UN/POP/EGM-CPD49/2015/16

ENGLISH ONLY

UNITED NATIONS EXPERT GROUP MEETING ON STRENGTHENING THE DEMOGRAPHIC EVIDENCE BASE FOR THE POST-2015 DEVELOPMENT AGENDA

Population Division
Department of Economic and Social Affairs
United Nations Secretariat
New York
5-6 October 2015

ESTIMATING POPULATION AND HEALTH QUANTITIES AND THEIR UNCERTAINTY FROM DATA OF LIMITED QUALITY[1]

Adrian E. Raftery*

[1] The opinions expressed in this paper are those of the author and do not necessarily reflect those of the United Nations or its Member States. This paper is reproduced as submitted by the author. This paper is being reproduced without formal editing.

[*] University of Washington.

Estimating Population and Health Quantities and their Uncertainty from Data of Limited Quality

Adrian E. Raftery
University of Washington

September 22, 2015

Abstract

Estimating past and current demographic and health quantities is hugely important. However, it is a difficult problem for countries without good vital registration systems, and there must rely on surveys and censuses that often have biases and substantial measurement error. I argue that Bayesian methods can be useful for this purpose, as they allow one to model, estimate and correct for biases and measurement errors, they allow multiple data sources to be combined, and they permit the incorporation of expert knowledge and information from other countries and contexts. They yield statements of uncertainty about the quantities being estimated. I briefly describe three problems for which Bayesian methods have been useful: estimating HIV prevalence in countries with generalized epidemics, estimating TFR in the absence of a vital registration system, and full Bayesian population reconstruction.

1 Introduction

The UN Population Division and other UN agencies such as UNAIDS face the problem of estimating current and past population and health quantities from data that are sparse and/or of poor quality. These include age-specific fertility, mortality and migration rates, and summaries of them such as the total fertility rate (TFR) and life expectancy at birth (e_0). They also include health quantities such as HIV prevalence.

This problem is particularly difficult for more than half of the world's countries that lack high-quality vital registration systems. In these countries, estimation relies crucially on surveys, which typically have biases and measurement errors and often are not carried out on a regular basis.

A major problem with data of this kind is the presence of systematic biases in a given direction, due for example to non-representative sampling, recall biases, and other systematic errors. An example is the estimation of HIV prevalence from ante-natal clinic (ANC) data, which has turned out to be systematically biased upwards.

A second problem is measurement error. A third source of variation is pure sampling variability in counts of vital events. While this is widely discussed in survey sampling textbooks, it is probably a small part of overall uncertainty for the kinds of problems we're discussing here, and indeed is often ignored in demographic applications, as an approximation.

A key point is that bias and measurement error are different and distinct. Biases need to be corrected. Different measurement error variances often lead to differential weighting of the different data sources.

I will argue that Bayesian approaches can have advantages for these problems. They allow one to model, estimate and correct for biases and measurement error variance explicitly. They allow multiple data sources to be combined, via the likelihood. They permit the incorporation of expert knowledge and information from other countries and contexts explicitly, via the prior distribution. They allow estimation for multiple countries simulataneously, taking advantage of their similarities, via Bayesian hierarchical models. And they automatically yield statements of uncertainty about the quantities being estimated.

I will briefly review three contexts where Bayesian methods have been found useful: estimation of HIV prevalence in countries with generalized epidemics, estimation of TFR in countries with sparse data, and full Bayesian population reconstruction. One lesson is that there is not a single Bayesian approach for all such problems, and that the model and approach need to be tailored to the context at hand.

2 Bayesian Approach

The Bayesian approach to statistical inference is based on the idea of representing uncertainty by probability distributions. It has become quite popular in many sciences in the past 20 years. There are many good textbooks, including Hoff (2009) and Gelman et al. (2013).

The basic idea is that inference about a quantity of interest, Q, is summarized by its *posterior distribution*, which is a probability distribution of it given all the available data and evidence. Bayes's theorem tells us that

$$p(Q|\text{Data}) \propto p(\text{Data}|Q)\ p(Q). \tag{1}$$

In equation (1), $p(Q|\text{Data})$ is the posterior distribution of Q which we seek, and $p(\text{Data}|Q)$ is the *likelihood* of the data given the truth Q, which takes account of bias and measurement error. Finally, $p(Q)$ is the *prior distribution* of Q, representing information about Q available before the data are collected, such as expert knowledge and information from other countries.

