Small area estimation of indigenous life expectancy in Australia

March 13, 2019

Abstract

Subnational mortality estimation, which is difficult enough when dealing with the whole population, becomes even more difficult when trying to make estimates for a group within the larger population, such as indigenous people. Bayesian statistical methods are a promising approach to such problems. We demonstrate the use of Bayesian methods, applied to the problem of deriving life tables for indigenous people in Australial, for subnational regions. We present early results, based on relatively aggregated data; by the time of the workshop, we expect to have much more disaggregated data and estimates. We describe how the models are set up, including an introduction to 'prior distributions', which allow users to encode expert judgements in a transparent and repeatable way. We also illustrate how model assumptions can be tested using 'replicate data'. All the calculations are carried out using open source *R* packages.

1 Introduction

Estimating how mortality varies across regions within a country can be challenging for traditional demographic methods. When the estimates are confined to specific groups within the population, the challenges become even greater. Bayesian statistics provides some promising avenues for addressing these challenges, however, as we hope to demonstrate using a project to derive life tables for indigenous people in small areas within Australia.

Estimating mortality rates for small areas requires a combination of robustness and flexibility. Once death counts have been disaggregated by age, sex, and geography, cell sizes become small, and randomness looms large. Direct approaches to mortality estimation, such as simply dividing observed deaths by observed population, become unreliable. Demographers have methods, such as model life tables, for dealing with noisy data. However, when used with complex datasets, these methods can smooth away genuine variation in underlying risks. The methods also typically require ad hoc adjustments, which makes them expensive to implement and difficult to replicate. And they typically do not yield measures of uncertainty.

Bayesian hierarchical models provide more satisfactor solutions to many of these problems (Congdon, 2010; Alexander et al., 2017; Bijak and Bryant, 2016). Most Bayesian hierarchical models smooth adaptively, in that they give greater weight to the raw data when cell counts are large, and greater weight to statistical predictors when the cell counts are small. The particular models that we present here include terms such as age effects and time effects that are substantially meaningful, and that can be specified in ways that capture demographic knowledge (Bryant and Zhang, 2018). Like most Bayesian models, our models yield detailed measures of uncertainty, including uncertainties for derived quantities such life expectancies. The methods have been used to produce official national, subnational, and ethnic life tables in New Zealand(Statistics New Zealand, 2015).

In this abstract, we present a model, and some illusrative results, for a highly aggregated public-domain dataset. By the time of the workshop, we will have mortality estimates based on much more disaggregated data from the Enhanced Mortality Database at the Australian Institute for Health and Welfare.

2 Data and Methods

Our dataset contains counts of deaths and population at risk, as reported by the Australian Bureau of Statistics[1]. The data covers the period 2010–2016, and is stratified by age at death, sex, indigenous status, state or territory, and year. The age groups are 0, 1-4, 5-14, ..., 65-74, 75+. The ABS omits the three Australian states or territories with the smallest number of indigenous people (Tasmania, Victoria, and the Australian Capital Territory) from the dataset.

[1]The data were downloaded from the database *Deaths, Year of registration, Indigenous status, Age at death, Sex, Five State/Territory* on the ABS website, on 1 March 2019

For indigenous deaths, the minimum cell size is 0, the maximum is 124, and the median is 17.

Figure 1: Direct estimates of mortality rates, 2016. Log scale. The points along the bottom of the graph represent values that were 0 on the original scale, and hence are undefined on the log scale.

Figure 1 show estimates of mortality rates in 2016, obtained using the traditional demographic approach of dividing death counts for each combination of the stratifying variables by the associated population at risk. It is clear that indigenous mortality rates are higher than non-indigenous rates across all states and territories. There is, however, too much random variation to permit confident answers to questions such as whether indigenous mortality rates are equal across all states and territories.

Let y_{asirt} denote deaths to people of age a, sex s, and indigenous status i,

in state or territory r, during year t. Similarly, let w_{asirt} denote the population at risk. We model deaths as draws from a Poisson distribution,

$$y_{asirt} \sim \text{Poisson}(\gamma_{asirt} w_{asirt}) \tag{1}$$

where γ_{asirt} is the mortality rate.

We model how the (log) mortality rate for each cell $asirt$ varies with age, sex, indigenous status, state or territory, and year,

$$\log \gamma_{asirt} \sim \text{N}(x_{asirt}\beta, \sigma^2). \tag{2}$$

Vector β contains a combination of main effects and interactions, which are listed in Table 1. Vector x_{asirt}, which is composed of 0s and 1s, assigns the appropriate elements of β to each cell in the classification.

