# Bayesian Demographic Estimation and Population Reconstruction

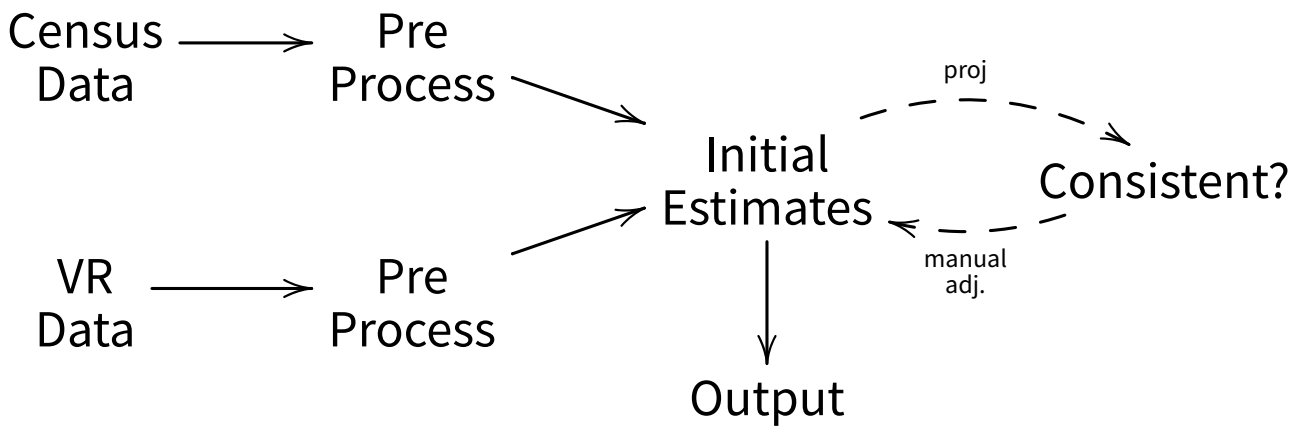
**UN Population Division Expert Group Meeting on**
***Methods for the World Population Prospects 2021 and Beyond***
**6–8 April 2020**

Mark Wheldon

United Nations, Population Division

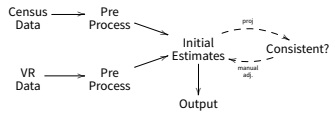
**Acknowledgements:** Joint work with Adrian Raftery, Sam Clark and Patrick Gerland
NICHD, Grants R01 HD054511 and K01 HD057246
BayesPop Working Group, CSSS, and CSDE at the UW
FHES at AUT

# WPP Estimation Process




- 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.

WPP Estimation Process




- 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.

- UN collects all available data from surveys, vital registration and censuses.
- Pre-processes them to get a single initial estimate for each age group and time period.
- Project and manually adjust to ensure consistency.
- Pre-processing requires use of “indirect estimation” methods taken from demography literature.

