Noname manuscript No.
(will be inserted by the editor)

Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records

Carl P. Schmertmann[1] · **Marcos R. Gonzaga**[2]

[1]Florida State University, Tallahassee FL USA
[2]Universidade Federal do Rio Grande do Norte, Natal RN Brazil

February 2, 2018

Abstract High sampling variability complicates estimation of demographic rates in small areas. In addition, many countries have imperfect vital registration systems, with coverage quality that varies significantly between regions. We develop a Bayesian regression model for small-area mortality schedules that simultaneously addresses the problems of small local samples and under-reporting of deaths. We combine a relational model for mortality schedules with probabilistic prior information on death registration coverage – derived from demographic estimation techniques such as Death Distribution Methods, and from field audits done by public health experts. We test the model on small-area data from Brazil. Incorporating external estimates of vital registration coverage though priors improves small-area mortality estimates by accounting for under-registration, and by automatically producing measures of uncertainty. Bayesian estimates show that when comparing mortality levels in small areas, noise often dominates signal. Differences in local point estimates

Carl P. Schmertmann
Center for Demography and Population Health, Florida State University, Tallahassee FL USA 32306-2240
Tel.: +1-850-644-7100
Fax: +1-850-644-8818
E-mail: schmertmann@fsu.edu

Marcos R. Gonzaga
Universidade Federal do Rio Grande do Norte, Natal, Brazil

of life expectancy are often small relative to uncertainty, even for relatively large areas in a populous country like Brazil.

1 Introduction

Small-area mortality estimation is a challenge for demographers and public health researchers, for two main reasons. First, there is the universal problem of small populations and high sampling variability in recorded deaths (Riggan et al. 1991; Bernardinelli and Montomoli 1992; Pletcher 1999). With low mortality rates and short periods of exposure, observed event/exposure ratios are very unstable and estimation of mortality patterns is difficult. In such situations models must fill the gap: smoothing procedures that use known, robust patterns in rate schedules must be combined with available data (e.g. Brass 1971; Wilmoth et al. 2012).

Second, in many countries or regions, vital registration is incomplete and some deaths go unrecorded in official statistics (Mathers et al. 2005; Frias et al. 2013). Demographers have proposed a variety of methods to estimate the completeness of death registration and adjust mortality estimates accordingly (Brass and others 1975; Preston et al. 1980; Preston and Hill 1980; Bennett and Horiuchi 1981, 1984; Hill 1987, 2007; Hill et al. 2009; Queiroz et al. 2013, 2017). However, most methods depend on approximate stability of the population's sex and age composition (Bhat 2002; Murray et al. 2010), an assumption that is not met in countries that have experienced recent, rapid demographic transitions. Migration between subnational areas can also make the methods' stability assumptions unlikely (Bhat 2002; Hill and Queiroz 2010; Bignami-Van Assche 2005). Finally, standard methods cannot provide uncertainty measures about the completeness of death records (Murray et al. 2010).

Demographers and statistical epidemiologists have improved models of small-area mortality schedules significantly in recent years. New approaches based on Bayesian models that "borrow strength" over different dimensions (age, time, space, and/or sex) allow estimation of life expectancies and mortality rates in regions with sparse, high-quality data (Congdon 2009; Ocaña Riola and Mayoral-Cortés 2010; Jonker et al. 2012; Stephens et al. 2013; Tsimbos et al. 2014; Alexander et al. 2017).

Although it is not directly related to small-area estimation, there is a growing literature on the use of Bayesian approaches in international demographic forecasting (e.g. Raftery et al. 2013, 2014; Gerland et al. 2014; Ševčíková et al.

2016) and in methods for constructing coherent sets of estimates over large numbers of administrative units (e.g. Alkema et al. 2011, 2013; You et al. 2015). The methods used in these studies have important features in common with our approach – in particular, partial pooling of information across related observational units, and incorporating information from external sources through prior distributions for parameters.

Statisticians have also addressed estimation in cases of incomplete or underreported data. Raftery (1988) proposed a Bayesian approach to the general problem of inferring the number of binomial trials from the number of successes. Moreno and Girón (1998) developed a model for estimating crime rates from imperfectly reported data, which is closely related to the model for mortality that we propose in this paper. Very recently, de Oliveira et al. (2017) have analyzed Brazilian infant mortality data with a model in which death reports may be censored in some geographic regions.

In this paper we propose a Bayesian regression model that specifically addresses the fundamental problems in small-area mortality estimation in countries with potentially defective registration. The model smooths age-specific mortality rates in small samples, while also accounting for uncertainty about the level of death registration. Bayesian regression produces estimates of small-area mortality rates and life expectancies, and of the uncertainty in those estimates.

Our model offers several advantages over existing non-Bayesian approaches to estimating age-specific mortality schedules in small areas with vital registration errors. It incorporates two sources of uncertainty: sampling variability and uncertainty about registration coverage. In addition, it uses a novel functional form for mortality schedules that stabilizes small-sample estimates without requiring strong assumptions about age-specific mortality patterns. Finally, estimated rate schedules from the model are continuous, smooth functions – unlike many existing corrections for under-registration that require different methods for infant, child, and adult deaths and may consequently have discontinuities in the estimated rate schedule.

Prior distributions for death coverage can include any available empirical information related to local data quality. In principle, prior information could include expert opinion, demographic estimates, or both. In our specific application to Brazilian data, prior information on age- and sex-specific vital registration comes mainly from standard demographic estimates, and from an intensive field audit conducted by public health researchers.

2 Modeling Strategy

2.1 A simple statistical model for deaths and registered deaths

For each age x within an area of interest, we observe exposure N_x and *registered* deaths R_x. The *true* number of deaths $D_x \geq R_x$ is not observed. We assume that true deaths have independent Poisson distributions at each age that depend on age-specific mortality rates μ_x:

$$D_x \sim Poisson(N_x \mu_x) \tag{1}$$

We further assume that each death is registered independently with an age-specific probability π_x, so that the total number of registered deaths at age x has a binomial distribution:

$$R_x \sim Binomial(D_x, \pi_x) \tag{2}$$

As shown in the Appendix, the distribution of registered deaths R_x implied by (1) and (2) is[1]

$$R_x \sim Poisson(N_x \mu_x \pi_x) \tag{3}$$

2.2 Identifiability of mortality rates

A distribution $R \sim Poisson(N\mu\pi)$ for registered deaths implies that the mortality rate is not identifiable from data on R and N, because all (μ, π) pairs that have the same product will have identical likelihoods $L(R|N, \mu, \pi) \propto e^{-N\mu\pi} (N\mu\pi)^R$. In other words, from the likelihood alone one cannot distinguish between situations of (high mortality, low registration) and situations of (low mortality, high registration).

In a classical, frequentist approach to mortality estimation this lack of identifiability is fatal. Unless the coverage probability π is known, there is no unique μ that maximizes the likelihood, and it is impossible to estimate

[1] The Appendix also demonstrates that a negative binomial distribution for total deaths implies a negative binomial distribution for registered deaths. A negative binomial model would be appropriate if the data exhibits *overdispersion* – i.e, higher variance than predicted by a Poisson model. With the Brazilian data that we use in this paper, extensive experimentation produced no evidence of overdispersion, and posterior distributions of mortality rates and life expectancies were virtually identical with Poisson and Negative Binomial specifications. We therefore use a standard Poisson distribution for D.

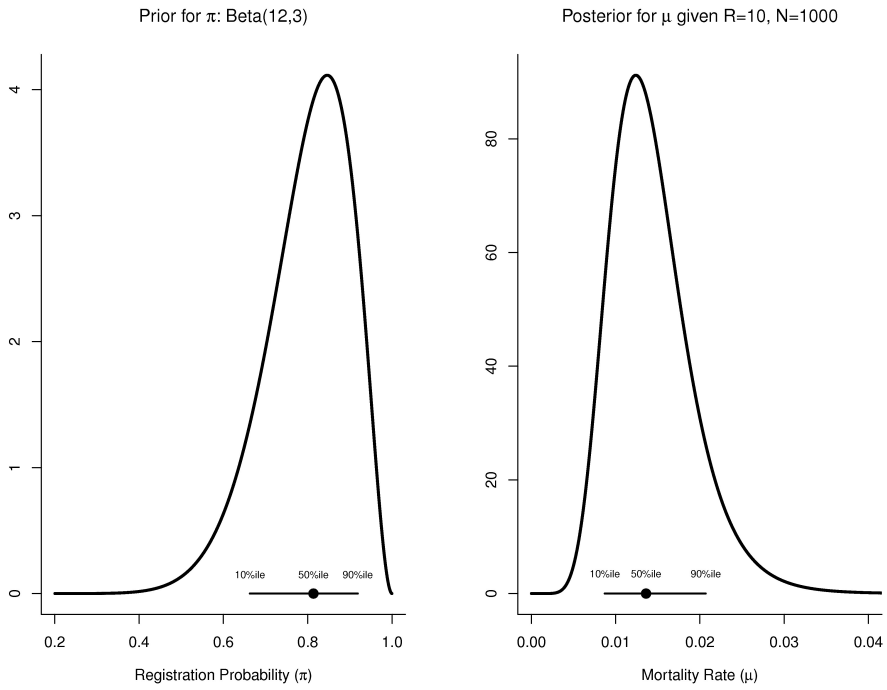


Fig. 1 Prior distribution for vital registration coverage; Posterior distribution of the mortality rate

the mortality rate from R and N. In contrast, a Bayesian approach allows the analyst to use probabilistic information about which coverage probabilities are more likely (expressed as a prior distribution $\pi \sim f_\pi$) to produce probabilistic statements about mortality rates μ, given R and N.

As an intuitive example, consider an age group in which we observe $N = 1000$ person-years of exposure and $R = 10$ registered deaths. Suppose that a small field audit in this location, conducted five years earlier, found that 12 out of 15 total deaths had been registered. From the field audit information it is reasonable to use $\pi \sim \ Beta(12, 3)$ as an expression of our prior knowledge about local death registration (Lynch 2007, pp 54–57). As seen in the left panel of Figure 1, this prior implies that the expected registration probability is $\pi = 0.80$, that there is a 80% probability that coverage is in the [0.66,0.92] range indicated by the solid bar, and so forth.

In a Bayesian approach, the mortality rate μ is treated as an uncertain value with a statistical distribution, and the prior for π adds probabilistic

information about coverage that allows us to infer which mortality rates are more and less likely, given exposure and a count of registered deaths. Based on the field audit information in our example, $\pi = 0.8$ is a very likely coverage level and $\pi = 0.4$ is very unlikely. Therefore, *a posteriori* (i.e. after observing $R = 10$ and $N = 1000$) one can say that $\mu = \frac{(10/1000)}{0.8} = .0125$ is a very plausible mortality rate, while $\mu = \frac{(10/1000)}{0.4} = .0250$ is much less plausible. The full posterior distribution combines the likelihood for R with the prior for π, producing a posterior distribution for μ that summarizes which mortality rates are more and which are less plausible given the observed data[2]:

$$P\left(\mu|R,N\right) = \int_0^1 L\left(R|N,\mu,\pi\right)\ f_\pi(\pi)\, d\pi \tag{4}$$

In our specific example, with likelihood $R \sim Poisson(N\mu\pi)$, prior $\pi \sim Beta(12,3)$, and data $(R,N) = (10,1000)$, this distribution is

$$P\left(\mu|\ R = 10, N = 1000\right) \propto \int_0^1 e^{-1000\mu\pi}(\mu\pi)^{10}\,\pi^{11}(1-\pi)^2\, d\pi \tag{5}$$

which appears in the right panel of Figure 1: the posterior median of the mortality rate is .014, and an 80% credible interval (10-90%ile) is [.009, .021].

