EUROPEAN POPULATION CONFERENCE (EPC) – BUDAPEST, 2014

Adult mortality in Catalonia in the 16th and 17th centuries

Francisco Villavicencio – villavicencio@demogr.mpg.de
Max Planck Institute for Demographic Research and University of Rostock, Germany
Fernando Colchero – colchero@imada.sdu.dk
Max-Planck Odense Centre on the Biodemography of Aging and University of Southern Denmark
Joana-Maria Pujadas-Mora – jpujades@ced.uab.es
Autonomous University of Barcelona and Centre for Demographic Studies, Spain
Anna Cabré – anna.cabre@uab.es
Autonomous University of Barcelona and Centre for Demographic Studies, Spain

Abstract

Due to the lack or incompleteness of census data and death records, little is known about mortality patterns of Catalonia in the 16th and 17th centuries. The purpose of this paper is to present a methodology to estimate adult life expectancy with unknown ages at death, using data extracted from historical marriage licenses records of the Diocese of Barcelona. The analysis method to be used is based on the Bayesian Survival Trajectory Analysis (BaSTA), a software package for estimating age-specific survival from capture-recapture/recovery data under a Bayesian framework that was originally designed to study the survival in wild animals with unknown ages and unknown ages at death. Marriage licenses records from the Diocese of Barcelona between 1451 and 1905 are conserved in set of 291 books known as *Llibres d'Esposalles*. The use of BaSTA is justified by the type of information registered in each marriage record and by the fact that the age of the grooms remains unknown. Over the 31-year period from 1598 to 1629, additional information concerning the spouses' parents is available, including a remark indicating if they were alive or not at the moment of their children's marriage. This enables the realization of a nominal record linkage among marriage records to reconstruct individual's lifespans: there will be several observations for each individual (the own marriage, and the marriages of the offspring) knowing if he or she was alive at each observation. Our model uses parametric laws of mortality as Gompertz-Makeham, and also model life tables that are introduced to select the best mortality pattern. Several simulations have been carried out, obtaining values of life expectancy at age 15 between 30 and 35 years.

1. Introduction

Due to the inexistence or incompleteness of census data and birth and death records, little is known about the population structure and mortality patterns of Catalonia in the 16th and 17th centuries and the use of indirect methods is imperative. The purpose of this paper is to present a methodology based on Bayesian probabilistic models to estimate adult life expectancy with unknown ages at death, using data extracted from historical marriage licenses records of the Diocese of Barcelona (Catalonia, Spain).

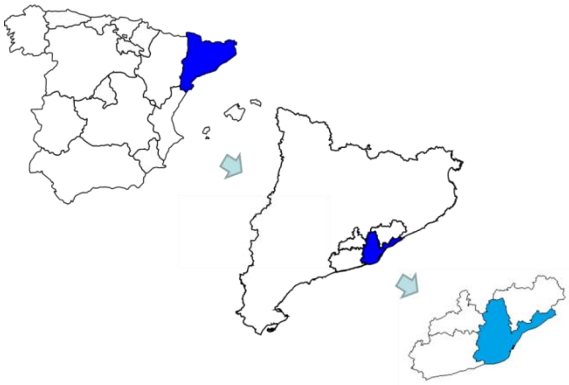


Figure 1. Catalonia, Spain – The Diocese of Barcelona – Main Deanship of Barcelona.

The first modern census in Spain –known as the *Floridablanca Census*- was carried out in 1787 and it was one of the most pioneering censuses in Europe at the time (Simon, 1996; Reher, 2000).[1] Prior to that date, enumerations of hearths and households since the beginning of the 14th century existed in Catalonia (Iglésies, 1982, 1992). However, as there were carried out for tax and military purposes, they omitted certain social groups, such as clergy and nobles, women were excluded, sociodemographic variables such as age or occupations were not considered, and there was a lack of continuity and territorial uniformity (Reher and Valero, 1995). Moreover, Parrish registers of baptism, communion, confirmation, marriage and deaths started in Catalonia approximately after the Council of Trent (1545-1563), but due to the poor maintenance of the archives in many localities and different wars that ravaged Catalonia in the 19th and 20th centuries, only a few are fully preserved nowadays.

In order to address this lack of specific data, historical registers of marriage licenses from the Diocese of Barcelona are a very rich data 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

[1] The *Floridablanca Census* is considered the first Spanish census of population. It was produced in Spain under the direction of the Count of Floridablanca (1728-1808), who was minister of Charles III (1716-1788) between 1785 and 1787 (Dopico and Rowland, 1990).

Barcelona Cathedral and known as *Llibres d'Esposalles*[2] (Baucells, 2002), which bring together information about more than 600,000 unions celebrated in over 250 parishes. All substantial information available in these marriage license books was used to create the Barcelona Historical Marriage Database (BHDM), a unique database covering a period of almost five hundred years, within the project *Five Centuries of Marriages* (Cabré and Pujadas-Mora, 2011).[3] A few previous researches have estimated the life expectancy of Catalonia in the 16th and 17th centuries reconstructing the population of some small Catalan parishes, as for example Muñoz-Pradas (1990) and Torrents (1993). However, they refer to small areas and a relative small number of individuals compared to the information available on the marriage license books.

The analysis method to be used is based on the Bayesian Survival Trajectory Analysis (BaSTA), a free open-source software package for estimating age-specific survival from capture-recapture/recovery data under a Bayesian framework that was originally designed to study the survival in wild animals with unknown age and unknown ages at death (Colchero and Clark, 2012; Colchero et. al., 2012). The use of BaSTA is justified by the type of information registered in every marriage record and by the fact that the age of the grooms remains unknown. For some periods, in addition to the usual information regarding the bridal couple (names, residence, occupation, etc.) supplementary information concerning the spouses' parents is available, which enables the realization of a nominal record linkage among marriage licenses, forming pedigrees. Especially relevant is the 31-year period from 1598 to 1629 (volumes 59 to 74) on which information about the spouses' parents include a remark indicating if they were alive or not at the moment of their children's marriage. Consequently, there will be several observations for each individual (the own marriage, and the marriages of the offspring) knowing if the individual was alive or not at each observation. This information is similar to the one available in capture-recapture studies for which BaSTA was originally designed: different observations of an individual but unknown exact age and age at death.