# Bayesian Population Reconstruction

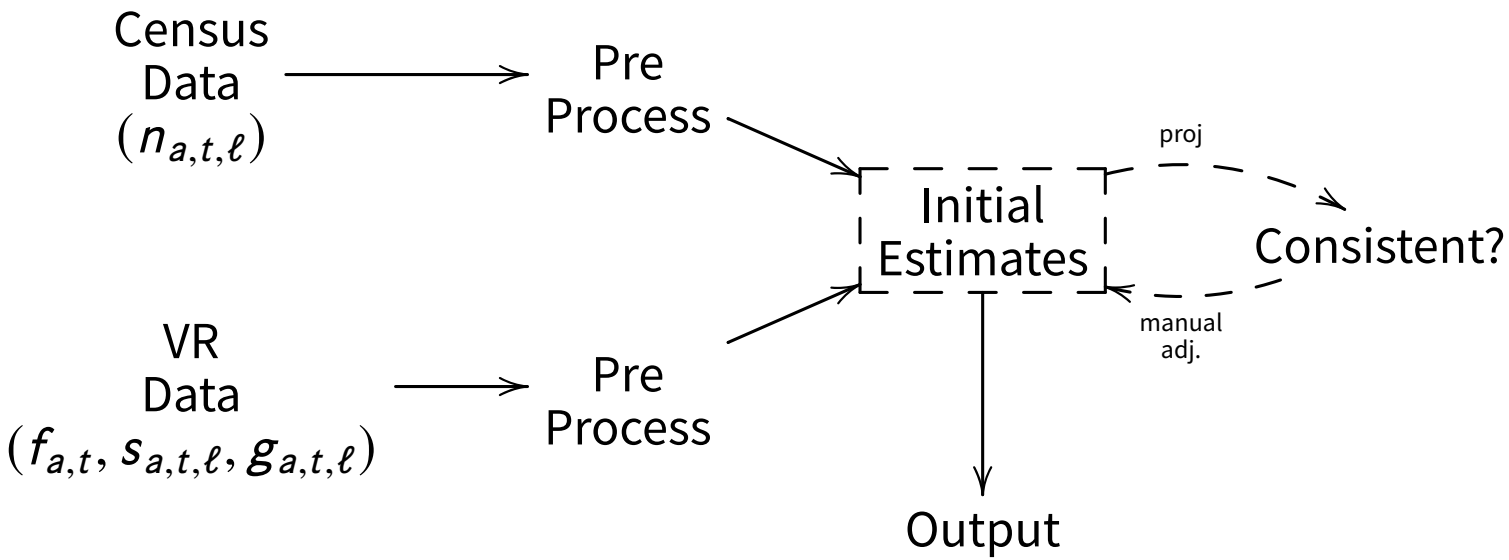
*Bayesian Population Reconstruction* was proposed as an improved method for reconstructing populations of the recent past.

## Goals

1. Quantify uncertainty probabilistically.
2. Estimate all parameters consistently.
3. Be easily replicable.
4. Use all reliable data and expert opinion.

# Bayesian Population Reconstruction

## Current UN Practice




## Bayesian Population Reconstruction

# Parameters and Notation

## Indices

- 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.
- Sex $\ell, F, M$.

## Parameters

- $f_{a,t}$: age-specific fertility rate.
- $s_{a,t}$: survival proportion.
  - I.e., proportion surviving from age $a - x$ to age $a$.
- $g_{a,t}$: net international migration.
- $n_{a,t}$: age-specific population count.

# Hierarchical Model

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

**I. Likelihood**

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

**II. Projection Model**

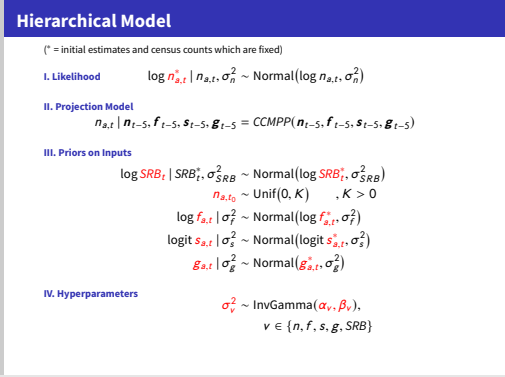
$$n_{a,t} \mid \boldsymbol{n}_{t-5}, \boldsymbol{f}_{t-5}, \boldsymbol{s}_{t-5}, \boldsymbol{g}_{t-5} = CCMPP(\boldsymbol{n}_{t-5}, \boldsymbol{f}_{t-5}, \boldsymbol{s}_{t-5}, \boldsymbol{g}_{t-5})$$

**III. Priors on Inputs**

$$\begin{aligned}
\log SRB_t \mid SRB^*_t, \sigma^2_{SRB} &\sim \text{Normal}\left(\log SRB^*_t, \sigma^2_{SRB}\right) \\
n_{a,t_0} &\sim \text{Unif}\left(0, K\right) \qquad , K > 0 \\
\log f_{a,t} \mid \sigma^2_f &\sim \text{Normal}\left(\log f^*_{a,t}, \sigma^2_f\right) \\
\text{logit}\, s_{a,t} \mid \sigma^2_s &\sim \text{Normal}\left(\text{logit}\, s^*_{a,t}, \sigma^2_s\right) \\
g_{a,t} \mid \sigma^2_g &\sim \text{Normal}\left(g^*_{a,t}, \sigma^2_g\right)
\end{aligned}$$

**IV. Hyperparameters**

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




The model has four levels. I shall start at level 3:

**3** The true vital rates and baseline population are given priors, conditional on the initial estimates and variance parameters which account for measurement error.

