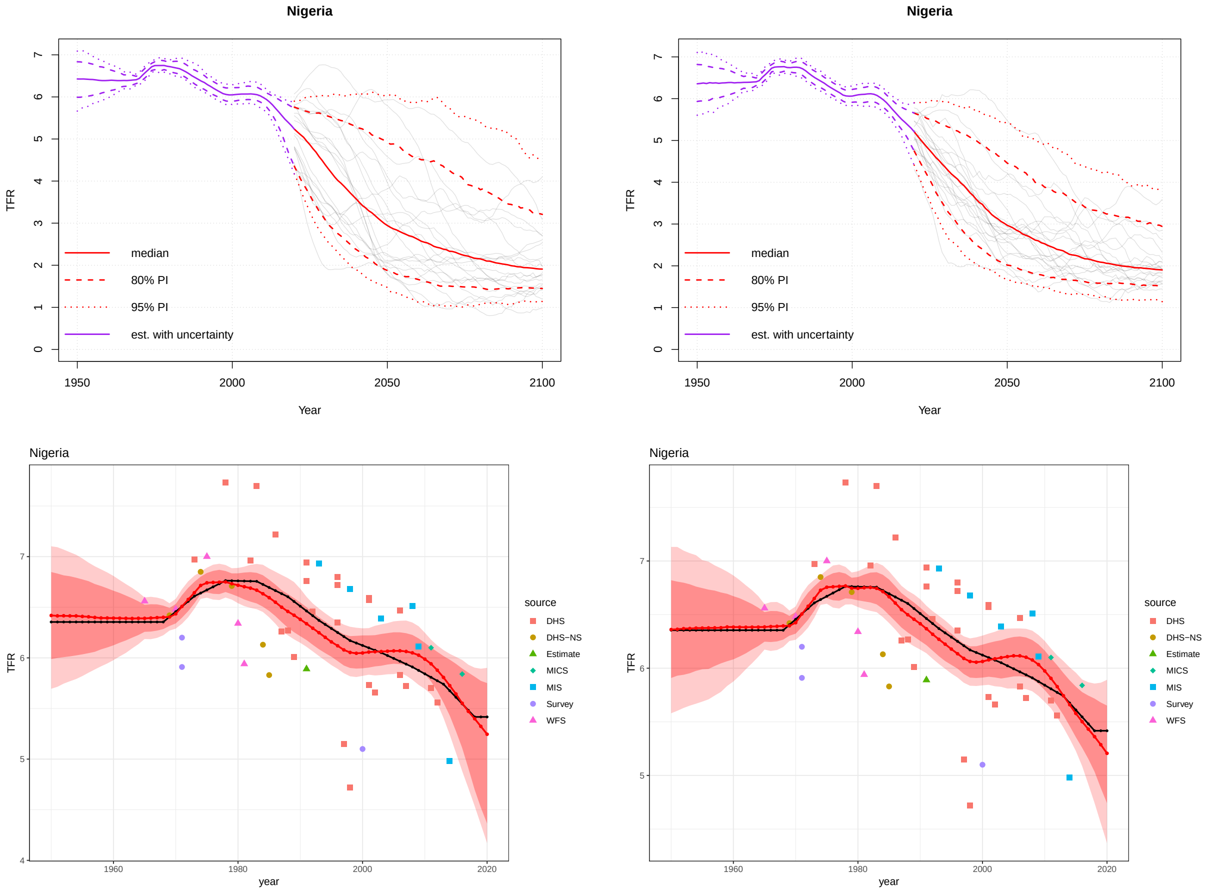


Figure 12: TFR prediction (top row) and estimation (bottom row) for Nigeria from an annual model with uncertainty with autoregressive component. Left column: without lower bound on σ_0. Right column: with `sigma0.min = 0.04`.

5.2. Recommendations

We have shown the flexibility of the new version of **bayesTFR** which can incorporate different variations of the TFR model as well as being compatible with the extant version of the model. As one of the key components in population projections currently adopted by the United Nations, this is a key step for migrating population projections from a five-year basis to an annual one. The package is designed to support UN analysts in this process, as well as to give other researchers and practitioners a tool to generate their own projections.

In addition to incorporating past uncertainty of TFR in the forecast, and performing annual-based projections, the package has introduced two other important components, namely the ability to specify VR data as unbiased, and the autoregressive component in phase II. In Section 5.1, we have described the reasoning behind these two new options, as well as for setting a lower bound on the standard deviation.