In the present paper we first analyze the main characteristics of the data set and the information available in the marriage records. Then, we focus on the procedure of marriage linkage that needs to be carried out in order to reconstruct the lifespan of the individuals. Third, we describe the Bayesian models that have been implemented based on the BaSTA package. Finally, we present some results using two models: Gompertz-Makeham law of mortality and model life tables.

[2] Catalan term whose translation into English would be *Marriage License Books*. Both expressions will be indistinctively used in the text.

[3] *Five Centuries of Marriages* is an Advanced Grant Project (2011-2016) funded by the European Research Council (ERC 2010-AdG_20100407) and directed by Professor Anna Cabré, director of the Centre for Demographic Studies and Professor of Human Geography at the Autonomous University of Barcelona, Spain. The research team includes researchers from the Autonomous University of Barcelona, the Center for Demographic Studies and the Computer Vision Center, with specific skills in historical databases and computer-aided recognition of ancient manuscripts.

2. Description of the data set

2.1 The Barcelona Historical Marriage Database (BHMD)

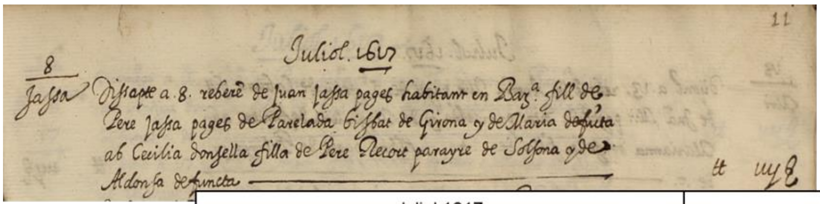
Although the first preserved volume of the licenses marriage books dates from 1451, there are reasons to believe that the custom of recording marriages in the Diocese of Barcelona existed before then as well (Baucells, 2002). According to Carreras-Candi (1913), the origin of the books comes from a privilege given by Pope Benedict XIII (1328-1423)[4] to the Barcelona Cathedral for its construction and subsequent maintenance when he visited the city in September 1409. The Pope granted the new Cathedral with the authority to levy a tax on every union celebrated in the Diocese, which were recorded in a centralized register until 1905. Each bridal couple had to pay a fee according to his socioeconomic status in an eight-step scale that went from free tax (Amore Dei) for people who couldn't afford it or were exempted from payment, to 12£, reserved to the high aristocracy (dukes, marquises, counts, viscounts).[5] Different taxes were applied to nobles, knights and lords, to the urban oligarchy and medical doctors, and to shopkeepers, royal notaries, merchants and masters of guilds. The vast majority of the Catalan society (90%) paid 4 *sous* (peasants, artisans and laborers).

In each marriage record there is information about the exact date when the tax was paid (supposedly – or close to- the date of marriage), the first names of the bridal couple, the surname of the groom, and other socioeconomic and geographical information such as marital status, occupation (groom) and origin or place of residence (groom). For some periods additional information concerning the spouses' parents is available, which enables a nominal record linkage among marriage records. Especially relevant is the continuous 31-year period from 1598 to 1629 (volumes 59 to 74) due to the quality and completeness of the data, including a remark whether the parents were alive or not at the moment of their children's marriage. Figure 2 shows the example of a marriage record that is illustrative, on which it can be noted that the mothers of both the groom and the bride were deceased.

For widowed brides there is also detailed information about the former husband (name, occupation and residence), something that is not the case for widowed men. Women at that time were not identified by themselves, they did not even have surnames and they were always referred to their fathers or their deceased husbands (divorce was not allowed). This might explain the greater availability of information about parents of single brides and former spouses of widowed brides, compared to grooms.

[4] Benedict XIII, born Pedro Martínez de Luna and also known as *Papa Luna*, was an Aragonese nobleman who became pope during the Western Schism (1378-1417).

[5] Taxes were paid in *sous*, where 1£ was equivalent to 20 *sous*. Although this scale was not the same in all volumes of the marriage license books (1451-1905), this is the range that corresponds to the period of study of the present research (1573-1629).



Juliol 1617	July 1617
Dissabte a 8 rebere(m) de Juan Jassa pages habitant en Bar(celon)a fill de/Pere Jassa pages de Parelada bisbat de Girona y de Maria defu(n)ta/a(m)b Cecilia donsella filla de Pere Recort parayre de Solsona y de/Aldonsa defuncta---4 sous	Saturday the 8th was received from Juan Jassa peasant from Barcelona son of Pere Jassa peasant from Peralada diocese of Girona and deceased Maria, and maiden Cecilia daughter of Pere Recort carder from Solsona and deceased Aldonsa---4 sous

Figure 2. Example of a marriage license from 1617 (Volume 69).

2.2 Standardization of nominal information

Marriages were recorded over a period of almost five hundred years by different scribes (usually a different scribe for each volume) with different handwritings. Texts are also confounded by the fact that Catalan spelling rules were not defined until the twentieth century (Bas, 1988). Therefore, in the building of the BHMD, a process of standardization of the data set has been carried out, including occupations, geographical locations, first names and surnames.

Occupations were codified consistent with the Historical International Standard of Classification of Occupations (HISCO) code, whereas geographical locations were grouped according to the current Spanish ZIP code. In the case of foreign locations (i.e. France) a special code was assigned in each case.

The standardization of first names and surnames was more challenging due to their particular nature. In Catalan, as in many languages, many family names have several variants as a result of dialectal differences and foreign influences, as well as an effect of misspellings and phonetic transcriptions. A general discussion about standardization problems of historical data can be found in Bloothooft (1994, 1998). In the building of the BHMD, in order to overcome these complications, a standardization of names was carried out, rectifying all misspellings according to current Catalan grammar rules. Moreover, accents were removed as well as articles, prepositions and conjunctions, in order to keep only the root of each name and facilitate the nominal linkage process. For example, family names like *Companys y Gispert* or *De la Sala* were transformed into *Companys Gispert* and *Sala*, respectively.

3. Record linkage

One of the goals of the project *Five Centuries of Marriages* is to develop methodological research on the linkage of identities and kinship (Cabré and Pujadas-Mora, 2011). In this context, researchers from the Computer Vision Center, together with demographers from the Center for Demographic Studies, have designed a software named *Search Offspring* that helps in the linking of marriage records based on a nominal linkage procedure. Using three key variables that are registered in almost 100% of the marriage records (first name and surname of the groom, and the first name of the bride) it is possible to link different marriages forming pedigrees. Thus, the first name of both bride and groom and the surname of the groom (marriage 1) can later be found in the role of parents (marriage 2) and form the key variables for record linkage. Figure 3 gives the relationship of the key variables schematically.