**4** The variance parameters are, themselves, given distributions at level 4.

**2** The priors on the vital rates and baseline counts induce a prior on subsequent counts through the deterministic projection model in level 2.

**1** Level 1 is a likelihood for the census counts which is parameterized by the projected counts and a variance parameter which accounts for measurement error.

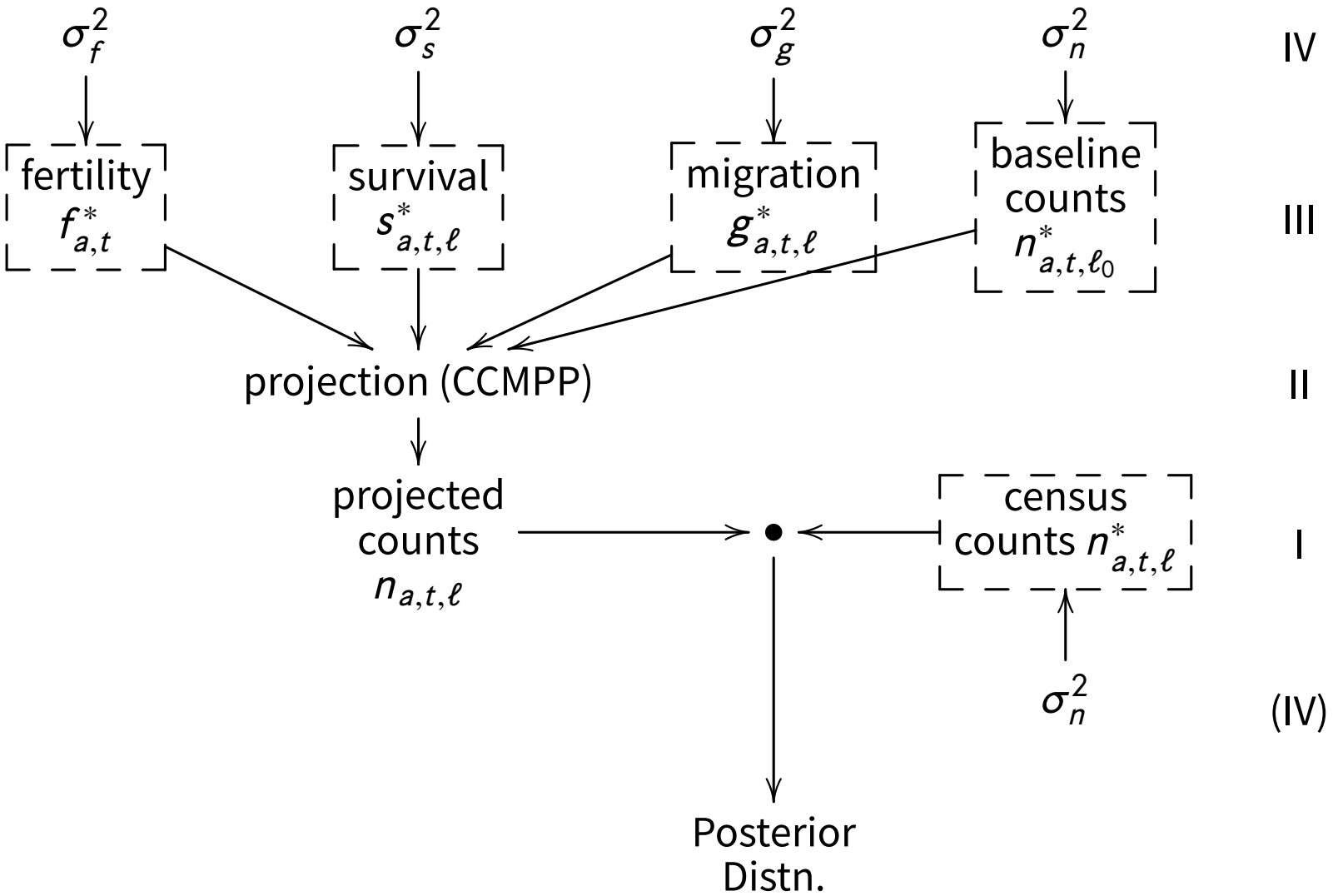
1. Here are the things we have data on
2. Here are the things we have to specify
3. Here are the things we do inference on