Based on our experiments and analysis, when using the annual model with uncertainty about the past in a production-like setting, i.e., if full convergence of the MCMC algorithm is desired, we recommend the following settings:

```
R> annual <- TRUE
R> nr.chains <- 3
R> total.iter <- 62000
R> thin <- 10
R> burnin <- 2000
R> iso.goodvr <- c(36, 40, 56, 124, 203, 208, 246, 250, 276, 300, 352, 372,
+    380, 392, 410, 428, 442, 528, 554, 578, 620, 724, 752, 756, 792, 826,
+    840)
R> m <- run.tfr.mcmc(output.dir = simu.dir.unc, nr.chains =
+    nr.chains, iter = total.iter, annual = annual, thin = thin,
+    uncertainty = TRUE, ar.phase2 = TRUE, iso.unbiased = iso.goodvr,
+    parallel = TRUE)
R> pred <- tfr.predict(sim.dir = simu.dir.unc, end.year = 2100,
+    burnin = burnin, nr.traj = 1000, uncertainty = TRUE)
```

The ISO codes listed include most European countries, Australia, Japan, South Korea, New Zealand and the United States. These countries have a long history of vital registration with coverage rates often around 99%, indicating that their observations have been of high quality.

You should expect a full simulation with these settings to run for several days. Thus, we recommend processing it by a batch script in the background, so that it can be left unattended.

6. Discussion

In this article, we have described the latest major update of the R package **bayesTFR**. This update significantly enriches the modeling framework in the previous version of the package, and gives analysts the flexibility to account for past TFR uncertainty, use annual data, and allow for an autoregressive model in phase II. Moreover, by making use of the vectorization nature of R (R Core Team 2023), the computational cost has been kept at a reasonable level while making the model more sophisticated. New functions for visualizing estimation results, as well as updated analysis tools will further support analysts in exploring the package outputs.

On the package development side, there are at least two major areas for future improvements. The first is modeling age-specific fertility rates with past uncertainty which is of interest to demographers. The second would be further vectorizing the MCMC process. If past uncertainty is included in the model, updating the estimates of TFR is the most time-consuming part of the process. Since we consider each past TFR per country and time period as a parameter, it adds over $14,000$ parameters in the annual case. Thus, the speed of the Metropolis-Hastings step for updating TFR plays a big role in determining the overall speed of the method. If past uncertainty is not included, updates of country-specific parameters dominate the computing time, and thus are subject to further optimization.

On the modeling side, there are also two obvious directions for improvement. First, instead of modeling the bias and standard deviation based on linear regression for each country separately, these could be folded into the process, giving a fully united probabilistic model. A pooled version could yield more robust estimates, especially given the small amount of data in some surveys. Another direction is related to the completeness of the VR data. The completeness of VR coverage is the most important factor for how precise the VR records are, and this is an important consideration for VR but not for other surveys. Due to the low bias of high quality vital registration systems, more research could be done on how to incorporate this information in the model.

Acknowledgments

This research was supported by NICHD grant no. R01 HD070936. Its content is solely the responsibility of the authors and does not necessarily represent the official views of NICHD or the United Nations. The authors are grateful to Patrick Gerland and Nathan Welch for helpful discussions and insightful comments.

References

Abel GJ, Barakat B, Samir KC, Lutz W (2016). “Meeting the Sustainable Development Goals Leads to Lower World Population Growth.” *Proceedings of the National Academy of Sciences of the United States of America*, **113**(50), 14294–14299. doi:10.1073/pnas.1611386113.

Alkema L, Raftery AE, Gerland P, Clark SJ, Pelletier F (2012). “Estimating Trends in the Total Fertility Rate with Uncertainty Using Imperfect Data: Examples from West Africa.” *Demographic Research*, **26**, 331–362. doi:10.4054/demres.2012.26.15.

Alkema L, Raftery AE, Gerland P, Clark SJ, Pelletier F, Buettner T, Heilig GK (2011). “Probabilistic Projections of the Total Fertility Rate for All Countries.” *Demography*, **48**, 815–839. doi:10.1007/s13524-011-0040-5.

Gelfand AE, Smith AF (1990). “Sampling-Based Approaches to Calculating Marginal Densities.” *Journal of the American Statistical Association*, **85**(410), 398–409. doi:10.1080/01621459.1990.10476213.

Gerland P, Biddlecom A, Kantorová V (2017). "Patterns of Fertility Decline and the Impact of Alternative Scenarios of Future Fertility Change in Sub-Saharan Africa." *Population and Development Review*, **43**, 21–38. doi:10.1111/padr.12011.

Liu P, Raftery AE (2020). "Accounting for Uncertainty about Past Values in Probabilistic Projections of the Total Fertility Rate for Most Countries." *The Annals of Applied Statistics*, **14**(2), 685–705. doi:10.1214/19-aoas1294.