In the full model, we use a parametric, relational system for log mortality rate schedules, as described in the next section. However, the main principle is illustrated by this simple example: we combine probabilistic prior knowledge about death registration with a statistical model that relates registered deaths, coverage, and exposure. The result is an *a posteriori* distribution for local mortality rates.

2.3 TOPALS relational model for mortality schedules

We model mortality by age with the TOPALS relational model (de Beer 2012; Gonzaga and Schmertmann 2016). In a TOPALS model the log mortality schedule is a sum of two functions: (1) a constant schedule (called the *standard*) that reflects basic age patterns, and (2) a parameterized, piecewise-linear function made up of straight-line segments between designated ages (called *knots*)

[2] For simplicity we leave the prior distribution of μ implicit in this introduction. By omitting an explicit prior we assume *a priori* that μ is equally likely to take any positive real value. The omitted (improper) prior is therefore $f_\mu(\mu) \propto I(\mu \geq 0)$, where $I()$ is a (0,1) indicator function. This yields a proper posterior distribution for (μ,π) and a proper marginal posterior for μ in equation (4).

that represents differences between the standard and the mortality schedule in the population of interest.

The vector of log rates over ages $x = 0 \dots 99$, $\boldsymbol{\lambda} \in \mathbb{R}^{100}$ in the TOPALS model is

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}^* + \mathbf{B}\,\boldsymbol{\alpha}$$

where $\boldsymbol{\lambda}^* \in \mathbb{R}^{100}$ is the standard schedule of log mortality rates (in our case, derived from national data for Brazil in 2010), $\mathbf{B}$ is a 100x7 matrix of fixed B-spline linear basis functions (de Boor 2001), and $\boldsymbol{\alpha} \in \mathbb{R}^7$ is a parameter vector.

The seven model parameters $\boldsymbol{\alpha} = (\alpha_0 \dots \alpha_6)'$ are the values of the piecewise-linear function at exact ages 0,1,10,20,40,70, and 100. For example, $\lambda_{40} = \lambda^*_{40} + \alpha_4$ and $\lambda_{70} = \lambda^*_{70} + \alpha_5$. Between knots the additive offsets to the standard schedule change linearly with age – for example, $\lambda_{50} = \lambda^*_{50} + \frac{2}{3}\alpha_4 + \frac{1}{3}\alpha_5$, $\lambda_{55} = \lambda^*_{55} + \frac{1}{2}\alpha_4 + \frac{1}{2}\alpha_5$, and $\lambda_{60} = \lambda^*_{60} + \frac{1}{3}\alpha_4 + \frac{2}{3}\alpha_5$. The space of possible mortality schedules in a TOPALS model is thus the set of curves that can be constructed by adding piecewise-linear functions to the standard log-rate schedule.[3]

The mortality rate at age x in a TOPALS model is

$$\mu_x(\boldsymbol{\alpha}) = \exp(\lambda^*_x + \mathbf{b}'_x\boldsymbol{\alpha})$$

where $\mathbf{b}'_x$ is the xth row of $\mathbf{B}$. Under the distributional assumptions outlined above, the log likelihood is

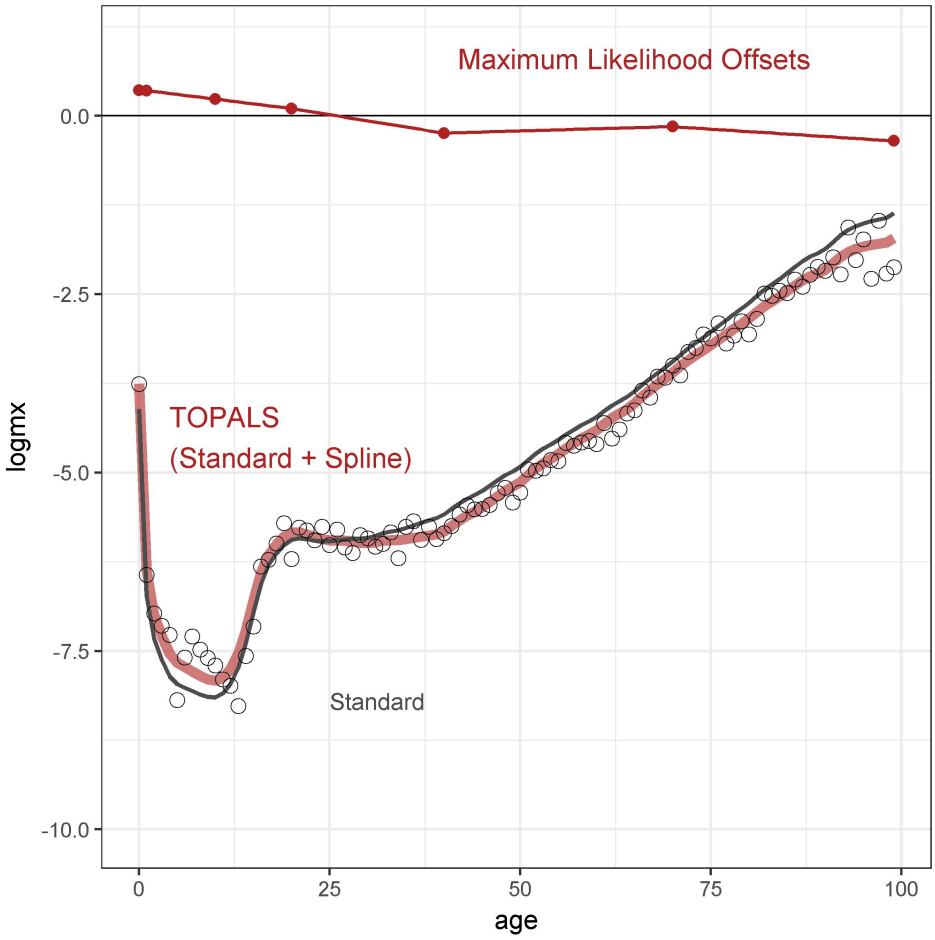
$$\ln L(\mathbf{R}\,|\,\mathbf{N},\,\boldsymbol{\alpha},\,\boldsymbol{\pi}) \;=\; c - \sum_x [N_x\pi_x \cdot \mu_x(\boldsymbol{\alpha})] \;+\; \sum_x [R_x \ln \mu_x(\boldsymbol{\alpha})] \qquad (6)$$

where c is a constant that does not vary with $\boldsymbol{\alpha}$.

Figure 2 illustrates a fitted TOPALS model, for males in the northern Brazilian state of Amapá in 2010. The figure shows a national standard $\boldsymbol{\lambda}^*$ for males (thin black line), observed $\ln(R_x/N_x)$ values from registered deaths and exposure in Amapá (open circles), the seven TOPALS offset parameters $\alpha_0 \dots \alpha_6$ (the vertical positions of the solid red dots), the linear spline offsets

[3] Gonzaga and Schmertmann (2016) show that this property makes the specific choice of a standard schedule $\boldsymbol{\lambda}^*$ far less important than in other relational models used in demography. Note also that this TOPALS model includes indirect standardization as a special case – in which all $\boldsymbol{\alpha}$ values are equal and the standard schedule is shifted up or down by the same amount at all ages.

Fig. 2 TOPALS model for males in Amapá state 2010. Open circles are observed $\ln(R_x/N_x)$ from registered deaths. Smooth dark line is national standard log mortality schedule. Heights of solid dots are maximum likelihood offsets $\alpha_0 \ldots \alpha_6$. Fitted TOPALS schedule is the sum of the standard schedule and the linear spline.



$\mathbf{B}\boldsymbol{\alpha}$ (thin red line connecting the offsets), and the fitted TOPALS model schedule (standard+spline, $\boldsymbol{\lambda}^* + \mathbf{B}\boldsymbol{\alpha}$, thick red line). In broad terms this fit suggests that Amapá's male mortality is higher than the standard at ages below 25, and lower at ages above 25. More subtly, the fit suggests that downward deviations from the standard are slightly larger at higher ages.

This particular fit for Amapá maximizes equation (6) over $\boldsymbol{\alpha}$ under the assumption of 100% death registration ($\pi_x = 1$, $R_x = D_x$) at all ages. It serves only to illustrate the components of a parametric TOPALS model. In

practice we relax the assumption of complete coverage, replacing it with prior distributions for π_x.

2.4 Model summary

Figure 3 summarizes our statistical approach to estimating mortality and life expectancy in a small area. We use information from other sources to develop priors for age-specific coverage. (In the hypothetical example above, for instance, a field audit suggested a beta distribution with $a = 12$ and $b = 3$). As indicated in the top right of Figure 3, we use a weak multivariate normal prior on $\boldsymbol{\alpha}$ (described in detail in section 4) to stabilize schedule estimates in very small populations.

Age-specific exposure $N_0 \ldots N_{99}$ and registered deaths $R_0 \ldots R_{99}$ are observed. The model combines the prior distributions for coverage and mortality parameters (f_π and f_α) with the Poisson likelihood $L(\mathbf{R}|\mathbf{N}, \boldsymbol{\alpha}, \boldsymbol{\pi}) = \prod_x L(R_x|N_x, \mu_x(\boldsymbol{\alpha}), \pi_x)$ to produce a posterior marginal distribution for $\boldsymbol{\alpha} = (\alpha_0 \ldots \alpha_6)'$, similar to Equation (4):

$$P(\boldsymbol{\alpha}|\mathbf{R}, \mathbf{N}) = \int L(\mathbf{R}|\mathbf{N}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \; f_\alpha(\boldsymbol{\alpha}) \; f_\pi(\boldsymbol{\pi}) \; d\boldsymbol{\pi} \tag{7}$$

In English, the left-hand side of (7) answers the question *"Given age-specific populations and registered deaths, combined with our uncertain knowledge of local death registration probabilities, which local mortality schedules are more and which are less plausible?"*. In practice, we answer by drawing a large number of random realizations $\boldsymbol{\alpha}_1^* \ldots \boldsymbol{\alpha}_K^*$ from the distribution in Equation (7) via Markov Chain Monte Carlo (MCMC) simulation. It is then an easy logical step to ask the same question about more and less plausible life expectancies e_0, via the simulated posterior distribution of $e_0(\boldsymbol{\alpha}_1^*) \ldots e_0(\boldsymbol{\alpha}_K^*)$

$$P(e_0 < c \,|\, \mathbf{R}, \mathbf{N}) \;\approx \frac{1}{K} \sum_k I\left[e_0(\boldsymbol{\alpha}_k^*) \;<\; c\right] \tag{8}$$

where $I[\,]$ is a (0,1) indicator function equal to 1 if the condition in brackets is true.