Table 1: Priors for main effects and interactions

Term	Prior
(Intercept)	Exchangeable with known variance
age	Local trend with covariates
sex	Exchangeable with known variance
region	Exchangeable
time	Local trend
indigenous	Exchangeable with known variance
age:sex	Exchangeable
age:indigenous	Exchangeable
sex:indigenous	Exchangeable
region:indigenous	Exchangeable
time:indigenous	Local level
age:sex:indigenous	Exchangeable

Each main effect and interaction in (2) is provided with a 'prior' distribution. Priors are a distinctive feature of Bayesian statistical analyses. They provide soft constraints on the estimates that reflect, in some way, our knowledge about the parameter in question, beyond what is contained in the data itself. We assume, for instance, that region effects follow an 'exchangeable' prior of the form

$$\beta_r^{\text{reg}} \sim \text{N}(0, \tau_{\text{reg}}^2). \tag{3}$$

By using this prior we are essentially stating that mortality levels across states and territories are similar, though not identical. The degree of similarity is governed by the parameter τ_{reg}. The prior for the time effect is a 'local trend' model (Prado and West, 2010, 119–120), which is a generalization of a random walk with drift. The age effect is given a similar prior: in statistical demography, smoothing over age is much like smoothing over time, so it makes sense to apply time series models to age effects.

In the course of building a model such as that of (1) and (2), we inevitably make simplifying assumptions. The model assumes, for instance, that, although

the level of mortality experienced by indigenous people varies from region to region, the underlying age pattern does not. In other words, the model does not include an age-indigenous-region interaction. Before we can accept the results from the model, we have to test to see that such assumptions are reasonable.

One such test is to generate 'replicate datasets' and compare these with the actual dataset. The replicate datasets are generated using the fitted values from our statistical model, except that the region effects and region-indigenous interactions are draw from their prior distributions. We compare the regional variation in age-patterns for these replicate datasets regional variation in the actual dataset. If our model is capturing regional variation adequately, then the partly synthetic data that we have generated should look similar to the actual data. If the model is not working well, then the replicate data and actual data should look different.

Our ultimate motivation for estimating mortality rates is to obtain reigonal life tables for indigenous people. An attractive feature of Bayesian methods in general is that obtaining estimates for derived quantities, such as life tables, is easy—including generating measures of uncertainty.

We carry out the estimation using our own open source *R* packages **dembase** and **demest**.

3 Results

Figure 2 shows modelled estimates of mortality rates in 2016. The vertical bars are 95% credible intervals: under the assumptions of the model, there is a 95% probability that the true mortality rate lies within the associated interval. The modelled estimates are substantially more regular and interpretable than the original direct estimates from Figure 1. Close inspection indicates that mortality rates rates for indigenous people do indeed vary across states and territories, even after smoothing away the random variation.

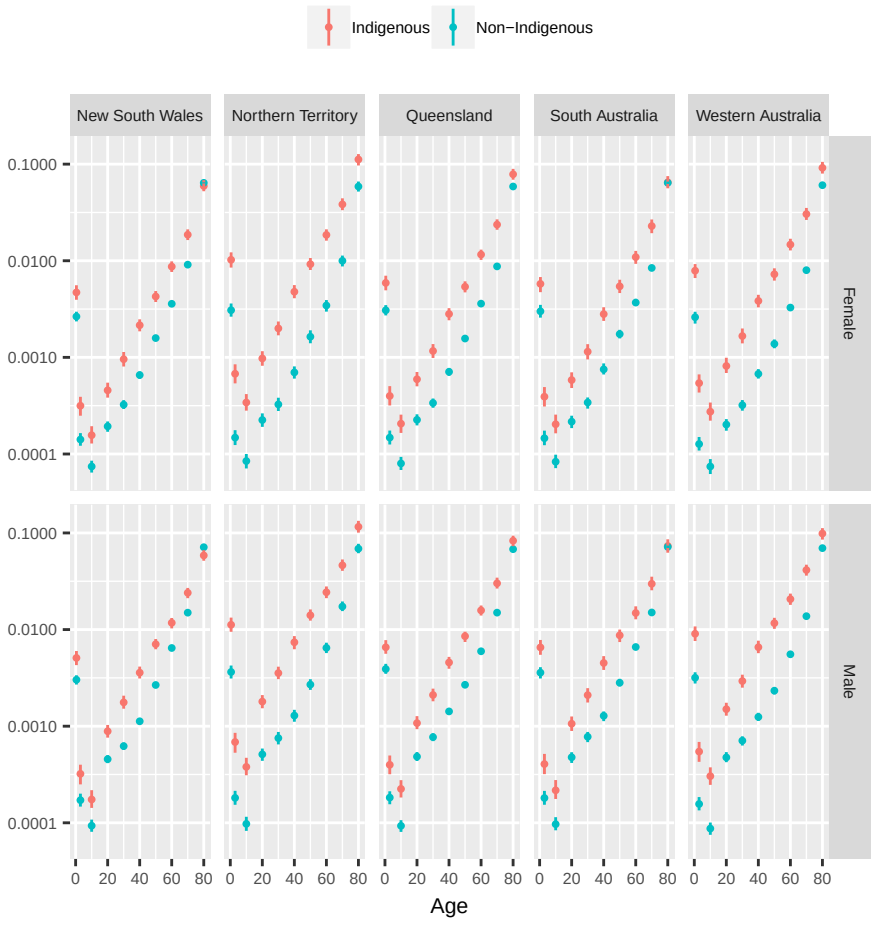

Figure 2: Modelled estimates of mortality rates, 2016, on a log scale. The vertical lines represent 95% credible intervals.