Typically, unknown parameters such as the bias and measurement error variance of measurements of Q, denoted by θ, are also estimated as part of the process. We then have

$$p(Q|\text{Data}, \theta) \propto p(\text{Data}|Q, \theta)\ p(\theta)\ p(Q). \tag{2}$$

Evaluating (1) or (2) directly can be hard, but they can often be approximated by Monte Carlo methods such as Markov chain Monte Carlo (MCMC), or importance sampling. In the case of (2), these yield a large set of values of the vector (Q, θ). The marginal posterior distribution of Q can then be approximated as the marginal distribution of the simulated values of Q.

Some of the advantages of the Bayesian approach in the present context are:

- It allows the incorporation of expert knowledge and information from other countries explicitly via the prior distribution.

- Multiple data sources can be incorporated directly via the likelihood. If there are m data sources (e.g. different surveys), so that $\text{Data} = (\text{Data}_1, \ldots, \text{Data}_m)$, then the corresponding likelihoods are simply multiplied together, so that

$$p(\text{Data}|Q, \theta) = p(\text{Data}_1|Q, \theta) \times \ldots \times p(\text{Data}_m|Q, \theta). \tag{3}$$

- It allows the simultaneous and consistent production of estimates for multiple countries taking advantage of their similarities, via a Bayesian hierarchical model. So far this has mostly been done for population projections rather than estimation or reconstruction (Alkema et al., 2011; Raftery et al., 2012, 2013), but there is great potential to do it for estimates as well.

- It allows the model to be easily specified in terms of directly interpretable quantities, rather than statistically convenient parameters that are hard to interpret.

These advantages are balanced by the fact that the model needs to be carefully specified and that estimation can take a lot of computer time.

3 Examples

I will now outline very briefly some practical applications of the Bayesian approach to estimating demographic and health quantities.

3.1 Estimating HIV Prevalence for Countries with Generalized Epidemics

There are about 40 countries defined by UNAIDS as having generalized HIV/AIDS epidemics, and for most of these countries, data on the epidemic are sparse and/or unrepresentative. There are two main kinds of data available for estimating HIV prevalence.

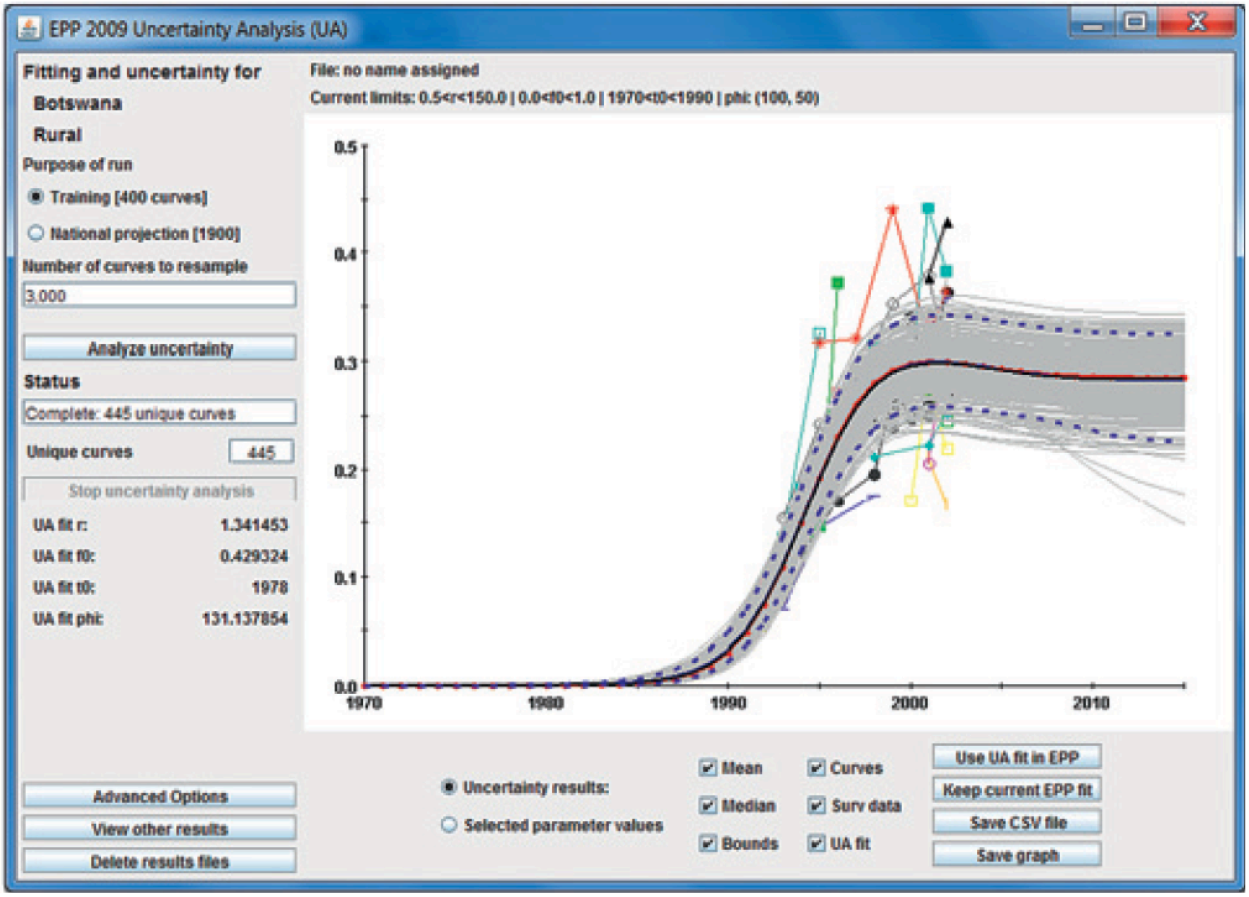
One kind of data consists of measured prevalences among pregnant women at ante-natal clinics (ANCs). These measurements can be frequent in time and so give an idea of trends, but they are clearly unrepresentative in being restricted to pregnant women and having poor geographic coverage. Also, early in the epidemic the ANCs sampled tended to be in the most severely affected areas, so that simply using the time series of estimated prevalences tends to underestimate the increase in prevalence (or give a false impression of a decrease).

