Bayesian Population Reconstruction

Training Course on Bayesian Population Projections: Theory and Practice
IUSSP IPC 2017, Cape Town, South Africa

Mark Wheldon

United Nations, Population Division
and
Auckland University of Technology, New Zealand

Acknowledgements: Joint work with Adrian Raftery, Sam Clark and Patrick Gerland
NICHD, Grants R01 HD054511 and K01 HD057246
BayesPop Working Group, CSSS, and CSDE at the UW
FHES at AUT

Background

How many people were there? How many births? How many deaths?

- Accurate estimates of
  - size
  - fertility
  - mortality
  - and migration rates

  of the recent past are essential for the study of population. (Forecasts will not be considered).
- They are essential for policy planning and evaluation and many other applications, e.g.,
  - provide a basis for forecasts and projections
  - historical studies
  - many more ...

Background

- Such estimates can be found in the United Nations Population Division's biennial *World Population Prospects (WPP)*, a comprehensive set of demographic statistics for all countries.
- Contains point estimates of ...
  - vital rates (fertilty and mortality rates)
  - net migration
  - population counts
- ... over the period 1950 to the present.
- Particularly important for countries without well-resourced official statistics systems of their own.

... However

Current methodology

1. Achieves consistency among estimates of different parameters in a manual, iterative process.
2. Produces estimates that can be analyst specific.
3. Does not include quantification of uncertainty.

Why We Need Uncertainty

Quantification of uncertainty is important because sources and reliability of data vary greatly among countries.

E.g., vital rates

- *More developed countries* have vital registration systems which achieve high coverage and low bias and measurement error variance.
- Many *less developed countries* rely on surveys (e.g., DHS) for vital rate data.
  - Coverage maybe incomplete.
  - Bias and measurement error variance can be high, can vary greatly from one source to another.
  - Expert knowledge is valuable.

Fragmentary Data

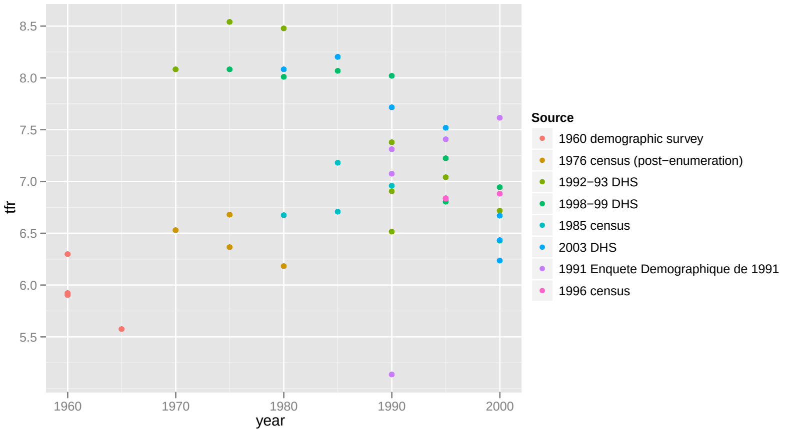


Figure: Estimates of Total Fertility Rates, Burkina Faso 1960–2000.

World news on NBCNEWS.com

Search nbcnews.com bing Search

A child is born and world population hits 7 billion

Doctor wonders 'whether there will be food, clean water, shelter, education and a decent life for every child'

Jump to video
India's growing population

Recommend 2.9k
Tweet 100
+1 16
Share 509

Below: Chart Video Discuss Related

msnbc.com staff and news serv
updated 10/31/2011 5:50:20 AM

Countries around the w
reaching 7 billion Mond
newborn infants symbo
that there may be too m
resources.

While demographers ar
population will reach th
Monday to symbolically

A string of festivities are
series of symbolic 7-bill

Fox News | Fox Business | uReport | Fox News Radio | Fox News Latino | Fox Nation | Fox News Insider | Login

FOX NEWS .com Fair & Balanced

Search

ON AIR NOW ›
7p et FOX Report w/ Shepard Smith
WATCH LIVE

On Air Personalities ›
8p et The O'Reilly Factor
It's the "No Spin Zone!"