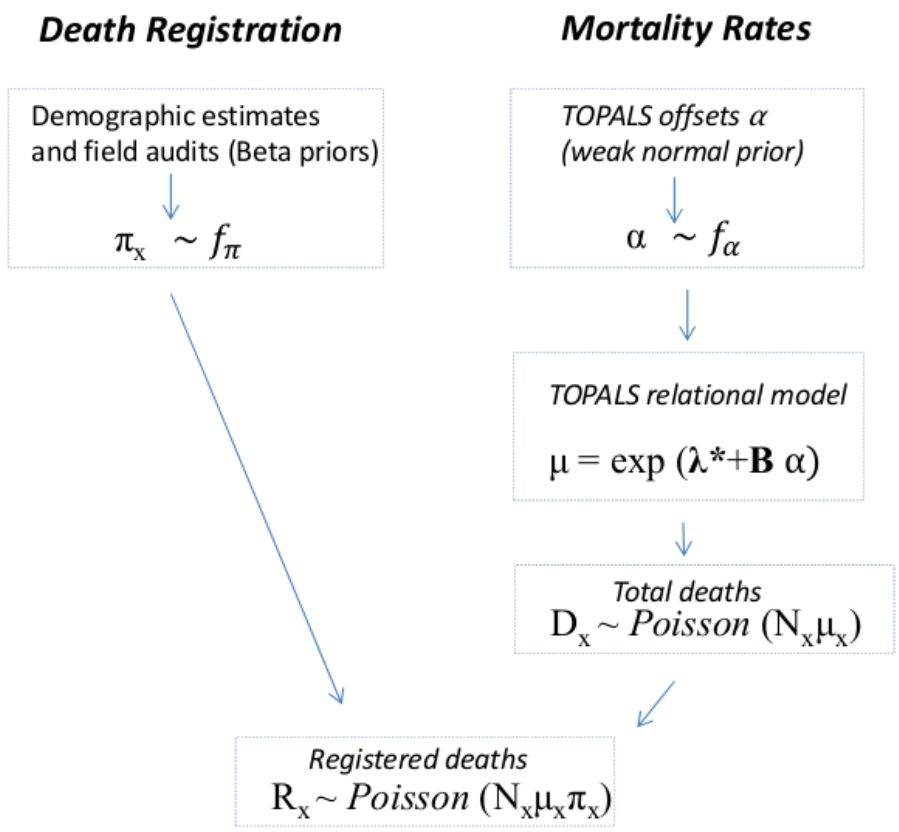


Fig. 3 An integrated coverage and mortality model. Registered deaths R_x and exposure N_x are observed. $\boldsymbol{\alpha}$ and π_x are uncertain parameters.

3 Data

We apply the model to small-area estimation in Brazil. Brazil has good census data, but incomplete vital registration. Death reporting has improved significantly over the past several decades, but there are still large regional differences in the quality of coverage (Paes 2005; de Mello Jorge et al. 2007; Instituto Brasileiro de Geografia e Estatística 2013; Queiroz et al. 2017). There are increasing demands for fine-grained mortality (and especially life expectancy) estimates, but the complications of small samples and differential coverage make this a difficult task. In short, Brazil is a good test case for a model that incorporates under-registration into small-area mortality estimation.

Population and deaths by sex and single-year age come from the Brazilian Demographic Census (2010) and from the Mortality Information System of the Ministry of Health (MS/SVS/CGIAE), respectively. The Mortality Information System is an official registry for deaths, with data collected from health providers(de Mello Jorge et al. 2007).

Population and registered deaths are available for 5565 Brazilian municipalities[4], 100 single-year ages $0 \ldots 99$, and 2 sexes. For each of the 1,113,000 combinations of (municipality, age, sex) we recorded the 2010 census population, the number of registered deaths over calendar years 2009–2011, and geographic identifiers. We used the 2010 census populations to estimate age- and sex-specific exposure over 2009–2011 for each (municipality, age, sex) combination. Details are in Gonzaga and Schmertmann (2016)

Brazilian census geography begins with 5 major regions (North, Northeast, Southeast, South, and Center-West) comprising 27 *states*[5] that cover the entire national territory. States are subdivided into *mesoregions* (137 total), mesoregions into *microregions* (558), and microregions into *municipalities* (5565). Our goal in this analysis is to estimate age- and sex-specific mortality schedules and life expectancies for all 27 states, and for each of the 558 microregions.[6]

4 Priors

4.1 Using death registration estimates from previous studies

Prior information about the likely levels and age patterns of death registration is essential for our model. In the case of Brazil, we identified six published estimates of death registration coverage, by state and sex, for 2010 (Queiroz 2012; Queiroz et al. 2013; Freire et al. 2005; Instituto Brasileiro de Geografia e Estatística 2013; Queiroz et al. 2017). These estimates all come from Death Distribution Methods (DDM), and refer mainly to deaths at ages 30+. The six estimates differ substantially in some cases. For example, for males 30+

[4] In Brazil, municipalities are the smallest areas responsible for registering vital events.

[5] For simplicity we call the Federal District that contains Brasília a *state*.

[6] It is important to note that even microregions are fairly large "small areas". With a single exception (the remote island of Fernando de Noronha had a total resident population of only 2630 in 2010), all had resident populations of at least twenty thousand in 2010. Rounded to the nearest thousand, the 10th , 50th , and 90th percentiles of microregional population were 63, 173, and 557 thousand, respectively. The largest microregion, metropolitan São Paulo, had a 2010 population of over 13 million.

in the Northeastern state of Maranhão (widely regarded as the state with the lowest vital registration coverage) the alternative coverage estimates were 50, 66, 75, 78, 78, and 97%. The range of these estimates suggests that in addition to being low, coverage levels for adult male deaths in Maranhão are uncertain.

At the *municipal* level, we use published estimates of death registration coverage from a research project known as *busca ativa*[7], which we translate as *field audit*. The field audit project (Szwarcwald et al. 2010; Frias et al. 2013) randomly selected 133 municipalities in Brazil's poorest regions, and compared the registered deaths in 2009–2010 with total deaths found for the same period from an exhaustive search of notary offices, clinics, official and unofficial cemeteries, and from interviews with health workers, midwives, funeral homes, pharmacies, and others. Based on correlations between municipal-level characteristics and levels of mortality coverage in the study areas, the field audit researchers estimated the likely probability of death registration in all 5565 Brazilian municipalities – for infant deaths (π_0) and for all deaths (which we call π_{all}). Field audit estimates can be aggregated to higher levels of geography – in particular, to microregion or state level – by using published information from the project on the estimated numbers of deaths in each municipality.

When aggregated to the 27 Brazilian states, field audit estimates of infant mortality coverage π_0 range from 65% (Maranhão) to 100% in several southern states. The range is naturally wider across the 558 microregions -- from 31% in some remote Amazonian locations to 100% in many southern microregions. Field audit estimates for death registration probabilities at all ages combined π_{all} range from 78 to 100% across states, and from 44 to 100% across microregions.

4.2 Prior Distributions for Coverage

To utilize the available coverage data in Brazil, we assume that in each small area there are distinct probabilities of death registration for three age intervals:

$$\pi_x = \begin{cases} \pi_0 & \text{if } x = 0 \\ \pi_1 & \text{if } x \in \{1 \dots 29\} \\ \pi_2 & \text{if } x \in \{30 \dots 99\} \end{cases} \tag{9}$$

[7] The complete name in Portuguese is *Busca ativa de óbitos e nascimentos no Nordeste e na Amazônia Legal* [Active search for deaths and births in the Northeast and the Amazonian administrative region]

that can be combined to produce an overall probability of registration

$$\pi_{\text{all}} = \; w_0\,\pi_0 + \; w_1\,\pi_1 + \; w_2\,\pi_2 \tag{10}$$

where w_x terms represent the proportion of all deaths that occur in the corresponding age group.[8]

Our choice of age groups was based in part on the availability of age ranges in the external coverage estimates – for example, DDM provides straightforward estimates of state-level π_2. However, our choice of age groups is also based on the demographic literature and on expert opinions. First, many Brazilian studies indicate that coverage of infant deaths π_0 is lower than coverage in any other age interval (Paes and Albuquerque 1999; Instituto Brasileiro de Geografia e Estatística 2013; Frias et al. 2013). This fits the pattern of other countries with defective vital records, in which infant deaths generally have the lowest coverage rates (Målqvist et al. 2008). Field audits also found high percentages of unreported infant deaths (Szwarcwald et al. 2010; Frias et al. 2013). Second, some Brazilian researchers suggest that external causes of death (such as homicides and transit accidents) have almost complete coverage in all Brazilian regions (Campos and Rodrigues 2004; Agostinho 2009). This is plausible, because deaths by violence and transit accident must be reported to the local health department not only by family or relatives, but also by the municipal coroner's office. Although the quality of information varies between regions, the notification procedures are determined by national law and the path of the death certificate from coroner's office to the Mortality Information System is identical in all Brazilian states (Borges et al. 2012). Deaths from external causes in Brazil have increased in the last decades, and these deaths are concentrated in the young adult age interval (de Mello Jorge et al. 1997; Souza et al. 2007; Matos et al. 2013). Taken together, this evidence suggests that coverage of infant deaths should not he higher than coverage of deaths at ages 1–29 : $\pi_0 \leq \pi_1$.

For infant deaths π_0, and for all deaths π_{all}, we build priors from field audit estimates. For every area, we aggregate the corresponding municipal field audit estimates, weighted by deaths, to calculate estimated coverage levels $\hat{\pi}_0$ and

[8] In practice we used identical weights for each region: $w = (.035, .109, .856)$ for males and $w = (.037, .047, .916)$ for females. These were calculated from national deaths over 2009-2011.

$\hat{\pi}_{all}$. The associated priors are

$$\pi_0 \sim Beta(\, K_0\, \hat{\pi}_0 \;,\; K_0\, [1-\hat{\pi}_0]\,) \qquad , \quad (K_0 - 5) \sim \text{exponential}(0.05) \tag{11}$$

$$\pi_{all} \sim Beta(\, K_{all}\, \hat{\pi}_{all} \;,\; K_{all}\, [1-\hat{\pi}_{all}])\,, \; (K_{all} - 5) \sim \text{exponential}(0.05) \tag{12}$$

where K_0 and K_{all} are hyperparameters representing (uncertain) levels of precision from the field audit estimates.[9]

For mortality coverage at ages 30+, which we denote π_2, external information consists of state-level estimates from the six DDM studies. We use the mean and variance of the DDM estimates to estimate the parameters of a beta distribution by the method of moments. The resulting prior is

$$\pi_{2,STATE} \sim Beta\,(\, K_2\, \phi_2 \;,\; K_2\, [1-\phi_2]\,) \tag{13}$$

where ϕ_2 is the mean of the six DDM estimates and K_2 is an estimated precision index.[10] For example, for Maranhão males the DDM estimates are (.50, .66, .75, .78, .78, .97), so we use $\pi_{2,Maranhão} \sim Beta(7.00 \times 0.74 \;,\; 7.00 \times 0.26)$. This prior distribution answers the question *"Given the DDM estimates, how plausible are different levels of coverage for males 30 and older in Maranhão?"* The answer is probabilistic: *a priori*, there is a 10% probability that the coverage level is below .52, a 50% probability that it is between .64 and .86 (the interquartile range), a 10% chance that it is above .92, and so on.[11]

Prior information for π_2 is at the state level, so when estimating coverage for substate areas like microregions the prior applies to the weighted registration probability

$$\pi_{2,STATE} \;=\; \gamma_a \pi_{2a} + \cdots + \gamma_z \pi_{2z} \sim Beta(\, K_2\, \phi_2 \;,\; K_2\, [1-\phi_2]\,) \tag{14}$$

[9] Hyperparameters K correspond to sample sizes in a field audit. Prior uncertainty about K represents uncertainty about the precision of the field audit estimates of π. Our (hyper)priors for K are fairly conservative: they imply that the most likely precision of the field audit estimates is equivalent to results from an audit slightly fewer than K=25 deaths in a region.