The second kind of data consists of national household surveys, such as DHS surveys, which are more representative, but are sparse in time, and sometimes nonexistent for a given country.

We developed a Bayesian model for this situation that assumed that prevalence followed the standard susceptible-infected-removed (SIR) pattern characteristic of infectious diseases, and allowed for measurement error in both the ANC and DHS data (Alkema et al., 2007). We assumed that the DHS data were unbiased (although measured with error), and we modeled the bias in the ANC data. For countries without DHS data, the bias in the ANC data couldn't be estimated directly, and for these countries we used a prior distribution of the bias based on experience in other countries. The method also incorporated a prior distribution of the SIR model outputs; the resulting variant of Bayesian inference is called Bayesian melding (Poole & Raftery, 2000).

MCMC methods didn't work well for this problem, so we developed an efficient importance sampling method, called Incremental Mixture Importance Sampling (IMIS) (Raftery & Bao, 2010). This is implemented in the R function `IMIS`.

The method worked better than non-Bayesian methods that UNAIDS had been applying before, such as maximum likelihood estimation and bootstrapping. It was adopted by UNAIDS after extensive evaluation by the UNAIDS Reference Group on Estimation, Projection and Modelling. It was incorporated into UNAIDS's EPP software, and is now part of the Spectrum package that UNAIDS uses and supports in member countries. Figure 1 shows an example probabilistic estimate of HIV prevalence over time for Botswana produced by UNAIDS's EPP software.



EPP 2009 Uncertainty Analysis (UA)

Fitting and uncertainty for

Botswana

Rural

Purpose of run

● Training [400 curves]

○ National projection [1900]

Number of curves to resample

3,000

Status

Complete: 445 unique curves

Unique curves 445

UA fit r: 1.341453

UA fit f0: 0.429324

UA fit t0: 1978

UA fit phi: 131.137854

File: no name assigned

Current limits: 0.5<r<150.0 | 0.0<f0<1.0 | 1970<t0<1990 | phi: (100, 50)

● Uncertainty results:

○ Selected parameter values

☑ Mean

☑ Median

☑ Bounds

☑ Curves

☑ Surv data

☑ UA fit

Figure 1: Posterior distribution of HIV prevalence over time in rural Botswana from Bayesian melding. The colored lines show estimates from ante-natal clinics. The grey lines show trajectories sampled from the posterior distribution. The solid black line shows the posterior median, and the blue dashed lines show the posterior 95% intervals. *Source:* Brown et al. (2010).