Home | Video | Politics | U.S. | Opinion | Entertainment | Tech | Science | Health | Travel | Lifestyle | World | Sports | On Air

Science Home | Archaeology | Dinosaurs | Planet Earth | Wild Nature | Air & Space | Natural Science

7 Billionth Person Born (Or Maybe More. Or Less. Who Knows?)

Published October 31, 2011 / FoxNews.com

Oct. 31, 2011: Newborn girl Yazuri Tarmeno -- one of several people worldwide named the 7 billionth born on the planet -- cries after being born at the Maternity hospital in Lima, Peru. (AP PHOTO/KAREL NAVARRO)

FOLLOW FOX NEWS SCITECH
Follow @fxnscitech 22.5K followers
Like 44,088 people like this.

RECOMMENDED VIDEOS
Man attempts to break sound barrier in daring skydive
Daredevil survives plunge from 71,581 feet

TRENDING IN SCIENCE

Goal

An improved method for reconstructing populations of the recent past that

1. Quantifies uncertainty probabilistically.
2. Estimates all parameters consistently.
3. Is easily replicable.
4. Uses all reliable data and expert opinion.

Notation & Parameters

- For age groups $[a, a+5)$, $a = 0, \ldots, 75$, $[80, \infty)$.
- Time periods $[t, t+5)$, $t = t_0, \ldots, T-5$.
- t_0 is the "baseline" year.
- We want to "reconstruct" the population over the period t_0 to T.

Input Parameters

- $f_{a,t}$: fertility rate.
  - Number of babies born per person-years lived by women age a.
- $s_{a,t}$: survival proportion.
  - Proportion surviving five more years.
  - A measure of mortality.
- $g_{a,t}$: net international migration.
- $n_{a,t}$: age-specific population count.

Parameters

Summary Parameters

- TFR_t: total fertility rate
  - $\text{TFR}_t \propto \sum_a f_{a,t}$
- $e_{0,t}$: life expectancy at birth
  - The average age at death under current age-specific mortality rates.
  - Function of $s_{a,t}$:
    $e_{0,t} = 5 \sum_{a=0}^{A} \prod_{i=0}^{a} s_{i,t} + \left(\prod_{i=0}^{A} s_{i,t} \right) \left(s_{A+5}/(1 - s_{A+5}) \right)$

Demographic Balancing Equation

- The fundamental relationship linking fertility, mortality, migration and population through time is:

  $$\text{count}_{t+5} = \text{count}_t + \text{births}_{[t,t+5)} - \text{deaths}_{[t,t+5)} + \text{net migration}_{[t,t+5)}$$

- Deterministic rule for population projection forward in time.
- Age-specific version called cohort component method of population projection (CCMPP).
- This is a discrete-time approximation to a continuous process; standard adjustments are used to improve accuracy.

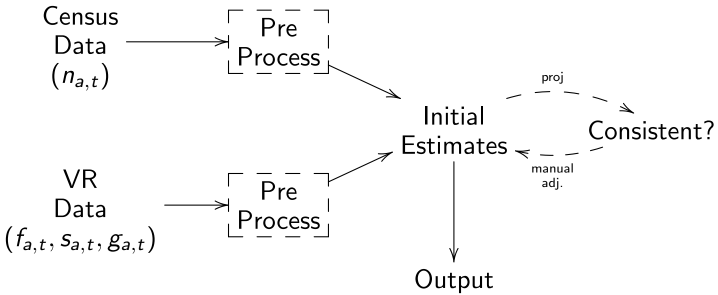
Comparing Counts

- Given
  - fertility, mortality and net migration over the period t_0 to T
  - population counts at time t_0,

  we can *project* to get population counts at $t_0 + 5, \ldots, T$.
- Census counts provide a second set of estimates in census years.
- These two counts can be compared. The difference can tell us something about data accuracy.