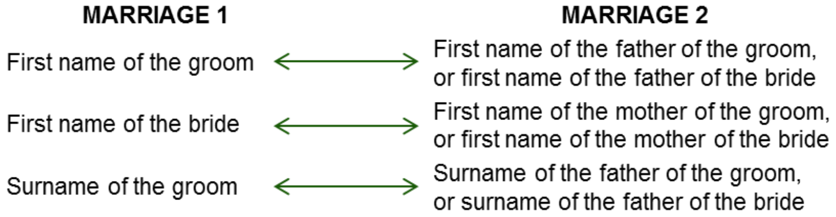

MARRIAGE 1
MARRIAGE 2
First name of the groom
First name of the father of the groom, or first name of the father of the bride
First name of the bride
First name of the mother of the groom, or first name of the mother of the bride
Surname of the groom
Surname of the father of the groom, or surname of the father of the bride

Figure 3. Key variables used in the record linkage.

3.1 Marriage linkage

The linkage program runs an algorithm that measures the Levenshtein distance between strings, computed as the minimum number of operations necessary to transform one string into another using three possible operations: replacement, insertion and deletion (Wagner and Fisher, 1974). The program includes certain particularities as, for example, it allows the replacement of two letters by a single one (*ph* by *f* or *ll* by *l*) with no additional cost. Moreover, the algorithm is adapted to Catalan grammar and includes a cost system that assigns a different cost to the substitution of certain pair of letters, like for example an *a* by an *e*, or an *o* by a *u*. Additionally, there might be false positives due to first names and surnames that are graphically very similar (only one or two letters changed), but etymologically different, like for example *Piera* and *Riera*, *Casals* and *Canals* or *Guell* and *Amell*. These cases are included in a table of exceptions.

The linkage program has been executed to identify the potential parents of the cohorts of children who got married in the 31-year period between 1598 and 1629. Using standardized names, each of the three key variables of a marriage (first name of the groom, first name of the bride and surname of the groom), treated as three strings $a_i, i = 1, 2, 3$, are compared with each of the three corresponding strings of another marriage ($b_i, i = 1, 2, 3$), returning an index of similarity $I(a, b)$ that is computed as follows. Let

$$sim(a_i, b_i) = \frac{total\ costs\ from\ a_i\ to\ b_i}{\max\left(length(a_i), length(b_i)\right)} \in [0,1] \tag{1}$$

be a measure of similarity between each pair of strings a_i and b_i, then

$$I(a,b) = 1 - \frac{1}{3}\sum_{i=1}^{3} sim(a_i, b_i) \in [0,1] \tag{2}$$

being $I(a,b) = 1$ if each of the three pairs of strings are identical. It has been established that a link is accepted if $I(a,b) \geq 0.85$, which means that the three pairs of strings compared are identical in at least an 85% on average. This threshold can be chosen by the user every time the application is executed.

Only two restrictions have been imposed in order to carry out the links: 1) minimum age at marriage of 15 years, 2) no pre-marital intercourse. Therefore, the minimum time difference between the date at marriage of an individual and his or her potential parents might be 189 months (15 years and 9 months). The *set of children* is composed by all marriage records that occurred between 1598 and 1629 as those are the volumes (59 to 74) on which the information about parents is most reliable. The *set of potential parents* is composed by all marriage records that occurred between 1573 and 1613 (volumes 47 to 67).

The process is divided in two, searching separately the potential parents of grooms and the potential parents of brides. Using the linkage program 13,742 links were found, from which 5,509 correspond to grooms and their potential parents, and 8,233 to brides. However, among all those links there were 5,291 over-links (38.1%) with multiple candidates, that is, cases where more than one marriage of possible parents with similar names has been found. Figure 4 shows the frequency of links according to the time difference between the marriages of children (groom or bride) and the marriages of their potential parents, distinguishing by single and multiple links.

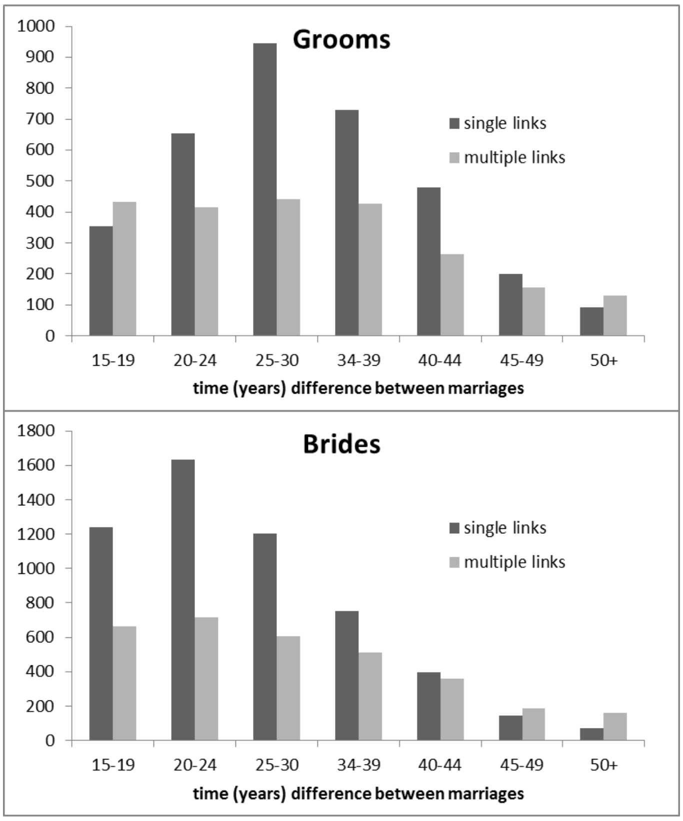


Figure 4. Frequency of links according to the time difference between marriages.

3.2 Reconstruction of the lifespans

The output of the linkage program has been deeply analyzed in order to reduce the number of multiple links. This complex process (which is described in another ongoing paper) takes into account all information available for each marriage record (occupation, geographical location, origin or tax paid) to determine the most likely link among multiple candidates.