3.2 Estimating TFR Without a Vital Registration System

Estimating past and present TFR is important for the UN's *World Population Prospects*, both for direct reporting of past population quantities, and as inputs to fertility projections. For countries with good vital registration systems this is easy, and high quality estimates are available directly from the vital statistics.

A majority of the world's countries do not have such vital registration systems, however, and for them estimating past and present TFR is much harder. Estimation needs to rely on surveys and censuses. Data on fertility in these countries come from multiple sources of uneven quality; the problems include limited coverage through time, bias, and measurement error. Moreover, this group includes the countries with high fertility, which are particularly important for understanding the population dynamics of their regions. The observations are typically collected retrospectively by either asking women about their births in a restricted period (e.g. the number of births in the last year before the survey/census) or their complete birth histories (birth of their first child, second child, etc.).

Alkema et al. (2012) proposed a multi-country method for estimating past and present TFR from such data. Observations were assumed to be equal to the true TFR plus a bias term, and with measurement error variances that varied across observations. Both bias and measurement error variance were assumed to be related to characteristics of the data source, such as whether it's a census, DHS survey or non-DHS survey, how far into the past the retrospective observations are, and whether the estimate is direct or uses demographic adjustement such as the Brass method. The relationship was estimated from the data. Uncertainty was assessed using the weighted likelihood bootstrap, which is a generalized Bayesian bootstrap. The method is automatic and replicable, and provides an assessment of uncertainty.

Figure 2 illustrates the method for Burkina Faso. Figure 2(a) shows the data. Figure 2(b) shows the bias-adjusted observations and compares them to the original observations; clearly the variation is a lot smaller. Figure 2(c) shows the estimates and their uncertainty.

3.3 Full Bayesian Population Reconstruction

I now describe a method called Bayesian population reconstruction, whose aim is to fully reconstruct past populations by age and sex from fragmentary data (Wheldon et al., 2010, 2013, 2015, 2016).

Bayesian population reconstruction reconciles two different estimates of population counts, those based on adjusted census counts (or similar data) and those derived by projecting initial estimates of the baseline population forward using initial estimates of vital rates. Adjusted

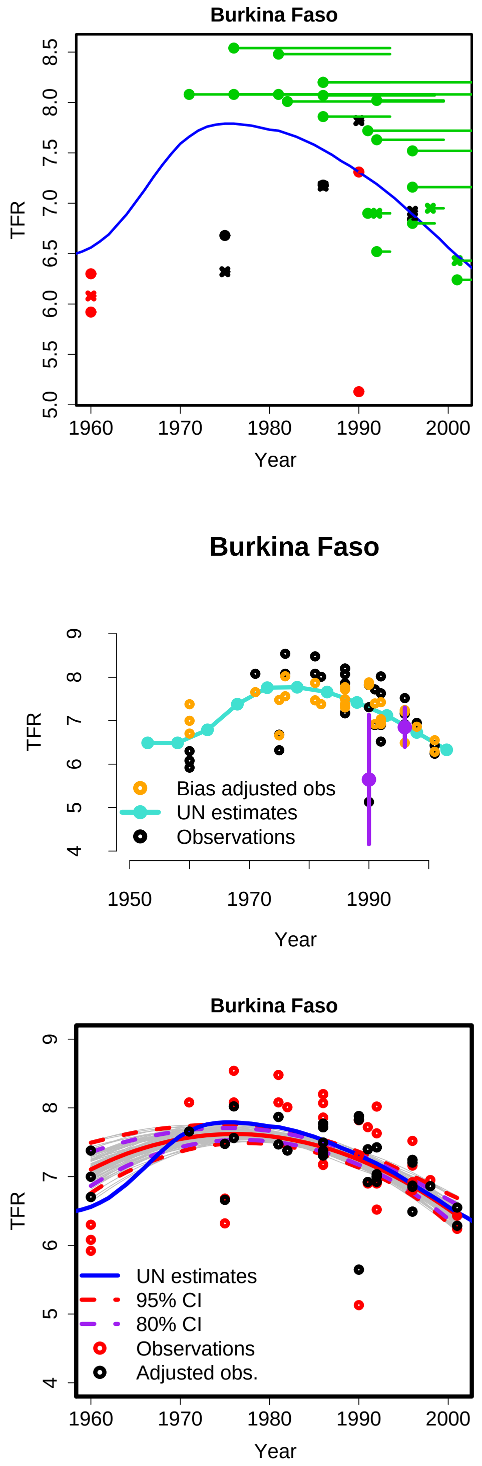


Figure 2: Estimating TFR with Uncertainty for Burkina Faso: (a) Direct observations (dots) and indirect observations (crosses) for different data sources. The green horizontal lines show DHS estimates and extend from the midpoint of the observation period to the year of data collection. Censuses are shown by black dots, and non-DHS surveys by red dots. The UN estimates are shown by the blue curve. (b) Original and bias-adjusted observations, with UN WPP 2006 estimates. The 95% confidence interval for the TFR is shown for two observations (vertical purple lines). (c) Median estimates and confidence intervals for the TFR.