## Inference

- Inference is based on the joint posterior distribution of the input parameters $(f_{a,t}, s_{a,t}, g_{a,t}, n_{a,t_0})$.

# Hierarchical Model: Key Relationships

(inputs are boxed)



# Measurement Error Variance

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

- The $\sigma_v^2$ are random. They reflect non-systematic error in the initial estimates.
- We specify $\alpha_v$ and $\beta_v$ 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.

**Measurement Error Variance**

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

- The $\sigma_v^2$ are random. They reflect non-systematic error in the initial estimates.
- We specify $\alpha_v$ and $\beta_v$ 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 ±p percent."*

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

We choose the hyperparameters using the MAE of the marginal prior distributions of the observables. For example, we use the joint marginal distribution of the vector of fertility rates, $(f_{1,1}, \ldots f_{A,T})$ which is

$$t_{AT}\left(\log \boldsymbol{f}^*, \beta_f/\alpha_f \mathbf{I}_{AT}, 2\alpha\right),\ \alpha \geq 1/2,\ \beta > 0$$

Then

$$\begin{aligned}
\text{MAE}\left(\log \boldsymbol{f}\right) &\equiv \text{E}\,|\log \boldsymbol{f} - \log \boldsymbol{f}^*| \\
&= \text{E}\left(\text{E}\left[\left(|\log \boldsymbol{f} - \log \boldsymbol{f}^*|\right)\left|\sigma_f^2\right.\right]\right) \\
&= \mathbf{1}_{AT} \cdot \text{E}\left(\sqrt{\frac{2}{\pi}}\sigma_f\right) \\
&= \mathbf{1}_{AT} \cdot \sqrt{\frac{2\beta}{\pi}} \cdot \underbrace{\frac{\Gamma(\alpha - 1/2)}{\Gamma(\alpha)}}_{h(\alpha)},\ \alpha \geq 1/2,\ \beta > 0
\end{aligned}$$

where $\mathbf{1}_{AT}$ is a length $AT$ vector of 1s.

# Example: Laos and New Zealand

- The method is most useful for countries with unreliable, fragmentary data (high uncertainty).
- However it also works for countries with very good data.

