Syddansk Universitet

Bayesian inference to estimate mortality from incomplete historical data

Villavicencio, Francisco ; Missov, Trifon I.; Colchero, Fernando

Publication date:
2016

Document version
Final published version

Citation for pulished version (APA):
Villavicencio, F., Missov, T. I., & Colchero, F. (2016). Bayesian inference to estimate mortality from incomplete historical data. Paper presented at Population Association of America 2016 Annual Meeting, Washington, D.C., United States.

Bayesian Inference to Estimate Mortality from Incomplete Historical Data

Francisco Villavicencio*[1,2], Trifon I. Missov[3], and Fernando Colchero[1,2]

[1]Dept. Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

[2]Max-Planck Odense Center on the Biodemography of Aging, Odense, Denmark

[3]Max Planck Institute for Demographic Research, Rostock, Germany

PAA 2016 Annual Meeting, Washington, D.C., March 31–April 2

Abstract

We present a Bayesian model to estimate adult mortality from historical datasets in which the ages of individuals are unknown. The motivation behind this model is to analyze data from the Barcelona Historical Marriage Database (BHMD), a dataset composed by marriage records in which the ages are not recorded, though some lower and upper times of birth and death can be extracted for many individuals. Here we test the model by applying it to a different historical database from which the exact times of birth and death are known, but that we intentionally "corrupt" in order to make it look like the data from the BHMD. The model uses a hierarchical Bayesian approach that requires the implementation of a Markov chain Monte Carlo (MCMC) algorithm to iteratively estimate mortality parameters and times of birth and death. Preliminary results show that the model is able to recover good estimates of the life expectancy and the ages at death.

---

*`villavicencio@imada.sdu.dk`

1 Introduction

We present a Bayesian model to estimate adult mortality from datasets in which the age of individuals are unknown. Our focus is on the analysis of human historical data, but the model can be easily extended to the study of other kind of human and non-human data in which the ages of individuals are unknown, or partially unknown. The motivation behind this model is to analyze data from the Barcelona Historical Marriage Database (BHMD), a dataset composed by marriage records in which the ages are not recorded.

Due to the lack or incompleteness of census data and death records, little is known about mortality patterns of Catalonia (Spain) in the 16^{th} and 17^{th} centuries. The first modern census conducted in Spain was was carried out in 1787 (Dopico and Rowland, 1990). Prior to this point, enumerations of hearths and households had been conducted in Catalonia since the mid fourteenth century (Feliu, 1999; Nadal and Giralt, 2000), but the lack of continuity hampered the study of these populations from a demographic perspective.

In this context, historical registers of marriage licenses from the diocese of Barcelona are a very rich source that offer interesting research opportunities. Between 1451 and 1905 those marriages were recorded in a set of 291 books conserved at the archive of the Barcelona cathedral (Baucells, 2002; Carreras Candi, 1913). The substantial volume of information available in these marriage license books—about 600,000 marriages celebrated in the Barcelona area from over 450 years—was used to create the Barcelona Historical Marriage Database (BHMD) (Cabré and Pujadas-Mora, 2011; Villavicencio et al., 2015).

The only temporal information available in the BHMD are the dates of the marriages. However, despite the ages of individuals remain unknown, the information about married couples and their parents allow to carry out a nominal record linkage among marriage records and reconstruct individuals' lifespans. This process, which is described in detail in Villavicencio et al. (2015), establishes for each individual some lower and upper times of birth based on the dates of the parents' marriages and the individuals' own marriage, as well as some lower and upper times of death based on the last time the individual is observed alive, and the first time he or she is mentioned as dead in one of the children's marriages. Figure 1 illustrates these ideas. These lifespan reconstruction was carried out only for the period 1573 to 1629, when the BHMD provides more reliable and complete data.