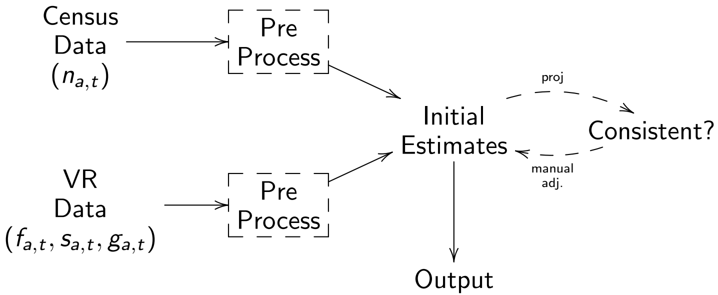
Population Reconstruction

Current UN Practice



- Gather all available data
- Pre-process data to get a single initial estimate for each age and time period.
- Compare projected with census counts; are they the same?
- Manually adjust if necessary.

Current UN Practice



- Pre-processing is needed to reduce parameter- and source-specific bias.
- Includes
  - Discarding some data completely.
  - Use of *indirect estimation techniques*.

Indirect Estimation

- Commonly used to correct for bias in surveys or provide estimates for ages and times when data are not available.
  - E.g., omission of births/deaths a long time ago, misplacement in time.
- Examples include
  - The P/F ratio method for estimating fertility.
  - Relational models for estimating mortality.
- Cause and magnitude of biases vary greatly among data sources, parameters, age groups and time.
- Bias-reduction via indirect estimation
  - Is source- and parameter-specific.
  - Requires detailed knowledge of data collection methods and possible sources of bias.
- No "one size fits all" approach: We do not attempt to replace this step.

Current UN Practice

Important Points

- Raw survey estimates are subject to biases, reduced by indirect estimation techniques.
- Pre-processing means that most of the available “data” points are already modeled.
- Consistency among the parameters is achieved in a manual, iterative fashion.
- No quantification of uncertainty.

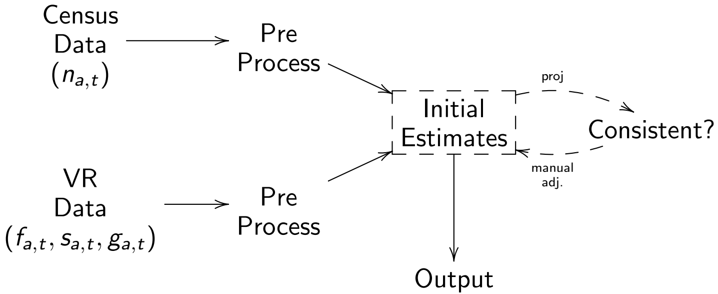
Bayesian Population Reconstruction

- Bayesian Population Reconstruction is a new method for reconstructing populations which is consistent with current UN practice.
- Pre-processed, initial estimates taken as bias-reduced input.
- Consistency among VR and census counts formalized by incorporating the CCMPP into a statistical model.
  - The parameters to be estimated include $\mathbf{n}_{t_0}$ and $\mathbf{f}_t$, $\mathbf{s}_t$, $\mathbf{g}_t$, $t = t_0, \ldots, T$.
  - The error variance is modeled statistically.

Females only (for now)

Current vs. Proposed

Current UN Practice



Proposed (Bayesian Population Reconstruction)

Hierarchical Model

(* = initial estimates and census counts which are fixed)

I. Likelihood

$$\log n^*_{a,t} \mid n_{a,t}, \sigma_n^2 \sim \text{Normal}\left(\log n_{a,t}, \sigma_n^2\right)$$

II. Projection Model

$$n_{a,t} \mid \mathbf{n}_{t-5}, \mathbf{f}_{t-5}, \mathbf{s}_{t-5}, \mathbf{g}_{t-5} = CCMPP(\mathbf{n}_{t-5}, \mathbf{f}_{t-5}, \mathbf{s}_{t-5}, \mathbf{g}_{t-5})$$