The final data set is compounded by 5,180 males and 4,988 females who first married between 1573 and 1613 and who got children that married between 1598 and 1629. In the linkage procedure, marriages of both singles and widowers have been considered. However, when reconstructing the lifespans, only individuals who have been identified in their first marriage are taken into account. Two reasons might justify this issue. First, as we are dealing with marriage data with unknown ages, our model requires a marriage age distribution. Whereas there is an extensive literature about first marriages age distribution (i.e. Coale and McNeil, 1972), the age distribution of second marriages is more challenging. Secondly, as has been already mentioned, former wives of widowed men who remarry are not mentioned and, consequently, a man who remarries cannot be identified with his previous marriage. Therefore, it might happen that the lifespan of a man is constructed twice (once taking information of his first marriage children and another with the information of his second marriage children) and there is no way to know that they belong to the same individual. Considering only individuals whose first marriage has been identified prevents from this lifespan duplicity.

Two more restrictions were imposed in the reconstruction of lifespans: 1) the maximum age an individual could marry was 50 years; 2) the maximum time an individual could leave after his or her first marriage was 85 years. Those two restrictions which are certainly very lax, are nonetheless necessary. The age of individuals is unknown and no information is available about their birth or death dates, so it is imperative to establish some upper and lower bounds of birth and death. For some individuals it will be possible to identify both their children and their parents; in those cases, the lower birth will be delimited by the date of their parent's marriage. Figure 5 sketches how lifespans are reconstructed starting with the date of marriage and using all information available about each individual to determine the lower and upper bounds of birth and death.

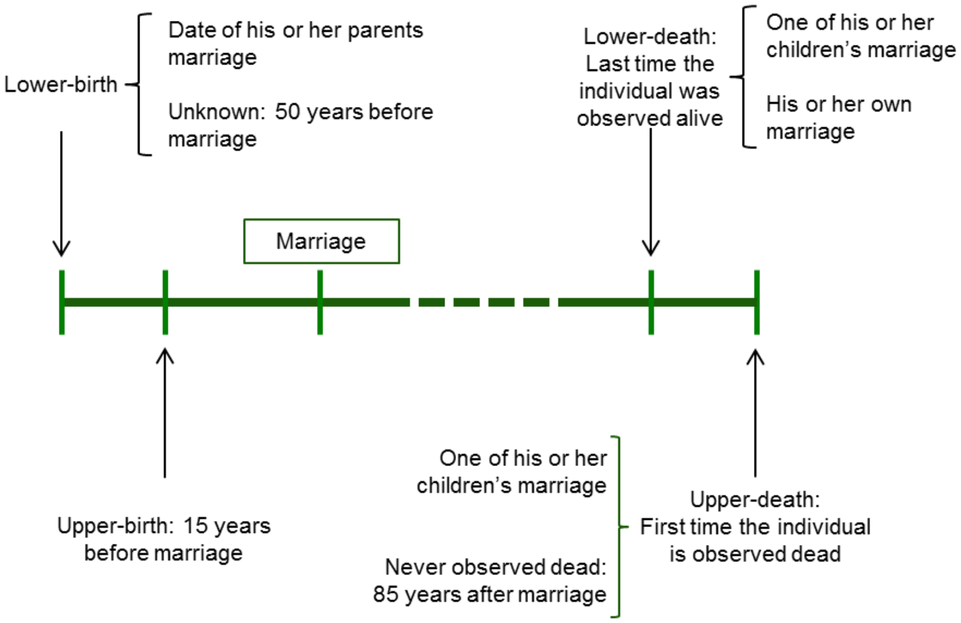

Lower-birth
Date of his or her parents marriage
Unknown: 50 years before marriage
Lower-death: Last time the individual was observed alive
One of his or her children's marriage
His or her own marriage
Marriage
Upper-birth: 15 years before marriage
One of his or her children's marriage
Never observed dead: 85 years after marriage
Upper-death: First time the individual is observed dead

Figure 5. Reconstruction of the lifespan of each individual.

We assume that all individuals are left-truncated at age at marriage. Those who are observed dead at least once are uncensored, while those who are never observed dead are censored. Table 1 summarizes the four different types of lifespans found. It is observed that the higher data availability about women's parents allows to better determining their lower birth, whereas more men are observed dead at least once (uncensored).

Type of lifespan	Males		Females	
Low birth known and censored	57	1.1%	144	2.9%
Low birth known uncensored	53	1.0%	71	1.4%
Low birth unknown censored	2,337	45.1%	2,839	56.9%
Low birth unknown uncensored	2,733	52.8%	1,934	38.8%
TOTAL	5,180	100.0%	4,988	100.0%
Assumptions in the reconstruction of lifespans				
Minimum age at marriage	15 years			
Maximum age at marriage	50 years			
Maximum time lived after marriage	85 years			

Table 1. Types of lifespans and assumptions

4. Methods: Bayesian inference

4.1 Bayesian Model

Broadly speaking, the main difference between Bayesian and classic statistics is that in the first case prior knowledge about the phenomenon of study is taken into account formally and explicitly, together with the sample. The Bayesian methodology might consist on three steps (Gelman et al., 2004):

1. Specify a full probability model, for all the observed and the unobserved quantities, which includes some previous knowledge (priors) about the parameters of the model.
2. Update the knowledge about the unknown parameters according to the observed data.
3. Evaluate the adjustment of the model to the data and its sensibility to changes in prior conditions.

Starting from the Bayes Theorem for conditional probabilities

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)} \tag{3}$$

let's define X as the *observed data* (i.e. ages at death), θ the *unknown parameters* (i.e. parameters of the mortality model) and $P(\theta)$ the *probability distribution of* θ. Then, the likelihood function $P(\theta|X)$ can be expressed as

$$P(\theta|X) = \frac{P(X|\theta)\, P(\theta)}{\int_{-\infty}^{+\infty} P(X|\theta)\, P(\theta) d\theta} \propto P(X|\theta)\, P(\theta) \tag{4}$$

The term $\int_{-\infty}^{+\infty} P(X|\theta)\, P(\theta) d\theta$ ($\sum_{\theta} P(X|\theta)\, P(\theta)$ in the discrete case) does not depend on θ, and with a fixed X it can be considered a constant, yielding to the *unnormalized posterior density* on the right-hand term of equation (4) (Gelman et al., 2004). $P(\theta)$ and $P(X|\theta)$ are usually referred in the literature as the *prior distribution* and the *sampling distribution*, respectively. Expressed in words, equation (4) indicates that the distribution of θ conditioned by the observed data is proportional to the product between the distribution of X conditioned to θ and the prior distribution of θ.