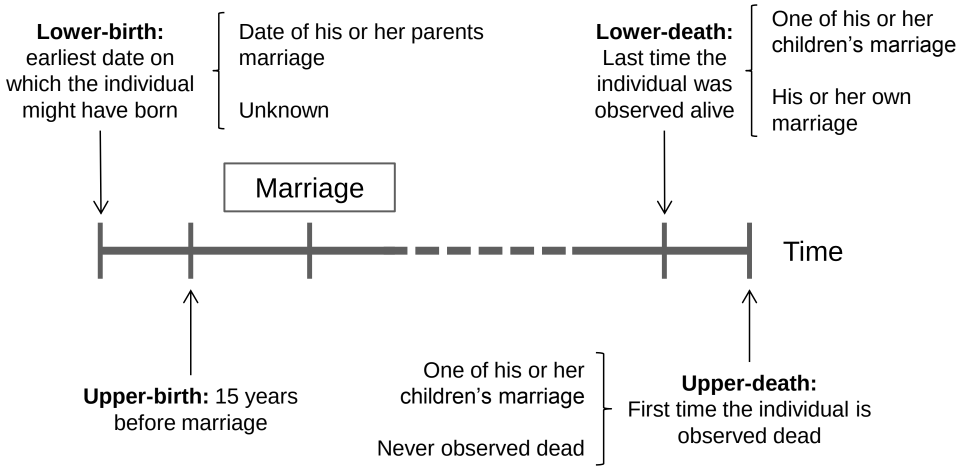

Figure 1: Reconstruction of the lifespan of each individual

Once the lifespans of 10,131 individuals were reconstructed (Villavicencio et al., 2015), a Bayesian approach is used to estimate adult age-specific mortality and life expectancy from data with unknown ages. The method is based on the Bayesian Survival Trajectory Analysis (`BaSTA`), a free, open-source R package for estimating age-specific survival from capture-recapture data under a Bayesian framework which was originally designed to study the survival of wild animals with unknown ages (Colchero and Clark, 2012; Colchero et al., 2012). The model has been extended for use in the analysis of this particular set of human historical data in which the ages of all the individuals are also unknown.

Nonetheless, the purpose of this paper is not to provide estimates of life expectancy and mortality curves of the Barcelona area in the 16$^{\text{th}}$ and 17$^{\text{th}}$ centuries, but to test the reliability of the Bayesian model. To that end, we use a different historical dataset from which we know the exact times of birth, marriage and death from all individuals. Then, we reconstruct the lifespans of those individuals following the scheme in Fig. 1, using all the available information about the individuals and their relatives. Next, we "corrupt" the data, ignoring the information about times of birth and death, and keeping only the same information available from the BHMD: times of marriage, and lower and upper bounds of birth and death. Finally, we test the model with these data and check if we are able to recover the correct adult mortality estimates and ages at death.

2 Data

In order to validate the model, we use the data generated by French demographer Louis Henry in his pioneering study about the population of France from 1670 to 1819. This data, which was made publicly available by the Institut National d'Études Démographiques (INED) in 2001 (Séguy, 2001), is the result of a long process of inventory, proofing, and archiving of documents collected during the historical demographic survey carried out by Louis Henry and his research team between 1958 and 1987. The survey made use of different sources from the pre-statistical era—mainly parish records—and benefited from a 1667 decree which regulated the registration of vital events in France (Séguy, 2001).

The data available in Séguy (2001) is divided into two categories: 1) anonymous data collecting exclusively marriages and deaths from 1740–1829, and 2) family records built through a process of family reconstitution from marriage records of 1670–1819.[1] Only the latter are used in our analysis due to its greater completeness. Baptism certificates were used to recover the birth dates of individuals born before 1670, and death certificates were used to recover the death dates of individuals dying after 1819, though not all information was available for all individuals (Séguy, 2001). As a result, these family records collect information about birth, marriage and death dates of individuals marrying between 1670 and 1819, and link individuals from different generations.

---

[1] This second category refers to what Louis Henry originally named *fiches de famille* (Séguy, 2001).

2.1 Reconstruction of Lifespans

In the process of "corrupting" the French data to reconstruct lifespans as in Fig. 1, we were able to obtain the lifespan of 9,101 men who married between 1700 and 1780. We limited our analysis to men data because it appear to be more complete and reliable. To correctly test the model, it was imperative to have information about times of birth, marriage and death, but also about the dates of marriage of the parents and the children in order to establish some lower and upper bounds of birth and death, as with the BHMD data. Therefore, given that marriage data was restricted to the time period 1670–1819, if only the lifespans of individuals married from 1700 onwards were considered, the previous marriages of their parents were more likely to be identified on the dataset, hence defining more accurate bounds of birth. Analogously, if only the lifespans of individuals married before 1780 were considered, those individuals were meant to be observed in their children's marriage in a later date before 1819, and more accurate bounds of death could be established. Otherwise, many more individuals would have remained unobserved after their marriage.

Certainly, it was not possible to establish narrow intervals of times of birth and death in those cases where the parents' or the children's marriages were not identified. In those situations, a wide and not restrictive interval of birth or death was assigned. Figure 2 sketches the different types of lifespans reconstructed depending on the available data. For instance, type *a)* is the case in which more information was available, in opposition to *h)*; types *a)* to *d)* correspond to individuals whose parents' marriage was identified, hence allowing to define narrower time intervals of birth, in contrast to *e)–h)*.

2.2 Digit Preference

The quality and completeness of the French data as a demographic historical database is beyond any doubt. Nonetheless, if one analyzes the reported ages at death or ages at marriage, a notorious effect of digit preference is observed. The tendency to round numeric values to certain preferred end-digits (usually 0 or 5) is something humans frequently do in