III. Priors on Inputs

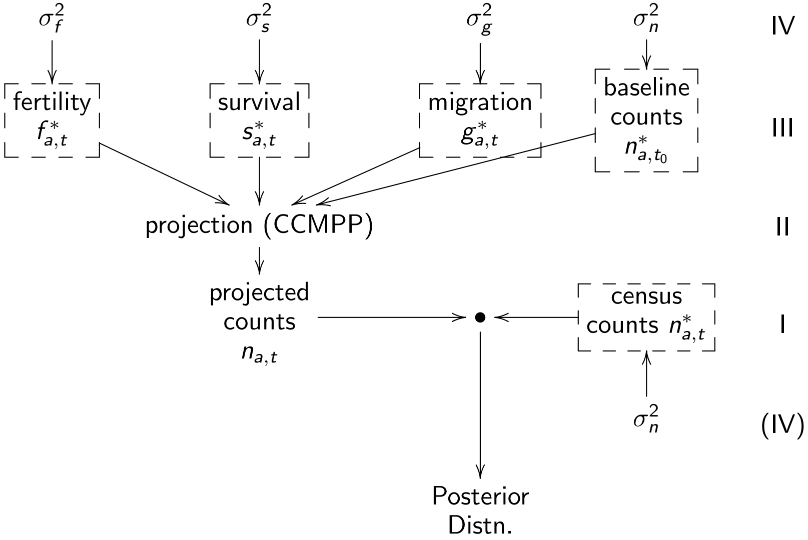
$$\log n_{a,t_0} \mid \sigma_n^2 \sim \text{Normal}\left(\log n^*_{a,t_0}, \sigma_n^2\right)$$
$$\log f_{a,t} \mid \sigma_f^2 \sim \text{Normal}\left(\log f^*_{a,t}, \sigma_f^2\right)$$
$$\text{logit}\, s_{a,t} \mid \sigma_s^2 \sim \text{Normal}\left(\text{logit}\, s^*_{a,t}, \sigma_s^2\right)$$
$$g_{a,t} \mid \sigma_g^2 \sim \text{Normal}\left(g^*_{a,t}, \sigma_g^2\right)$$

IV. Hyperparameters

$$\sigma_v^2 \sim \text{InvGamma}(\alpha_v, \beta_v), \quad v \in \{n, f, s, g\}$$

Hierarchical Model: Key Relationships

(inputs are boxed)



Measurement Error Variance

$$\sigma_v^2 \sim \mathrm{InvGamma}(\alpha_v, \beta_v), \;\; v = f, s, g, n$$

- The σ_v^2 are random. They reflect non-systematic error in the initial estimates.
- We specify α_v and β_v using expert opinion
  - possibly informed by empirical estimates of measurement error,
  - or elicited as mean absolute relative error (MARE) through statements like

    "With probability 90 percent, the initial estimates are accurate to within approximately $\pm p$ percent."

- This is a flexible approach: some data sources have information about their accuracy, but others do not.

Summary

Inputs

- Bias-reduced initial estimates of
  - Age-specific fertility and mortality rates.
  - Net international migration.
  - Population counts.
- Expert opinion about measurement error variance (based on data if available).

Outputs

- Joint posterior distribution of the inputs.
- Gives 95% credible intervals for input parameters and transformations.

Model Checking

Questions

- Are credible intervals calibrated? I.e., do the 95% intervals contain the truth 95% of the time?
- Are posterior medians closer to the truth than the initial estimates?

Method

- Simulate an hypothetical population.
- Generate initial estimates under the model.
- Run reconstruction many times.

