Foundations of Data Science
©American Institute of Mathematical Sciences
Volume **1**, Number **2**, June **2019**
doi:10.3934/fods.2019008
pp. **177–196**

LEVELS AND TRENDS IN THE SEX RATIO AT BIRTH AND MISSING FEMALE BIRTHS FOR 29 STATES AND UNION TERRITORIES IN INDIA 1990–2016: A BAYESIAN MODELING STUDY

Fengqing Chao*

Institute of Policy Studies, Lee Kuan Yew School of Public Policy
National University of Singapore
1C Cluny Road, House 5259599, Singapore

Ajit Kumar Yadav

International Institute for Population Sciences
Govandi Station Road, Mumbai, Maharashtra 400088, India

Abstract. The sex ratio at birth (SRB) has risen in India and reaches well beyond the levels under normal circumstances since the 1970s. The lasting imbalanced SRB has resulted in much more males than females in India. A population with severely distorted sex ratio is more likely to have prolonged struggle for stability and sustainability. It is crucial to estimate SRB and its imbalance for India on state level and assess the uncertainty around estimates. We develop a Bayesian model to estimate SRB in India from 1990 to 2016 for 29 states and union territories. Our analyses are based on a comprehensive database on state-level SRB with data from the sample registration system, census and Demographic and Health Surveys. The SRB varies greatly across Indian states and union territories in 2016: ranging from 1.026 (95% uncertainty interval [0.971; 1.087]) in Mizoram to 1.181 [1.143; 1.128] in Haryana. We identify 18 states and union territories with imbalanced SRB during 1990–2016, resulting in 14.9 [13.2; 16.5] million of missing female births in India. Uttar Pradesh has the largest share of the missing female births among all states and union territories, taking up to 32.8% [29.5%; 36.3%] of the total number.

1. **Introduction.** The sex ratio at birth (SRB; ratio of male to female live births) varies roughly between 1.03 and 1.07 [11, 12] for most of human history. However, since the 1970s, SRBs have risen in several countries in Asia and Eastern Europe [5, 7, 8, 9, 14, 15, 22, 23, 24, 25, 26, 27, 28, 30, 36, 38, 40]. The SRB imbalance is largely drive by the co-existence of three factors that resulted in sex-selective abortion [23, 24]: persisting strong son preference, accessibility and affordability of prenatal sex diagnosis and abortion [3, 20, 21, 39, 51], and fertility decline which leads to smaller family sizes. As a result in such population, sex-selective abortion could allow people to not only avoid large families but also have male offspring.

The imbalanced SRB can lead to severe and prolonged consequences in both demographic and social aspects. The cause of imbalanced SRB, i.e. the practice of

2010 *Mathematics Subject Classification.* Primary: 62H12; Secondary: 91D20.
Key words and phrases. Bayesian hierarchical model, sex ratio at birth, missing female births, India, subnational estimation.

* Corresponding author: Fengqing Chao.

sex-selective abortion, is considered as prenatal gender discrimination and violates baby girls' rights of survival, participation and development. Imbalanced SRB is one of the main factors that leads to the phenomenon of "missing women" [47] with a society having much more males than females. The off-balanced sex structure within a population could exert an indirect effect on demographic issues such as marriage pressure on adult males [35]. This would also lead to increased levels of antisocial behaviour and violence, for instance, and may eventually affect long-term stability and social sustainable development [6, 10, 29, 30].

India has a long history of imbalanced SRB due to sex-selective abortion. In India, prenatal diagnosis (PD) became available soon after abortion was legalized in 1971. PD was introduced in India as a method for detecting fetal abnormalities but was soon used for prenatal sex selection [3, 51]. Since then, the combination of PD and abortion has been widely used for the systematic elimination of females fetuses [37]. First amniocenteses in the 1970s were openly advertised and extensively used in urban areas for sex-selective abortions [34]. The results of the 1981 census already showed skewing of the sex ratio among children between 0 and 6 years, and there were concerns about the sex imbalance in the population [34]. The sharp increase in the child sex ratio since the 1970s is a direct result of the widespread practice of the sex-selective abortions in India [4, 19, 20, 39, 48, 51].

It is crucial to monitor the SRB in India on state level in addition to the national level. India is one of the most populous countries in the world. Its persisting imbalanced SRB on national level has been discussed in multiple studies [12, 23, 25, 49]. At the same time, India is highly heterogeneous across states and union territories in demography. To date, assessments of SRB and corresponding imbalance for Indian states and union territories have largely relied on direct reporting from census, surveys, or sample registration system [31, 32, 33, 45]. Estimation of the degree of SRB imbalance for Indian states and union territories is complicated by the amount of uncertainty associated with SRB observations due to data quality issues, sampling and stochastic errors. An up-to-date systematic analysis for SRB for Indian states and union territories over time using all available data with reproducible methods for estimation is urgently needed.

To fill the research void, we produce model-based SRB annual estimates from 1990 to 2016 for 29 states and union territories in India (Telangana state is not considered since it was established in 2014; Adaman and Nicobar Islands and Pondicherry have not been included because they are very small). Our analyses are based on a comprehensive database on state-level SRB with data from the sample registration system (SRS), census, and Demographic and Health Surveys. We implement a Bayesian hierarchical model to estimate SRB in India on state-level to account for the varying levels of uncertainties associated with observations. Our results imply great disparity of SRB in India across states, over geographic locations and over time. We identify 18 states and union territories with SRB imbalance during 1990–2016. The total number of missing female births for the 18 identified state and union territories during 1990–2016 is estimated to be 14.9 million (95% uncertainty interval [13.2; 16.5]), and almost one third are from Uttar Pradesh.

The structure of the paper is as follows: in Section 2, we summarize the database compiled for model fitting and the calculation of sampling and stochastic errors for observations. In Section 3, we describe the Bayesian hierarchical model that we develop to estimate the state-level SRB over time. In Section 4, we summarize the miscellaneous computations after model fitting, including the calculation of missing

female births and identification of states and union territories with SRB imbalance. We assess the model performance via validation exercise and the approach is explained in Section 5. The estimation and validation results are presented in Section 6. Finally, we summarize the main contributions and study limitations in Section 7.

2. **Data.** Subnational-level data on births by sex are recorded in the sample registration system (SRS), the 2011 India Census and Demographic and Health Surveys (DHS). The SRS provides data on an annual basis, while the 2011 India Census provides information for the previous 24 months and DHS for longer retrospective periods on full birth histories asked to women of reproductive ages.

Table 1 provides an overview of the data sources in the database. There are 937 data points from 29 Indian states and union territories. In total, there are 2,530 state-years of information in the database. On average, 87.2 state-years of data are available for each of the 29 Indian states and union territories.

Data Source Type	**Series Name (Series Period)**	**Total # Obs. (Range # Obs. per state/union territory)**
Census	Census (2011)	29 (1–1)
DHS		598 (4–55)
	DHS (1992–1993)	101 (1–11)
	DHS (1998–1999)	107 (1–12)
	DHS (2005–2006)	100 (1–11)
	DHS (2015–2016)	290 (1–21)
SRS	SRS	310 (3–17)
Total		**937 (5–73)**

TABLE 1. **Observations by source type.** The series periods are in brackets after the series names, they refer to the periods when the fieldwork of the surveys are conducted. The total number of observations are the sum of observations from the 29 Indian states and union territories. DHS: Demographic and Health Surveys. SRS: Sample Registration System.

The SRB database as summarized in Table 1 is based on several steps of data quality checking and pre-processing. The detailed steps in mathematical formulas are in Appendix and are summarized below. We first calculate the sampling error for DHS data series (Appendix A) and the stochastic error for SRS data (Appendix B), for each 1-year observation period. We then merge the observations from DHS and SRS data series based on the variance of log-transformed SRB (Appendix C). The merge of observations is in order to generate observations with associated uncertainty at a reasonable and controlled level. After merging, we exclude DHS observations with reference dates beyond 20 years prior the survey date. This exclusion criterion is to remove observations with potentially larger recall errors and truncation for older women compared to the recent reference period.

3. **Methods.** We develop a Bayesian hierarchical model which captures state-level difference together with a data model to take into account the data quality. The model overview is given in Appendix D. The detailed explanations of the model and the joint density of posterior distribution are in the rest of this section.