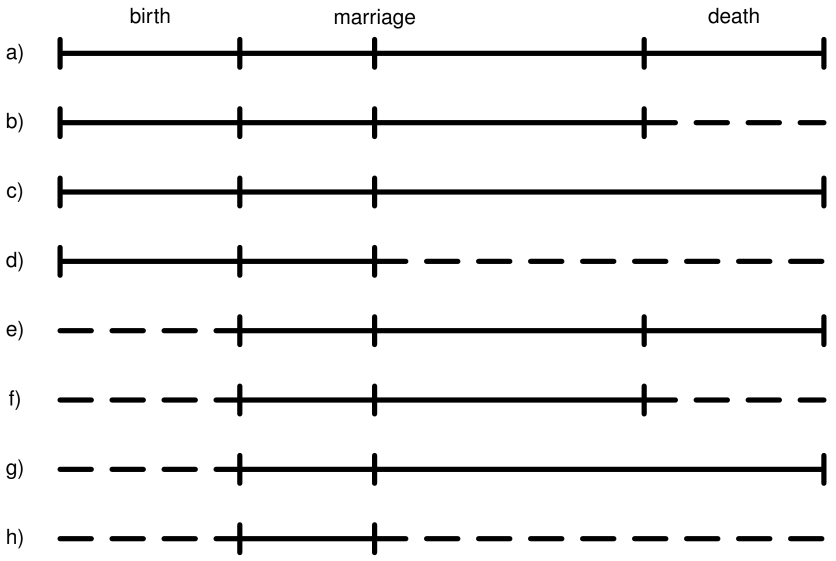

Figure 2: Types of lifespans reconstructed: Types *a)* to *d)* correspond to individuals whose parents' marriage was identified, in contrast to *e)–h)*. Types *a)*, *c)*, *e)*, and *g)* are individuals who were observed dead at least once in their children's marriage, in contrast to *b)*, *d)*, *f)*, and *h)*. Types *a)* and *e)* are individuals whose dead can be placed between that marriages of two of their children; *b)* and *f)* are individuals who were alive when last observed in one of their children's marriage; *c)* and *g)* are individuals who were already dead when first observed after their marriage; and *d)* and *h)* are individuals never observed again after their marriage

their daily life, and it might have a strong effect when reporting data. The digit preference—or age heaping when the reported numbers refer to ages—has been largely observed in demographic and biomedical studies, not only in ages, but also with weight, height or blood measurements (De Lusignan et al., 2004; Myers, 1940; Rowland, 1990).

Different techniques have been developed to deal with digit preference, among which we have chosen to adapt the method devised by Camarda et al. (2008). Their approach suggests that the observed frequencies are "the outcome of a misreporting process that transforms a smooth, but latent, age distribution into observed data" (Camarda et al., 2008, p. 388). Therefore, they propose a technique to estimate the unobserved latent distribution, and provide estimates for the misreported proportions. While adapting this method to our data, we carried out a smoothing process of the distribution of ages at marriage and ages at death.

Figure 3 shows the effect of the smoothing; the histograms correspond to all men who reached age 15 and got married in 1700–1780. It can be observed that the digit preference leaded to unusual heaps at the preferred ages. Less notorious heaps were noticed in the distribution of ages at marriage, though a smoothing was also pertinent.

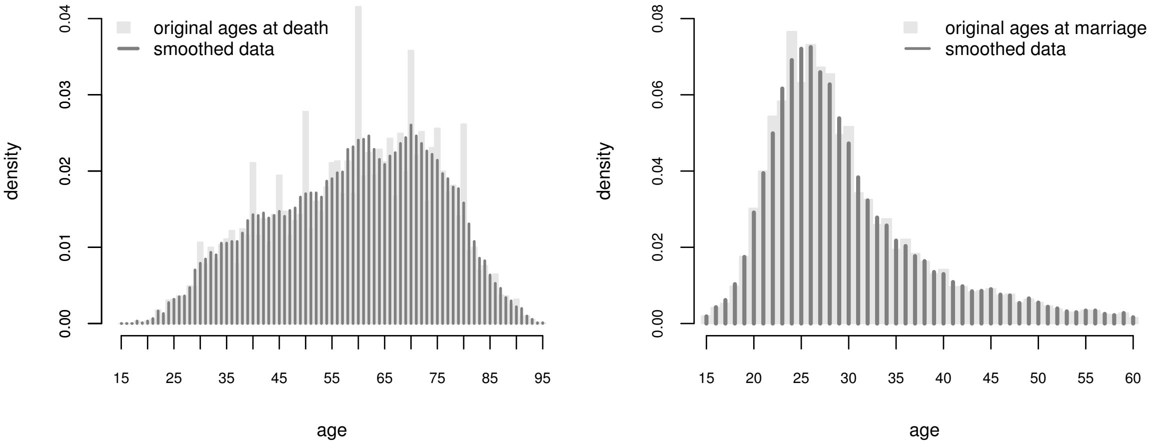

Figure 3: Histograms of ages at death before and after the smoothing (left panel), and histograms of ages at marriage before and after the smoothing (right panel). All men from the French data who reached age 15 and got married in 1700–1780. Different scales are used in both graphs

Camarda et al. (2008) assume that the misreporting observations can only be moved to immediate neighboring digits that are one step away. Hence, using the estimates for the misreported proportions provided by their model, we redistributed the ages at death and ages at marriage of our population by updating the year of birth. Given that the French family records are based on marriage records and the subsequent union formation, we considered that the times of birth were more likely to be incorrect due to a misreported age at the time of marriage or death.