[10] Denoting the mean and variance of DDM estimates as $\bar{x}$ and s^2, the method of moments estimators (cf. Glen and Leemis 2017, pp. 227–228) are $\phi_2 = \bar{x}$ and $K_2 = \frac{\bar{x}(1-\bar{x})}{s^2} - 1$.

[11] Note that priors based on *busca ativa* estimates are constructed from a single coverage estimate for each region, by adding a hyperparameter for the estimate's unknown precision. In contrast, priors from DDM estimates are based on multiple estimates per region, and use the variance of those estimates as an index of (im)precision. A third alternative, which we do not use here, is to choose beta distribution parameters ϕ and K so that available estimates are all in a specified range of prior probability – for example, a 90% probability that $\pi \in [min(DDM), max(DDM)]$.

where γ_i is the proportion of state deaths at ages 30+ that occur in substate area $i \in \{a \dots z\}$ and π_{2i} is the death registration probability in area i.[12]

Finally, we use qualitative information from the literature, by adding a prior that completely rules out any triples of local coverage probabilities that do not match our assumed order:

$$\pi_0 \leq \pi_2 \leq \pi_1 \tag{15}$$

That is, we insist that in every local area infant mortality coverage cannot be higher than coverage at other ages, and that coverage of deaths at ages 1–29 cannot be lower than at other ages.

4.3 Prior distribution for TOPALS parameters

TOPALS parameters $\boldsymbol{\alpha} \in \mathbb{R}^7$ determine the shape and level of the age-specific mortality schedule. We use a vague prior for $\boldsymbol{\alpha}$, so that local death and exposure data are the primary determinants of rate estimates, via the likelihood function. Our prior for $\boldsymbol{\alpha}$ is multivariate normal distribution derived from two principles: (1) each α_i component should have very similar prior probabilities over a wide range of possible values, and (2) very large differences between consecutive components $\alpha_i - \alpha_{i-1}, i = 1 \dots 6$ are unlikely.

Based on those principles, we use the prior distribution $\alpha \sim N(\mathbf{0}, \Sigma)$ with covariance matrix[13]

$$\Sigma = \begin{bmatrix} 3.11 & 2.71 & 2.39 & 2.15 & 1.97 & 1.86 & 1.80 \\ 2.71 & 2.80 & 2.47 & 2.22 & 2.03 & 1.92 & 1.86 \\ 2.39 & 2.47 & 2.62 & 2.35 & 2.16 & 2.03 & 1.97 \\ 2.15 & 2.22 & 2.35 & 2.56 & 2.35 & 2.22 & 2.15 \\ 1.97 & 2.03 & 2.16 & 2.35 & 2.62 & 2.47 & 2.39 \\ 1.86 & 1.92 & 2.03 & 2.22 & 2.47 & 2.80 & 2.71 \\ 1.80 & 1.86 & 1.97 & 2.15 & 2.39 & 2.71 & 3.11 \end{bmatrix}.$$

The marginal priors for each α_i component are uninformative about levels, which are measured on the log-mortality rate scale. For example, the infant mortality offset $\alpha_0 \sim N(0, 3.11)$ *a priori*, so there is a 57% prior probability

[12] Because we have only state-level prior information about death registration in this age group, we can only assess the prior probability of a *set* of substate coverage levels $(\pi_{2a} \dots \pi_{2z})$, by looking at whether their weighted average is likely.

[13] This prior distribution arises from two lines in the *Stan* programming language. From our first principle (diffuse marginal distributions for each α_i) we add $\boldsymbol{\alpha} \sim$ *normal(0,4)* to the model. From the second principle (small differences between consecutive parameter values) we add $\alpha_i - \alpha_{i-1} \sim$ *normal(0,sqrt(0.5))* – as in Gonzaga and Schmertmann (2016). These statements in *Stan* both represent changes to the log prior density of any proposed $\boldsymbol{\alpha}$ vector, which together yield this specific multivariate normal distribution. The results that we report in this paper are extremely insensitive to the choice of priors for $\boldsymbol{\alpha}$.

that it falls in [-1.39,+1.39]. This corresponds to a very vague prior assumption about a region's infant mortality rate: it says that there is a 57% probability that infant mortality rates are between one-fourth and four times the rate in the standard schedule, and a 43% prior probability that rates might be even more extreme. The correlation structure in Σ is also only weakly informative about differences between α components; it serves mainly to stabilize estimates in extremely small populations with very few deaths, by giving slightly higher prior probabilities to simple up-and-down shifts in the standard schedule (which occur when $\alpha_0 = \alpha_1 = \cdots = \alpha_6$ and differences between consecutive αs are all zero).

5 Results

We implemented the full model in *Stan* (Carpenter et al. 2017), a language that allows MCMC sampling from complex posterior distributions. For both sexes, we estimated posterior distributions of mortality parameters $\boldsymbol{\alpha}$, complete log mortality schedules $\boldsymbol{\lambda}(\boldsymbol{\alpha}) = \boldsymbol{\lambda}^* + \mathbf{B}\boldsymbol{\alpha}$, and life expectancy $e_0[\boldsymbol{\lambda}(\boldsymbol{\alpha})]$ for all 27 Brazilian states and all 558 microregions. We call this the *adjusted* model. In order to learn about the effects of under-registration of deaths, we also estimated the same posterior distributions under the (incorrect) assumption of 100% registration ($\pi_0 = \pi_1 = \pi_2 = 1$). We call this second version the *unadjusted* model. We focus here on summary results for male life expectancy. Complete results, together with data and code, are available on the paper's companion website at `http://mortality-subregistration.schmert.net/`.

5.1 Effects of imperfect mortality coverage π on estimated state-level life expectancy

Figure 4 illustrates posterior distributions of male life expectancy for two states: the small state of Amapá (total population $\approx$ 0.6 million), and the Federal District that contains Brasília ($\approx$ 2.5 million). The figure includes posterior densities before any adjustment for under-registration (black lines on the right in each panel), and after adjustment (blue lines on the left).

Life expectancy is a complicated nonlinear function of the underlying $\boldsymbol{\alpha}$ parameters, but a Bayesian approach to estimation permits easy calculation of uncertainty in e_0 for both the unadjusted and adjusted estimates, via equation

(8). Figure 4 shows that even if death registration were complete (right-hand densities), there would still be considerable uncertainty about state-level life expectancy because of sampling variability. Sampling uncertainty is naturally greater for Amapá, which has a smaller population, but is non-trivial even for the Federal District, which has a male population of over one million.

Adjusting for under-registration lowers life expectancy estimates. The difference can be very large in areas where coverage is poor. A probabilistic model allows us to estimate the magnitude of these effects. In Amapá, for example, plausible corrections for under-registration of deaths reduce male e_0 by approximately three years, as illustrated by difference between the unadjusted and adjusted posterior medians in the top panel of Figure 4. The corresponding adjustment is very small for the Federal District, where death registration is nearly complete.

In addition to lowering the mean, uncertainty about vital registration coverage also increases uncertainty about life expectancy. This effect is visible for both states, but it is larger for the state with lower coverage levels (Amapá). As with the decrease in means, the fact that the variance will increase is qualitatively obvious. However, a carefully constructed probabilistic model allows demographers to estimate the increase in uncertainty caused by imperfect vital registration.

Figure 5 shows estimated 2010 male life expectancies for all 27 states, disaggregated by region. It reports the Bayesian model's death registration adjustments and distributional information using the abbreviated format on the horizontal axes of Figure 4. Figure 4's results for Amapá (AP) and the Federal District (DF) appear on the fourth line from the top and the eleventh line from the bottom, respectively, of Figure 5

For each state, Figure 5 includes the unadjusted male life expectancy (e_0) calculated directly from registered deaths (open circles) and the adjusted estimates and 80% posterior interval from our partial-coverage model (two-letter abbreviations and shaded bars). Horizontal distances between open circles and state abbreviations in Figure 5 illustrate median adjustments in life expectancy due to unregistered deaths.

Accounting for under-registration of deaths leads to large downward adjustments in life expectancy in the North and Northeast regions, and to small or negligible downward adjustments elsewhere. The most extreme adjustment is for the state of Maranhão (MA) in the Northeastern region. Direct use of death registration data would suggest that Maranhão has Brazil's highest

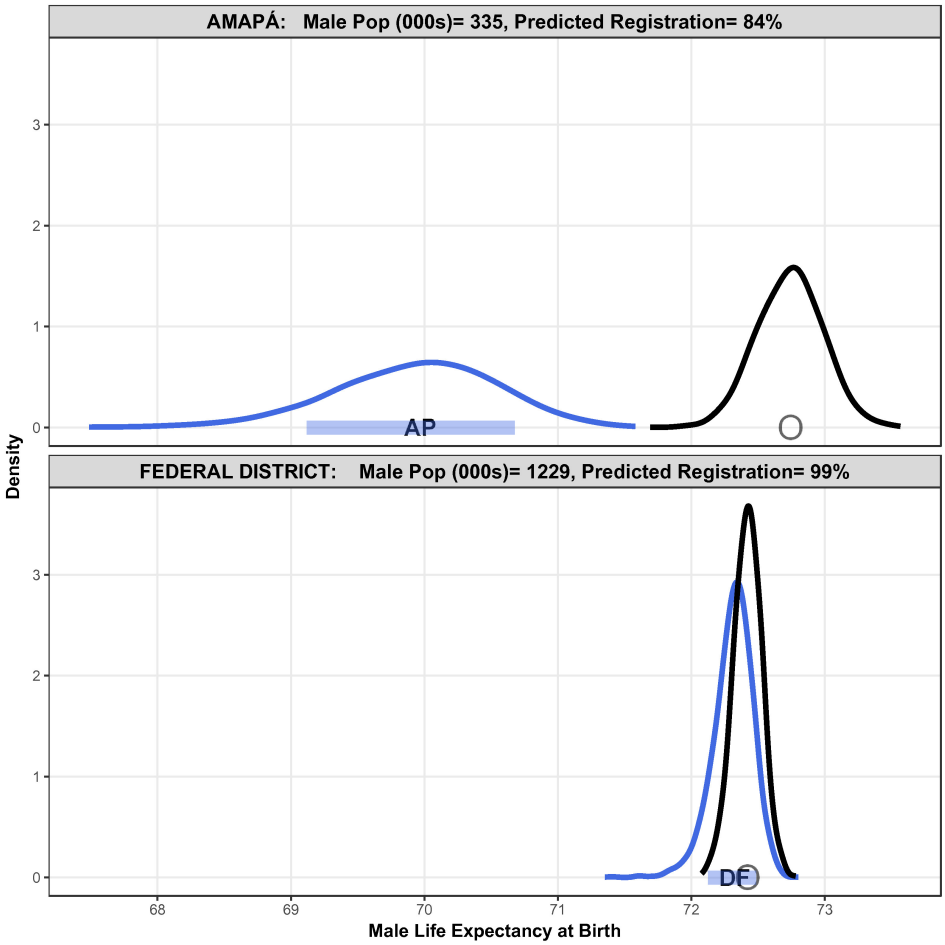


Fig. 4 Male 2010 Life Expectancy in two Brazilian states, before and after adjustment for under-registration of deaths. Black curves to the right represent the posterior density under the assumption of perfect death registration ($\pi = 1$ at all ages). Blue curves to the left include the priors for (π_0, π_1, π_2) developed in the text. Horizontal bars represent 80% posterior intervals (10–90%ile) for the adjusted distributions. Open circles at unadjusted medians; state abbreviations at adjusted posterior medians.

male life expectancy (74.3 years). In contrast, adjusted Bayesian estimates for Maranhão have a median of 70.4 years, almost four years lower. Northern states such as Pará (PA: 72.2 → 68.6 years), Amapá (AP: 72.7 → 70.0), and Amazonas (AM: 72.3 → 69.8) also require large downward adjustments in e_0 to account for likely levels of unregistered deaths .