3.1. **Data model.** The India state-level SRB observations are indexed by $i = 1, \cdots, n$. r_i denotes the i-th observed SRB in state $s[i]$ and year $t[i]$. $R_{s,t}$ is the outcome of interest, the true SRB in state s in year t. r_i is modeled on log-scale. Let $V_{s,t} = \log(R_{s,t})$, we assume:

$$v_i = V_{s[i],t[i]} + \delta_i,$$

where $v_i = \log(r_i)$ and δ_i is the error term for SRB observations on the log-scale. The distribution of the error term δ_i follows a standard approach that has been used to estimate various population indicators like under-5 child mortality [1, 2, 13], and sex ratio at birth [12]. In the data model, we account for the variations in biases and nonsampling errors across different data collection methods (SRS vs non-SRS). The model also takes account the error differences across data sources as explained further below.

3.1.1. *SRS data.* The error distribution for observations from SRS is given by:

$$\delta_i \sim \mathcal{N}(0, \sigma_i^2), \text{ for } i \in \mathcal{I}_1,$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. σ_i^2 is the stochastic variance and is calculated as described in Appendix B. SRS is usually considered having a high data quality and complete birth record since it is the registration system directly from the government. Hence, the only source of uncertainty for SRS data are assumed to be the stochastic variance σ_i^2. $\mathcal{I}_1 = \{i = 1, \cdots, n | y[i] = \text{SRS}\}$ denotes the set of indexes for SRS observations where $y[i]$ is the source type for the i-th observation (shown in Table 1).

3.1.2. *Non-SRS data.* For observations from DHS series or census, the error is assumed to be normally distributed as:

$$\begin{aligned} \delta_i | \phi_i^2 &\sim \mathcal{N}(0, \phi_i^2), \text{ for } i \in \mathcal{I}_2, \\ \phi_i^2 &= \sigma_i^2 + \omega_{y[i]}^2, \\ \omega_y &\overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 2), \text{ for } y = 1, \cdots, z, \end{aligned}$$

where $\mathcal{I}_2 = \{i = 1, \cdots, n | y[i] \neq \text{SRS}\}$ refers to the set of indexes for non-SRS observations. $\mathcal{U}(a, b)$ denotes a continuous uniform distribution with lower and upper bounds at a and b. $z = 2$ is the total number of data source types for non-SRS observations, i.e. census and DHS. The variance term ϕ_i^2 is modeled as the sum of sampling variance σ_i^2 and non-sampling variance $\omega_{y[i]}^2$. The sampling variance σ_i^2 is given as explained in Appendix A. The non-sampling variance $\omega_{y[i]}^2$ is modeled by data source type to account for the errors that are not possible to quantify. For observations from DHS series or census, the non-sampling variance ω_y^2 is crucial to be taken into account since it is the uncertainty resulted from the errors/mistakes made during the fieldwork or data collection process: e.g. data recoding errors by the interviewers, non-response errors. Non-sampling errors are unavoidable and not possible to eradicate.

3.2. **SRB model.** We model the log-transformed SRB $V_{s,t}$ for an Indian state or union territory s, year t as a sum of two components:

$$\begin{aligned} V_{s,t} &= a0 + P_{s,t}, \text{ for } s = 1, \cdots, k, \text{ for } t = 0, \cdots, h, \\ a0 &\sim \mathcal{N}(\mu_{a0}, \sigma_{a0}^2). \end{aligned}$$

$\exp\{\mu_{a0}\}$ is the national SRB baseline for India estimated to be 1.053 from study [12]. σ_{a0} is set to be 0.002. $k = 29$ refers to the number of Indian states and union territories included in this study. $t = 0$ is equivalent to the year 1990 and $t = h$ refers to the year 2016.

$P_{s,t}$ is the divergence of $V_{s,t}$ from $a0$. $P_{s,t}$ is estimated by an autoregressive time series process of order one (AR(1)) with a normal distribution. It has a state-specific level parameter b_s following a Student t-distribution with mean at 0, a global variance parameter σ_b^2 and degree of freedom of 3. It is modeled to differ across Indian states and union territories to incorporate SRB differences due to demographic heterogeneity on state level in India. For each Indian state and union territory $s = 1, \cdots, k$:

$$\begin{aligned} P_{s,0}|b_s, \rho, \sigma_\epsilon &\sim \mathcal{N}(b_s, \sigma_\epsilon^2/(1-\rho^2)), \\ P_{s,t} &= b_s + \rho \cdot (P_{s,t-1} - b_s) + \epsilon_{s,t}, \text{ for } t = 1, \cdots, h, \\ b_s|\sigma_b &\sim t(0, \sigma_b^2, \nu = 3), \\ \epsilon_{s,t}|\sigma_\epsilon &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \text{ for } t = 1, \cdots, h. \end{aligned}$$

The fluctuation term $\epsilon_{s,t}$ follows an i.i.d. normal distribution with mean at zero and global variance parameter σ_ϵ^2. For $s = 1, \cdots, k$, and $t = 1, \cdots, h$, the conditional distribution for $P_{s,t}$ given the starting point of time series $P_{s,0}$ is:

$$P_{s,t}|P_{s,0} = b_s(1-\rho^t) + \rho^t P_{s,0} + \sum_{m=0}^{t-1} \epsilon_{s,t-m}\rho^m,$$

$$P_{s,t}|P_{s,0}, b_s, \rho, \sigma_\epsilon \sim \mathcal{N}\left(b_s(1-\rho^t) + \rho^t P_{s,0}, \frac{\sigma_\epsilon^2}{1-\rho^2}(1-\rho^{2t})\right).$$

Mutually independent priors are assigned to ρ, σ_ϵ and σ_b:

$$\begin{aligned} \rho &\sim \mathcal{U}(0,1), \\ \sigma_\epsilon &\sim \mathcal{U}(0, 0.01), \\ \sigma_b &\sim \mathcal{U}(0, 0.02). \end{aligned}$$

The prior for ρ is chosen such that the model only allows the AR(1) series to converge back to the state-specific level b_s from one direction rather than fluctuating up and down.

3.3. **Posterior distribution.**

3.3.1. *Likelihood.* For the i-th observation where data type $y[i]$ is SRS, its likelihood on log-scale up to proportion is:

$$p(v_i|V_{s[i],t[i]}) \propto \exp\left\{\frac{V_{s[i],t[i]}^2 - 2v_i V_{s[i],t[i]}}{2\sigma_i^2}\right\}.$$

Since $V_{s,t} = a0 + P_{s,t}$, the likelihood is:

$$p(v_i|a0, P_{s[i],t[i]}) \propto \exp\left\{\frac{2v_i(a0 + P_{s[i],t[i]}) - (a0 + P_{s[i],t[i]})^2}{2\sigma_i^2}\right\}, \text{ for } i \in \mathcal{I}_1.$$

For the i-th observation where data type $y[i]$ is non-SRS (i.e. DHS and census), its likelihood on log-scale up to proportion is:

$$p(v_i|V_{s[i],t[i]},\omega_{y[i]}) \propto \frac{1}{\sqrt{\sigma_i^2+\omega_{y[i]}^2}} \exp\left\{-\frac{(v_i-V_{s[i],t[i]})^2}{2(\sigma_i^2+\omega_{y[i]}^2)}\right\}.$$

Similarly, the likelihood can be written as:

$$p(v_i|a0,P_{s[i],t[i]},\omega_{y[i]}) \propto \frac{1}{\sqrt{\sigma_i^2+\omega_{y[i]}^2}} \exp\left\{-\frac{(v_i-a0-P_{s[i],t[i]})^2}{2(\sigma_i^2+\omega_{y[i]}^2)}\right\}, \text{ for } i\in\mathcal{I}_2.$$

3.3.2. *Posterior density.* The joint posterior density for all parameters and hyper parameters up to proportion is:

$$\begin{aligned}
&p(V_{1:k,0:h},\omega_{1:z},a0,P_{1:k,0:h},b_{1:k},\rho,\sigma_\epsilon|v_{1:n}) \propto \\
&\frac{\sigma_b^{3k}(1-\rho^2)^{k(h-1)/2}\prod_{i\in\mathcal{I}_2}(\sigma_i^2+\omega_{y[i]}^2)^{-1/2}}{\sigma_\epsilon^{k(h-1)}\left[\prod_{t=1}^h(1-\rho^{2t})^{k/2}\right]\prod_{s=1}^k(3\sigma_b^2+b_s^2)^2} \times \exp\left\{\frac{2\mu_{a0}a0-a0^2}{2\sigma_{a0}^2}\right\} \times \\
&\exp\left\{\sum_{i\in\mathcal{I}_1}\frac{2v_i(a0+P_{s[i],t[i]})-(a0+P_{s[i],t[i]})^2}{2\sigma_i^2}-\sum_{i\in\mathcal{I}_2}\frac{(v_i-a0-P_{s[i],t[i]})^2}{2(\sigma_i^2+\omega_{y[i]}^2)}\right\} \times \\
&\exp\left\{-\sum_{s=1}^k\sum_{t=1}^h\frac{(1-\rho^2)(P_{s,t}-b_s+b_s\rho^t-\rho^tP_{s,0})^2}{2\sigma_\epsilon^2(1-\rho^{2t})}-\sum_{s=1}^k\frac{\sigma_\epsilon^2(P_{s,0}-b_s)^2}{2(1-\rho^2)}\right\}.
\end{aligned}$$

