Web appendix for levels and trends in sex ratio at birth in provinces of Pakistan from 1980 to 2020 with scenario-based missing female birth projections to 2050: a Bayesian modeling approach

Fengqing Chao*[1], Muhammad Asif Wazir[2], and Hernando Ombao[1]

[1]Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[2]UNFPA-Pakistan, Pakistan

October 18, 2021

*Corresponding author (FC); Email: fengqing.chao@kaust.edu.sa

The views expressed in this article are those of the authors and do not necessarily reflect the views of the United Nations Population Fund (UNFPA).

1 List of abbreviations

CV	coefficient of variation
DHS	Demographic and Health Survey
MICS	Multiple Indicator Cluster Survey
SRB	sex ratio at birth

2 Data

2.1 Sampling errors in the DHS data

Both Demographic and Health Survey (DHS) and Multiple Indicator Cluster Survey (MICS) provide individual-level data with the full birth history of each woman of reproductive age interviewed during the survey field-work period. We calculated the sampling error in the log-transformed sex ratio at birth (SRB) obtained from the DHS and MICS data series using the jackknife method [1, 2, 3]. For a certain DHS or MICS data series, let U denote the total number of clusters (based on the cluster/primary sampling unit numbers in the survey data [4]). The u-th partial prediction of SRB is determined by the following equation:

$$r_{-u} = \frac{\sum_{n=1}^{N} \mathbb{I}_n(x_n = \text{male}; d_n \neq u) w_n}{\sum_{n=1}^{N} \mathbb{I}_n(x_n = \text{female}; d_n \neq u) w_n}, \text{ for } u \in \{1, \ldots, U\},$$

where n indexes the live births in each state-survey-year; N is the total number of live births; and x_n, d_n, and w_n are the sex, cluster number, and sampling weight for the n-th live birth, respectively. The sampling weight of each birth w_n is extractable from the survey data and reflect the survey sampling design [4]. We define $\mathbb{I}_n(\cdot) = 1$ if the condition inside the brackets is true and $\mathbb{I}_n(\cdot) = 0$ otherwise. The u-th pseudo-value estimate of the SRB on the log-scale is:

$$\begin{aligned} \log(r)_u^* &= U \log(r') - (U-1)\log(r_{-u}), \text{ where} \\ r' &= \frac{\sum_{n=1}^{N} \mathbb{I}_n(x_n = \text{male}) w_n}{\sum_{n=1}^{N} \mathbb{I}_n(x_n = \text{female}) w_n}. \end{aligned}$$

The sampling variance is:

$$\begin{aligned} \sigma^2 &= \frac{\sum_{u=1}^{U} (\log(r)_u^* - \overline{\log(r)_u^*})^2}{U(U-1)}, \text{ where} \\ \overline{\log(r)_u^*} &= \frac{1}{U} \sum_{u=1}^{U} \log(r)_u^*. \end{aligned}$$

In the DHS or MICS data, the annual log-transformed SRB observations are merged such that the coefficient of variation (CV) for log-transformed SRB is below 0.1 or the merged period reaches five years [5]. For a certain DHS/MICS data series, let $\{t_n, t_{n-1}, \cdots, t_1\}$ be years with recorded births from recent to past. The merge starts from the most recent year t_n and is performed by the following algorithm:

The above described sampling error and merging observation periods are computed for each DHS and MICS data series.

Merging process of DHS and MICS data

1: **for** $t \in \{t_n, t_{n-1}, \cdots, t_1\}$ **do**
2: **if** $t = t_n$ **then**
3: Compute σ as explained in above. Compute CV$= \sigma / \log(r)^*_u$
4: **if** CV < 0.1 **or** $t_n - t_{n-1} > 1$ **or** $t_{n+1} - t_n = 5$ **then**
5: **stop** and move to the previous time point
6: **else**
7: Repeat step 3–5 based on births from t_n and t_{n-1}

3 Bayesian model for provincial SRB estimation and projection

3.1 Notations

Table 1 summarizes the notations and indexes used in this study. $\mathcal{N}(\mu, \sigma^2)$ refers to a normal distribution with mean μ and variance σ^2. $\mathcal{U}(a, b)$ denotes a continuous uniform distribution with lower and upper bounds at a and b respectively.

