Statistical methods to compare mortality for a group with non-divergent populations. Application to Spanish regions.

A. Debón[1], F. Martínez-Ruiz[2], F. Montes[3]

[1]Centro de Gestión de la Calidad y del Cambio.
Universidad Politécnica de Valencia. Spain

[2] Oficina de Estadística. Ayuntamiento de Valencia. Spain

[3] Dpt. d'Estadística i I. O.. Universitat de València. Spain

Abstract

This paper develops a model to compare geographical differences in the mortality of related regions, taking into account the interdependence between them. Additionally, the model allows us to provide an adequate solution for studying the mortality of a group pertaining to a larger population. It should therefore be possible to improve the mortality analysis for the regions in a country by taking into account the patterns within that country. Using official data from the Spanish National Institute of Statistics (Instituto Nacional de Estadística, INE), we applied a modification of the Lee-Carter model to Spanish regions. The results of this model were then compared with other similar models such as the logit Brass, Li and Lee and Russolillo-Giordano-Haberman. One interesting feature of our model is its simplicity, as the comparison of mortality patterns is accomplished by means of a simple index.

Keywords: Brass model, Lee-Carter model, Geographical index, Geostatistics.

1 Introduction

The increasing demand for information about the evolution and demographic characteristics of the population has required a great deal of adaptation by the state organisms responsible for official statistics, which have been compelled to offer more and more information and to put new analysis and demographic projection techniques into practice. Mortality tables are a traditional product of public statistics that the Spanish National Institute of Statistics (INE) has included in its ordinary publications for decades, and for which a new calculation method has recently been introduced which provides greater precision (Spanish National Institute of Statistics (INE), 2009).

1. Andalucia
2. Aragon
3. Asturias
4. Baleares
5. Canarias
6. Cantabria
7. Castilla y la Mancha
8. Castilla y León
9. Cataluña
10. Comunidad Valenciana
11. Extremadura
12. Galicia
13. Madrid
14. Murcia
15. Navarra
16. Pais Vasco
17. La Rioja

Figure 1: Administrative structure of Spain (Autonomous regions).

The purpose of our study is to deal with the measurement of the phenomenon of mortality at the level of individual regions, attempting to provide the most adequate solution for studying the mortality of a group pertaining to a larger population. In this case there are problems associated with the incidence of any demographic phenomenon in a small population, such as the randomness of results, possible inconsistencies between data about deaths for individual regions in relation to the total for the whole country, etc.. The methodology of mortality tables that this article proposes was developed with the aim of establishing an operating procedure which permits the comparison of mortality tables at a national level with those at a regional level, in particular the administrative region s of Spain (called Autonomous Communities). It also permits the construction of tables for Autonomous Communities from those for the whole country. This study arises from the growing interest in obtaining ever more exact and integrated measurements of the conditions and expecttaion of life, as well as current projections of their future trends. Due to the administrative structure of Spain (see Fig. 1) it is of particular interest to have mortality tables broken down to these territorial levels.

Considering life tables for Autonomous Communities, a recent study by Goerlich and Pinilla (2009) uses the methodology described in the protocol (Wilmoth et al., 2007) of the Human Mortality Database to produce regional life tables and compares them by means of indicators such as life expectancy, disability-free life expectancy and the Gini coefficient. All these indicators share the advantage of summarising information about mortality independent of the age structure. However, the comparison of Autonomous Communities varies according to the year in which the indicators were obtained, and is therefore dynamic.

Our proposal is a model with an age effect, a period effect and a a geographical effect (Autonomous Community) to which a random error is added. This model permits the comparison over different years whit just one index.

This article is structured as follows: Section 2 introduces four models adapted to measuring the influence of region on mortality: the classical Brass model (Brass, 1971), models derived from the original Lee-Carter model (Li and Lee, 2005; Russolillo et al., 2010) and our new model, also a derivation of Lee-Carter. Section 3 presents the results of the application of these four models to the analysis of mortality data for Spain as a whole and for the Autonomous Communities for the period 1991-2008. The results which provide the distinct models for the period 1991-2008 are compared by means of their goodness-of-fit and residuals. This is of particular interest, as we have not found evidence of any other similar study on mortality data in Spain. Finally, Section 4 establishes the conclusions to be drawn from the results in the previous section.

2 A review of models

We consider a set of mortality rates in the form of dynamic life tables for different territories. We wish to produce smoother estimates, $\hat{q}_{xti}$, of the true but unknown mortality probabilities q_{xti} from the set of crude mortality rates, $\dot{q}_{xti}$, for each age x and year t in each region i. The crude rate at age x is typically based on the corresponding number of deaths recorded, d_{xti}, relative to those initially exposed to risk, E_{xti}.