4. **Post-modeling calculation.**

4.1. **Computing and estimation.** We obtain posterior samples of all the model parameters and hyper parameters using a Markov chain Monte Carlo (MCMC) algorithm, implemented in the open source software `R 3.5.1` [44] and `JAGS 4.1.0` [42] (Just another Gibbs Sampler), using `R`-packages **R2jags** [50] and **rjags** [41]. Results are obtained from 12 chains with a total number of 2,000 iterations in each chain, while the first 10,000 iterations are discarded as burn-in, and thinning for every 2 iterations. The final posterior sample size for each parameter is 4,000. Convergence of the MCMC algorithm and the sufficiency of the number of samples obtained are checked through visual inspection of trace plots and convergence diagnostics of Gelman and Rubin [18], implemented in the **coda** `R`-package [43].

4.2. **Calculation of missing female births.** The estimated and expected female live births for an Indian state or union territory s year t, denoted as $B_{s,t}^F$ and $B_{s,t}^{FE}$ respectively, are obtained as follows:

$$B_{s,t}^F = \frac{B_{s,t}}{1+R_{s,t}},$$
$$B_{s,t}^{FE} = \frac{B_{s,t}-B_{s,t}^F}{\exp\{a0\}}.$$

The annual number of missing female births (AMFBs) for an Indian state or union territory s in year t is defined as:

$$B_{s,t}^{F*} = B_{s,t}^{FE} - B_{s,t}^F.$$

The cumulative number of missing female births (CMFBs) for period t_1 to t_2 in an Indian state or union territory s is defined as the sum of AMFBs from the year t_1 up to the year t_2:

$$Z^{F*}_{s,[t_1,t_2]} = \sum_{t=t_1}^{t_2} B^{F*}_{s,t}.$$

4.3. **Identifying Indian states/union territories with outlying missing female births.** An Indian state or union territory is identified to have outlying missing female births (i.e. SRB imbalance) if its AMFB in at least one year since 1990 is above zero for more than 95% of the posteriors samples. That is:

$$\sum_{t=0}^{h} \mathbb{I}_t \left\{ \sum_{g=1}^{G} \mathbb{I}_g \left[\left(B^{F*}_{s,t} \right)^{(g)} > 0 \right] / G > 95\% \right\} \geq 1,$$

where $\left(B^{F*}_{s,t} \right)^{(g)}$ is the g-th posterior sample of the AMFB for Indian state or union territory s in year t. $\mathbb{I}_n(\cdot) = 1$ if the condition inside brackets is true and $\mathbb{I}_n(\cdot) = 0$ otherwise.

5. **Model validation.** To test the performance for the model, we leave out data points after a certain survey year [2]. Normally, the left-out observation is around 20% of the total observations. However, given the scarcity of the data for state-level SRB in India, we leave out data after survey year 2015 which is the most recent survey year. 33.1% of the total observations are left out. After leaving out data, we fit the model to the training data set, and obtain point estimates and uncertainty intervals that would have been constructed based on available data set in the survey year selected.

For each log-transformed left-out observation $v_j = \log(r_j)$, we simulate its predictive probability distribution (PPD) $\left\{ v_j^{(g)} | g = 1, \cdots, G \right\}$. Let $v_j^{(g)}$ be the g-th simulated PPD for v_j, it is simulated as:

$$v_j^{(g)} \sim \mathcal{N}\left(a0^{(g)} + P^{(g)}_{s[j],t[j]}, \sigma_j^2 \right), \text{ for } j \in \mathcal{I}_1,$$
$$v_j^{(g)} \sim \mathcal{N}\left(a0^{(g)} + P^{(g)}_{s[j],t[j]}, \sigma_j^2 + \left(\omega^{(g)}_{y[j]} \right)^2 \right), \text{ for } j \in \mathcal{I}_2.$$

$s[j]$, $t[j]$ and $y[j]$ refer to the Indian state and union territory, the reference year and the data source type for the j-th left-out observation r_j. Let $\widetilde{r}_j$ denote the posterior median of the PPD for $\exp\{v_j\}$:

$$\widetilde{r}_j = \text{median} \left\{ \exp\left\{ v_j^{(g)} \right\} | g = 1, \cdots, G \right\}.$$

We calculate median errors and median absolute errors for the j-th left-out observations, where error is defined as:

$$e_j = r_j - \widetilde{r}_j.$$

Coverage is given by:

$$\frac{1}{J} \cdot \sum_{j=1}^{J} \mathbb{I}_j(r_j \geq l_j) \cdot \mathbb{I}_j(r_j \leq u_j),$$

where J refers to the total number of left-out observations, and l_j and u_j correspond to the lower and upper bounds of the 95% prediction interval for the left-out

observation r_j. The validation measures are calculated for 1000 sets of left-out observations, where each set consists one randomly selected left-out observation from each Indian state or union territory. The reported validation results are based on the mean of the outcomes from the 1000 sets of left-out observations.

For the median estimates based on full data set and training data set, error is defined as:

$$e_{s,t} = \widehat{R}_{s,t} - \widetilde{R}_{s,t},$$

where $\widehat{R}_{s,t}$ is the posterior median for state s in year t based on the full data set, and $\widetilde{R}_{s,t}$ is the posterior median for the same state-year based on the training data set. Coverage is computed in a similar manner as for the left-out observations, based on the lower and upper bounds of the 95% uncertainty interval of $\widetilde{R}_{s,t}$ from the training data set.

6. **Results.**

6.1. **Sex ratio at birth for Indian states/union territories.**

The levels of SRB in India varies across states and union territories in 2016 (Figure 1 and Table 2). In 2016, the highest SRB is estimated in Haryana at 1.181 (95% uncertainty interval [1.143; 1.217]), followed by Punjab at 1.156 [1.115; 1.194] and Uttarakhand at 1.152 [1.116; 1.189]. The lowest SRB among the 29 Indian states and union territories are estimated in Mizoram at 1.026 [0.971; 1.087], Chhattisgarh at 1.035 [1.001; 1.070] and Kerala at 1.038 [1.006; 1.069]. In 2016, SRB in 13 states and union territories are significantly above the national SRB baseline value 1.053 (Table 2): Andhra Pradesh, Assam, Bihar, Delhi, Gujarat, Haryana, Jammu and Kashmir, Jharkhand, Madhya Pradesh, Punjab, Rajasthan, Uttar Pradesh, Uttarakhand.

Between 1990 and 2016, the changes in SRB are not significantly different from zero for all states and union territories. 13 states have increases in their SRB point estimates between 1990 and 2016. The largest increases in SRB point estimates between 1990 and 2016 are in Andhra Pradesh at 0.035 [-0.018; 0.087] and in Uttarakhand at 0.034 [-0.043; 0.109]. Meanwhile, the greatest decreases in SRB point estimates during the same period are in Punjab with a decrease at -0.054 [-0.119; 0.008] and in Himachal Pradesh at -0.029 [-0.097; 0.037].

Geographically, there is a large amount of heterogeneity in SRB across the Indian states and union territories in 1990 and in 2016 (Figure 2). In general, the highest SRB are concentrated in most of the north-western states and union territories and the lowest SRB are mostly estimated to be in the southern part of India in both 1990 and 2016 but with exceptions. In 1990, the SRB are the highest in northern states and union territories Punjab at 1.210 [1.157; 1.264], Haryana at 1.181 [1.130; 1.233] and Jammu and Kashmir at 1.150 [1.087; 1.215]. SRB become lower as the states and union territories are further in the south except for Chhattisgarh. Chhattisgarh has one of the lowest SRBs in 1990 (1.038 [0.985; 1.094]) but is surrounded by states and union territories with much higher SRB. Comparing to the 1990 SRB geographic distribution, the state-level SRB in 2016 are more divided towards the low and high ends: SRB point estimates is lower than 1.05 in seven states and union territories, increased from two states and union territories back in 1990; while four states and union territories have their SRB higher than 1.14 in 1990 and 2016.