3 Bayesian Model

Bayesian methodology typically consists of three steps (Gelman et al., 2004): 1) A full probability model (likelihood function), including some prior knowledge of the phenomenon of study, must be specified; 2) knowledge about the unknown parameters based on the observed data is updated, and a posterior distribution of the parameters is obtained; and 3) the adjustment of the model to the data is evaluated. These three steps are usually repeated several times in an iterative procedure in order to obtain the best estimates.

Following this general pattern, our model not only estimates the unknown parameters, but also the unknown times of birth and death, given that the ages of individuals are unknown. The model has been implemented in R (R Core Team, 2015), and the code is optimized to take advantage of parallel computing using the `snowfall` R package (Knaus, 2013).

3.1 The Likelihood Function

Since the BHMD consists of marriage records, the model is conditioned on reaching the minimum age at marriage, that we assume to be 15 years.[2] Therefore, all of the individuals from the BHMD—and from the "corrupted" French data—are truncated at their age at marriage, and in the estimation of adult mortality, our model not only requires a distribution of ages at death, but also a distribution of ages at marriage.

The most well-known age at marriage distribution for historical populations in the demographic literature is the Coale-McNeil model (Coale and McNeil, 1972; Trussell, 1976). This distribution is based on a double exponential function whose coefficients were obtained by fitting the Swedish female data from the 18th century. Nevertheless, we chose to use a gamma distribution for the ages at marriage as it showed to be more flexible and provided

---

[2] Both the BHMD and the French data suggest that some individuals might have been married at younger ages. As a matter of fact, in the fourteenth century the canon law established a minimum age at marriage of 12 years for women and 14 for men, limits that were generally accepted by the Catholic Church in the following centuries (Gaudemet, 1987). However, age 15 has been commonly used as a dividing line between childhood and adulthood in historical demography, as it is the case in the Eurasia Project on Population and Family History (Bengtsson et al., 2004, Chap. 2). Moreover, using the standard age of 15—instead of 12 or 14—allows for a better comparison of our results with existing abridged 5-year age life tables that can be found in the literature, as for example the regional model life tables by Coale and Demeny (Coale et al., 1983).

a better fit. This distribution is given by

$$g(z|\phi) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}(z-15)^{\alpha-1}e^{-(z-15)/\beta} \ , \tag{1}$$

where $z \geq 15$ refers to the age at marriage, $\Gamma(\alpha)$ denotes the Gamma function, and $\phi = (\alpha, \beta)$ is the vector of the shape and the scale parameters to be estimated, with $\alpha, \beta > 0$.

Several parametric mortality models were tested, and the one that presented a better fit was the Gompertz-Makeham law of mortality (Makeham, 1867), which is defined by the equation

$$\mu(x|\theta) = c + ae^{b(x-15)} \ , \tag{2}$$

where $x \geq 15$ represents the age at death, and $\theta = (a, b, c)$ is the vector of parameters to be estimated using the Bayesian inference, with $a, c > 0$ and $b \in \mathbb{R}$. The "Makeham term" c accounts for the risk of death from all causes which do not depend on age (Thatcher et al., 1998), and helps providing a good fit in those populations with a high mortality regime, as it is the case in many historical populations.

From the mortality hazard defined in (2), a corresponding survival function can be derived,

$$S(x|\theta) = e^{-\int_{15}^{x}\mu(s|\theta)ds} = \exp\left[\frac{a}{b}\left(1 - e^{b(x-15)}\right) - c(x-15)\right] \ , \tag{3}$$

as well as the probability density function of ages at death

$$\begin{aligned} f(x|\theta) &= \mu(x|\theta)S(x|\theta) \\ &= \left(c + ae^{b(x-15)}\right)\exp\left[\frac{a}{b}\left(1 - e^{b(x-15)}\right) - c(x-15)\right] \quad \text{for } x \geq 15. \end{aligned} \tag{4}$$

As a result, the likelihood function of the ages at death and marriage with respect to the parameters for an individual j can be expressed as

$$L(x_j, z_j|\theta, \phi) = \frac{f(x_j|\theta)}{S(z_j|\theta)}g(z_j|\phi) \ , \quad \text{for } j = 1, \ldots, n \ , \tag{5}$$

where n is the population size, and the term in the denominator accounts for the truncation at the age at marriage, given that all individuals need to survive to their marriage to be observed in our population.

3.2 Estimation of the Target Parameters

As we previously mentioned, our main goal is to test if our Bayesian model is able to recover the adult mortality estimates of a population from which we have full information, but that we intentionally "corrupted" ignoring the exact times of birth and death. To that end, we used the observed ages at death $\mathbf{x}$ and the ages at marriage $\mathbf{z}$ from the French data to estimate some target parameters θ_T and ϕ_T by maximum likelihood estimation. We used the likelihood function $L(\theta|\mathbf{x}) = \prod_{j=1}^{n} f(\theta|x_j)/S(\theta|z_j)$ for the mortality parameters, and $L(\phi|\mathbf{z}) = \prod_{j=1}^{n} g(\phi|z_j)$ for the age at marriage parameters, obtaining the maximum likelihood estimators $\theta_T = (0.00082, 0.08233, 0.00984)$ and $\phi_T = (3.40320, 4.28594)$. It needs to be noted that these estimated parameters—as well as the ones presented in the results of Sect. 4—assume survivorship to age 15.