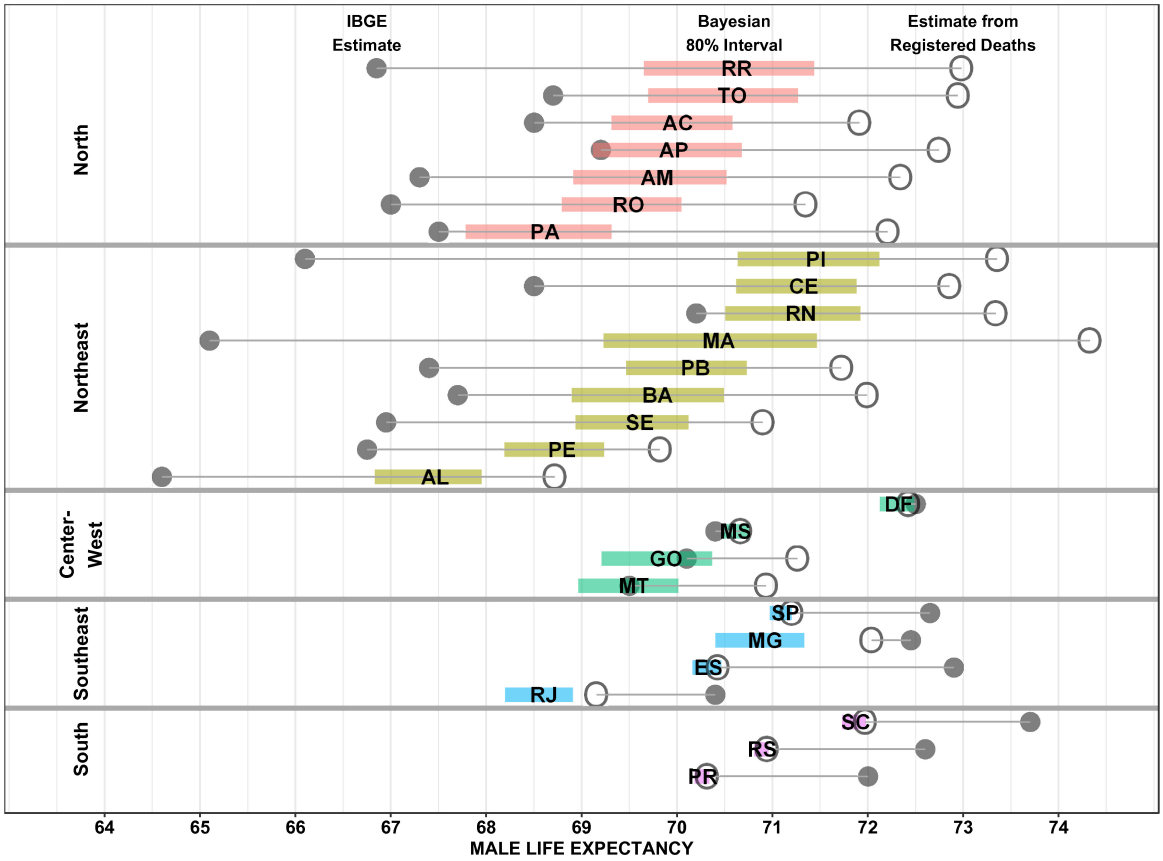


Fig. 5 Male life expectancy by state, Brazil 2010. Unadjusted estimates from deaths registered by the Mortality Information System (SIM) are open circles. Official state-level estimates from Brazil's statistical agency (IBGE) are solid circles. Shaded bars represent 80% posterior probability intervals after under-registration adjustment in the Bayesian coverage model; IBGE state abbreviations (listed at `http://schmert.net/BayesBrass/brazilian_regions_and_states.txt`) appear at posterior medians.

5.2 Comparison to Brazil's official state-level estimates

Bayesian posterior distributions for e_0 come from a range of plausible, state-specific registration coverage levels. They are therefore useful as comparative benchmarks for the official state-level estimates from IBGE, Brazil's census bureau. IBGE uses a complex, multistep procedure to correct for under-registration of deaths (Instituto Brasileiro de Geografia e Estatística 2013). Their procedure combines different indirect methods for mortality and coverage estimation, and differs by region. As a result of this complexity, researchers (including us) have been unable to replicate the official state-level estimates.

A comparison to adjusted and unadjusted means from vital registration data is therefore useful for understanding the official state-level estimates.

Filled circles in Figure 5 correspond to IBGE estimates for male e_0 in each state. Thus, when a filled circle falls to the left of the open circle the IBGE estimate is equivalent to assuming that deaths are underregistered. A filled circle to the right of an open circle is equivalent to assuming *over*-registration of deaths.

Figure 5 shows that official life expectancy estimates for Southern and Southeastern states (bottom panels) are implausibly high. Comparison to registered death data indicates that IBGE estimates are plausible only if the vital registration system substantially overcounts deaths in these states.[14]

Figure 5 also shows that the IBGE estimates for states in Brazil's North and Northeast regions are implausibly low. For almost all Northern and Northeastern states, the official e_0 estimate for males falls far below the 10%ile of the adjusted posterior distribution. In other words, likely combinations of vital registration coverage and mortality schedules in these states produce life expectancies much higher than the official estimates.

5.3 Signal and noise in state life expectancy estimates

Comparisons between areas are often important for public policy purposes, including allocation of health expenditures and other resources. Realistic estimates of the uncertainty of life expectancy highlight potential problems with allocating resources based on the most likely point estimates. This is true even for areas with large populations. As noted earlier, states are not especially small areas: the least populous Brazilian state (Roraima) has nearly half a million residents. Even so, after adjusting for likely under-registration of deaths, it can be difficult to determine which of a pair of states has the higher life expectancy.

Figure 6 shows 80% credible intervals for each state's male e_0 in the full model, including likely under-registration of deaths. As in the previous plot, the horizontal bars span the interval between the 10th and 90th percentile of the posterior distribution. Solid dots indicate the posterior medians. States are sorted in order of posterior medians, which we use as best-guess point

[14] The high estimates for Southern and Southeastern life expectancy probably result from IBGE's *compatibilization* step (Instituto Brasileiro de Geografia e Estatística 2013, Tables 6 and 13), in which they adjust national totals by removing deaths from these two regions.

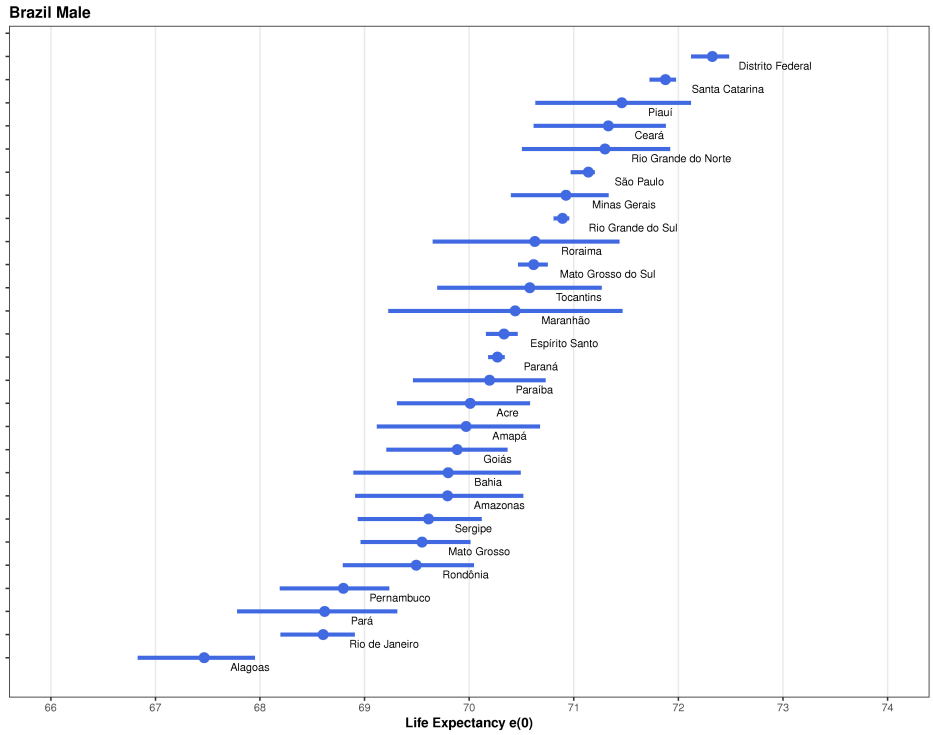


Fig. 6 Male Life Expectancy by state (medians and 80% posterior intervals), Brazil 2010

estimates. The Federal District has the highest estimate (72.3), Santa Catarina has the second-highest (71.9), and so forth. Some states, such as Rio Grande do Sul (8th line from the top, median=70.9) or Paraná (14th, 70.3) have very large populations and almost perfect vital registration. Consequently there is very high certainty about e_0 in those states. In contrast, some of Brazil's sparsely-populated northern states such as Roraima (9th, 70.6) or Amazonas (20th, 69.8) have smaller populations and much less certain coverage levels. In these cases we are much less certain about which state actually has the higher life expectancy.

Samples from the posterior distribution allow us to make probabilistic statements about the ranking of areas, which could be important for allocation of health resources. For example, the posterior medians of e_0 in Amazonas and Roraima are 69.8 and 70.6 respectively, but in 730 of 4000 samples from the posterior in equation (8) the Amazonas life expectancy was higher. Thus, even though the point estimate for Amazonas life expectancy is almost one year lower, the *a posteriori* estimate is $P[e_0^{AM} > e_0^{RO}] \approx .18$.

Analysis of posterior samples also allows probabilistic statements about which states have the lowest and highest life expectancies. For example, Alagoas had the lowest $e_0(\boldsymbol{\alpha})$ value in 3592 of 4000 samples, so there is approximately a 90% probability that it has a lower male life expectancy than all other states. Other candidates for lowest male e_0 are Pará ($p_{lowest} = 0.06$) and Pernambuco ($p_{lowest} = 0.01$). Analogous calculations show that the Federal District ($p_{highest} = 0.91$), Piauí ($p_{highest} = 0.05$), and Rio Grande do Norte ($p_{highest} = 0.02$) are the top candidates for highest male life expectancy.