Symbol	Description
	Index
i	Indicator of the ith SRB observation across all province-years, $i \in \{1, \ldots, 531\}$.
t	Indicator of year, $t \in \{1980, \ldots, 2050\}$.
p	Indicator of provinces of Pakistan, $p \in \{1, \ldots, 7\}$.
	Unknown Parameters
$\Theta_{p,t}$	Model fitting to the true SRB in Pakistan province p in year t.
$\Phi_{p,t}$	Province-year-specific multiplier for capturing the natural fluctuation in SRBs around the national baseline b in Pakistan province p in year t.
$\alpha_{p,t}$	SRB imbalance in Pakistan province p in year t.
t_{0p}	Start year of SRB inflation in Pakistan province p.
δ_p	Indicator of the presence ($\delta_p = 1$) or absence ($\delta_p = 0$) of SRB inflation in Pakistan province p.
ξ_p	Maximum level of SRB inflation in Pakistan province p.
λ_{1p}	Period length of the increase stage of the sex ratio transition in Pakistan province p.
λ_{2p}	Period length of the stagnation stage of the sex ratio transition in Pakistan province p.
λ_{3p}	Period length of the decrease stage of the sex ratio transition in Pakistan province p, which returns the SRB to the national SRB baseline.
ω	Non-sampling error.
	Known Quantities
r_i	The ith SRB observation.
σ_i	Sampling error for the ith SRB observation (computed in Section 2.1).
b	Baseline level of SRB over the whole of Pakistan [6], where $b = 1.063$.
ρ	Autoregressive Indicator of $\Phi_{p,t}$, where $\rho = 0.9$ [6, 7].
σ_ϵ	Standard deviation of distortion parameter for $\Phi_{p,t}$, where $\sigma_\epsilon = 0.004$ [6, 7].

Table 1: **Summary of notations used in this study.**

3.2 Model of sex ratio at birth by Pakistan province

The model is based on the model described in [8] with modifications to allow the model to better address the data quality and availability of provincial SRB data in Pakistan. The outcome of interest $\Theta_{p,t}$, namely, the SRB in Pakistan province p in year t, is modeled as follows:

$$\begin{aligned}
\Theta_{p,t} &= b\Phi_{p,t} + \delta_p \alpha_{p,t}, \\
\log(\Phi_{p,t}) &\sim \mathcal{N}(0, (1-\rho^2)/\sigma_\epsilon^2), \text{if } t = 1980, \\
\log(\Phi_{p,t}) &= \rho \log(\Phi_{p,t-1}) + \epsilon_{p,t}, \text{if } t \in \{1981, \cdots, 2020\}, \\
\epsilon_{p,t} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2),
\end{aligned}$$

where $b = 1.056$ is the SRB baseline level of all Pakistan, estimated from national SRB observations in Pakistan before the reference year 1970 [6, 7]. $\Phi_{p,t}$ follows an AR(1) time series model on the log scale, which captures the natural fluctuations of SRB in each province over time. The values of ρ and σ_ϵ ($\rho = 0.9$ and $\sigma_\epsilon = 0.004$) were not estimated but were borrowed from a previous study [6, 7], which robustly estimated the parameters from an extensive national SRB database.