3.3 The Hierarchical Model

In our model, the ages at death and ages at marriage—which depend on the times of birth, marriage and death—are unknown, and therefore need to be estimated. Let's define the variables

$$t_B := \text{times of birth}$$

$$t_M := \text{times of marriage}$$

$$t_D := \text{times of death}$$

from which only the exact times of marriage t_M are known, whereas t_B and t_D are bounded by some upper and lower bounds of birth and death that vary among individuals depending on the available information about each of them (see Figs. 1 and 2). In order to simplify the notation, t_B, t_M and t_D are conceptualized as vectors with length equal to the population size, such that each of their cells corresponds to a single individual.

On the other hand, the estimation of the marriage parameters ϕ has not yet been included in the model, and these parameters remain fixed along the simulations. As a result, our goal is to estimate the mortality parameters, and the exact times of birth and death of each individual. Accordingly, the likelihood function in (5) can be rewritten as

$$L(t_B, t_D|\theta) = \frac{f(t_D - t_B|\theta)}{S(t_M - t_B|\theta)} g(t_M - t_B) \ , \tag{6}$$

from where it is clear that the ages at death are $x = t_D - t_B$, and the ages at marriage $z = t_M - t_B$.

In order to estimate the mortality parameter and the exact times of birth and death, we use a hierarchical approach that requires the implementation of a Markov chain Monte Carlo (MCMC) algorithm using Metropolis-within-Gibbs sampling (Gelfand and Smith, 1990; Metropolis et al., 1953). At each iteration, the joint posterior is divided into two sections: 1) The estimation of mortality parameters of our parametric mortality model; and 2) the estimation of the unknown times of birth and death, which will depend on the respective lower and upper bounds of birth and death. To assess the ergodicity to those estimations, the model runes several parallel iterations of the MCMC algorithm.

Initialization of the model Previous to the initialization of the model, a Bayesian hierarchical framework requires including some prior knowledge of the phenomenon under study (Gelman et al., 2004). In our case, this is to determine some prior mortality parameters θ_P; this prior values reflect our belief on the phenomenon of study, and ideally should not have a strong effect on the outcome of the analysis (Colchero et al., 2012). In each simulation, our model selects the prior parameters randomly, forcing them to be close, but different, to the target parameters θ_T.

Next, the model randomly selects some initial parameters to be used as starting point in the iterative procedure. These initial parameters cannot have any effect on the outcome of the model. Figure 4 shows the example of four convergent simulations with initial parameters chosen from a wide range of values.

Finally, the model selects some initial times of birth and death for each individual, that depend on their respective lower and upper bounds of birth and death. Let's denote by t_B^i the vector of times of birth in the i^{th} iteration, t_D^i the vector of times of death in the i^{th} iteration, l the vector of all lower bounds of birth, and u the vector of all upper bounds of birth. Then, the initial vector of times of birth t_B^0 is sampled from a uniform distribution, such that $t_B^0 \sim U(l, u)$. The initial vector of times of death t_D^0 is computed analogously.

Step 1: Estimation of the mortality parameters At each iteration i, a new set of parameters θ^i is sampled from a normal distribution with mean θ^{i-1}, the vector parameter of the previous step (Metropolis-within-Gibbs sampling). Formally,

$$\theta^i \sim N(\theta^{i-1}, \Sigma_\theta) \ , \tag{7}$$

where Σ_θ is the "jump" covariance matrix among parameters.[3] Then, applying a Bayesian approach, the conditional posterior distribution of the new mortality parameters is given by

$$p(\theta^i | t_B^{i-1}, t_D^{i-1}) \propto L(t_B^{i-1}, t_D^{i-1} | \theta^i) p(\theta^i | \theta_P) \ , \tag{8}$$

where t_B^{i-1} and t_D^{i-1} are the vectors of times of birth and death from the previous iteration, $L(t_B^{i-1}, t_D^{i-1} | \theta^i)$ is the likelihood function defined in (6), and $p(\theta^i | \theta_P)$ accounts for the prior distribution of the mortality parameters, which is assumed to be normally distributed with mean θ_P.

---

[3] We introduce the covariance matrix Σ_θ for mathematical convention, because θ is a three dimensional vector. However, given that the three parameters are independent, Σ_θ is simply a 3×3 matrix with variances on the diagonal and zeros elsewhere.

The new mortality parameters are accepted according to an acceptance rate, given by

$$\min\left\{1, \frac{p(\theta^i|t_B^{i-1}, t_D^{i-1})}{p(\theta^{i-1}|t_B^{i-1}, t_D^{i-1})}\right\} , \tag{9}$$

where the value on the denominator $p(\theta^{i-1}|t_B^{i-1}, t_D^{i-1})$ is the posterior distribution of the mortality parameters from the previous iteration. If the value in (9) is larger than a sampled random number between 0 and 1, the new parameter vector θ^i is accepted; otherwise, it is discarded and $\theta^i = \theta^{i-1}$. This process is carried out at each iteration, creating a sequence of estimated parameters, as shown in Fig. 4.