5.4 Microregion-level estimates

We also used the Bayesian TOPALS model to estimate adjusted posterior distributions of $\boldsymbol{\alpha}$ and $e_0(\boldsymbol{\alpha})$ separately by sex for all 558 Brazilian microregions. Results for males appear in Figure 7, which displays posterior medians, and in Figure 8, which displays the widths of each microregion's 80% posterior probability interval.

The point estimates in Figure 7 show that high-mortality (low-e_0) regions for males tend to be concentrated along the Atlantic coast, particularly in large cities. There are also pockets of high mortality in scattered areas of northern and western Brazil. Point estimates of mortality are especially low in parts of southern Brazil, and in the states of Piauí and Ceará.

Some of the point estimates in Figure 7 are much more reliable than others, however. Figure 8 shows a strong north-south gradient in the precision of male e_0 estimates. Posterior distributions (analogous the adjusted densities for states in Figure 4) are notably narrower in southern microregions. As a result, we can be much more certain about small-area life expectancies in Brazil's South and Southeast than in other regions.

Differences in the uncertainty of these small-area estimates arise for two reasons. First, at any given level of vital registration coverage, microregions with larger populations will have more reliable mortality estimates because of larger sample sizes. Many regions in northern and western Brazil are very sparsely populated, so that even if registration coverage were known exactly local estimates would still be subject to considerable sampling error. Second, the level of underregistration of deaths is *also* uncertain in many small areas. The Bayesian model accounts for this extra source of uncertainty, resulting in wider posterior distributions for small areas with more uncertain vital registration coverage.

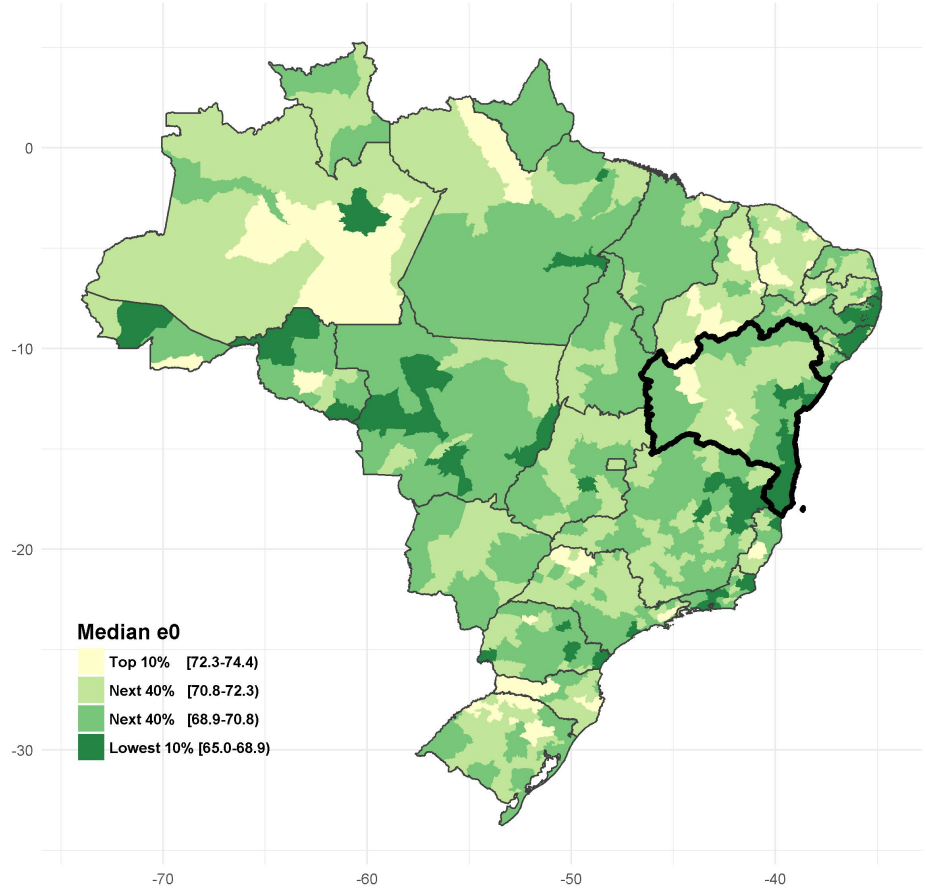


Fig. 7 Posterior medians of male life expectancy at birth, Brazilian microregions 2010. Darker colors indicate higher mortality and lower e_0. The state of Bahia is highlighted with a thick border.

5.5 Signal and noise in microregion life expectancy estimates

We have already highlighted the difficulty of ranking or comparing state-level e_0 estimates. For smaller areas, the signal-to-noise ratio changes for the worse, and comparisons become even more difficult. As an example, Figure 9 shows Bayesian adjusted estimates of male e_0 for the 32 microregions in the state of Bahia (highlighted with a thick border in Figures 7 and 8). At this level of geography, rankings of areas and differences in estimated life expectancy are largely overwhelmed by uncertainty. For example, Livramento do Brumado

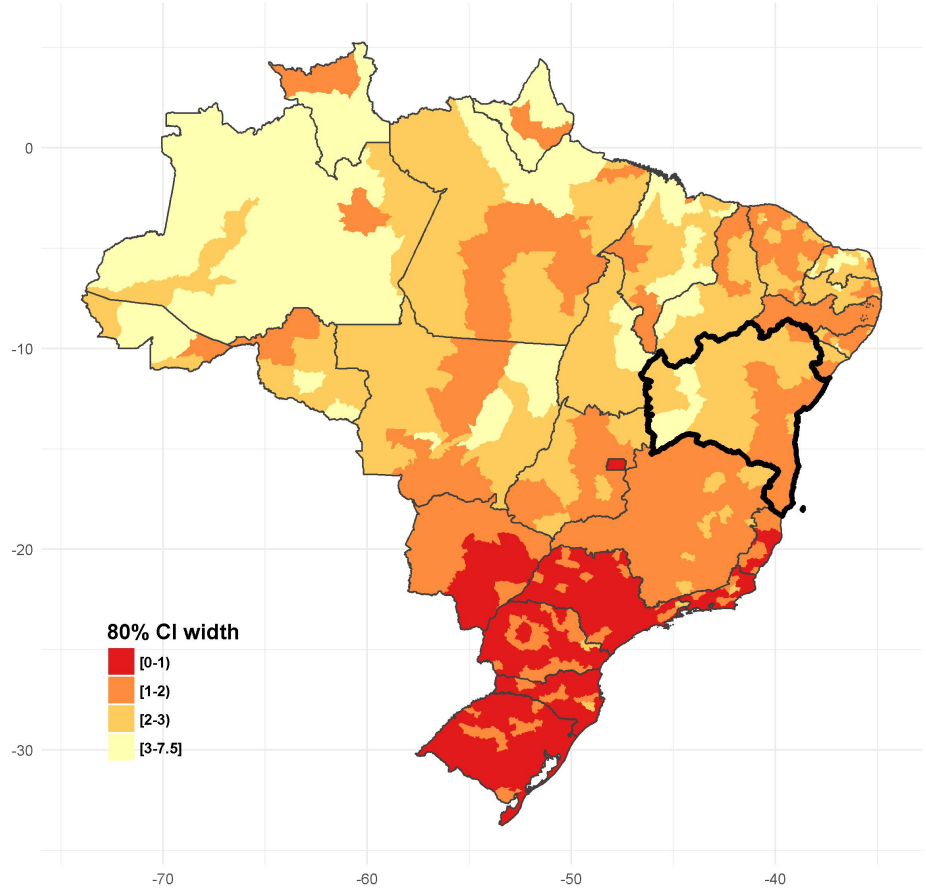


Fig. 8 Uncertainty of male life expectancy at birth, as measured by the width of the 80% posterior probability interval, Brazilian microregions 2010. Darker colors indicate more certain estimates. The state of Bahia is highlighted with a thick border.

(1st line of Figure 9) has the highest estimated male e_0, with a posterior median of 73.3 years. But there is only a 37% posterior probability that this microregion has the highest male life expectancy: it ranked #1 in 1479 of 4000 posterior samples from the posterior in equation (7) . Cotegipe (2nd line, posterior median of 73.0) has a 29% chance of being #1, and Jeremoabo (3rd line, 72.2) has a 9% chance, and so on.

Consider comparing Livramento do Brumado (location #1, on the top line of Figure 9, estimated e_0=73.3 years) and Boquira (location #11, 11th line, 71.6 years). Repeated sampling from the posterior distribution (local priors +

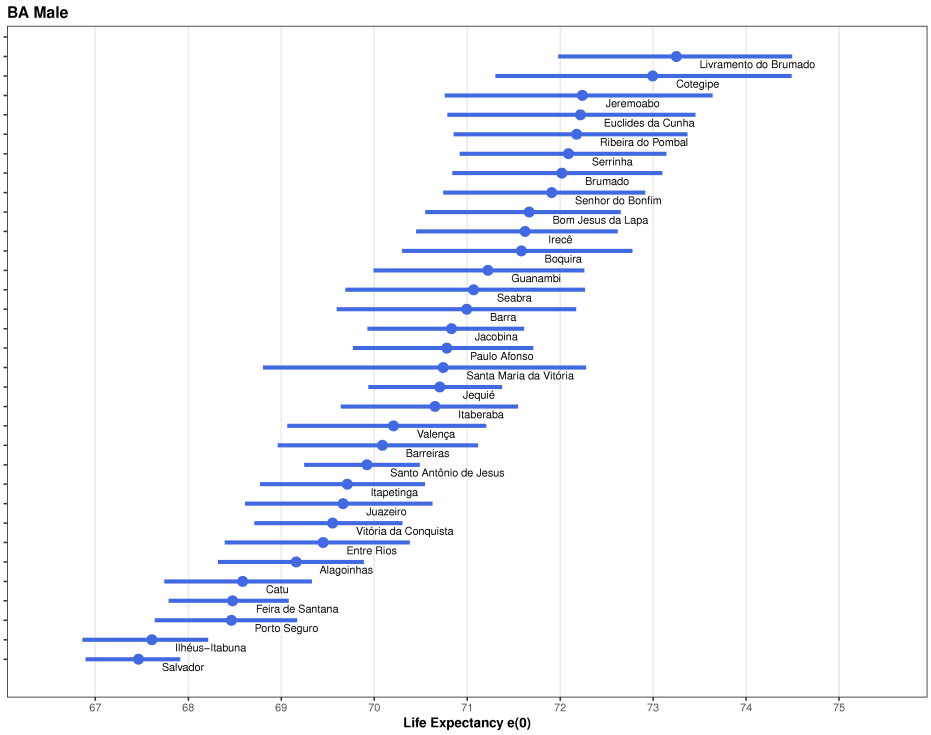


Fig. 9 Male Life Expectancy (median and 80% posterior intervals) for 32 microregions in Bahia

local data) produces a large number of plausible (e_0^1, e_0^{11}) pairs for male life expectancies in these two places, illustrated in Figure 10. Uncertainty is a result of small sample sizes, very small numbers of registered deaths at some ages, and imprecise estimates of local death registration probabilities. Although the point estimate for e_0^1 is nearly two years higher than the estimate for e_0^{11}, there is still considerable uncertainty about which region has the higher life expectancy. Given data and priors about death registration, there is an estimated 11% posterior probability that male life expectancy is actually higher in Boquira than in Livramento do Brumado ($e_0^{11} - e_0^1 > 0$ in 447 of 4000 samples), and even a 3% probability that it is more than one year higher ($(e_0^{11} - e_0^1 > 1$ in 107 of 4000 samples).