census counts are raw counts which have been processed to reduce common biases such as undercount and age heaping.

Initial point estimates of the input parameters are derived from data. Baseline population estimates come from adjusted census counts, and fertility and mortality estimates come from surveys such as the DHS and from vital registration data. The model defines a joint prior distribution over these parameters which is parameterized by the initial point estimates and standard deviations. Typically, the initial point estimates serve as the marginal medians of the prior distribution. The standard deviations represent measurement uncertainty about the point estimates. These distributions induce a probability distribution on the population counts at the end of each projection step within the period of reconstruction.

This yields a Bayesian hierarchical model for all available data which is estimated by MCMC. It provides a full joint posterior distribution of all age- and sex-specific vital rates and population counts over the reconstruction period. Hence it also provides a full posterior distribution of derived quantities such as the TFR, e_0 and the net migration rate. It is implemented in the R package `popReconstruct`.

A limitation of the method is that it requires the full data on which population estimates are to be based, back into the past. This is currently not readily available for many countries, although the data do exist for most countries. The UN Population Division's Gates-funded data project aims to rectify this situation, so Bayesian population reconstruction may be applicable to many countries in the future. So far it has been applied to about a dozen countries for which the full data are available.

Figure 3 shows the results of Bayesian population reconstruction for some demographic quantities of interest for Laos, a country with no vital registration system and few data, and New Zealand, a country with a good vital registration system and excellent data. These include statistically principled confidence intervals for these quantities based on all available data, possibly for the first time. For TFR, the average posterior 95% interval width is 0.3 children for Laos, and 0.03 for New Zealand, with similar results for mortality.

For migration, the method provdes posterior estimates and posterior intervals for net migration, shown in Figure Figure 3(d). For New Zealand, the prior medians were based on international passenger surveys. The posterior medians track these fairly well, but the prior medians are outside the posterior intervals. This shows that even in a country like New Zealand, with some of the best migration data, estimating net migration is challenging. This underlines the importance of assessing uncertainty about the estimates.