Step 2: Estimation of times of birth and death Similarly to the mortality parameters, at each iteration new times of birth and death are proposed for all individuals, also using Metropolis-within-Gibbs sampling (Gelfand and Smith, 1990; Metropolis et al., 1953). In this case, however, new values are sampled from a truncated normal distribution, depending on the respective lower and upper bounds of birth and death. Formally,

$$t_B^i \sim N_{[l,u]}(t_B^{i-1}, \Sigma_t) , \tag{10}$$

where t_B^{i-1} is the vector of times of birth from the previous iteration $i-1$, and t_B^i is the proposed new times of birth, each of which truncated to lie between the respective values from vectors l and u. The new times of death t_D^i are computed analogously.

Next, the conditional posterior distribution of times of birth and death is computed as

$$p(t_B^i, t_D^i|\theta^i) \propto p(\theta^i|t_B^i, t_D^i)p(t_B^i, t_D^i|\theta_P) , \tag{11}$$

where $p(t_B^i, t_D^i|\theta_P)$ accounts for the prior age distribution of the population. This prior age distribution depends on the prior mortality parameters θ_P, and is the quotient between the survivorship and the life expectancy at age 15, since the model is conditioned on reaching that age:

$$p(t_B^i, t_D^i|\theta_P) = \frac{S(t_D^i - t_B^i|\theta_P)}{e(15|\theta_P)} \; . \tag{12}$$

Besides, note that the first term on the right-hand side of (11) is

$$p(\theta^i|t_B^i, t_D^i) = f(\theta^i|t_D^i - t_B^i)g(t_M - t_B^i) = f(t_D^i - t_B^i|\theta^i)g(t_M - t_B^i) \; , \tag{13}$$

which is equal to the likelihood function defined in (6), but without the truncation term on the denominator. Whereas truncation at the age at marriage needs to be considered when estimating the mortality parameters, it is not necessary when estimating times of birth and death and the parameters remain invariant. Again, new times of birth and death are accepted according to an acceptance rate analogous to the one described in (9).

As a final remark, at the end of Step 1 the posterior distribution of times of birth and death defined in (11) is updated with the new parameters. Similarly, at the end of Step 2, the posterior distribution of the parameters defined in (8) is also updated with the new times of birth and death.

4 Results

We present the results for four parallel simulations of 10,000 iterations each, and with prior parameters $\theta_P = (0.00097, 0.07864, 0.00674)$. A burn-in period of 2,000 iterations is used to discard the initial estimates in order to consider only the samples of the chain that converge.

Figure 4 shows the sequence of parameters for each of the four simulations. Note that different initial parameters are selected in each case, but all four sequences converge to the same values, which proves the ergodicity of the model. Table 1 shows the estimated parameters, computed by mixing the values obtained in the four sequences, after the burn-in.

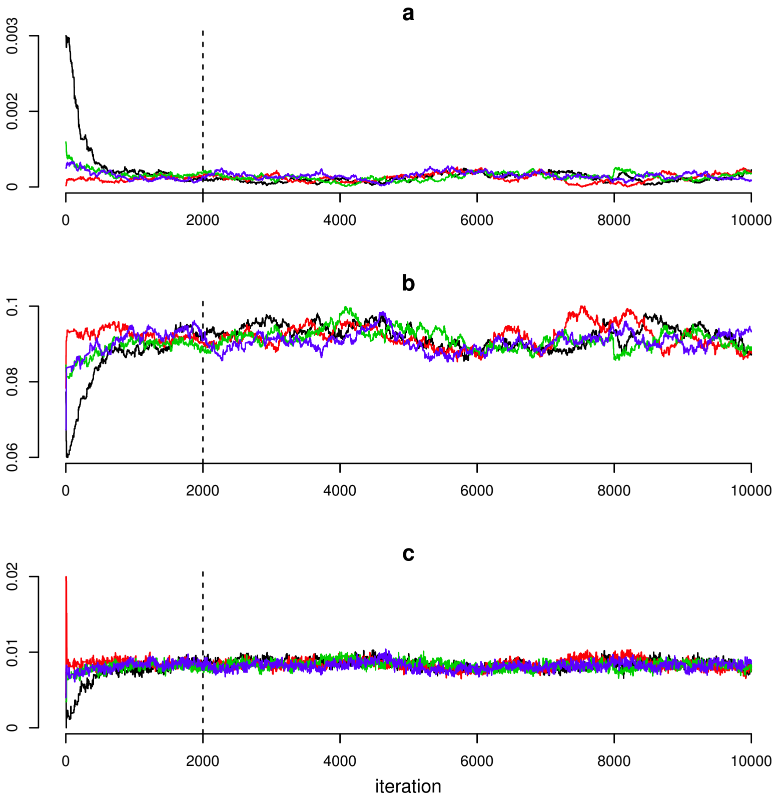

Figure 4: Convergence of the mortality parameters from four parallel simulations of 10,000 iterations each. The dashed vertical lines indicate the burn-in period. Men married in 1700–1780 from the French data