4.2 Bayesian Survival Trajectory Analysis (BaSTA)

The functioning of BaSTA is thoroughly described in Colchero and Clark (2012) and Colchero et al. (2012). This section highlights its main characteristics and how it has been adapted to historical data.

Capture-recapture/recovery (CRR) studies are based on the repeated sampling of a population in which individuals are first marked and released, and at each subsequent occasion, they are either recaptured, not detected, or recovered dead. Observations consist of longitudinal individual histories of recapture, ideally bounded by times of birth and death. However, the nature of studying wild animals entails that there is a large proportion of individuals whose ages or ages at death are unknown because they were born before the study started or because they died after the study had ended. As a result, in some cases many records are discarded because of left-truncation (unknown age) or right-censoring (unknown age at death). In this context, BaSTA is an R package developed as an alternative approach that combines estimation of survival parameters and imputation of unknown states (i.e. times of birth and death) within a Bayesian hierarchical framework (Colchero and Clark, 2012).

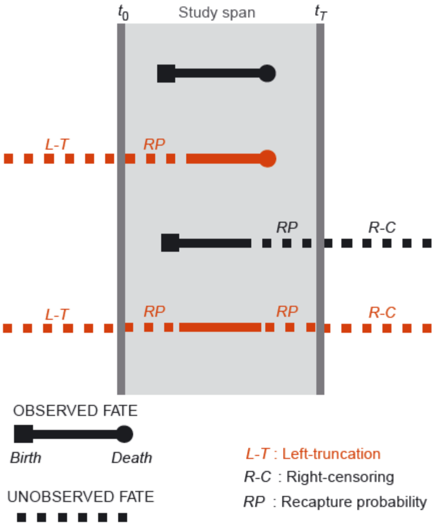


Figure 6. Types of capture histories in wild animals studies, from Colchero et al. (2012).

The present research focuses in those individuals who got married and that at least one of their descendants got married too. The ages of individuals are unknown: as has been sketched in Figure 5, there are only upper and lower bounds of birth, depending on the marriage date (and in some cases in the marriage of the parents), and upper and lower bounds of time of death, depending on marriages of the corresponding descendants. Accordingly, individuals are observed at irregular intervals, which makes a difference with capture-recapture methods where repeated samplings of populations are made in regular time-intervals, and a probability of recapture is included in the model.

The BaSTA algorithm runs one or multiple Markov Chain Monte Carlo (MCMC) algorithms to estimate mortality parameters and ages at death. Specifically, the conditions for posterior simulation are computed by Metropolis-within-Gibbs sampling (Colchero and Clark, 2012). MCMC methods are a class of algorithms for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. The Metropolis Algorithm can be understood as the rejection method used for generating steps in a Markov chain and keep only good samples according to an acceptance rate (Beichl and Sullivan, 2000).

BaSTA runs as many MCMC iterations as indicated by the user, and a burn-in period is also required. Typically, initial samples are not completely valid because the Markov Chain has not stabilized, and the burn-in period allows to determine how many of the initial samples will be discarded in order to consider only the samples of the chain that converge. In each step, the prior conditions are substituted by the new ones provided that they are better estimates; otherwise the same values are kept in the following step (Colchero et al, 2012).

Within a Bayesian framework, two models are developed in the present research. In the first model mortality is fitted through the Gompretz-Makeham law of mortality, while the second one uses Coale and Demeny model life tables (Coale et al., 1983) as input to determine which model life table might better describe the mortality patterns in Catalonia in the 16th and 17th century.

4.3 The parametric model

In a parametric model, the force of mortality (or hazard rate) is commonly defined as

$$\mu(x|\theta) = \lim_{dx \to 0} \frac{P(x \leq X < x + dx | X \geq x, \theta)}{dx} \tag{5}$$

where x is age, X a random variable of ages at death and θ the corresponding survival parameters. From the force of mortality, the survival function $S(x|\theta)$ can be easily derived,

$$S(x|\theta) = P(X \geq x|\theta) = \exp\left[-\int_0^x \mu(t|\theta)\, dt\right] \tag{6}$$

as well as the probability density function of ages at death

$$f(x|\theta) = P(x \leq X < x + dx|\theta) = \mu(x|\theta)\, S(x|\theta) \tag{7}$$

In our model we assume that all individuals are left-truncated at their corresponding age at marriage (z_i). Thus, the probability of dying at age $X_i = x$ is conditioned on surviving to z_i and can be expressed as

$$P(X_i = x|X_i > z_i, \theta) = \frac{P(X_i = x|\theta)}{P(X > z_i|\theta)} = \frac{f(X_i|\theta)}{S(z_i|\theta)} \tag{8}$$

where $S(z|\theta)$ describes survivorship to age at marriage. As the age at marriage is unknown, (8) requires an age at first marriage distribution. To that end, we use the model presented by Coale and McNeil (1972), whose coefficients are derived from the nineteenth-century Swedish data:

$$g(z) = .1946\ \exp\left[-.174\,(z - 6.06) - e^{-.2881\,(z-6.06)}\right] \tag{9}$$

Function $g(z)$ has its origin at 0, which represents the earliest age at which a consequential number of first marriages occur. In our case, 15 years.