Model Checking

Results

- Coverage of 95% Credible Intervals
  - Achieved coverages were close to 0.95, as desired.
- Improvement Over Initial Estimates
  - On average, the truth was closer to the posterior medians than the initial estimates.
- Further Model Checks
  - We also compared our model with a more flexible version where σ_n varied by age and time.
  - Separate variances seemed unnecessary.
  - Results were unchanged.

Developing and Developed Countries

Developing and Developed Countries

- The new method is most useful for countries with unreliable, fragmentary data where uncertainty is high.
- I also show that it works for countries with very good data by reconstructing the female population of New Zealand (1961–2006),
- and fragmentary data by reconstructing the female population of Laos (1985–2005).
- The reconstruction periods are determined by the available data.
- Population sizes are similar (counts in millions):

	1961	1985	2005
Laos	—	1.8	2.8
New Zealand	1.2	1.7	2.1

- Data quality and availability are very different.

Initial Estimates: Laos

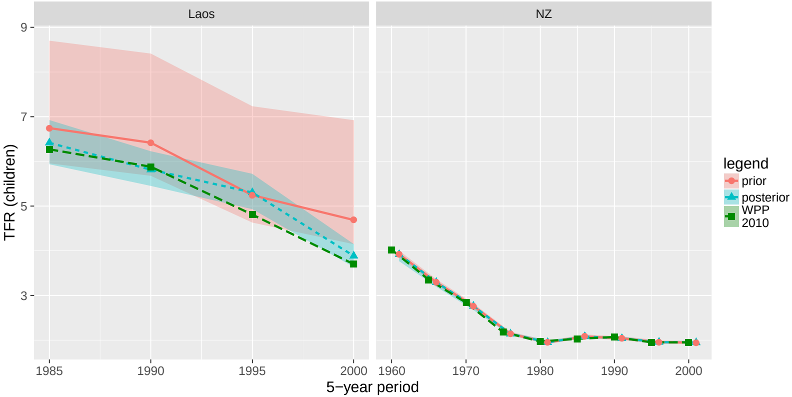
Parameter	Sources	Methods	Rel. Error. (%)
Pop count	Censuses 1985, 1995, 2005	Adjustments for undercount	10
Fertility	Birth history surveys	Indirect methods, smoothing	10
Survival	Birth history surveys	Indirect methods, model life tables	10
Migration	None	Centered at zero, large rel. err.	20

Initial Estimates: New Zealand

Parameter	Sources	Methods	Rel. Error. (%)
Pop count	Censuses, 5-yearly, 1961–2006	Post-enumeration survey	1
Fertility	Vital registration	Coverage study	1
Survival	Vital registration	Coverage study	1
Migration	Immigration Cards		5