Table 1: Estimated mortality parameters: Mean values ($\hat{\theta}$), standard errors (SE), and 95% credible intervals (CI)

	mean ($\hat{\theta}$)	SE	95% CI
$\hat{a}$	0.00042	0.00007	$(0.00028, 0.00057)$
$\hat{b}$	0.09062	0.00302	$(0.08560, 0.09715)$
$\hat{c}$	0.01096	0.00065	$(0.00972, 0.01226)$

Note that from the target parameters $\theta_T = (0.00082, 0.08233, 0.00984)$, only the Makeham term c falls into the 95% credible intervals of the estimated parameters. However, even if the estimated $\hat{a}$ is much lower than its target, and the estimated $\hat{b}$ higher than its target, the survival curve obtained from the target parameters fits into the 80% credible interval of the survival curve formed by the estimated parameters, as shown in Fig. 5. Moreover, the

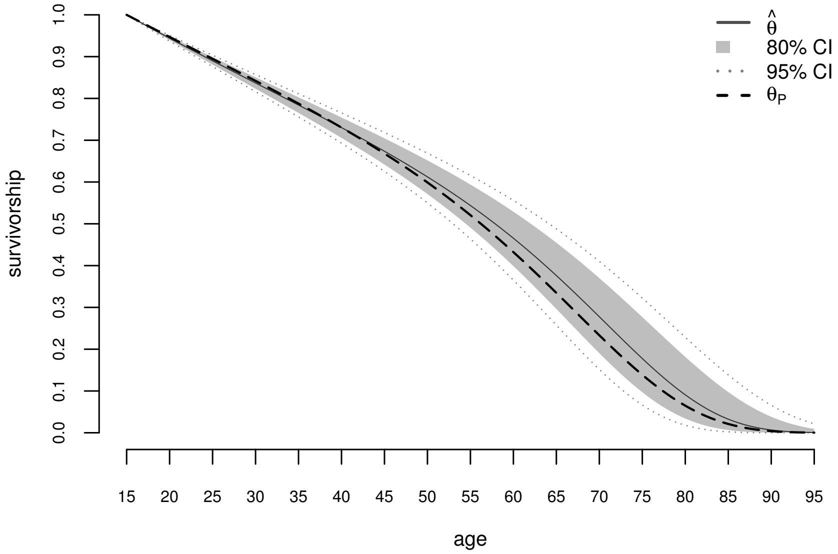

Figure 5: Survivorship from a Gompertz-Makeham law of mortality: Estimated parameters ($\hat{\theta}$), 80% and 95% confidence intervals, and target parameters (θ_P)

model is also able to recover good estimates of life expectancy: whereas the observed life expectancy at age 15 was $e_{15} = 38.60$, the estimated from the model is $\widehat{e}_{15} = 39.75$.

Figure 6 shows the traces of the ages at death obtained for six random individuals. Each graph refers to a single individual: In gray scale are the four traces of ages at death estimated among the 10,000 iterations in each of the four parallel simulations. The black dotted lines are the actual ages at death that the model aims to recover. Note that the initial ages at death are different in each simulation, but on average they all converge to values that are close to the known ages at death. The degree of variability depends on the corresponding lower and upper bounds of birth and death of each individual.

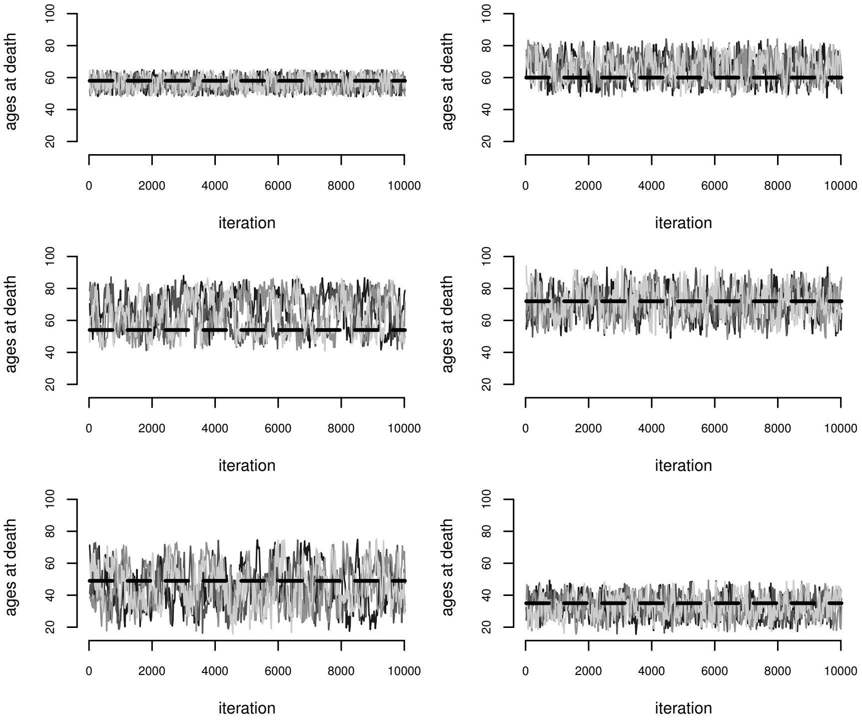

Figure 6: Traces of the estimated ages at death for six random individuals among the four different simulations. The black dotted lines are the actual ages at death

5 Discussion

The development of a methodology to estimate mortality patterns from incomplete demographic data is a novel and fastly growing discipline. Applying methods which were originally designed for biodemographic studies is challenging and it opens the possibility to adapt the model to similar demographic data from other countries and periods.