Neal RM (2003). "Slice Sampling." *The Annals of Statistics*, **31**, 705–741. doi:10.1214/aos/1056562461.

Raftery AE, Alkema L, Gerland P (2014). "Bayesian Population Projections for the United Nations." *Statistical Science*, **29**, 58–68. doi:10.1214/13-sts419.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ševčíková H, Alkema L, Liu P, Raftery AE, Fosdick B, Gerland P (2023). ***bayesTFR**: Bayesian Fertility Projection*. R package version 7.3-2, URL https://CRAN.R-project.org/package=bayesTFR.

Ševčíková H, Alkema L, Raftery AE (2011). "**bayesTFR**: An R Package for Probabilistic Projections of the Total Fertility Rate." *Journal of Statistical Software*, **43**(1), 1. doi:10.18637/jss.v043.i01.

Ševčíková H, Raftery AE (2016). "**bayesPop**: Probabilistic Population Projections." *Journal of Statistical Software*, **75**(5), 1–29. doi:10.18637/jss.v075.i05.

Ševčíková H, Raftery AE, Gerland P (2018). "Probabilistic Projection of Subnational Total Fertility Rates." *Demographic Research*, **38**, 1843. doi:10.4054/demres.2018.38.60.

United Nations (2015). *World Population Prospects: The 2015 Revision*. United Nations, New York.

United Nations (2019a). "World Fertility Data 2019." URL https://www.un.org/en/development/desa/population/publications/dataset/fertility/wfd2019.asp.

United Nations (2019b). *World Population Prospects: The 2019 Revision*. United Nations, New York. doi:10.18356/13bf5476-en.

United Nations Population Division (2020). ***wpp2019**: World Population Prospects 2019*. R package version 1.1-1, URL https://CRAN.R-project.org/package=wpp2019.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. URL https://ggplot2.tidyverse.org/.

A. Prior distributions

Here we provide a full description of the Bayesian hierarchical model, which was summarized in the main text for annual model. Level 1 is used if `uncertainty = TRUE`:

$$\begin{aligned}
\text{Level 1: } & y_{c,t,s} \mid f_{c,t} \sim \mathcal{N}(f_{c,t} + \delta_{c,s}, \rho^2_{c,s}) \,, \\
& \mathsf{E}[\delta_{c,s}] = \boldsymbol{x}_{c,s}\boldsymbol{\beta} \,, \\
& \mathsf{E}[\rho_{c,s}] = \boldsymbol{x}_{c,s}\boldsymbol{\gamma} \,; \\
\text{Level 2: } & \text{Phase I: } f_{c,t} = f_{c,t-1} + \varepsilon_{c,t} \,, \\
& \text{Phase II: } f_{c,t} = f_{c,t-1} - d_{c,t-1} \,, \\
& \texttt{ar.phase2 = FALSE} : d_{c,t} = g_{c,t} + \varepsilon_{c,t} \,, \\
& \texttt{ar.phase2 = TRUE} : d_{c,t} - g_{c,t} = \phi(d_{c,t-1} - g_{c,t-1}) + \varepsilon_{c,t} \,, \\
& \text{Phase III: } f_{c,t} = \mu_c + \rho_c(f_{c,t-1} - \mu_c) + \varepsilon_{c,t} \,, \\
& \boldsymbol{\theta}_c = (\Delta_{c1}, \Delta_{c2}, \Delta_{c3}, \Delta_{c4}, d_c) \\
& \varepsilon_{c,t} \sim \mathcal{N}(0, \sigma^2_{c,t}) \,, \\
& g(f_{c,t} \mid \boldsymbol{\theta}_c) = - \frac{d_c}{1 + \exp\left(-\frac{2\ln(9)}{\Delta_{c1}}\left(f_{c,t} - \sum_i \Delta_{ci} + 0.5\Delta_{c1}\right)\right)} \\
& \qquad + \frac{d_c}{1 + \exp\left(-\frac{2\ln(9)}{\Delta_{c3}}\left(f_{c,t} - \Delta_{c4} - 0.5\Delta_{c3}\right)\right)}
\end{aligned}$$

The country-specific variance, $\sigma_{c,t}$, varies according to the phase and the current fertility level, as follows:

$$\begin{aligned}
& \sigma_{c,t} = c_{1975}(t)\left(\sigma_0 + (f_{c,t} - S)(-aI_{f_{c,t}>S} + bI_{f_{c,t}<S})\right) \text{ for } t \text{ is in phase II.} \\
& c_{1975}(t) = cI_{t\leq 1975} + I_{t>1975} \,.
\end{aligned}$$