Results

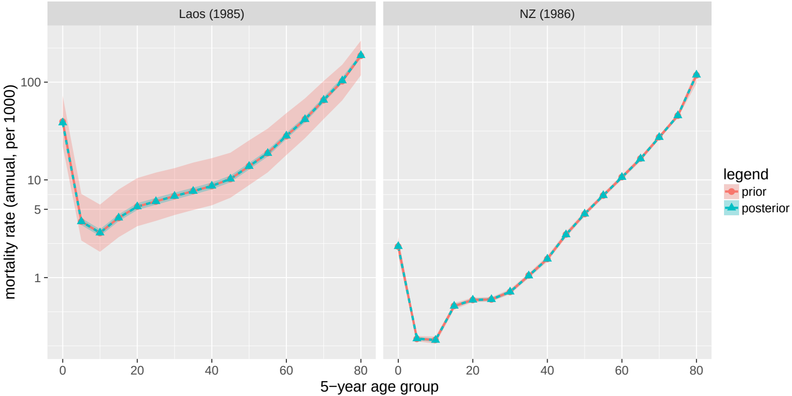
Total Fertility Rate



95% Posterior Interval Half-widths

	Mean Half-Width
Laos	0.4
NZ	0.04

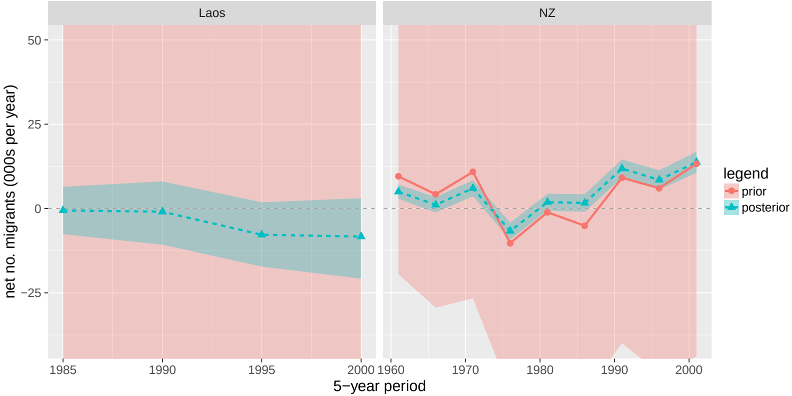
Age Specific Mortality Rate



95% Posterior Interval Half-widths

	Mean Half-Width
Laos	2
NZ	0.5

Total Net Migration



95% Posterior Interval Half-widths

	Counts (000s)
	Mean Half-Width
Laos	9.5
NZ	2.6

Developing and Developed Countries

Summary

- Posterior credible intervals were much more narrow for New Zealand than Laos, reflecting higher data quality.
- The method works for countries with very good data as well as those with fragmentary and unreliable data.

Two-Sex Reconstruction

Two-Sex Reconstruction

- Requires
  - two-sex CCMPP,
  - two-sex hierarchical model.
- As a consequence of including males, two sex reconstruction allows estimation of sex ratios such as
  - sex ratio at birth (SRB),
  - sex ratios of mortality.

Two-Sex CCMPP

- We use *female dominant* projection where births depend on female fertility.
- For ages 5+, projection is done separately for each sex with sex-specific mortality, migration and population counts

$$s_{a,t,\ell},\ g_{a,t,\ell},\ n_{a,t,\ell},\ \ell = F, M.$$

- The total number aged $[0, 5)$ is calculated from the number of births, b_t, which is a function of the CCMPP inputs $(f_{a,t}, s_{a,t,F}, g_{a,t,F})$.
- The birth count is shared among sexes using sex ratio at birth (SRB_t)then subjected to mortality and migration.
- All-sex births are computed then decomposed because SRB is a parameter of interest.

Two-Sex Hierarchical Model

I. Likelihood $\log n^*_{a,t,\ell} \,|\, n_{a,t,\ell}, \sigma^2_n \sim \text{Normal}\left(\log n_{a,t,\ell}, \sigma^2_n\right)$

II. Projection Model

$$n_{a,t,\ell} \,|\, \mathbf{n}_{t-5,\cdot}, \mathbf{f}_{t-5}, \ldots, SRB_{t-5} = Proj(\mathbf{n}_{t-5,\cdot}, \mathbf{f}_{t-5,\cdot}, \mathbf{s}_{t-5,\cdot}, \mathbf{g}_{t-5,\cdot}, SRB_{t-5})$$

III. Priors on Inputs

$$\begin{aligned}
\log SRB_t \,|\, SRB^*_t, \sigma^2_{SRB} &\sim \text{Normal}\left(\log SRB^*_t, \sigma^2_{SRB}\right) \\
\log n_{a,t_0,\ell} \,|\, \sigma^2_n &\sim \text{Normal}\left(\log n^*_{a,t_0,\ell}, \sigma^2_n\right) \\
\log f_{a,t} \,|\, \sigma^2_f &\sim \text{Normal}\left(\log f^*_{a,t}, \sigma^2_f\right) \\
\text{logit}\, s_{a,t,\ell} \,|\, \sigma^2_s &\sim \text{Normal}\left(\text{logit}\, s^*_{a,t,\ell}, \sigma^2_s\right) \\
g_{a,t,\ell} \,|\, \sigma^2_g &\sim \text{Normal}\left(g^*_{a,t,\ell}, \sigma^2_g\right)
\end{aligned}$$