2.1 The Logit method

A classical model proposed by Brass (1971) establishes a functional relationship between the survival functions l_{xti}, corresponding to regional mortality tables i, and l_{xte} corresponding to a standard table, e. The initial value of $l_{0ti} = 1$. The relationship is given by

$$\text{logit}(l_{xti}) = \alpha_{ti} + \beta_{ti}\text{logit}(l_{xte}) \quad i = 1, 2, \ldots, r, \tag{1}$$

where r is the total number of regions and the logit transformation for each age x, period t and region i is,

$$\text{logit}(l_{xti}) = 0.5 \ln\left(\frac{1 - l_{xti}}{l_{xti}}\right), \ \forall x > 0$$

From the survival functions estimated by the model,

$$\hat{l}_{xti} = \frac{1}{1 + 2\widehat{\text{logit}(l_{xti})}},$$

the number of deaths is calculated by means of the difference between the survivors in an age group and the previous one, and the probability of death as the ratio between the number of deaths and the number of survivors. The forecast of mortality is carried out by modelling the time series for parameters using an $ARIMA$ model. Murray et al. (2003) generalize the principle underlying Brass's approach, postulating that there is some transformation of the survival

3

function such that all transformed survival functions are linear functions of each other. A detailed discussion on this model can be found in Booth (2006).

Brass's model is the most classic of those that we are going to present, and despite having been widely used in demographical and actuarial contexts, it contains some weaknesses that should be pointed out:

1. It assumes that there is a linear relationship between the logits of the survival functions.

2. It makes predictions from separately adjusted ARIMA models; the underlying assumption being that the two series corresponding to α_t and β_t are independent.

2.2 Lee-Carter models

The Lee-Carter model, developed in Lee and Carter (1992), consists of adjusting the following function to the mortality measurement,

$$m_{xt} = \exp(a_x + b_x k_t + \epsilon_{xt}), \tag{2}$$

or its equivalent

$$\ln(m_{xt}) = a_x + b_x k_t + \epsilon_{xt}, \tag{3}$$

applied to its logarithm transformation. This is an age-period (AP) model, as the double subscript refers to the age, x, and to the year or unit of time, t. In (2) and (3), a_x and b_x are age-dependent parameters and k_t is a specific mortality index for each year or unit of time. The errors ϵ_{xt}, with a zero average and variance σ_ϵ^2, reflect the historical influences of each specific age that are not captured by the model.

There are many commonly used link functions, and their choice can be somewhat arbitrary as Haberman and Renshaw (2008) point out. Cossette et al. (2007) using, for example, the complementary log-log (cloglog) link. The rationale for choosing these links is the fact that as we work with mortality tables whose contents are raw estimates of q_{xt}, the application of model (3) could produce estimates greater than 1 (Lee, 2000). Our choice is the logit of death probability q_{xt},

$$\log\left(\frac{q_{xt}}{1 - q_{xt}}\right) = a_x + b_x k_t + \epsilon_{xt}. \tag{4}$$

Debón et al. (2008) show how to obtain the parameter estimation by means of the SVD method (Lee and Carter, 1992), generalized linear models GLM (Currie et al., 2004) and maximum-likelihood ML (Brouhns et al., 2002). This model presents a problem of identifiability, as given a solution of (3), (a_x, b_x, k_t), any transformation of the type $(a_x, b_x/c, ck_t)$ or $(a_x + cb_x, b_x, k_t - c)$, $\forall c$, is also a solution. In order to avoid this difficulty and to get a single solution, some constraints must be imposed on the parameters. Although Lee and Carter

(1992) propose the normalization $\sum_x b_x = 1$ and $\sum_t k_t = 0$, we propose other constraints.

2.2.1 Li and Lee model

Li and Lee (2005) propose a variant of the original Lee-Carter model for estimating mortality in countries that form part of a group, instead of considering them individually. Though Li and Lee proposed this model for different countries, here it is applied to the different Autonomous Communities in Spain. Their model is expressed by,

$$log(m_{xti}) = a_{xi} + b_x k_t + b_{xi} k_{ti} + \epsilon_{xti}. \tag{5}$$

The rationale for this model is that to avoid long-term divergence in mean mortality forecasting for a group, a necessary and sufficient condition is that all the populations in the group have the same b_x and the same drift term for k_t (Li and Lee, 2005). The a_x are estimated separately for each individual population in this group, a_{xi} for country i, since they do not cause long-term divergence and hence do not need to be the same for each population. The authors propose the adjustment of the model by means of SVD.

We have adapted this model to the logit(q_{xti}),

$$\text{logit}(q_{xti}) = a_{xi} + b_x k_t + b_{xi} k_{ti} + \epsilon_{xti},$$

whose adjustment by means of SVD is as follows.

- The estimate a_{xi} is

$$\hat{a}_{xi} = \frac{\sum_t \text{logit}(\hat{q}_{xti})}{T},$$

  with T number of years.