The main message of Figures 9 and 10 lies in the very high overlap between the posterior distributions of many of the microregions. These are not especially small areas: Livramento do Brumado was the least populous microregion

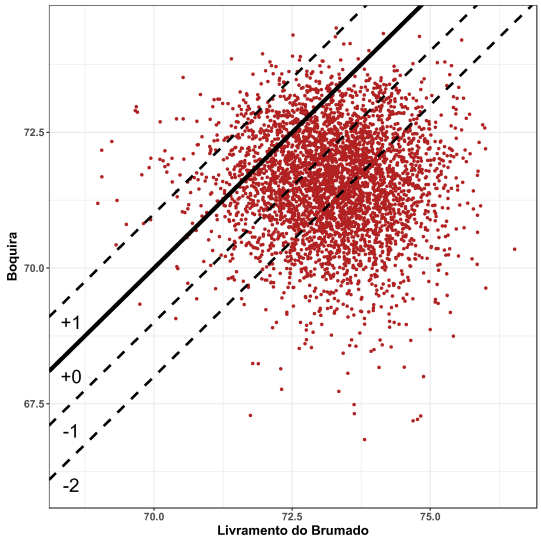


Fig. 10 Sample from joint posterior distribution of male life expectancies in two microregions, Bahia 2010. Diagonal lines labeled by difference in life expectancies

in Bahia, but it had a total population of 97,786 in 2010, and the median total population of Bahian microregions was slightly below 290,000.

Despite these fairly large areas, however, uncertainty dominates most pairwise comparisons. It is clear that at this geographic level researchers and policy makers should not rely on point estimates to distinguish high- and low-mortality areas – especially if differences in best-guess estimates of median e_0 are less than one year. That result applies even more strongly to smaller areas such as municipalities.

6 Conclusion

A Bayesian model is a natural approach to estimating small-area mortality when the vital registration system is imperfect. In this paper we show that it is straightforward to combine age-specific mortality rates and coverage probabilities in a unified model.

The TOPALS relational model for age-specific rates is a useful component in Bayesian modeling for small areas. Because it is flexible but low dimensional, the TOPALS parametric model for age-specific mortality schedules makes it possible to estimate rates and life expectancies even for very small populations.

It is illuminating to consider correction for under-registration as a statistical, rather than arithmetic, process. An explicitly statistical model naturally emphasizes the uncertainty of results. Brazilian microregions are not especially small areas, but nevertheless the uncertainty in life expectancy estimates is often much greater than the differences between local point estimates.

One of our main substantive findings concerns the signal-to-noise ratio in small-area estimates. The levels of posterior uncertainty in microregional life expectancy estimates for Brazil make it clear that attempts to estimate mortality at even smaller levels of geography (for example, for the 417 municipalities in Bahia that are subregions of those displayed in Figure 9) face serious and possibly insurmountable difficulties. Even with good statistical methods for estimating mortality in very small populations, realistic assessment of uncertainty suggests that it may be extremely difficult to draw meaningful distinctions about the mortality of those populations. Demographers should stay humble.

Appendix: Statistical distribution of registered deaths

A generalized Poisson distribution for a random count variable Y, using a mixture of heterogeneous risks, is (Greene 1997, pp 939–940)

$$P(Y=k) = \int_0^\infty \frac{e^{-\lambda z}\,(\lambda z)^k}{k!}\, g(z)\, dz$$

where z is a multiplicative risk factor with density $g(z)$ over positive real numbers. This mixture model $Y \sim PoissonMix(\,\lambda,\, g\,)$ describes the distribution of count variable $Y \in \{0, 1, 2, \ldots\}$ in terms of a scalar parameter λ and a density function $g(\,)$. It generalizes the Poisson distribution by allowing the mean and variance of Y to differ. In particular, it provides a framework for modeling overdispersion ($V(Y) > E(Y)$), which is often observed in count data.

The mixture model includes the standard Poisson distribution as a limiting case: as the distribution $g(z)$ approaches a constant at $z = 1$, Y's distribution approaches a Poisson with $E(Y) = V(Y) = \lambda$. It also includes the negative binomial distribution: if $g(z)$ is a gamma density with $E(z) = 1$ and $V(z) = \frac{1}{\theta}$, then Y has a negative binomial distribution with $E(Y) = \lambda$ and $V(Y) = \lambda + \frac{\lambda^2}{\theta}$. Other $\{\lambda, g(\,)\}$ mixtures yield other discrete distributions for Y.

Suppose that total deaths in a population follow a distribution in this generalized family, so that the probability of D deaths is

$$P(D) = \int_0^\infty \frac{e^{-\lambda z}\,(\lambda z)^D}{D!}\, g(z)\, dz$$

If deaths are registered independently with probability π, then

$$P(R|D) = \frac{D!}{R!(D-R)!}\pi^R\,(1-\pi)^{D-R} \quad \text{for } R \in \{0, 1, 2, \ldots D\}$$

and the joint probability of a pair of integers (R, D) is

$$P(R, D) = \int_0^\infty \frac{e^{-\lambda z}\,(\lambda z)^D}{R!(D-R)!}\,\pi^R\,(1-\pi)^{D-R}\; g(z)\, dz \quad \text{for } D \in \{0, 1, 2, ...\} \text{ and } R \in \{0, 1, 2, \ldots D\}$$

In terms of registered deaths R and unregistered deaths $U = D - R$ the same expression is

$$P(R, U) = \int_0^\infty \frac{e^{-\lambda z}\,(\lambda z)^{R+U}}{R!\,U!}\,\pi^R\,(1-\pi)^U\; g(z)\, dz \quad \text{for } R \in \{0, 1, 2, ...\} \text{ and } U \in \{0, 1, 2, \ldots\}$$

The marginal probability of R registered deaths is therefore

$$\begin{aligned}
P(R) &= \sum_{U=0}^{\infty} \left[\int_0^{\infty} \frac{e^{-\lambda z} (\lambda z)^{R+U}}{R!\, U!} \pi^R (1-\pi)^U \; g(z)\, dz \right] \\
&= \int_0^{\infty} \left[\sum_{U=0}^{\infty} \frac{e^{-\lambda z} (\lambda z)^{R+U}}{R!\, U!} \pi^R (1-\pi)^U \right] g(z)\, dz \\
&= \int_0^{\infty} \frac{e^{-\lambda z} (\lambda z)^{R}}{R!} \pi^R \left[\sum_{U=0}^{\infty} \frac{(\lambda z)^U}{U!} (1-\pi)^U \right] g(z)\, dz \\
&= \int_0^{\infty} \frac{e^{-\lambda z} (\lambda z)^{R}}{R!} \pi^R \left[e^{+\lambda z (1-\pi)} \right] g(z)\, dz \\
&= \int_0^{\infty} \frac{e^{-\lambda z} (\lambda z)^{R}}{R!} \pi^R \left[e^{+\lambda z (1-\pi)} \right] g(z)\, dz \\
&= \int_0^{\infty} \frac{e^{-\lambda \pi z} (\lambda \pi z)^{R}}{R!} \; g(z)\, dz
\end{aligned}$$

The distribution of registered deaths R therefore has exactly the same mathematical form as the marginal distribution of total deaths D, except that parameter λ is replaced with $\lambda\pi$. That is

$$\left.\begin{array}{l} D \sim PoissonMix(\,\lambda,\, g\,) \\ R \sim Binom(D, \pi) \end{array}\right\} \quad \Rightarrow \quad R \sim PoissonMix(\,\lambda\pi,\, g\,)$$

This general proof applies to special cases where $D \sim Poisson$ or $D \sim NegBinom$, as well as to other Poisson mixtures. Most importantly for this paper, it demonstrates that if total deaths have a Poisson distribution with expected value $\lambda = N\mu$, then registered deaths also have a Poisson distribution, with expected value $\lambda\pi = N\mu\pi$.

References

Agostinho, C. 2009. Estudo sobre a mortalidade adulta, para Brasil entre 1980 e 2000 e Unidades da Federação em 2000: uma aplicação dos métodos de distribuição de mortes. *Belo Horizonte: Faculdade de Ciências Econômicas, UNiversidade Federal de Minas Gerais*.

Alexander, Monica, Emilio Zagheni, and Magali Barbieri. 2017. A Flexible Bayesian Model for Estimating Subnational Mortality. *Demography*. doi:10.1007/s13524-017-0618-7. https://doi.org/10.1007/s13524-017-0618-7.

Alkema, Leontine, Adrian E. Raftery, Patrick Gerland, Samuel J. Clark, François Pelletier, Thomas Buettner, and Gerhard K. Heilig. 2011. Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography* 48 (3): 815–839. doi:10.1007/s13524-011-0040-5. https://doi.org/10.1007/s13524-011-0040-5.

Alkema, Leontine, Vladimíra Kantorová, Clare Menozzi, and Ann Biddlecom. 2013. National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. *The Lancet* 381 (9878): 1642–1652. doi:10.1016/S0140-6736(12)62204-1. http://linkinghub.elsevier.com/retrieve/pii/S0140673612622041.

Bennett, Neil G., and Shiro Horiuchi. 1981. Estimating the Completeness of Death Registration in a Closed Population. *Population Index* 47 (2): 207–221. http://www.jstor.org/stable/2736447.

Bennett, Neil G., and Shiro Horiuchi. 1984. Mortality Estimation from Registered Deaths in Less Developed Countries. *Demography* 21 (2): 217–233. http://www.jstor.org/stable/2061041.

Bernardinelli, Luisa, and Cristina Montomoli. 1992. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine* 11 (8): 983–1007. doi:10.1002/sim.4780110802. http://dx.doi.org/10.1002/sim.4780110802.

Bhat, PN Mari. 2002. Completeness of India's sample registration system: an assessment using the general growth balance method. *Population studies* 56 (2): 119–134.

Bignami-Van Assche, S. 2005. Province-specific mortality in China 1990-2000. [Unpublished] 2005. Presented at the 2005 Annual Meeting of the Population Association of America Philadelphia Pennsylvania March 31-April 2 2005..

Borges, Doriam, Dayse Miranda, Thais Duarte, Fernanda Novaes, Kryssia Ettel, Tatiana Guimarães, and Thiago Ferreira. 2012. Mortes violentas no Brasil: uma análise do fluxo de informações. *Rio de Janeiro: LAV/UERJ*.

Brass, W. 1971. Mortality models and their uses in demography. *Transactions of the Faculty of Actuaries* 33: 123–142.

Brass, William, and et al. 1975. Methods for estimating fertility and mortality from limited and defective data. *Methods for estimating fertility and mortality from limited and defective data.*.

Campos, Nelson Otávio Beltrão, and RN Rodrigues. 2004. Ritmo de declínio nas taxas de mortalidade dos idosos nos estados do Sudeste, 1980-2000. *Revista Brasileira de Estudos de População* 21 (2): 323–42.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (1): 1–32. doi:10.18637/jss.v076.i01. https://www.jstatsoft.org/index.php/jss/article/view/v076i01.