The binary identifier of the sex ratio transition, δ_p, follows a Bernoulli distribution:

$$\begin{aligned}
\delta_p|\pi_p &\sim \mathcal{B}(\pi_p), \text{ for } p \in \{1, \cdots, 7\}, \\
\text{logit}(\pi_p)|\mu_\pi, \sigma_\pi &\sim \mathcal{N}(\mu_\pi, \sigma_\pi^2), \text{ for } p \in \{1, \cdots, 7\}.
\end{aligned}$$

To ensure that the probability parameter π_p lies in the interval [0, 1], we use the logit-transformed π_p follows a hierarchical normal distribution with a global mean and variance μ_π and σ_π^2, respectively.

$\alpha_{p,t}$ refers to the province-specific SRB imbalance process. It is modeled by a trapezoidal function that represents the increasing, stagnation, and decreasing stages of the sex ratio transition (Figure 1).

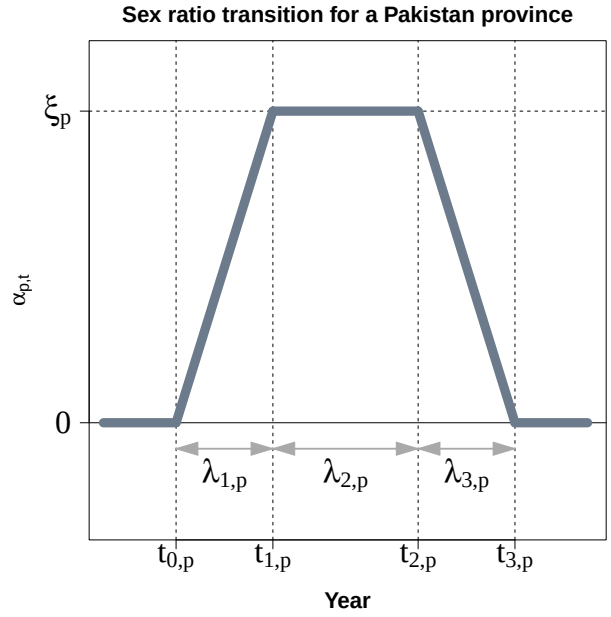

Figure 1: **Representation of the SRB inflation process in a Pakistan province.**

$$\alpha_{p,t} = \begin{cases} (\xi_p/\lambda_{1p})(t-\gamma_p), & t_{0p} < t < t_{1p} \\ \xi_p, & t_{1p} < t < t_{2p} \\ \xi_c - (\xi_c/\lambda_{3p})(t-t_{2p}), & t_{2p} < t < t_{3p} \\ 0, & t < t_{0p} \text{ or } t > t_{3p} \end{cases}$$

$$t_{1p} = t_{0p} + \lambda_{1p}, t_{2p} = t_{1p} + \lambda_{2p}, t_{3p} = t_{2p} + \lambda_{3p}.$$

The start year of the SRB inflation t_{0p} is modeled by a continuous uniform prior distribution with a lower bound at 1970 and upper bound at 2050, respectively. The province-specific period lengths of the three stages of the SRB inflation (λ_{1p}, λ_{2p} and λ_{3p}) are assigned with informative priors (see Section 3.4 for prior specifications).

3.3 Data quality model

r_i is the i-th observed SRB in province $p[i]$ in year $t[i]$, where i indexes all SRB observations across the provinces over time. r_i is assumed to follow a normal distribution on the log scale with mean of $\log(\Theta_{p[i],t[i]})$ (explained above) and variance of σ_i^2:

$$\log(r_i)|\Theta_{p[i],t[i]} \sim \mathcal{N}(\log(\Theta_{p[i],t[i]}), \sigma_i^2 + \omega^2), \text{ for } i \in \{1, \cdots, 531\},$$

where σ_i^2 is the sampling error variance of $\log(r_i)$, which reflects the uncertainty in log-scaled SRB observations because of the survey sampling design. σ_i^2 is calculated using a jackknife method (see Section 2.1). ω^2 is the non-sampling error variance representing the uncertainty contributed by non-responses, recall errors, and data input errors. ω^2 is immeasurable and is estimated using the model.