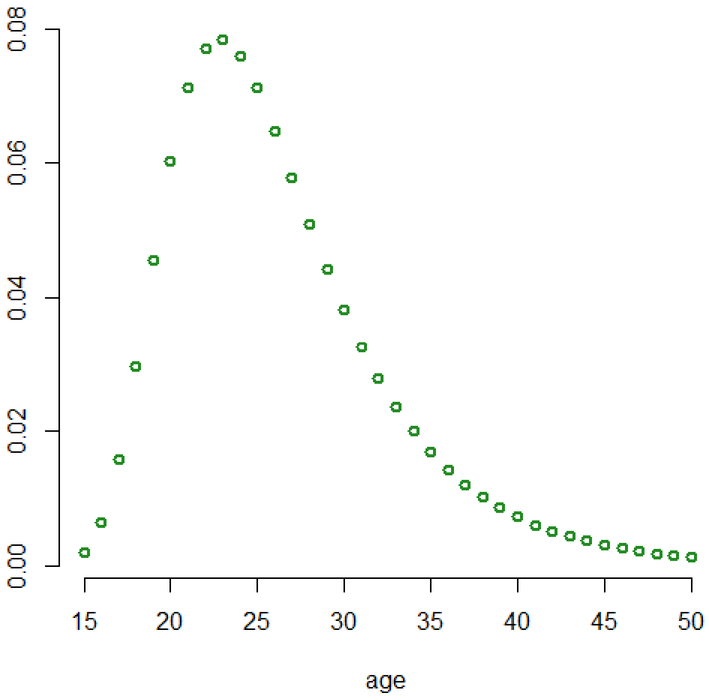


Figure 7. Coale-McNeil age at first marriage distribution.

4.3.1 Initialization of the model

A Bayesian framework requires including some prior knowledge of the phenomenon under study. Prior values reflect our belief in the values of the parameters and ideally they will have no strong effect on the outcome of the analysis (Cochero et al., 2012). In a parametric mortality model, for example, that would be to choose some prior parameters θ^P which might be extracted from the literature. In our study, we derived prior parameters by fitting a Gompertz-Makeham curve to the mortality rates presented by Muñoz-Pradas (1990) in his study of the population of Penedès in the late 17th century, a rural area 60 km away from Barcelona. Next, we determine a prior age at death distribution, which is straightforward using the prior parameters θ^P and the mortality model selected. Then, the algorithm divides the posterior of the joint distribution in two sections: 1) estimation of survival parameters θ; 2) estimation of the unknown ages at death X.

4.3.2 Posterior distribution of the survival parameters

Once the priors have been defined, the algorithm continues with the following steps.

1. For each individual k, we define the following variables:
    a. Date of reaching age 15 b_k. So far, the date at marriage registered in the marriage books.
    b. Date of d_k. So far, the last date the individual is observed in one of his or her children's marriage, regardless of whether the individual has been observed dead at least once (uncensored) or not (censored).
    c. Age at death, computed as the time lived after reaching age 15: $X_k = d_k - b_k$.
2. Select some initial parameters θ of the mortality model to be used as starting point in the iterative procedure. We assume that each of the parameters of the model are normally distributed, with a mean equal to the corresponding prior, allowing to build the probability density function of the parameters,

$$h\left(\theta_j \middle| \theta^P{}_j\right) \sim N\left(\, \theta^P{}_j, \sigma\right), j = 1 \dots \dim \theta^P \tag{10}$$

3. Using equations (4), (8) and (10), we build the likelihood of the initial parameters conditioned to the current ages at death,

$$L(\theta|X) \propto \prod_k \frac{f(X_k|\theta)}{S(z_k|\theta)} \prod_j h\left(\theta_j \middle| \theta^P{}_j\right) \tag{11}$$

4.3.3 Posterior distribution of the ages at death

The likelihood of the posterior distribution of ages at death will depend on the initial model parameters and the ages at marriage. The age at marriage is computed as the difference between the date of marriage and the date of reaching age 15: $z_k = m_k - b_k$.

$$L(X|\theta, z) \propto \prod_k f(\theta|X_k)\; g(z_k|X_k)\; c(X_k|\theta^P) \tag{12}$$

Where $f(\theta|X_k)$ is the probability density function of ages at death, $g(z_k|X_k)$ the probability density function of ages at first marriage, and $c(X_k|\theta^P)$ the prior distribution of ages at death, that depends on the prior parameters.

4.3.4 MCMC and jumps

Once the model has been initialized with the prior and initial parameters, and the starting posteriors, the algorithm runs the MCMC algorithm. At each iteration, the procedure is carried out in two steps: sampling of the new model parameters and sampling of the ages at death. To that end, the jump values define the standard deviation of the distribution from which the MCMC algorithm draws the next value of the chain.

a) Sample model parameters

For each parameter of the model, new parameters are sampled centered on the previous values (Metropolis sampling)

$$\theta_j{}^{t+1} \sim N\big(\,\theta_j{}^t, \sigma_\theta\big), j = 1 \dots \dim \theta^P \tag{13}$$

Given the current ages at death and an acceptance probability

$$P\big(\theta_j{}^{t+1}\big|\theta_j{}^t\big) = \min\left\{1, \frac{\prod_k \frac{f(X_k|\theta^{t+1})}{S(z_k|\theta^{t+1})} h\big(\theta_j{}^{t+1}\big|\theta^P{}_j\big)}{\prod_k \frac{f(X_k|\theta^t)}{S(z_k|\theta^t)} h\big(\theta_j{}^t\big|\theta^P{}_j\big)}\right\} \tag{14}$$

The new parameter value is accepted if its acceptance probability (14) is higher than a uniformed random number.

b) Sample of ages at death

The sampling of ages at death is carried out computing new birth and death dates. The new values are centered on the old ones using a normal truncated distribution taking into account the upper and lower bounds of birth and death. Denoting lb as the lower bound of birth and ub as the upper bond of birth,

$$b_k{}^{t+1} \sim N_{lb \le b \le ub}\big(\,b_k{}^t, \sigma_b\big) \tag{15}$$

And analogously for the times of death. Then, denoting $x_k{}^{t+1}$ as the new age at death, for each individual k an acceptance probability is computed, and

$$P(x_k{}^{t+1}|x_k{}^t) = \min\left\{1, \frac{f\big(\theta\big|X_k{}^{t+1}\big)\, g\big(z_k{}^{t+1}\big|X_k{}^{t+1}\big)\, c\big(X_k{}^{t+1}\big|\theta^P\big)}{f\big(\theta\big|X_k{}^t\big)\, g\big(z_k{}^t\big|X_k{}^t\big)\, c\big(X_k{}^t\big|\theta^P\big)}\right\} \tag{16}$$

Once again, for each individual the new times of birth and death are accepted if the acceptance probability is higher than a uniformed random number.