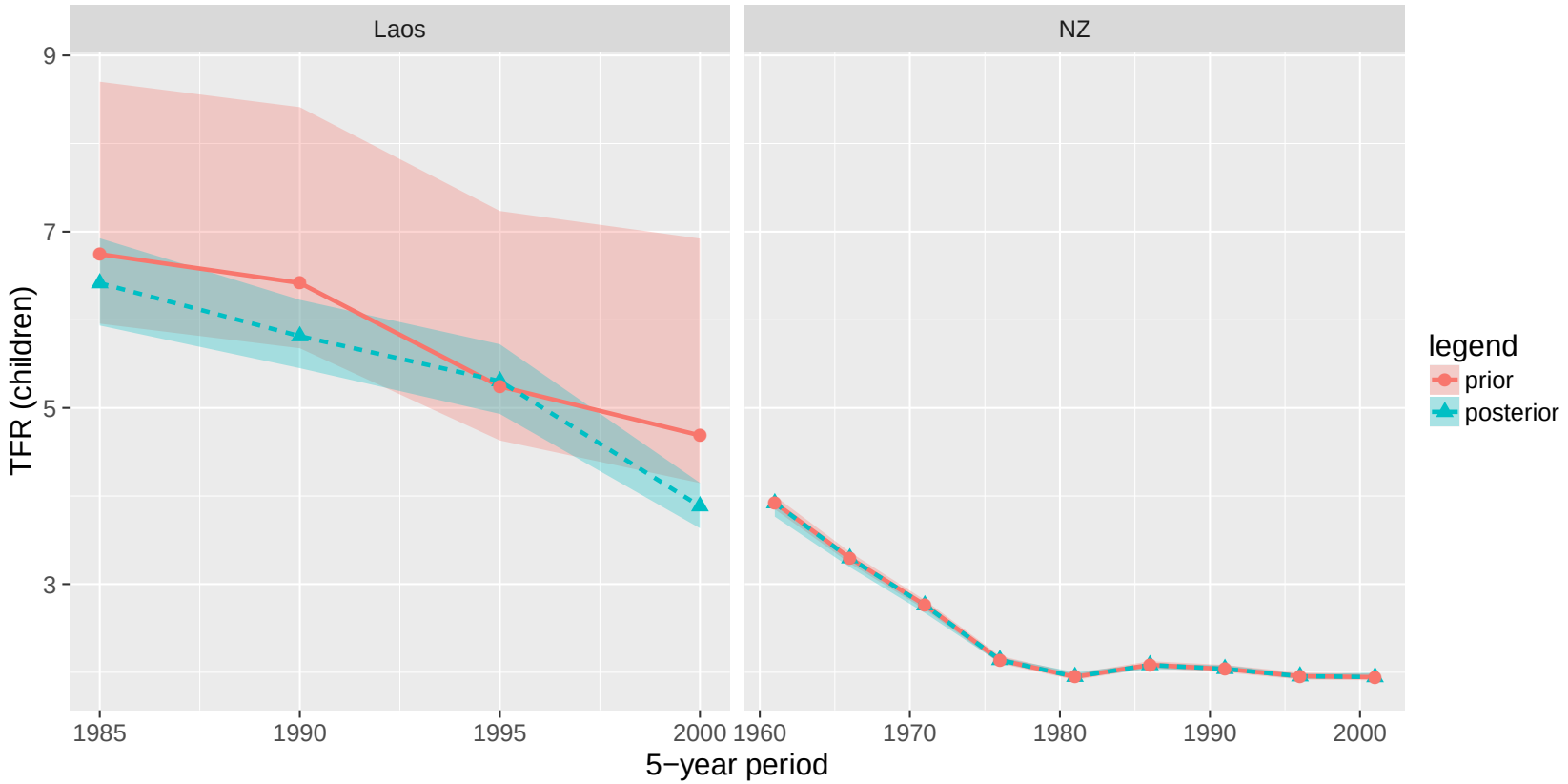
**Reconstructing the female populations of Laos (1985–2005) and New Zealand (1961–2006)**

- 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.