The results shown in Figure 3 suggest that our model is able to generate realistic geographic variability in age-profiles. Replicate data generated from

the model look much the same as actual data.

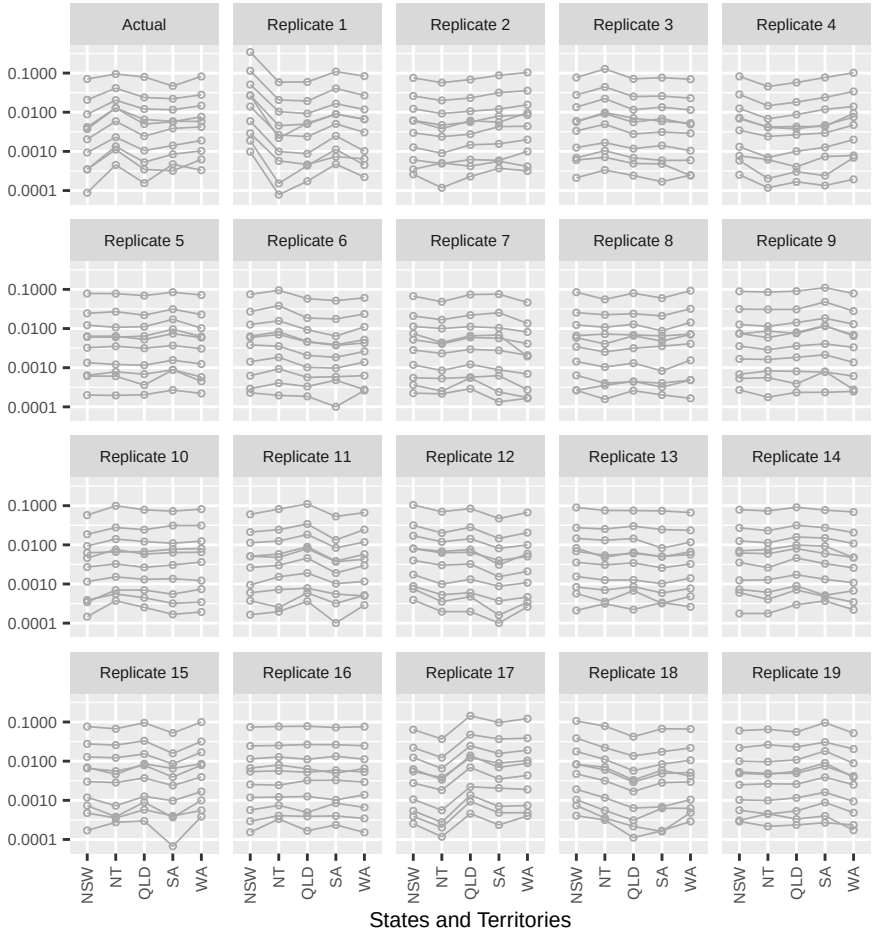

Figure 3: Direct estimates of mortality rates by age and state or territory for indigenous females: the actual dataset and 19 replicate datasets. Each line represents a different age group. Deaths are aggregated over the whole period 2010–2016.

Based on the modelled mortality rates, we generate life tables for each combination of sex, indigenous status, region, and year. Life expectancies at birth from these life tables are shown in Figure 4, along with 95% credible intervals. Although indigenous life expectancies are lower than non-indigenous ones in all states and territories, the difference is particularly pronounced in the Northern Territory.

Figure 4: Estimates of life expectancy, with 95% credible intervals.

4 Discussion

Between now and the time of the workshop, we expect to considerably expand on the work shown here. With more disaggregated data, we will use single years of age, higher maximum ages, and smaller geographical units. Modelling becomes more difficult as the data become sparser, requiring greater attention to the specific form of the priors. We also intend to carry out much more model checking and sensitivity analyses, including more use of replicate data. One particular form of sensitivity analysis will be estimating models with indigenous people only, without any form of pooling with the rest of the population.

The methods and software we are developing are very general, in that they place no restrictions on the dimensions that are included in the model, and allow for a wide choice of priors. They can be also be used for mortality forecasting, using exactly the same specifictions that are used for estimation.

References

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041.

Bijak, J. and Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population studies*, 70(1):1–19.

Bryant, J. and Zhang, J. L. (2018). *Bayesian Demographic Estimation and Forecasting*. CRC Press.

Congdon, P. (2010). *Applied Bayesian Hierarchical Models*. CRC Press, Boca Raton.

Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*. CRC Press, Boca Raton.

Statistics New Zealand (2015). New Zealand period life tables: Methodology for 2012–14.