IV. Hyperparameters

$$\begin{aligned}
\sigma^2_v &\sim \text{InvGamma}(\alpha_v, \beta_v), \\
&\quad v \in \{n, f, s, g, SRB\}
\end{aligned}$$

Sex Ratios in India, 1971–2001

In most human populations

- There are slightly more males at ages below $\sim$30, more females age ages above $\sim$30.
- Female life expectancy is higher than male.
- Sex ratio at birth is 1.04–1.06 males per female.

In India

- It is believed that males outnumbered females even at older ages.
- Estimates of female life expectancy are lower than male.
- Since 1970s, some estimates of SRB in some areas have been above 1.06.

Why?

- Thought to be due to a cultural preference for sons over daughters manifested in
  - higher female mortality pre-1970s
  - high SRBs post-1970s

Sex Ratios in India, 1971–2001

Bayesian Reconstruction will allow us to answer probabilistic questions about sex ratios.

Research Questions

What is the probability that

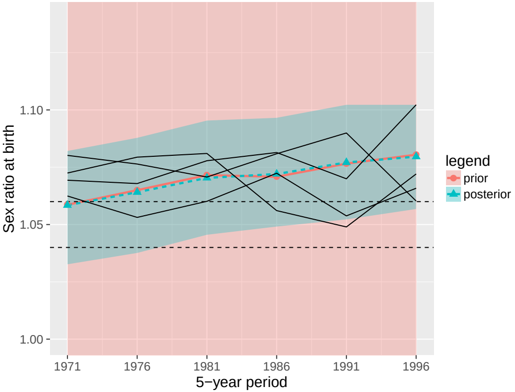
i. SRB was unusually high between 1971 and 2001?

ii. The SRB increased in India since 1971?

Initial Estimates

Parameter	Sources	Methods	Rel. Error. (%)
Pop count	Censuses, 5-yearly, 1971–2001	Adjustments for undercount	10
Fertility	Indian SRS, Nat. Fam. Health Sur.	Indirect methods, smoothing	10
Survival	"	Indirect methods, model life tables	10
Migration	None	Centered at zero, large rel. err.	20
SRB	Indian SRS, Nat. Fam. Health Sur.	Indirect methods, smoothing	10

Sex Ratio at Birth: Level



Pr(SRB > 1.06)

1971	1976	1981	1986	1991	1996
0.44	0.66	0.83	0.86	0.93	0.96

Sex Ratio at Birth: Trend

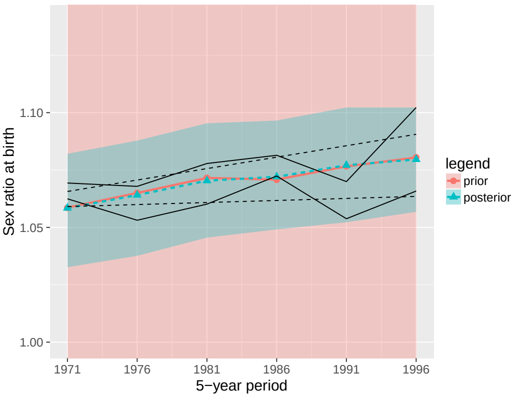
- The trend in SRB between 1971 and 2001 in the posterior was summarized by:
  1. the slope coefficient of the linear model (OLS slope)

$$\mathrm{SRB}_i = \alpha_i + \beta_i \cdot \mathrm{time} + \epsilon_{i,\mathrm{time}}$$

  2. the simple difference

$$\mathrm{SRB}_{i,1996} - \mathrm{SRB}_{i,1971}$$

Sex Ratio at Birth: Trend



Sex Ratio at Birth: Trend

Pr(SRB Increased)

Statistic values greater than zero indicate an increase.