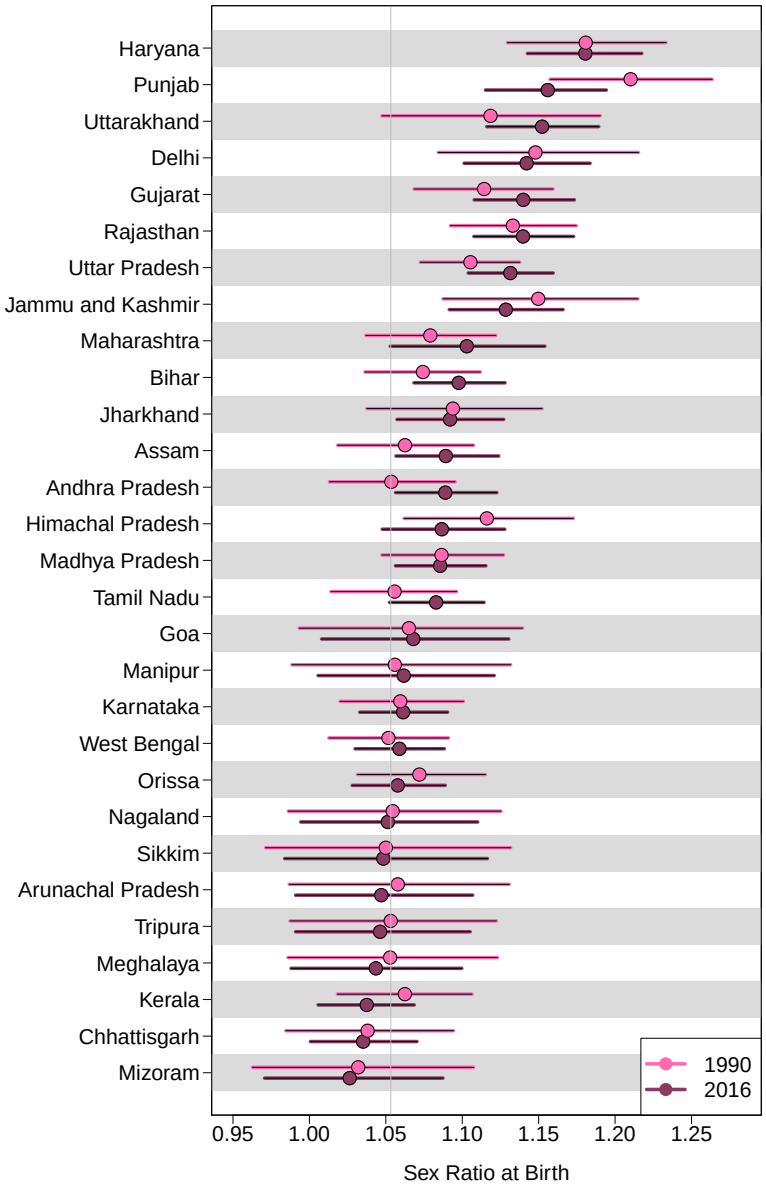

FIGURE 1. **SRB point estimates and uncertainty for all Indian states/union territories in 1990 and 2016.** States/union territories are ordered by decreasing point estimates of SRB for the year 2016. Dots are point estimates. Horizontal lines are 95% uncertainty intervals. The vertical line indicates the SRB baseline level for the whole India at 1.053.

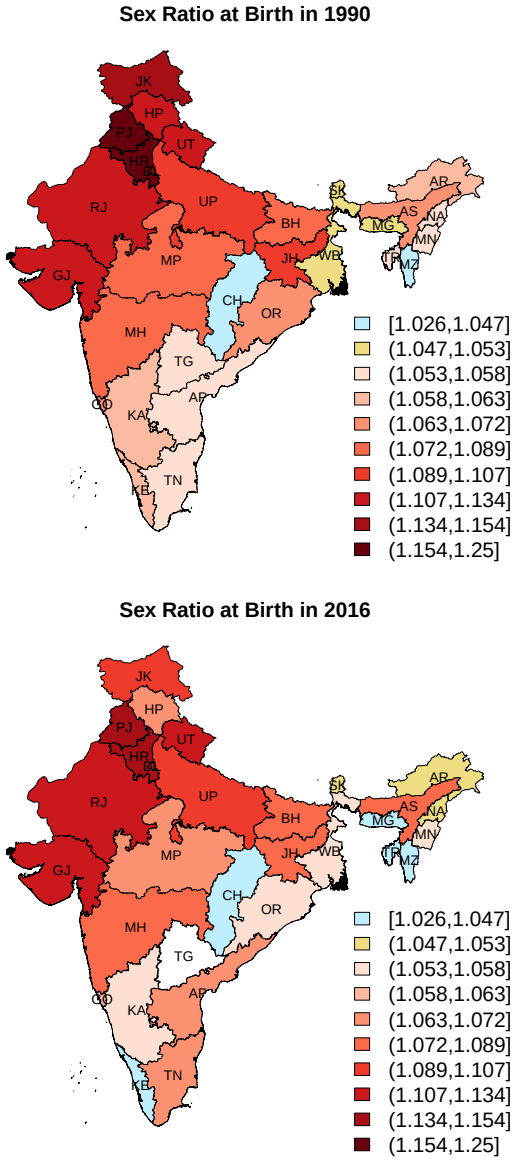

FIGURE 2. **SRB point estimates in 1990 and 2016 for 29 states and union territories in India.** Top: SRB in 1990. Bottom: SRB in 2016. State and union territory names are: Andhra Pradesh (AP); Arunachal Pradesh (AR); Assam (AS); Bihar (BH); Chhattisgarh (CH); Delhi (DL); Goa (GO); Gujarat (GJ); Haryana (HR); Himachal Pradesh (HP); Jammu and Kashmir (JK); Jharkhand (JH); Karnataka (KA); Kerala (KE); Madhya Pradesh (MP); Maharashtra (MH); Manipur (MN); Meghalaya (MG); Mizoram (MZ); Nagaland (NA); Orissa (OR); Punjab (PJ); Rajasthan (RJ); Sikkim (SK); Tamil Nadu (TN); Telangana (TG); Tripura (TR); Uttar Pradesh (UP); Uttarakhand (UT); West Bengal (WB). In 1990, TG is estimated together with AP. In 2016, TG is not estimated.

6.2. **Imbalanced sex ratio at birth for Indian states/union territories.** In total, 18 Indian states and union territories are identified to have imbalanced SRB during 1990–2016. Table 2 lists the year in which these identified states and union territories have the maximum SRB during 1990–2016 and values of their maximum SRB. The years with maximum SRB in the 18 states and union territories range from 1999 in both Jharkhand and Madhya Pradesh to 2015 in four states and union territories: Andhra Pradesh, Assam, Tamil Nadu, and Uttarakhand. SRB in 13 states and union territories reaches their state-level maxima before 2010 and five states and union territories reaches their local maximum values after 2010. As Andhra Pradesh, Assam, Tamil Nadu, and Uttarakhand have their maximum SRB estimated in 2015, the most recent year with data, the SRB in these states and union territories are possible to become more imbalanced in the near future. Among the rest 14 states and union territories, the SRB in 11 of them are in the midst of converging back to national SRB baseline, while the SRB in Himachal Pradesh, Karnataka, Kerala, and Maharashtra have converged back to the national SRB baseline value 1.053 by 2016 (i.e. not statistically significantly different from 1.053).

Among the 18 states and union territories, the state-level SRB maxima range from 1.075 [1.056; 1.095] in Karnataka to 1.250 [1.218; 1.282] in Punjab. Haryana is the other state with maximum SRB point estimate above 1.200 besides Punjab, and is estimated to be 1.226 [1.199; 1.255]. Except for Punjab and Haryana, the state-level SRB maxima are significantly above 1.100 in eight states and union territories: Bihar at 1.134 [1.116; 1.153], Delhi at 1.160 [1.125; 1.197], Gujarat at 1.161 [1.136; 1.186], Himachal Pradesh at 1.137 [1.104; 1.172], Jammu and Kashmir at 1.171 [1.137; 1.207], Rajasthan at 1.161 [1.141; 1.182], Uttat Pradesh at 1.152 [1.136; 1.169] and Uttarakhand at 1.154 [1.123; 1.186].