4.3.5 Output

The outputs of this model are converged sequences of estimates for the model parameters $\hat{\theta}$ and the ages at death $\hat{X}$ (depending on birth and death dates $\hat{b}$ and $\hat{d}$). From these converged sequences, mean values and 95% credible intervals can be computed (Colchero and Clark, 2012)

4.4 The life tables model

The model life tables' model is much simpler. It consists in selecting a set of model life tables as input. In our case we use a set of 52 Coale and Demeny model life tables from the four models (or regions) North, South, West and East, corresponding to populations with a maximum life expectancy at birth of 50 years (UN, 2010). The survivorship of each of these 52 life tables are randomly ordered by columns in a single table.

The algorithm starts selecting a random initial life table l_{ini}. At each iteration, instead of sampling the mortality parameters, the algorithm selects randomly a new life table l_{new} from the set of 52 model life tables and evaluates how the current ages at death X_i fit with l_{new}. If the fit is better, the new life table is selected, otherwise it keeps the old one. Next, the algorithm samples the new times of birth and death in the same way as in the parametric model.

5. Results

In this section we present some preliminary results using different initial values. To ensure that parameter estimates derived from MCMC routines converge appropriately, it is necessary to run several simulations from over-dispersed initial parameter values (Gelman et al., 2004).

5.1 Parametric model

For the parametric model, we use the well-known Gompertz-Makeham law of mortality described by the following mortality rate function (Thatcher et al., 1998)

$$\mu(x) = c + a \cdot e^{bx} = c + e^{b_1 x + b_0} \qquad (17)$$

Two scenarios are selected for the initial parameters of the mortality model, one with a high mortality and another with a low mortality. The prior parameters are the same in both cases (extracted from Muñoz-Pradas, 1990) and in each case the initial parameters are the same for both males and females. We observe that in both cases the algorithm converges to very similar parameter values, obtaining similar life expectancies at age 15, especially for women, whereas for men there is less than a one year difference between the two scenarios. **Error! Reference source not found.** summarizes these results.

Scenario	Parameters	Initial values	Sex	e_{15}	Standard error	2.5% quantile	97.5% quantile
Low mortality	b_0 b_1 c	-10 0.15 0.0001	**Men**	33.48227	0.263925	32.95069	33.95508
			Women	35.43174	0.270413	34.87962	35.95865
High mortality	b_0 b_1 c	-5 0.20 0.01	**Men**	32.67805	0.313753	32.05921	33.26154
			Women	35.24361	0.271597	34.72749	35.76525

Table 2. Makeham law of mortality. Initial and final parmeters

The following two graphs show the final survivorship schedules obtained for males and females, in both cases with the same priors but different initial parameters.

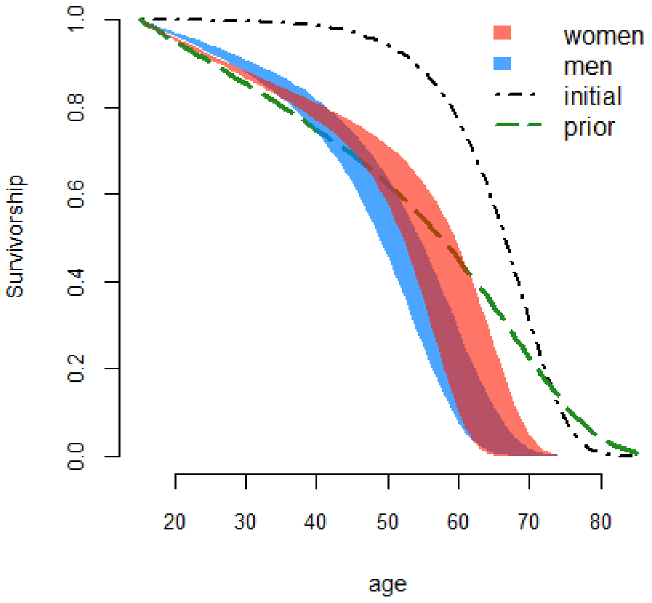


Figure 8. Survivorship (95% confidence intervals) from initial parameters corresponding to low mortality

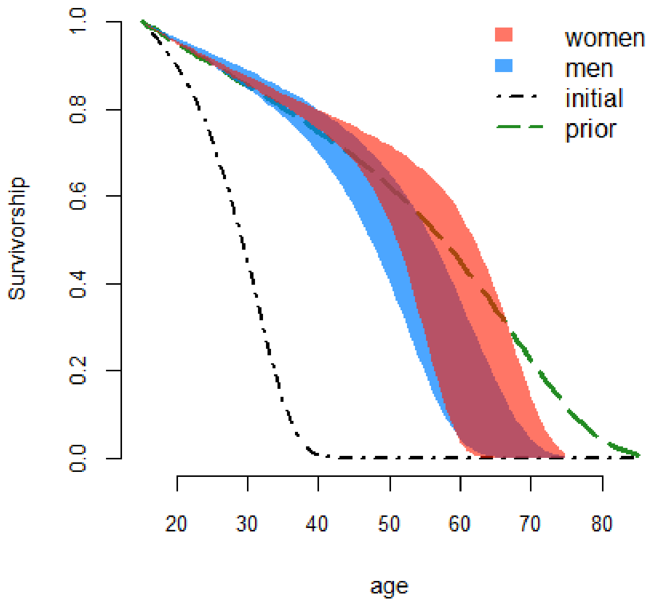


Figure 9. Survivorship (95% confidence intervals) from initial parameters corresponding to high mortality

5.2 Life tables model

Several runs of the life tables model have been carried out, using a different initial life table at each time. In all simulations, for both sexes the algorithm converges to the life table with a lower life expectancy at age 15 among the Coale and Demeny life tables, which corresponds to the life table with a life expectancy at birth of 20 years from the West model. This life table match with a life expectancy at age 15 of 31.15 years for females and 30.91 for males (UN, 2010), slightly lower than the values obtained in the parametric model. Nevertheless, even though this values are also plausible, the shape of the survival curve is notoriously different to the one obtained in the parametric model. Moreover, the fact that the algorithm converges to the life table with the lowest life expectancy at age 15 suggests that maybe Coale and Demeny life tables are insufficient to describe the mortality patterns of Catalonia in the 16th and 17th century and that it would be useful to include more model life tables with a higher mortality in the model.