Statistic	95% CI	Prob > 0
OLS slope	$[-0.00034, 0.0021]$	0.93
$\text{SRB}_{1996} - \text{SRB}_{1971}$	$[-0.011, 0.054]$	0.92

Conclusions

Research Questions

What is the probability that

i. SRB was unusually high between 1971 and 2001?
   - $\Pr(\text{SRB} > 1.06) = \{0.93\ (1991\text{–}1996),\ 0.96\ (1996\text{–}2001).\}$

ii. The SRB increased in India since 1971?
   - $\Pr(\text{an increase}) = \{0.93\ (\text{OLS}), 0.92\ (\text{Diff})\}$.

Life Expectancy at Birth

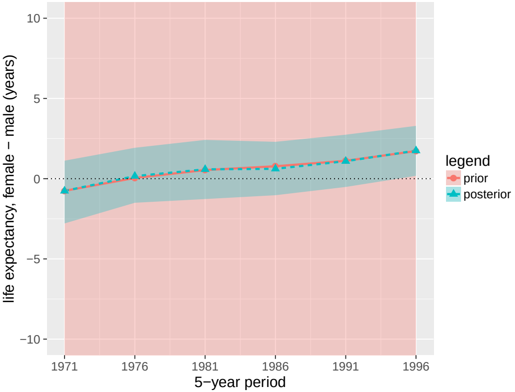
Research Questions

What is the probability that

i. Female life expectancy was lower than male life expectancy between 1971 and 2001?

ii. The gap between male and female life expectancy decreased since 1971?

Life Expectancy at Birth



Probability female life expectancy < male

1971	1976	1981	1986	1991	1996
0.79	0.42	0.26	0.22	0.09	0.01

Life Expectancy at Birth

- Repeating the trend analysis done for SRB:
  - Statistic values greater than 0 indicate an increase in the difference over the reconstruction period.

Statistic	95% CI	Prob > 0
OLS slope (LEB $\sim$ time)	[0.0066, 0.1721]	0.98
$\text{LEB}_{1996} - \text{LEB}_{1971}$	[0.01, 0.17]	0.98

Conclusions

Research Questions

What is the probability that

i. Female life expectancy was lower than male life expectancy between 1971 and 2001?
   - $\Pr(\text{fem } e_0 \ < \ \text{male}) = 0.79$ in 1971.

ii. The gap between male and female life expectancy reduced since 1971?
   - $\Pr(\text{an increase}) = \{0.98\ (\text{OLS}), 0.98\ (\text{Diff})\}$.

Summary

- Bayesian Population Reconstruction can provide probabilistic answers to meaningful questions about transformations of the input parameters, as well as the parameters themselves.

Summary

Summary

Bayesian Population Reconstruction is a new method for reconstructing population structures of the past.

- **Inputs**
  - Bias-reduced initial estimates of age-specific vital rates, net international migration and population counts.
  - Expert knowledge and/or data about measurement error variance.
- **Output**
  - Joint posterior distribution over all parameters.
- **Improvements**
  - Quantitative estimates of uncertainty, expressed probabilistically.
  - Trends and uncertainty are estimated coherently.
  - Estimation requires a single run over the inputs and requires no subjective adjustments.
- **Features**
  - Works across a range of data quality contexts.
  - Provides two-sex reconstructions.
  - Transformations of parameters can also be studied.

(*Refs:* Wheldon et al. (2013, 2016, 2015); *popReconstruct* on GitHub)

References

Wheldon, M. C., Raftery, A. E., Clark, S. J., and Gerland, P. (2013), "Reconstructing Past Populations With Uncertainty From Fragmentary Data." *Journal of the American Statistical Association*, 108, 96–110.

— (2015), "Bayesian reconstruction of two-sex populations by age: estimating sex ratios at birth and sex ratios of mortality," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178, 977–1007.

— (2016), "Bayesian population reconstruction of female populations for less developed and more developed countries," *Population Studies*, 70, 21–37.