6.3. **Missing female births for Indian states/union territories.** Table 3 summarizes the results of the annual number of missing female births (AMFB) and cumulative number of missing female births (CMFB) across the 18 Indian states and union territories identified with imbalanced SRB during 1990–2016. For the whole India by summing up numbers from the 18 states and union territories, the total CMFB during 1990–2016 is 14.9 [13.2; 16.5] million. The average AMFB during 1990–2000 for all India is 461 [378; 544] thousand and increased to 612 [551; 672] thousand during 2001–2016.

Uttar Pradesh has the largest contribution to the national CMFB during 1990–2016, taking up to 32.8% [29.5%; 36.3%]. When looking into different time periods, the CMFB from Uttar Pradesh increased its share of total CMFB from 30.9% [23.6%; 38.1%] during 1990–2000 to 33.8% [30.9%; 37.0%] during 2001–2016. The increased contribution to the total CMFB from Uttar Pradesh is mainly due to the increased AMFB overtime. The average AMFB in Uttar Pradesh is 142 [101; 184] thousand during 1990–2000, and increased to an average of 207 [185; 229] thousand during 2001–2016. Consequently, the CMFB in Uttar Pradesh is estimated to be 4.9 [4.3; 5.5] million from 1990 to 2016.

Rajasthan and Bihar are another two states and union territories with CMFB during 1990–2016 significantly above one million, estimated at 1.8 [1.5; 2.0] million and 1.6 [1.2; 1.9] million respectively. Their contributions of the national CMFB during 1990–2016 are 11.8% [10.2%; 13.6%] for Rajasthan and 10.6% [8.4%; 12.6%] for Bihar. Both Rajasthan and Bihar have increases in the point estimates of their average AMFB from period 1990–2000 to period 2001–2016: 56 [39; 73] thousand to

State/ Union Territory	SRB 1990	SRB 2016	SRB change 1990–2016	Maximum SRB Year	Maximum SRB Value
Andhra Pradesh§	1.054 [1.013; 1.095]	1.089 [1.056; 1.123]	0.035 [-0.018; 0.087]	2015	1.089 [1.061; 1.117]
Arunachal Pradesh	1.058 [0.987; 1.131]	1.047 [0.991; 1.107]	-0.010 [-0.087; 0.063]		
Assam§	1.063 [1.018; 1.107]	1.089 [1.057; 1.124]	0.027 [-0.029; 0.083]	2015	1.090 [1.061; 1.121]
Bihar§	1.074 [1.036; 1.112]	1.098 [1.068; 1.128]	0.023 [-0.024; 0.072]	2002	1.134 [1.116; 1.153]
Chhattisgarh	1.038 [0.985; 1.094]	1.035 [1.001; 1.070]	-0.003 [-0.067; 0.060]		
Delhi¶§	1.148 [1.084; 1.215]	1.142 [1.101; 1.184]	-0.006 [-0.082; 0.067]	2003	1.160 [1.125; 1.197]
Goa	1.065 [0.993; 1.139]	1.068 [1.008; 1.130]	0.003 [-0.075; 0.077]		
Gujarat¶§	1.114 [1.068; 1.159]	1.140 [1.108; 1.173]	0.026 [-0.029; 0.083]	2000	1.161 [1.136; 1.186]
Haryana¶§	1.181 [1.130; 1.233]	1.181 [1.143; 1.217]	-0.000 [-0.064; 0.063]	2000	1.226 [1.199; 1.255]
Himachal Pradesh¶	1.116 [1.062; 1.173]	1.087 [1.048; 1.128]	-0.029 [-0.097; 0.037]	2001	1.137 [1.104; 1.172]
Jammu and Kashmir¶§	1.150 [1.087; 1.215]	1.128 [1.091; 1.166]	-0.021 [-0.096; 0.049]	2003	1.171 [1.137; 1.207]
Jharkhand§	1.094 [1.037; 1.152]	1.092 [1.057; 1.127]	-0.002 [-0.070; 0.063]	1999	1.101 [1.061; 1.142]
Karnataka	1.059 [1.020; 1.101]	1.061 [1.033; 1.091]	0.002 [-0.049; 0.052]	2004	1.075 [1.056; 1.095]
Kerala	1.062 [1.018; 1.106]	1.038 [1.006; 1.069]	-0.025 [-0.080; 0.029]	2002	1.085 [1.063; 1.109]
Madhya Pradesh§	1.086 [1.047; 1.127]	1.086 [1.056; 1.115]	-0.001 [-0.051; 0.047]	1999	1.092 [1.071; 1.115]
Maharashtra	1.079 [1.037; 1.122]	1.103 [1.053; 1.154]	0.024 [-0.040; 0.087]	2011	1.111 [1.073; 1.150]
Manipur	1.056 [0.989; 1.132]	1.062 [1.006; 1.121]	0.005 [-0.070; 0.080]		
Meghalaya	1.053 [0.986; 1.123]	1.043 [0.988; 1.100]	-0.009 [-0.084; 0.063]		
Mizoram	1.032 [0.963; 1.107]	1.026 [0.971; 1.087]	-0.005 [-0.080; 0.067]		
Nagaland	1.054 [0.986; 1.125]	1.051 [0.994; 1.110]	-0.003 [-0.080; 0.070]		
Orissa	1.072 [1.031; 1.115]	1.058 [1.028; 1.089]	-0.014 [-0.066; 0.037]		
Punjab¶§	1.210 [1.157; 1.264]	1.156 [1.115; 1.194]	-0.054 [-0.119; 0.008]	2000	1.250 [1.218; 1.282]
Rajasthan¶§	1.133 [1.092; 1.175]	1.140 [1.108; 1.173]	0.007 [-0.044; 0.058]	2004	1.161 [1.141; 1.182]
Sikkim	1.050 [0.971; 1.132]	1.048 [0.984; 1.117]	-0.001 [-0.079; 0.072]		
Tamil Nadu	1.056 [1.014; 1.096]	1.083 [1.052; 1.114]	0.027 [-0.024; 0.080]	2015	1.083 [1.058; 1.109]
Tripura	1.053 [0.987; 1.122]	1.046 [0.991; 1.105]	-0.007 [-0.081; 0.068]		
Uttar Pradesh¶§	1.105 [1.073; 1.138]	1.131 [1.104; 1.160]	0.026 [-0.016; 0.069]	2002	1.152 [1.136; 1.169]
Uttarakhand§	1.118 [1.047; 1.190]	1.152 [1.116; 1.189]	0.034 [-0.043; 0.109]	2015	1.154 [1.123; 1.186]
West Bengal	1.052 [1.013; 1.091]	1.059 [1.030; 1.088]	0.007 [-0.042; 0.056]		

TABLE 2. **SRB results by Indian state/union territory.** Point estimates and 95% uncertainty intervals (where apply) for (i) SRB in 1990 and 2016; (ii) change of SRB between 1990 and 2016; (iii) the year in which the maximum SRB is (only for states/union territories identified with imbalanced SRB during 1990–2016); and (iv) the maximum SRB during observation period (only for states/union territories identified with imbalanced SRB during 1990–2016). Numbers in brackets are 95% uncertainty intervals. ¶: SRB in 1990 is significantly different from national SRB baseline value 1.053. §: SRB in 2016 is significantly different from national SRB baseline value 1.053. States/union territories are in alphabetic order.