3.4 Priors

The following informative priors are assigned to the province-level parameters related to the sex ratio transition: the maximum level of SRB inflation ξ_p, and the period lengths for the increasing, stagnation and decreasing stages of the sex ratio transition (λ_{1p}, λ_{2p} and λ_{3p}, respectively). The means of prior distributions are taken from a systematic study [8] which modeled the sex ratio transition of multiple countries, including Pakistan. The standard deviations of prior distribution are set such that the CV (defined as the ratio between mean and standard deviation) is 0.1. The informative priors assist the provincial level modeling of the sex ratio transition in Pakistan by exploiting of the corresponding information at the national level. For $p \in \{1, \cdots, 7\}$, we have:

$$\begin{aligned} \xi_p &\sim \mathcal{N}(0.06, 0.006^2), \\ \lambda_{1p} &\sim \mathcal{N}(11.0, 1.1^2), \\ \lambda_{2p} &\sim \mathcal{N}(7.6, 0.8^2), \\ \lambda_{3p} &\sim \mathcal{N}(16.1, 1.6^2). \end{aligned}$$

The start year of the SRB inflation t_{0p} is modeled as a continuous uniform prior distribution with a lower bound at 1970 and upper bound at 2050, respectively. For $p \in \{1, \cdots, 7\}$, we have:

$$t_{0p} \sim \mathcal{U}(1970, 2050). \tag{1}$$

Vague priors are assigned to the parameters related to the Indicator that detects sex ratio transitions and the standard deviation of the start year.

$$\begin{aligned} \text{inverse-logit}(\mu_\pi) &\sim \mathcal{U}(0, 1), \\ \sigma_\pi &\sim \mathcal{U}(0, 2), \\ \sigma_{t0} &\sim \mathcal{U}(0, 10). \end{aligned}$$

3.5 Scenario-based simulated projections for SRB inflation

The province-specific SRB imbalance process $\delta_p \alpha_{p,t}$ is simulated using posterior samples from the model. The simulated $\delta_p \alpha_{p,t}$ is added to the projected $\Theta_{p,t}$ for different starting years of the SRB inflation in each province.

For $g \in \{1, \ldots, G\}$, the gth simulated SRB inflation based on ${\alpha'_{p[j],t[j]}}^{(g)}$ and $\delta_p^{(g)}$. ${\alpha'_{p[j],t[j]}}^{(g)}$ is the gth simulated SRB imbalance process, with the start year of inflation fixed at $t_0 \in \{2021, \cdots, 2050\}$. ${\alpha'_{p[j],t[j]}}^{(g)}$ and $\delta_p^{(g)}$ are simulated for a "new" province, without considering any province-specific data, following the model specification for these parameters. ${\alpha'_{p[j],t[j]}}^{(g)}$ and $\delta_p^{(g)}$ are simulated using the posterior samples of all parameters and their related hyper-parameters, but with the start year parameter t_{0p} fixed at t_0 for ${\alpha'_{p[j],t[j]}}^{(g)}$.

4 Model validation

The performance of the inflation model was evaluated by two approaches: 1) out-of-sample validation and 2) one-province simulation.

4.1 Out-of-sample validation

We leave out 13% of the data points since the data collection year 2018 instead of reference year, which has been used for assessing model performance of demographic indicators largely based on survey data [9, 10, 11, 12]. After leaving out the data, we fit the model to the training dataset, and obtain point estimates and credible intervals that would have been constructed from the available dataset in the selected survey year.

We calculate the median errors and median absolute errors in the left-out observations. The errors are defined as $e_j = y_j - \widetilde{y}_j$, where $\widetilde{y}_j$ refers to the posterior median of the predictive distribution based on the training dataset for the jth left-out observation y_j. The coverage is given by $1/J \sum \mathbb{I}[y_j \geq l_j] \mathbb{I}[y_j \leq u_j]$, where J refers to the number of left-out observations, and l_j and u_j correspond to the lower and upper bounds, respectively, of the 95% prediction interval of the jth left-out observation y_j. The validation measures are calculated for 1000 sets of left-out observations where each set contains one randomly selected left-out observation from each Pakistan province. The reported validation results are based on the mean outcomes of the 1000 sets of left-out observations.