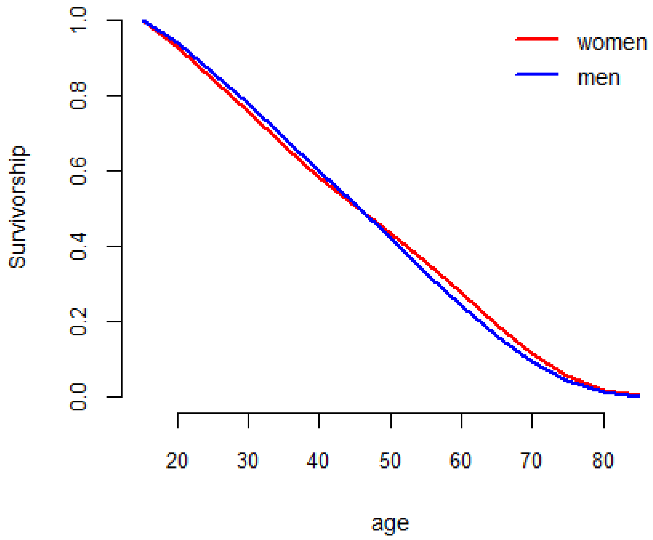


Figure 10. Model West with life expectancy at birth of 20 years

6. Conclusions

This is an innovative project in the field of historical demography that will bring light into the study of mortality in Catalonia in the 16th and 17th centuries, and in this first approach BaSTA seems to be a very interesting and useful tool. The fact of applying methods that were originally designed for biodemographic studies is challenging and opens up the possibility of developing a methodology that could be adapted to similar historical data from other countries and periods.

Two models have been implemented, and the parametric one appears to be more useful and flexible. Preliminary results for both models are plausible, although more research is being carried out in order to implement other methods using BaSTA, as for example the relational model from William Brass (1975).

References

Bas, Jordi (1988). *Els cognoms Catalans i la seva història* [*Catalan surnames and their history*]. Barcelona: Cap Roig, pp. 254.

Baucells, Josep. (2002), "*Esposalles* de l'arxiu de la catedral de Barcelona: un fons documental únic (1451-1905)" ["Marriage licenses from the archives of the Barcelona Cathedral: a unique document fund (1451-1905)"]. *ARXIUS. Butlletí del Servei d'Arxius*, 35, 1-2.

Beichl, Isabel and Sullivan, Francis (2000). "The Metropolis Algorithm". *Computing in Science and Engineering*. 2, 65-69.

Bloothooft, Gerrit (1994). “Corpus-Based Name Standardization”. *History and Computing*. 6.3, pp. 159-167.

Bloothooft, Gerrit (1998), “Assessment of Systems for Nominal Retrieval and Historical Record Linkage”. *Computers and the Humanities*. 32, pp. 39-56.

Brass, William (1975). *Methods for Estimating Fertility and Mortality from Limited and Defective Data*. Chapel Hill, North Carolina, Carolina Population Center.

Cabré, Anna; Pujadas-Mora, Joana M. (2011). “Five Centuries of Marriages (5CofM). A project of historical demography in the Barcelona area”. *Papers de Demograpfia* 2011, **368**, Universitat Autònoma de Barcelona.

Carreras-Candi, Francesc (1913). “Les obres de la Catedral de Barcelona. 1298-1445” [“The Works of the Barcelona Cathedral. 1298-1445”]. *Boletín de la Real Academia de Buenas Letras*. Barcelona: Vol. 49, pp. 22-30.

Coale, Ansley J. and McNeil, D.R. (1972). “The distribution by age of the frequency of first marriage in a female cohort”. *Journal of the Amercian Statistical Association*. Vol. 67, No. 340, pp. 743-749.

Coale, Ansley J.; Demeny, Paul; Vaughan, Barbara (1983). *Regional model life tables and stable populations*. Second edition. New York: Academic Press, 495 pp.

Colchero, Fernando; Jones, Owen R; Rebke, Maren (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* 2012, **3**, 466-470.

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

Dopico, Fausto and Rowland, Robert (1990). “Demografía del Censo de Floridablanca. Una aproximación” [“Demographics from the Floridablanca Census. An approach”]. *Revista de Historia Económica*. Año VIII, nº 3, pp. 591-618.

Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Rubin, Donald B. (2004), *Bayesian Data Analysis.* 2nd Edition. Chapman & Hall/CRC. Boca Raton, FL, USA.

Iglésies, Josep (1982). *El Fogatge de 1553. Estudi i transcripció* [*The Hearth’s Enumeration of 1553. Study and transcription*]. Barcelona: Ed. Rafel Dalmau.

Iglésies, Josep (1992). *El Fogatge de 1497. Estudi i transcripció* [*The Hearth’s Enumeration of 1497. Study and transcription*]. Barcelona: Ed. Rafel Dalmau.

Muñoz-Pradas, Francesc (1990). *Creixement demographic, mortalitat, nupcialitat al Penedès. Segles XVII-XIX* [*Population growth, mortality, nuptiality in Penedès. 17th to 19th centuries*]. PhD Thesis. Autonomous University of Barcelona, Spain.

Reher, David and Valero, Ángeles (1995). *Fuentes de información demográfica en España* [*Sources of demographic data in Spain*]. Madrid: Centro de Investigaciones Sociológicas.

Reher, David (2000). “La investigación en Demografía histórica: pasado, presente y futuro” [“Research in Historical Demography: Past, Present and Future”]. *Boletín de la Asociación de Demografía Histórica,* XVIII, II, 2000, pp. 15-78.

Simon, Antoni (1996). *La població catalana a l'edat moderna. Deu estudis* [*Catalan population at the Modern Ages. Ten Studies*]. Universitat Autónoma de Barcelona, Mongrafies Manuscrits.

Thatcher, A. Roger; Kannisto, Väinö; Vaupel, James W. (1998). "The force of mortality at ages 80 to 120". *Odense Monographs of Population Aging 5*. Odense: Odense University Press.

Torrents, Àngels (1993). *Transformacions demogràfiques en un municipi industrial català: Sant Pere de Riudebitlles, 1608-1935* [*Demographic changes in a Catalan industrial municipality: Sant Pere de Riudebitlles, 1608-1935*]. PhD Thesis. University of Barcelona, Spain.

UN – United Nations, Department of Economic and Social Affairs, Population Division. http://esa.un.org/wpp/Model-Life-Tables/download-page.html (Updated June 2010).

Wagner, Robert A. & Fischer, Michael J. (1974). "The String-to-String Correlation Problem". *Journal of the ACM*. 21(1), pp. 168-173.