India & State/ Union Territory	Average AMFB (,000) 1990–2000	Average AMFB (,000) 2001–2016	CMFB (,000) 1990–2016	Proportion of National CMFB (%) 1990–2000	Proportion of National CMFB (%) 2001–2016	Proportion of National CMFB (%) 1990–2016
India	461 [378; 544]	612 [551; 672]	14,861 [13,239; 16,465]	100	100	100
Andhra Pradesh	2 [-16; 22]	21 [12; 29]	359 [85; 632]	0.6 [0.0; 4.5]	3.4 [2.0; 4.7]	2.4 [0.6; 4.1]
Assam	2 [-7; 11]	8 [3; 13]	146 [14; 279]	0.4 [0.0; 2.2]	1.3 [0.5; 2.0]	1.0 [0.1; 1.9]
Bihar	45 [21; 68]	67 [54; 80]	1,567 [1,202; 1,918]	9.8 [4.8; 14.3]	10.9 [9.0; 12.9]	10.6 [8.4; 12.6]
Delhi	10 [5; 15]	12 [9; 14]	295 [213; 379]	2.1 [1.1; 3.3]	1.9 [1.5; 2.4]	2.0 [1.4; 2.6]
Gujarat	41 [26; 56]	46 [39; 53]	1,187 [974; 1,402]	9.0 [5.8; 12.3]	7.5 [6.4; 8.7]	8.0 [6.6; 9.4]
Haryana	32 [25; 38]	33 [30; 37]	885 [778; 990]	6.9 [5.3; 9.0]	5.4 [4.8; 6.2]	6.0 [5.2; 6.8]
Himachal Pradesh	4 [2; 6]	3 [2; 4]	88 [52; 125]	0.9 [0.4; 1.4]	0.4 [0.3; 0.6]	0.6 [0.3; 0.8]
Jammu and Kashmir	8 [4; 11]	9 [7; 11]	234 [176; 292]	1.7 [0.9; 2.6]	1.5 [1.2; 1.8]	1.6 [1.2; 2.0]
Jharkhand	12 [0; 24]	14 [7; 20]	352 [156; 550]	2.6 [0.0; 5.2]	2.2 [1.2; 3.3]	2.4 [1.1; 3.6]
Karnataka	5 [-8; 18]	6 [0; 12]	145 [-41; 334]	1.1 [0.0; 3.9]	0.9 [0.0; 1.9]	1.0 [0.0; 2.2]
Kerala	5 [-3; 12]	0 [-3; 3]	52 [-56; 156]	1.0 [0.0; 2.6]	0.0 [0.0; 0.5]	0.3 [0.0; 1.0]
Madhya Pradesh	27 [8; 46]	28 [18; 38]	742 [470; 1,031]	5.9 [1.9; 9.8]	4.6 [3.1; 6.0]	5.0 [3.3; 6.7]
Maharashtra	26 [-2; 54]	43 [15; 71]	975 [318; 1,613]	5.6 [0.0; 11.2]	7.0 [2.6; 11.1]	6.5 [2.3; 10.4]
Punjab	35 [29; 42]	29 [25; 32]	846 [746; 949]	7.7 [6.0; 9.8]	4.7 [4.1; 5.3]	5.7 [5.0; 6.6]
Rajasthan	56 [39; 73]	71 [62; 80]	1,755 [1,495; 2,004]	12.1 [8.5; 16.1]	11.6 [10.2; 13.2]	11.8 [10.2; 13.6]
Tamil Nadu	4 [-10; 18]	9 [3; 15]	191 [-13; 382]	0.9 [0.0; 3.7]	1.5 [0.5; 2.5]	1.3 [0.0; 2.5]
Uttar Pradesh	142 [101; 184]	207 [185; 229]	4,873 [4,262; 5,498]	30.9 [23.6; 38.1]	33.8 [30.9; 37.0]	32.8 [29.5; 36.3]
Uttarakhand	5 [1; 8]	7 [5; 9]	160 [86; 234]	1.0 [0.1; 1.8]	1.1 [0.8; 1.5]	1.1 [0.6; 1.6]

TABLE 3. **Results for number of missing female births, for 18 Indian state/union territories with imbalanced SRB.** Point estimates and 95% uncertainty intervals for (i) the average annual number of missing female births (AMFB) in thousands for periods 1990–2000 and 2001–2016; (ii) the cumulative number of missing female births (CMFB) in thousands for period 1990–2016; (iii) the proportion of state-level CMFB to the national CMFB for periods 1990–2000, 2001–2016 and 1990–2016. Numbers in brackets are 95% uncertainty intervals. Proportions may not sum up to 100% due to rounding. States/union territories are ordered alphabetically.

71 [62; 80] thousand for Rajasthan, and 45 [21; 68] thousand to 67 [54; 80] thousand for Bihar.

For the rest 15 states and union territories besides Uttar Pradesh, Rajasthan, and Bihar, 12 have increases in the point estimates of their average AMFB from period 1990–2000 to period 2001–2016: Andhra Pradesh, Assam, Delhi, Gujarat, Haryana, Jammu and Kashmir, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Tamil Nadu, and Uttarakhand. Specifically, the average AMFB during period 2001–2016 is more than ten times of that during period 1990–2000 in Andhra Pradesh, four times in Assam, and more than twice in Tamil Nadu.

6.4. **Validation results.** The validation results indicate good calibrations of the model. In the out-of-sample validation, observations obtained from the year 2015

onward are left out. There are 310 left-out observations for the out-of-sample validation, consisting 33.1% of the total observations.

Table 4 summarizes the results related to the left-out observations for the validation exercise. Median errors and median absolute errors are close to zero for left-out observations. The coverage of 95% and 80% prediction intervals are symmetrical. The coverages are higher than expected for both the 95% and 80% prediction intervals and hence the model prediction is conservative.

Table 5 shows the comparison results between estimates obtained based on the full dataset and estimates based on the training set. Median errors and the median absolute errors are close to zero. The proportions of updated estimates that fall below the uncertainty intervals constructed based on the training set are within the expected values for both 95% and 80% uncertainty intervals.

Median error	0.005
Median absolute error	0.035
Below 95% prediction interval (%)	0.4
Above 95% prediction interval (%)	0.3
Expected (%)	**2.5**
Below 80% prediction interval (%)	3.0
Above 80% prediction interval (%)	2.8
Expected (%)	**10**

TABLE 4. **Validation results for left-out observations.** Error is defined as the difference between a left-out observation and the posterior median of its predictive distribution.

	1995	**2005**	**2015**
Median error	0.003	-0.001	0.003
Median absolute error	0.003	0.002	0.003
Below 95% uncertainty interval (%)	0	0	0
Above 95% uncertainty interval (%)	0	0	0
Expected proportions (%)	**$\leq$2.5**	**$\leq$2.5**	**$\leq$2.5**
Below 80% uncertainty interval (%)	0	0	0
Above 80% uncertainty interval (%)	0	0	0
Expected proportions (%)	**$\leq$10**	**$\leq$10**	**$\leq$10**

TABLE 5. **Differences in SRB estimates in selected observation years based on training set and full dataset.** Error is defined as the differences between an estimate based on full dataset and training set. The proportions refer to the proportions (%) of states/union territories in which the median SRB estimates based on the full dataset fall below or above their respective 95% and 80% uncertainty intervals based on the training set.

7. **Discussion.** To our best knowledge, this is the first study providing annual estimates and uncertainties of state-level SRB in India from 1990 to 2016 for 29 states and union territories. Although the imbalanced SRB and missing female births in India were well analysed in many studies before on the national level [12, 23, 25, 49], it is the first Bayesian modeling study to assess the SRB imbalance

on India state level and the resulting missing female births based on all available data. Our study shows that there exist great disparities in SRB across Indian states and union territories and over time. Among the 29 states and union territories included in this study, we identify 18 of them to have imbalanced SRB with various levels of state-level maximum SRB and the year in which local maxima occurred. It is crucial to conduct in-depth analyses on state level to address these differences.

Our study shows that it is necessary and important to provide subnational estimates for prime demographic indicators for countries with great population heterogeneity such as India. Our estimate for the number of missing female birth during 1990–2016 for the whole India is slightly lower than (but not statistically significantly different from) the result from the most recent cross-country systematic assessment of SRB [12], which is 14.9 [13.2; 16.5] million in our study versus 16.6 [12.1; 21.0] million in [12]. The difference between the two national estimates is a direct result as the different data sources used for state-level and national-level studies. For the case of India, there are more national-level observations with longer time series than subnational observations per state and union territory. In addition and importantly, our estimate for the whole India is based on aggregation of state-level estimates and hence the heterogeneity across the Indian states and union territories are accounted in the national estimates. As discussed in a previous study [46], the national aggregates would be different based on different subgroups for India.

The main limitation of this study is the lack of state-level SRB data in the early period of sex ratio transition for the SRB imbalance. Only 4.2% (39 out of 937) of the state-level SRB observations are with reference years before 1980 and none is before 1970. Given that the sex ratio transition has begun in part of India during the 1970s, it is challenging to model the beginning of the sex ratio transition on state level in India given the limited data before 1980. In addition, ideally we would model the SRB baseline values differ across Indian states and union territories. Since no state-level data are available before 1970, we use the national SRB baseline level for India from [12] instead. Future research should focus on collecting and generating state-level and other subnational-level SRB data. The importance of subnational-level data with high quality will increase in order to better estimate and track the sex ratio transition in India on the subnational level and provide in-time data-driven evidence for policy planning.