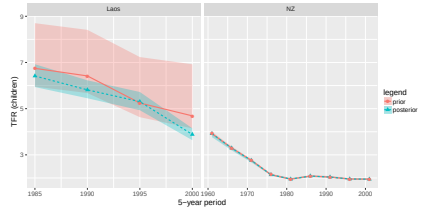
# Total Fertility Rate




## 95% Posterior Interval Half-widths

| | Mean Half-Width |
|---|---|
| Laos | 0.4 |
| NZ | 0.04 |

**Total Fertility Rate**

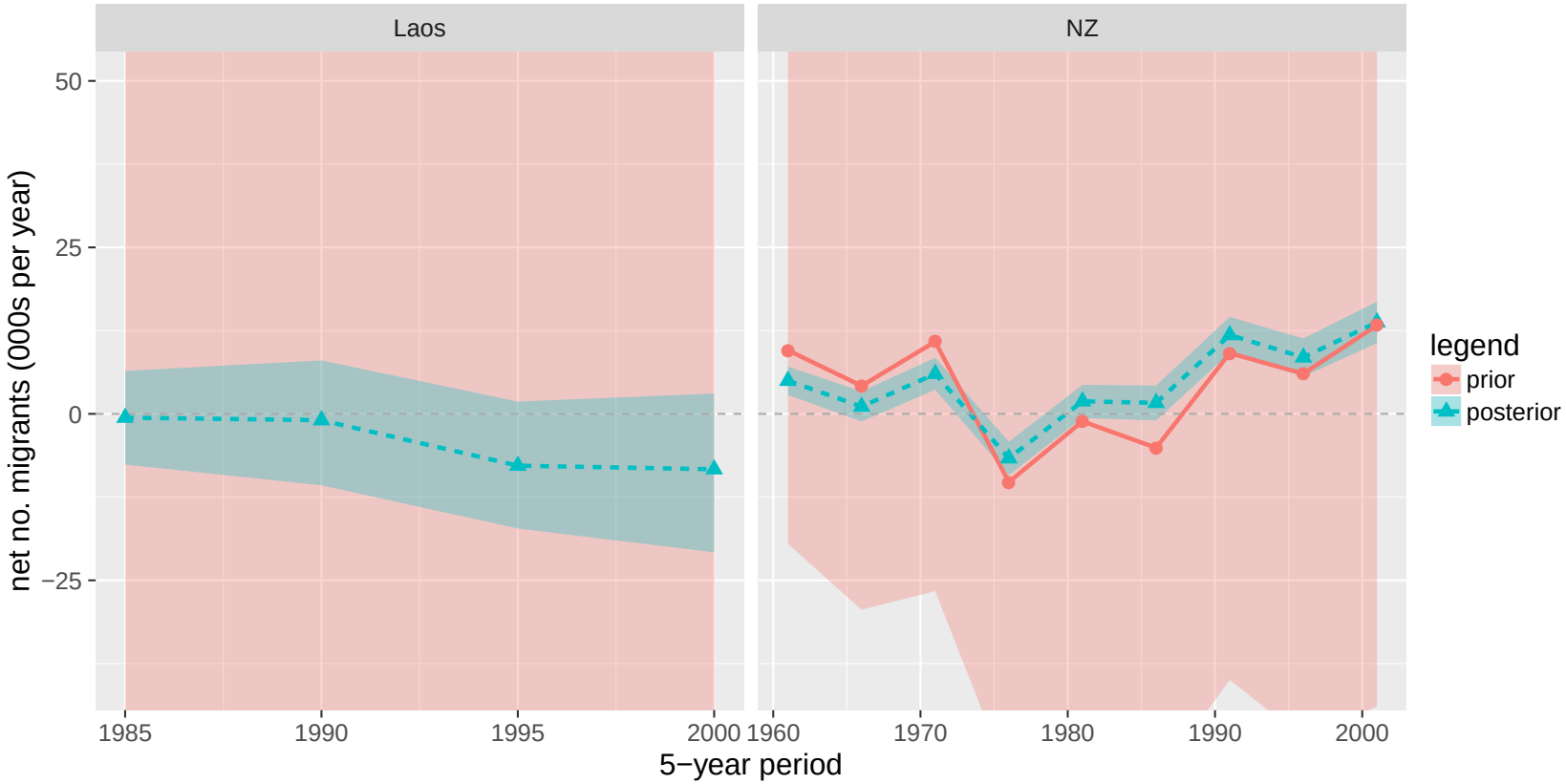



**95% Posterior Interval Half-widths**

| | Mean Half-Width |
|---|---|
| Laos | 0.4 |
| NZ | 0.04 |

- These plots show prior and posterior medians and the limits of 95% credible intervals.
  - The prior on TFR is in red, the posterior is in blue.
  - The initial estimates correspond to the prior median, shown by the red line.
- TFR in Laos was much higher than in New Zealand.
- Posterior uncertainty is quantified using the mean half-width of the credible intervals:
  - Mean half-width for New Zealand was about 1/10th that of Laos.
- I show the WPP 2010 estimates for comparison
  - Very similar data were used but our methods are different.
  - The WPP 2010 estimates are different but not too different!

# Total Net Migration




## 95% Posterior Interval Half-widths

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

# 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.

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

# Use in WPP

- Computation speed needs to be fast—could be a challenge with single-year age and time.
- Needs to expand beyond span of most census collections, back to 1950.
  - Some countries have only one or two censuses, or rely on admin. data, household surveys.
- Inputs and/or measurement error could be modelled (but see first point).
- Extensive testing required.

# 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.