The country-level parameters, $\{U_c, \rho_c, \mu_c, \gamma_{ci}, \Delta_{c4}, d_c\}$, are specified as follows:

$$\text{Level 3: } U_c \begin{cases} = f_{c,\tau} & \tau_c \geq 1950 \\ \sim U(\min\{5.5, \max_t\{f_{c,t}\}\}, 8.8) & \tau_c < 1950 \end{cases}$$

$$\begin{aligned}
& \phi_c = \log\left(\frac{d_c - 0.05}{0.5 - d_c}\right) \,, \\
& \phi_c \sim \mathcal{N}(\chi, \psi^2) \,, \\
& \Delta'_{c4} = \log\left(\frac{\Delta_{c4} - 1}{2.5 - \Delta_{c4}}\right) \,, \\
& \Delta'_{c4} \sim \mathcal{N}(\Delta_4, \delta_4^2) \,, \\
& p_{ci} = \frac{\Delta_{ci}}{U_c - \Delta_{c4}} \text{ for } i = 1, 2, 3 \,, \\
& p_{ci} = \frac{\exp(\gamma_{ci})}{\sum_j \exp(\gamma_{cj})} \,, \\
& \gamma_{ci} \sim \mathcal{N}(\alpha_i, \delta_i^2) \,, \\
& \mu_c \sim \mathcal{N}(\bar{\mu}, \sigma^2_\mu) \,, \\
& \rho_c \sim \mathcal{N}(\bar{\rho}, \sigma^2_\rho) \,;
\end{aligned}$$

where τ_c is the starting year of phase II for country c.

The hyperparameters are $\{s_\tau, \sigma_0, a, b, S, c, \sigma_\epsilon, \chi, \psi, \Delta_4, \delta_4, \boldsymbol{\alpha}, \boldsymbol{\delta}, \bar{\mu}, \sigma_\mu, \bar{\rho}, \sigma_\rho\}$. Some of these refer to level 2 and some to level 3. The prior distribution of these hyperparameters is as follows (ϕ is used if `ar.phase2 = TRUE`):

$$\begin{aligned}
\text{Level 4: } 1/s_\tau^2 &\sim \text{Gamma}(1, 0.4^2)\,,\\
\sigma_0 &\sim U[0.002, 0.6], \quad \text{recommended } U[0.04, 0.6]\,,\\
a &\sim U[0, 0.2]\,,\\
b &\sim U[0, 0.2]\,,\\
S &\sim U[3.5, 6.5]\,,\\
c &\sim U[0.8, 2]\,,\\
\sigma_\epsilon &\sim U[0, 0.5]\,,\\
\chi &\sim \mathcal{N}(-1.5, 0.6^2)\,,\\
1/\psi^2 &\sim \text{Gamma}(1, 0.6^2)\,,\\
\Delta_4 &\sim \mathcal{N}(0.3, 1)\,,\\
1/\delta_i^2 &\sim \text{Gamma}(1, 1) \text{ for } i = 1, 2, 3, 4\,,\\
\alpha_1 &\sim \mathcal{N}(-1, 1)\,,\\
\alpha_2 &\sim \mathcal{N}(0.5, 1)\,,\\
\alpha_3 &\sim \mathcal{N}(1.5, 1)\,,\\
\bar{\mu} &\sim U[0, 2.1]\,,\\
\sigma_\mu &\sim U[0, 0.318]\,,\\
\bar{\rho} &\sim U[0, 1]\,,\\
\sigma_\rho &\sim U[0, 0.289]\,,\\
\phi &\sim U[0, 1]\,.
\end{aligned}$$

B. List of unbiased VR countries

In this article, we are assuming that the following countries have nearly perfect VR histories:

Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Korea, Latvia, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, the United Kingdom, the United States.

Affiliation:

Peiran Liu
Department of Statistics
University of Washington
Seattle, WA, United States of America
E-mail: prliu@uw.edu

Hana Ševčíková
Center for Statistics and the Social Sciences
University of Washington
Box 354322
Seattle, WA 98195-4322, United States of America
E-mail: hanas@uw.edu

Adrian E. Raftery
Department of Statistics and Sociology
University of Washington
Box 354320
Seattle, WA 98195-4320, United States of America
E-mail: raftery@uw.edu
URL: https://www.stat.washington.edu/raftery/

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
March 2023, Volume 106, Issue 8 *Submitted:* 2020-09-25
doi:10.18637/jss.v106.i08 *Accepted:* 2022-10-10