Our estimation model for state-level SRB in India can be easily extended to any other subnational dimensions of India for estimating SRB and its imbalance as long as the data for corresponding subpopulation are available. Other subgroup dimensions would be defined by urban or rural residence, and/or by mothers' education levels. Our analysis can also be applied to other countries with emerging sex ratio transition (for example Nepal [17]) to offer reproducible and timely estimates of SRB and the severity of its imbalance.

Appendices. For simplicity, all notations in the Appendix A and B refer to observations or births from a certain Indian state/union territory, a certain data series and reference year (i.e. year of birth).

Appendix A. **Sampling errors for DHS data.** DHS provides individual-level data with the full birth history for each women at reproductive age interviewed during the survey fieldwork period. We calculate the sampling error for log-transformed

SRB for DHS data series using the jackknife method [16]. Let U denote the total number of clusters. The u-th partial prediction of SRB is given by:

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

where n indexes the live births in each state-survey-year, N is the total number of live births. x_n is the sex for the n-th live birth. d_n is the cluster number for the n-th live birth. w_n is the sampling weight for the n-th live birth. The u-th pseudo-value estimate of the SRB on log-scale is:

$$\log(r)_u^* = U \cdot \log(r') - (U-1) \cdot \log(r_{-u}), \text{ where}$$

$$r' = \frac{\sum_{n=1}^{N} \mathbb{I}_n(x_n = \text{male}) \cdot w_n}{\sum_{n=1}^{N} \mathbb{I}_n(x_n = \text{female}) \cdot w_n}.$$

The sampling variance is:

$$\sigma^2 = \frac{\sum_{u=1}^{U} (\log(r)_u^* - \overline{\log(r)_u^*})^2}{U(U-1)},$$

where $\overline{\log(r)_u^*} = \frac{1}{U} \sum_{u=1}^{U} \log(r)_u^*$.

Appendix B. **Stochastic errors for SRS data.** SRS does not provide information on individual-level full birth histories. It provides the total number of births registered in a certain year with other basic birth information like sex and place of births. For observations from SRS, a Monte Carlo simulation is used to approximate the stochastic variance. For a state-year, the g-th simulated number of male live births $B_m^{(g)}$ is obtained as follows:

$$B_m^{(g)} \overset{\text{i.i.d.}}{\sim} \mathcal{B}(B, p_m), \text{ for } g = 1, \ldots, G,$$

where $\mathcal{B}(n, p)$ denotes a binomial distribution with n independent trails and probability of success p. G is the total number of simulations, B is the total number of live births as observed in SRS data, and p_m is the observed proportion of male live births. The corresponding g-th simulation for SRB is given by:

$$r^{(g)} = \frac{B_m^{(g)}}{B - B_m^{(g)}}, \text{ for } g = 1, \ldots, G.$$

The stochastic variance for SRB on log-scale is:

$$\sigma^2 = \frac{\sum_{g=1}^{G} \left(\log(r^{(g)}) - \overline{\log(r)}\right)^2}{G-1},$$

where $\overline{\log(r)} = \frac{1}{G} \sum_{g=1}^{G} \log(r^{(g)})$.

Appendix C. **Merge observation period.** For DHS and SRS data, the annual log-transformed SRB observations are merged such that the sampling or stochastic error is below 0.05.

For a certain data series (either one of the DHS or the SRS series), let $\{t_n, t_{n-1}, \cdots, t_1\}$ be the years with recorded births from recent to past. The merge starts from the most recent year t_n:

Merging process for DHS and SRS 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 Appendix A for DHS data and Appendix B for SRS data
4: **if** $\sigma < 0.05$ **or** $t_n - t_{n-1} > 1$ **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}

Appendix D. Model overview.

D.1. **Notations.** Table 6 summarizes the notations and indexes used in Section 3.

Symbol	Description
i	Indicator for observation, $i = 1, \ldots, n$, where $n = 937$.
t	Indicator for year, $t = 0, \ldots, h$. $t = 0$ refers to year 1990 and $t = h$ refers to year 2016.
s	Indicator for Indian state/union territory, $s = 1, \ldots, k$, where $k = 29$.
y	Indicator for data source type, $y = 1, \cdots, z$, where $z = 2$.
$\mathcal{I}_1$	$\mathcal{I}_1 = \{i = 1, \cdots, n \vert y[i] = \text{SRS}\}$ denotes the set of indexes for SRS observations.
$\mathcal{I}_2$	$\mathcal{I}_2 = \{i = 1, \cdots, n \vert y[i] \neq \text{SRS}\}$ refers to the set of indexes for non-SRS observations.
v_i	The i-th SRB observation on log-scale.
δ_i	The i-th error term for v_i.
σ_i^2	The i-th stochastic variance (if data is SRS) or sampling variance (if data is non-SRS) for v_i.
ω_y^2	The non-sampling variance parameters with non-SRS data source type for $y = 1, \cdots, z$.
$V_{s,t}$	The model fitting for the true SRB for state/union territory s in year t on log-scale.
$P_{s,t}$	The difference between $V_{s,t}$ and $a0$ for state/union territory s in year t.
$a0$	The baseline level parameter of SRB for the whole India on the log-scale.
ρ	Autoregressive parameter in AR(1) time series model for $P_{s,t}$.
σ_ϵ^2	Variance of distortion parameter in AR(1) time series model for $P_{s,t}$.
b_s	The state-specific level parameters for $s = 1, \cdots, k$ in AR(1) time series model for $P_{s,t}$.
σ_b^2	The variance parameter for b_s.

TABLE 6. **Notation summary.**

D.2. **Data model.**

$$\begin{aligned} v_i &= V_{s[i],t[i]} + \delta_i, \text{ for } i = 1, \cdots, n, \\ \delta_i &\sim \mathcal{N}(0, \sigma_i^2), \text{ for } i \in \mathcal{I}_1, \\ \delta_i|\phi_i^2 &\sim \mathcal{N}(0, \phi_i^2), \text{ for } i \in \mathcal{I}_2, \\ \phi_i^2 &= \sigma_i^2 + \omega_{y[i]}^2, \text{ for } i \in \mathcal{I}_2. \end{aligned}$$

D.3. **SRB model.**

$$\begin{aligned} V_{s,t} &= a0 + P_{s,t}, \text{ for } s = 1, \cdots, k, \text{ for } t = 0, \cdots, h, \\ P_{s,0}|b_s, \rho, \sigma_\epsilon &\sim \mathcal{N}(b_s, \sigma_\epsilon^2/(1-\rho^2)), \text{ for } s = 1, \cdots, k, \\ P_{s,t} &= b_s + \rho \cdot (P_{s,t-1} - b_s) + \epsilon_{s,t}, \text{ for } s = 1, \cdots, k, \text{ for } t = 1, \cdots, h, \\ b_s|\sigma_b &\sim t(0, \sigma_b^2, \nu = 3), \text{ for } s = 1, \cdots, k, \\ \epsilon_{s,t}|\sigma_\epsilon &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \text{ for } s = 1, \cdots, k, \text{ for } t = 1, \cdots, h. \end{aligned}$$

D.4. **Prior distributions.** Mutually independent priors are assigned to hyper-parameters:

$$\begin{aligned} \omega_y &\overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 2) \text{ for } y = 1, \cdots, z, \\ a0 &\sim \mathcal{N}(\mu_{a0}, \sigma_{a0}^2), \\ \rho &\sim \mathcal{U}(0, 1), \\ \sigma_\epsilon &\sim \mathcal{U}(0, 0.01), \\ \sigma_b &\sim \mathcal{U}(0, 0.02). \end{aligned}$$

where $\mu_{a0} = 1.053$ and $\sigma_{a0} = 0.002$.

Acknowledgments. We would like to thank the Reviewers and the Editors for their insightful comments and helpful discussions.

REFERENCES

[1] L. Alkema, F. Chao, D. You, P. Jon and C. C. Sawyer, 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*, **2** (2014), e521–e530.

[2] L. Alkema and J. New, Global estimation of child mortality using a Bayesian B-spline bias-reduction model, *The Annals of Applied Statistics*, **8** (2014), 2122–2149.

[3] G. N. Allahbadia, The 50 million missing women, *Journal of Assisted Reproduction and Genetics*, **19** (2002), 411–416.

[4] G. Aravamudan, *Disappearing Daughters: The Tragedy of Female Foeticide*, Penguin Books, New Delhi, 2007.