Congdon, Peter. 2009. Life expectancies for small areas: a Bayesian random effects methodology. *International Statistical Review* 77 (2): 222–240.

de Beer, Joop. 2012. Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research* 27: 543–592. doi:10.4054/DemRes.2012.27.20. http://www.demographic-research.org/volumes/vol27/20/.

de Boor, C. 2001. *A practical guide to splines. Applied mathematical sciences*. Springer. https://books.google.com/books?id=m0QDJvBI_ecC.

de Mello Jorge, Maria Helena Prado, Vilma Pinheiro Gawryszewski, and MDRDDO Latorre. 1997. Análise dos dados de mortalidade. *Revista de saúde pública* 31: 5–25.

de Mello Jorge, MHP, Ruy Laurenti, and Sabina Léa Davidson Gotlieb. 2007. Análise da qualidade das estatísticas vitais brasileiras: a experiência de implantação do SIM e do SINASC. *Ciência e Saúde Coletiva* 12 (3): 643–654.

de Oliveira, Guilherme Lopes, Rosangela Helena Loschi, and Renato Martins Assunção. 2017. A random-censoring poisson model for underreported data. *Statistics in Medicine*. doi:10.1002/sim.7456. sim.7456. `http://dx.doi.org/10.1002/sim.7456`.

Freire, Flávio H, Everton C Lima, Bernardo L Queiroz, Marcos R Gonzaga, and FH Souza. 2005. Mortality estimates and construction of life tables for small areas in brazil, 2010. [Unpublished] 2015. Presented at the 2015 Annual Meeting of the Population Association of America San Diego CA.

Frias, Paulo Germano de, Celia Landmann Szwarcwald, Paulo Roberto Borges de Souza Junior, Wanessa da Silva de Almeida, and Pedro Israel Cabral Lira. 2013. Correcting vital information: estimating infant mortality, Brazil, 2000-2009. *Revista de saúde pública* 47 (6): 1048–1058.

Gerland, Patrick, Adrian E. Raftery, Hana Ševčíková, Nan Li, Danan Gu, Thomas Spoorenberg, Leontine Alkema, Bailey K. Fosdick, Jennifer Chunn, Nevena Lalic, Guiomar Bay, Thomas Buettner, Gerhard K. Heilig, and John Wilmoth. 2014. World population stabilization unlikely this century. *Science* 346 (6206): 234–237. doi:10.1126/science.1257469. http://science.sciencemag.org/content/346/6206/234.

Glen, Andrew G., and Lawrence M. Leemis, eds. 2017. *Computational probability applications. International series in operations research & management science*. Cham: Springer.

Gonzaga, Marcos Roberto, and Carl Paul Schmertmann. 2016. Estimating age-and sex-specific mortality rates for small areas with TOPALS regression: an application to Brazil in 2010. *Revista Brasileira de Estudos de População* 33 (3): 629–652.

Greene, William H. 1997. *Econometric Analysis*, 3. edn. Upper Saddle River, NJ: Prentice Hall.

Hill, Kenneth. 2007. Methods for measuring adult mortality in developing countries: A comparative review. The global burden of disease 2000 in aging populations. Research paper no. 01.13. 2002 [online] Available from http://www. hsph. harvard. edu/burdenofdisease/p ublications/p.

Hill, Kenneth, and Bernardo Queiroz. 2010. Adjusting the general growth balance method for migration. *Revista Brasileira de Estudos de População* 27 (1): 7–20.

Hill, Kenneth, Danzhen You, and Yoonjoung Choi. 2009. Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data errors. *Demographic Research* 21: 235–254.

Hill, Kenneth H. 1987. Estimating census and death registration completeness. In *Asian and Pacific population forum/East-West Population Institute, East-West Center*, Vol. 1, 8–13.

Instituto Brasileiro de Geografia e Estatística, ed. 2013. *Tábuas abreviadas de mortalidade por sexo e idade: Brasil, grandes regiões e unidades da federação, 2010. Estudos e pesquisas. Informação demográfica e socioeconômica*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística - IBGE.

Jonker, Marcel F, Frank J Van Lenthe, Peter D Congdon, Bas Donkers, Alex Burdorf, and

Johan P Mackenbach. 2012. Comparison of Bayesian random-effects and traditional life expectancy estimations in small-area applications. *American journal of epidemiology*.

Lynch, Scott M. 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-71265-9_3.

Mathers, Colin D, Doris Ma Fat, Mie Inoue, Chalapati Rao, and Alan D Lopez. 2005. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the world health organization* 83 (3): 171–177.

Matos, Karla, Martins de Godoy, and Christine Baccarat. 2013. Mortalidade por causas externas em crianças, adolescentes e jovens: uma revisão bibliográfica. *Espaço para a Saúde-Revista de Saúde Pública do Paraná* 14 (1/2): 82–93.

Moreno, Elías, and Javier Girón. 1998. Estimating with incomplete count data A Bayesian approach. *Journal of Statistical Planning and Inference* 66 (1): 147–159. doi:10.1016/S0378-3758(97)00073-6. http://linkinghub.elsevier.com/retrieve/pii/S0378375897000736.

Målqvist, Mats, Leif Eriksson, Nguyen Thu Nga, Linn Irene Fagerland, Dinh Phuong Hoa, Lars Wallin, Uwe Ewald, and Lars-\AAke Persson. 2008. Unreported births and deaths, a severe obstacle for improved neonatal survival in low-income countries; a population based study. *BMC international health and human rights* 8 (1): 4.

Murray, Christopher JL, Julie Knoll Rajaratnam, Jacob Marcus, Thomas Laakso, and Alan D Lopez. 2010. What can we conclude from death registration? Improved methods for evaluating completeness. *PLoS Med* 7 (4): 1000262.

Ocaña Riola, Ricardo, and José María Mayoral-Cortés. 2010. Spatio-temporal trends of mortality in small areas of Southern Spain. *BMC Public Health* 10 (1): 1. http://www.biomedcentral.com/1471-2458/10/26.

Paes, Neir Antunes. 2005. Avaliação da cobertura dos registros de óbitos dos estados brasileiros em 2000. *Revista de Saúde Pública* 39 (6): 882–890.

Paes, Neir Antunes, and Marconi Edson Esmeraldo Albuquerque. 1999. Avaliação da qualidade dos dados populacionais e cobertura dos registros de óbitos para as regiões brasileiras. *Rev Saúde Pública* 33 (1): 33–43.

Pletcher, Scott D. 1999. Model fitting and hypothesis testing for age-specific mortality data. *Journal of Evolutionary Biology* 12 (3): 430–439.

Preston, Samuel, and Kenneth Hill. 1980. Estimating the completeness of death registration. *Population studies* 34 (2): 349–366.

Preston, Samuel, Ansley J Coale, James Trussell, and Maxine Weinstein. 1980. Estimating the completeness of reporting of adult deaths in populations that are approximately stable. *Population Index*.

Queiroz, Bernardo L. 2012. Estimativas do grau de cobertura e da esperança de vida para as unidades da federação no Brasil entre 2000 e 2010. [Unpublished] 2012. Presented at the XVIII Encontro de Estudos de População da ABEP, Aguas de Lindóia..

Queiroz, Bernardo L, Everton C Lima, Flávio H Freire, and Marcos R Gonzaga. 2013. Adult mortality estimates for small areas in Brazil, 1980–2010: a methodological approach. *The Lancet* 381: 120.

Queiroz, Bernardo Lanza, Flávio Henrique Miranda de Araújo Freire, Marcos Roberto Gonzaga, and Everton Emanuel Campos de Lima. 2017. Completeness of death-count coverage and adult mortality (15q45) for Brazilian states from 1980 to 2010. *Revista Brasileira de Epidemiologia* 20: 21–33. `http://www.scielo.br/scielo.php?script=sci_arttext&`

pid=S1415-790X2017000500021&nrm=iso.

Raftery, Adrian E. 1988. Inference for the Binomial N Parameter: A Hierarchical Bayes Approach. *Biometrika* 75 (2): 223–228. http://www.jstor.org/stable/2336170.

Raftery, Adrian E., Jennifer L. Chunn, Patrick Gerland, and Hana Ševčíková. 2013. Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography* 50 (3): 777–801. doi:10.1007/s13524-012-0193-x. https://doi.org/10.1007/s13524-012-0193-x.

Raftery, Adrian, Nevena Lalic, Patrick Gerland, Nan Li, and Gerhard Heilig. 2014. Joint probabilistic projection of female and male life expectancy. *Demographic Research* 30: 795–822. doi:10.4054/DemRes.2014.30.27. http://www.demographic-research.org/volumes/vol30/27/.

Riggan, Wilson B, Kenneth G Manton, John P Creason, Max A Woodbury, and Eric Stallard. 1991. Assessment of spatial variation of risks in small populations. *Environmental health perspectives* 96: 223.

Souza, Maria de Fátima Marinho de, Cynthia Gazal-Carvalho, Deborah Carvalho Malta, Airlane Pereira Alencar, Marta Maria Alves da Silva, Otaliba Libânio de Morais Neto, and et al.. 2007. Análise da mortalidade por homicídios no Brasil. *Epidemiologia e Serviços de Saúde* 16 (1): 7–18.

Stephens, Alexandre S, Stuart Purdie, Baohui Yang, and Helen Moore. 2013. Life expectancy estimation in small administrative areas with non-uniform population sizes: application to Australian New South Wales local government areas. *BMJ open* 3 (12): 003710.

Szwarcwald, CL, OL Morais Neto, PG Frias, PRB de Souza Júnior, JJC Escalante, RB de Lima, and RC Viola. 2010. Busca ativa de óbitos e nascimentos no Nordeste e na Amazônia Legal: estimação das coberturas do SIM e do SINASC nos municípios brasileiros. *Ministério da Saúde (BR). Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Saúde Brasil.*

Tsimbos, Cleon, Stamatis Kalogirou, and Georgia Verropoulou. 2014. Estimating spatial differentials in life expectancy in greece at local authority level. *Population, Space and Place* 20 (7): 646–663.

Ševčíková, Hana, Nan Li, Vladimíra Kantorová, Patrick Gerland, and Adrian E. Raftery. 2016. Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections. In *Dynamic Demographic Analysis*, ed. Robert Schoen, 285–310. Cham: Springer. DOI: 10.1007/978-3-319-26603-9_15. ISBN 978-3-319-26603-9.

Wilmoth, John, Sarah Zureick, Vladimir Canudas-Romo, Mie Inoue, and Cheryl Sawyer. 2012. A flexible two-dimensional mortality model for use in indirect estimation. *Population Studies* 66 (1): 1–28. doi:10.1080/00324728.2011.611411. http://www.tandfonline.com/doi/abs/10.1080/00324728.2011.611411.

You, Danzhen, Lucia Hug, Simon Ejdemyr, Priscila Idele, Daniel Hogan, Colin Mathers, Patrick Gerland, Jin Rou New, and Leontine Alkema. 2015. Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *The Lancet* 386 (10010): 2275–2286. doi:10.1016/S0140-6736(15)00120-8. http://linkinghub.elsevier.com/retrieve/pii/S0140673615001208.