For the point estimates obtained from the full and training datasets, we define the errors in the true SRB as $e(\Theta)_{p,t} = \widehat{\Theta}_{p,t} - \widetilde{\Theta}_{p,t}$, where $\widehat{\Theta}_{p,t}$ is the posterior median in province p in year t obtained from the full dataset, and $\widetilde{\Theta}_{p,t}$ is the posterior median in the same province-year obtained from the training dataset. Similarly, the error in the sex ratio transition process with probability is defined as $e(\alpha\delta)_{p,t} = \widehat{\alpha}_{p,t}\widehat{\delta}_p - \widetilde{\alpha}_{p,t}\widetilde{\delta}_p$. The coverage is computed similarly to the left-out observations and is based on the lower and upper bounds of the 95% credible interval of $\widetilde{\Theta}_{p,t}$ from the training dataset.

4.2 One-province simulation

We assess the inflation model performance in a one-province simulation. In each of the seven Pakistan provinces, we consider all data points as the test data and simulate the SRB using the posterior samples of the global parameters obtained from the sex ratio transition model (using the full dataset).

The gth simulated SRB $\Theta_{p,t}^{(g)}$ in province p in year t, and the gth simulated SRB $\Theta_{p[j],t[j]}^{(g)}$ for the jth left-out data point in province $p[j]$ in year $t[j]$ are obtained as follows for $g \in \{1, \ldots, G\}$: $\Theta_{p,t}^{(g)} = b\Phi_{p,t}^{(g)} +$

$\alpha_{p,t}^{(g)}\delta_p^{(g)}$, where the simulated $\Phi_{p[j],t[j]}^{(g)}$, $\alpha_{p[j],t[j]}^{(g)}$ and $\delta_p^{(g)}$ refer to a "new" province. This simulation follows the model specifications of these parameters without considering any province-specific data. $\alpha_{p,t}$ and δ_p are simulated using the posterior samples of all parameters and their related hyper-parameters. After generating the simulated values, we calculate results as described for the out-of-sample validation (Section 4.1).

5 Validation and simulation results

Table 2 summarizes the results of the left-out SRB observations in the out-of-sample validation exercise and one-country simulation. The median errors and median absolute errors are nearly zero in the left-out observations. The coverages of the 95% and 80% prediction intervals are more conservative than expected. The wider-than-expected prediction interval in the left-out observations can be primarily attributed to larger uncertainty in more recent observations.

Table 3 compares the model estimates obtained from the full dataset and the training set in the out-of-sample validation exercise. Here we examined the model estimates of the true SRB $\Theta_{r,t}$ and the inflation process with country-specific probability $\delta_r\alpha_{r,t}$. The median errors and the median absolute errors are close to zero.

In summary, the validation results indicate reasonably good calibrations and prediction power of the inflation model with conservative credible intervals.

	Validation Out-of-Sample	**Simulation**
# province in test dataset	6	8
Median error	0.020	-0.003
Median absolute error	0.047	0.071
Below 95% prediction interval (%)	0.0	0.2
Above 95% prediction interval (%)	0.0	3.2
Expected (%)	**2.5**	**2.5**
Below 80% prediction interval (%)	0.0	7.6
Above 80% prediction interval (%)	8.0	9.2
Expected (%)	**10**	**10**

Table 2: **Validation and simulation results for left-out SRB observations.** Error is defined as the difference between a left-out SRB observation and the posterior median of its predictive distribution. SRB observations with data collection years since 2018 are left out. Numbers in the parentheses after the proportions indicate the average number of left-out observations fall below or above their respective 95% and 80% prediction intervals.