[5] I. Attané and C. Z. Guilmoto, *Watering the Neighbour's Garden: The Growing Demographic Female Deficit in Asia*, Committee for International Cooperation in National Research in Demography, Paris, 2007.

[6] J. Banister, Shortage of girls in China today, *Journal of Population Research*, **21** (2004), 19–45.

[7] S. Basten and G. Verropoulou, "Maternity migration" and the increased sex ratio at birth in Hong Kong SAR, *Population Studies*, **67** (2013), 323–334.

[8] J. Bongaarts, The implementation of preferences for male offspring, *Population and Development Review*, **39** (2013), 185–208.

[9] J. Bongaarts and C. Z. Guilmoto, How many more missing women? Excess female mortality and prenatal sex selection, 1970–2050, *Population and Development Review*, **41** (2015), 241–269.

[10] Y. Cai and W. Lavely, China's missing girls: Numerical estimates and effects on population growth, *China Review*, **3** (2003), 13–29.
[11] A. Chahnazarian, Determinants of the sex ratio at birth: Review of recent literature, *Biodemography and Social Biology*, **35** (1988), 214–235.
[12] F. Chao, P. Gerland, A. R. Cook and L. Alkema, 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*, **116** (2019), 9303–9311.
[13] F. Chao, D. You, P. Jon, L. Hug and L. Alkema, National and regional under-5 mortality rate by economic status for low-income and middle-income countries: a systematic assessment, *The Lancet Global Health*, **6** (2018), e535–e547.
[14] M. Das Gupta, Z. Jiang, B. Li, Z. Xie, W. Chung and H. Bae, Why is son preference so persistent in East and South Asia? A cross-country study of China, India and the Republic of Korea, *The Journal of Development Studies*, **40** (2003), 153–187.
[15] G. Duthé, F. Meslé and J. Vallin, High sex ratios at birth in the Caucasus: Modern technology to satisfy old desires, *Population and Development Review*, **38** (2012), 487–501.
[16] B. Efron, *The Jackknife, the Bootstrap, and other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982.
[17] M. D. Frost, M. Puri and P. R. A. Hinde, Falling sex ratios and emerging evidence of sex-selective abortion in Nepal: evidence from nationally representative survey data, *BMJ open*, **3** (2013), e002612.
[18] A. Gelman and D. B. Rubin, Inference from iterative simulation using multiple sequences, *Statistical Science*, **7** (1992), 457–472.
[19] S. M. George, Millions of missing girls: From fetal sexing to high technology sex selection in India, *Prenatal Diagnosis*, **26** (2006), 604–609.
[20] S. M. George, Sex selection/determination in India: Contemporary developments, *Reproductive Health Matters*, **10** (2002), 190–192.
[21] D. Goodkind, *Sex-selective Abortion, Reproductive Rights, and the Greater Locus of Gender Discrimination in Family Formation: Cairo's Unresolved Questions*, University of Michigan, Population Studies Center, 1997.
[22] D. Goodkind, Child underreporting, fertility, and sex ratio imbalance in China, *Demography*, **48** (2011), 291–316.
[23] C. Z. Guilmoto, *Sex Imbalances at Birth: Current Trends, Consequences and Policy Implications*, UNFPA Asia and Pacific Regional Office, Bangkok, Thailand, 2012.
[24] C. Z. Guilmoto, The sex ratio transition in Asia, *Population and Development Review*, **35** (2009), 519–549.
[25] C. Z. Guilmoto, Skewed sex ratios at birth and future marriage squeeze in China and India, 2005–2100, *Demography*, **49** (2012), 77–100.
[26] C. Z. Guilmoto, Son preference, sex selection, and kinship in Vietnam, *Population and Development Review*, **38** (2012), 31–54.
[27] C. Z. Guilmoto, X. Hoàng and T. N. Van, Recent increase in sex ratio at birth in Viet Nam, *PLoS One*, **4** (2009), e4624.
[28] C. Z. Guilmoto and Q. Ren, Socio-economic differentials in Birth Masculinity in China, *Development and Change*, **42** (2011), 1269–1296.
[29] T. Hesketh and J. Min, The effects of artificial gender imbalance: Science & Society Series on Sex and Science, *EMBO reports*, **13** (2012), 487–492.
[30] V. M. Hudson and A. M. Den Boer, *Bare Branches: The Security Implications of Asia's Surplus Male Population*, MIT Press, Cambridge, Mass., 2004.
[31] P. Jha, M. A. Kesler, R. Kumar, F. Ram, U. Ram, L. Aleksandrowicz, D. G. Bassani, S. Chandra and J. K. Banthia, Trends in selective abortions of girls in India: analysis of nationally representative birth histories from 1990 to 2005 and census data from 1991 to 2011, *The Lancet*, **377** (2011), 1921–1928.
[32] P. Jha, R. Kumar, P. Vasa, N. Dhingra, D. Thiruchelvam and R. Moineddin, Low male-to-female sex ratio of children born in India: national survey of 1·1 million households, *The Lancet*, **367** (2006), 211–218.
[33] R. Kaur, S. S. Bhalla, M. K. Agarwal and P. Ramakrishnan, *Sex Ratio at Birth – The Role of Gender, Class and Education*, United Nations Population Fund, New Delhi, 2017.
[34] Lancet India Correspondent, Misuse of amniocentesis, *The Lancet*, **321** (1983), 812–813.

[35] S. Li, Imbalanced sex ratio at birth and comprehensive intervention in China, in *4th Asia Pacific Conference on Reproductive and Sexual Health and Rights* (Hyderabad, India, Oct 29-31, 2007), United Nations Population Fund, 2007.
[36] T. Lin, The decline of son preference and rise of gender indifference in Taiwan since 1990, *Demographic Research*, **20** (2009), 377–402.
[37] K. Madan and M. H. Breuning, Impact of prenatal technologies on the sex ratio in India: An overview, *Genetics in Medicine*, **16** (2014), 425–432.
[38] F. Meslé, J. Vallin and I. Badurashvili, A sharp increase in sex ratio at birth in the Caucasus. Why? How?, in *Watering the neighbour's garden: The growing demographic female deficit in Asia* (eds. I. Attané and C. Z. Guilmoto), Committee for International Cooperation in National Research in Demography, (2007), 73–88.
[39] N. Oomman and B. R. Ganatra, Sex selection: The systematic elimination of girls, *Reproductive Health Matters*, **10** (2002), 184–188.
[40] C. B. Park and N. H. Cho, Consequences of son preference in a low-fertility society: Imbalance of the sex ratio at birth in Korea, *Population and Development Review*, **21** (1995), 59–84.
[41] M. Plummer, *Rjags: Bayesian Graphical Models Using MCMC*, 2011. Available from: http://CRAN.R-project.org/package=rjags.
[42] M. Plummer, JAGS: A program for analysis of bayesian graphical models using gibbs sampling, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria, (2003), Available from: http://mcmc-jags.sourceforge.net/.
[43] M. Plummer, N. Best, K. Cowles and K. Vines, CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News*, **6** (2006), 7–11. Available from: https://cran.r-project.org/package=coda.
[44] R. Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. Available from: https://www.R-project.org/.
[45] T. K. Roy and A. Chattopadhyay, Daughter discrimination and future sex ratio at birth in India, *Asian Population Studies*, **8** (2012), 281–299.
[46] K. C. Samir, M. Wurzer, M. Speringer and W. Lutz, Future population and human capital in heterogeneous India, *Proceedings of the National Academy of Sciences*, **115** (2018), 8328–8333.
[47] A. Sen, Missing women, *British Medical Journal*, **304** (1992), 587–588.
[48] B. R. Sharma, N. Gupta and N. Relhan, Misuse of prenatal diagnostic technology for sex-selected abortions and its consequences in India, *Public Health*, **121** (2007), 854–860.
[49] O. P. Sharma and C. Haub, Sex ratio at birth begins to improve in India, *Population Reference Bureau*, 2008. Available from: http://www.prb.org/Publications/Articles/2008/indiasexratio.aspx.
[50] Y. Su and M. Yajima, *R2jags: A Package for Running Jags from R*, 2015. Available from: http://CRAN.R-project.org/package=R2jags.
[51] S. L. Tandon and R. Sharma, Female foeticide and infanticide in India: an analysis of crimes against girl children, *International Journal of Criminal Justice Sciences*, **1** (2006), 1–10.

E-mail address: chao.fengqing@gmail.com
E-mail address: ajitkumaryadav1989@gmail.com