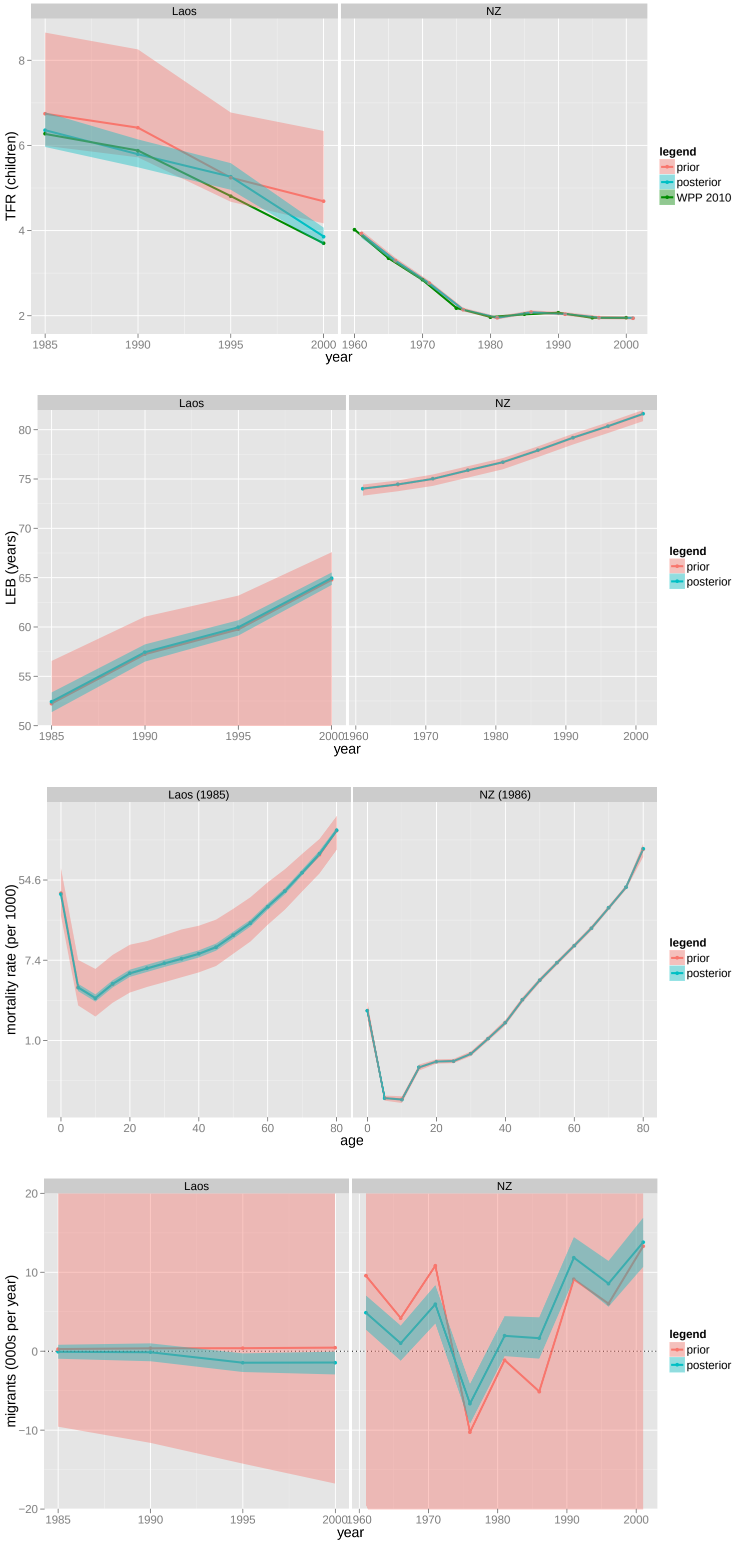


Figure 3: Bayesian Population Reconstruction for Laos and New Zealand: (a) TFR; (b) life expectancy at birth; (c) age-specific mortality rates; (d) net migration. *Source:* Wheldon et al. (2016).

4 Conclusion

I have discussed the estimation of past and present demographic and health quantities from sparse data with biases and substantial measurement error. I have suggested that Bayesian approaches can be useful in this context, and I have shown three substantial examples where Bayesian methods have provided solutions. The specific Bayesian method varies by application.

While Bayesian estimation of demographic and health quantities appears to be potentially useful, it is at an early stage, and further experience will be important in refining the methods.

Acknowledgements: This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) under NIH Grants R01 HD054511 and R01 HD070936.

References

Alkema, L., Raftery, A. E., & Clark, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *Annals of Applied Statistics*, *1*, 229–248.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., & Pelletier, F. (2012). Estimating the total fertility rate from multiple imperfect data sources and assessing its uncertainty. *Demographic Research*, *26*, 331–362.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., & Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography*, *48*, 815–839.

Brown, T., Bao, L., Raftery, A. E., Salomon, J. A., Baggaley, R. F., Stover, J., & Gerland, P. (2010). Modelling HIV epidemics in the antiretroviral era: the UNAIDS Estimation and Projection package 2009. *Sexually Transmitted Infections*, *86*, i3–i10.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. New York: Springer.

Poole, D. & Raftery, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, *95*, 1244–1255.

Raftery, A. E. & Bao, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics*, *66*, 1162–1173.

Raftery, A. E., Chunn, J. L., Gerland, P., & Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, *50*, 777–801.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., & Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, *109*, 13915–13921.

Wheldon, M. C., Raftery, A. E., Clark, S. J., & Gerland, P. (2010). A Bayesian model for estimating population size and demographic parameters with uncertainty. Paper presented at the Annual Meeting of the Population Association of America, Dallas, Texas. http://paa2010.princeton.edu/papers/100654.

Wheldon, M. C., Raftery, A. E., Clark, S. J., & Gerland, P. (2013). Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association*, *108*, 96–110.

Wheldon, M. C., Raftery, A. E., Clark, S. J., & Gerland, P. (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)*, in press.

Wheldon, M. C., Raftery, A. E., Clark, S. J., & Gerland, P. (2016). Bayesian population reconstruction of female populations for less developed and more developed countries. *Population Studies*, in press.