Model Validation	$\Theta_{p,t}$			$\delta_p \alpha_{p,t}$		
(Out-of-Sample)	**1995**	**2005**	**2015**	**1995**	**2005**	**2015**
Median error	0.001	0.001	0.002	0.000	0.000	0.000
Median absolute error	0.001	0.001	0.004	0.000	0.000	0.000
Below 95% credible interval (%)	0.0	0.0	0.0	0.0	0.0	0.0
Above 95% credible interval (%)	0.0	0.0	0.0	0.0	0.0	0.0
Expected (%)	**$\leq$2.5**	**$\leq$2.5**	**$\leq$2.5**	**$\leq$2.5**	**$\leq$2.5**	**$\leq$2.5**
Below 80% credible interval (%)	0.0	0.0	0.0	0.0	0.0	0.0
Above 80% credible interval (%)	0.0	0.0	0.0	0.0	0.0	0.0
Expected (%)	**$\leq$10**	**$\leq$10**	**$\leq$10**	**$\leq$10**	**$\leq$10**	**$\leq$10**

Table 3: **Validation results of estimates based on the training set.** Error define the differences between the model estimates (i.e. $\Theta_{p,t}$ or $\delta_p \alpha_{p,t}$) obtained from the full and training datasets, and proportions refer to the proportions (%) of countries in which the median estimates from the full dataset fall below or above their respective 95% and 80% credible intervals respectively, in the training set.

References

[1] Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician. 1983;37(1):36–48.

[2] Efron B, Tibshirani RJ. An introduction to the bootstrap. In: An introduction to the bootstrap. CHAPMAN & HALLI/CRC; 1994. .

[3] International ICF. Demographic and Health Survey Sampling and Household Listing Manual. Calverton, Maryland, U.S.A.: MEASURE DHS; 2012. p. 78–79. Available from: https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

[4] Verma V, Lê T. An analysis of sampling errors for the demographic and health surveys. International Statistical Review/Revue Internationale de Statistique. 1996:265–294.

[5] Pedersen J, Liu J. Child mortality estimation: appropriate time periods for child mortality estimates from full birth histories. PLoS medicine. 2012;9(8).

[6] Chao F, Gerland P, Cook AR, Alkema L. Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. Proceedings of the National Academy of Sciences. 2019;116(19):9303–9311.

[7] Chao F, Gerland P, Cook AR, Alkema L. Web Appendix Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. 2019. DOI: 10.6084/m9.figshare.12442373. Available at https://www.pnas.org/content/pnas/suppl/2019/04/10/1812593116.DCSupplemental/pnas.1812593116.sapp.pdf.

[8] Chao F, Gerland P, Cook AR, Alkema L. Global estimation and scenario-based projections of sex ratio at birth and missing female births using a Bayesian hierarchical time series mixture model. Ann Appl Statist. 2021;15(3):1499–1528.

[9] Alkema L, Wong MB, Seah PR. Monitoring progress towards Millennium Development Goal 4: A call for improved validation of under-five mortality rate estimates. Statistics, Politics and Policy. 2012;3(2).

[10] Alkema L, Chao F, You D, Pedersen J, Sawyer CC. National, regional, and global sex ratios of infant, child, and under-5 mortality and identification of countries with outlying ratios: a systematic assessment. The Lancet Global Health. 2014;2(9):e521–e530.

[11] Chao F, You D, Pedersen J, Hug L, Alkema L. National and regional under-5 mortality rate by economic status for low-income and middle-income countries: a systematic assessment. The Lancet Global Health. 2018;6(5):e535–e547.

[12] Chao F, You D, Pedersen J, Hug L, Alkema L. Web appendix National and regional under-5 mortality rate by economic status for low-income and middle-income countries: a systematic assessment. 2018. DOI: 10.6084/m9.figshare.12442244. Available at https://ars.els-cdn.com/content/image/1-s2.0-S2214109X18300597-mmc1.pdf.