Further developments of the model will include the estimation of the age at marriage parameters ϕ. Moreover, we wish to carry out a systematic sensitivity analysis on the diagnosis of the model. Preliminary tests suggest that the model is more sensitive to available data rather than the prior or the initial parameters of the simulations. For instance, if 50% of the exact times of birth and 50% of the exact times of deaths were known, the model is able to retrieve more accurate parameter estimates, such that the target parameters lie on the 95% confidence intervals. Certainly, those are some of the issues that we will analyze in depth in the near future.

References

Baucells, J. (2002). Esposalles de l'arxiu de la catedral de Barcelona: Un fons documental únic (1451–1905). *Butlletí del Servei d'Arxius 35*, 1–2.

Bengtsson, T., C. Campbell, and J. Z. Lee (2004). *Life under pressure: Mortality and living standards in Europe and Asia, 1700-1900.* Cambridge, USA: The MIT Press.

Cabré, A. and J. M. Pujadas-Mora (2011). Five centuries of marriages (5CofM): A project of historical demography in the Barcelona area. *Papers de Demografia 368.*

Camarda, C. G., P. H. Eilers, and J. Gampe (2008). Modelling general patterns of digit preference. *Statistical Modelling 8*(4), 385–401.

Carreras Candi, F. (1913). Les obres de la catedral de Barcelona (1298–1445). *Boletín de la Real Academia de Buenas Letras 49*, 22–30.

Coale, A. J., P. Demeny, and B. Vaughan (1983). *Regional model life tables and stable populations* (2 ed.). Studies in Population. New York, US: Academic Press.

Coale, A. J. and D. R. McNeil (1972). The distribution by age of the frequency of first marriage in a female cohort. *Journal of the American Statistical Association 67*(340), 743–749.

Colchero, F. and J. S. Clark (2012). Bayesian inference on age-specific survival for censored and truncated data. *Journal of Animal Ecology 81*, 139–249.

Colchero, F., O. R. Jones, and M. Rebke (2012). BaSTA: An R package for Bayesian estimation of age-specific survival from incomplete mark-recapture/recovery data with covariates. *Methods in Ecology and Evolution 3*, 466–470.

De Lusignan, S. d., J. Belsey, N. Hague, and B. Dzregah (2004). End-digit preference in blood pressure recordings of patients with ischaemic heart disease in primary care. *Journal of Human Hypertension 18*(4), 261–265.

Dopico, F. and R. Rowland (1990). Demografía del censo de Floridablanca. Una aproximación. *Revista de Historia Económica / Journal of Iberian and Latin American Economic History (Second Series) 8*(3), 591–618.

Feliu, G. (1999). La demografia baixmedieval catalana: estat de la qüestió i propostes de futur. *Revista d'Història Medieval 10*, 13–43.

Gaudemet, J. (1987). *Le mariage en Occident: les moeurs et le droit.* Paris: Ed. du Cerf. English edition: Gaudemet, J. (2002) Marriage in the western world: Morals and the law (trans: Roche, I.). Paris: University of Notre Dame Press.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian data analysis* (2 ed.). Chapman and Hall/CRC.

Knaus, J. (2013). *snowfall: Easier cluster computing (based on snow).* R package version 1.84-6.

Makeham, W. M. (1867). On the law of mortality. *Journal of the Institute of Actuaries (1866–1867) 13*(6), 325–358.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Myers, R. J. (1940). Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America 41*, 395–415.

Nadal, J. and E. Giralt (2000). *Immigració i redreç demogràfic: els francesos a la Catalunya dels segles XVI i XVII.* Vic, Spain: Eumo Editorial.

R Core Team (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rowland, M. L. (1990). Self-reported weight and height. *American Journal of Clinical Nutrition 52*(6), 1125–1133.

Séguy, I. (2001). *La population de la France de 1670 à 1829: l'enquête Louis Henry et ses données.* Paris, France: Ined.

Thatcher, A. R., V. Kannisto, and J. W. Vaupel (1998). The force of mortality at ages 80 to 120. Volume 5 of *Monographs on Population Aging.* Odense, Denmark: Odense University Press.

Trussell, T. J. (1976). A refined estimator of measures of location of the age at first marriage. *Demography 13*(2), 225–233.

Villavicencio, F., J. P. Jordà, and J. M. Pujadas-Mora (2015). Reconstructing lifespans through historical marriage records of Barcelona from the sixteenth and seventeenth centuries. In G. Bloothooft, P. Christen, K. Mandemakers, and M. Schraagen (Eds.), *Population Reconstruction*, Chapter 10, pp. 199–216. Springer International Publishing Switzerland.