- The values b_x and k_t, common factors for all populations, are estimated by singular value decomposition (SVD) applied to the matrix,

$$\frac{\sum_i (\text{logit}(\hat{q}_{xti}) - \hat{a}_{xi})}{r},$$

  where r is the number of regions. After b_x and k_t have been incorporated into the model, if the variance reduction obtained with this reduced model is small, then the procedure continues to incorporate the specific term for each region, otherwise the procedure stops.

- The specific factors for the ith population, b_{xi} and k_{ti}, are obtained using the first-order vectors derived from applying the SVD to the residual matrix of the common factor model,

$$\log\left(\frac{q_{xti}}{1 - q_{xti}}\right) - \hat{a}_{xi} - b_x k_t, \; i = 1, \ldots, r.$$

5

Then $\hat{k}_{ti}$ and $\hat{b}_{xi}$ are the respective first-right and first-left singular vectors for $i = 1, ..., r$.

Finally, the solution is denoted as $(\hat{a}_{xi}, \hat{b}_x, \hat{k}_t, \hat{b}_{xi}, \hat{k}_{ti})$.

Generalized non-linear models can be used as an alternative method for estimating the model (5), as Turner and Firth (2006) do for the force of mortality. Due to the large number of parameters, this approach produces computational problems.

The last step, prediction beyond the period under observation, is carried out by adjusting two $ARIMA$ models to the time parameters series. So the prediction for year $t_n + s$ is

$$\text{logit}(\hat{q}_{x,t_n+s,i}) = \hat{a}_{xi} + \hat{b}_x \hat{k}_{t_n+s} + \hat{b}_{xi} \hat{k}_{(t_n+s)i}.$$

As in the case of Brass's model, this model also has some weaknesses:

1. It cannot be estimated for regions with $q_x = 0$ or $q_x = 1$.

2. It has a large number of parameters and, in particular, the specific factor in every region, $b_{xi}k_{ti}$, can produce excessive divergences between projections for the different regions.

3. The a_{xi} that it provides behave erratically for regions with small populations.

4. It does not provide any index for the comparison of regions.

In order to avoid the divergences indicated in item 2 and to compare the estimations obtained with the models that follow, we have used a reduced version of this model that suppresses the specific term.

That is,

$$\text{logit}(\hat{q}_{xti}) = a_{xi} + b_x k_t + \epsilon_{xti}, \qquad (6)$$

the problem of identifiability is solved with the following restrictions, $k_{t_0} = 0$ and $b_0 = 1$. Model (6) is called the common factor model by Li and Lee (2005). In this model, trends are assumed to be common to all the regions, which is reasonable given that the focus is on the different regions of the same country. Moreover, we propose a small improvement to the adjustment of this model that consists of obtaining the estimations b_x and k_t for maximum likelihood once the values of the a_{xi} have been fixed with an offset term. These values are easily obtained with the package gnm, providing the solution for SVD as an initial point. This considerably improves the adjustment.

The interpretation of the parameters is as follows,

1. a_{xi} coefficients describe the average shape of the age profile in the region i.

2. The evolution of b_x gives an idea of how fast the ratios decrease in response to changes in k_t for the country.

3. The values k_t represent the national tendency of mortality during the period.

The listed weaknesses of Brass's model, and those for this last one, lead us to consider two other models that avoid them and, in particular, permit comparison between different regional units, adding only one new parameter for each of them.

2.2.2 Russolillo-Giordano-Haberman (RGH) model

By analogy with the Lee-Carter models and in order to obtain a parsimonious model, Russolillo et al. (2010) propose adding a factor index that specifically modifies mortality for each member of the group. The proposed model is,

$$\log\left(\frac{q_{xti}}{1 - q_{xti}}\right) = a_x + b_x k_t I_i + \epsilon_{xti}. \tag{7}$$

The problem of identifiability is solved by setting $k_{t_0}=0$, for some t_0, and $b_0=1$ and $I_1=1$. Then, for t_0 we can write (7) in the form,

$$\text{logit}(q_{xt_0i}) = a_x + b_x k_{t_0} I_i = a_x,$$

and a_x's are the logit of the probabilities of death for each age and for each region during the period t_0. Consequently,

$$q_{xt_0i} = \frac{\exp a_x}{1 + \exp a_x}.$$

Comparing the odds of the probabilities of death in any period t with those of the reference period t_0, for any region i, we have

$$\log(OR) = \log\frac{q_{xti}/(1 - q_{xti})}{q_{xt_0i}/(1 - q_{xt_0i})} = a_x + b_x k_t I_i - a_x = b_x k_t I_i. \tag{8}$$

In addition for $I_1=1$, $\log\left(\frac{q_{xt1}}{1-q_{xt1}}\right) = a_x + b_x k_t$.

The following is the interpretation of the parameters,

1. a_x coefficients describe the average shape of the age profile for the period t_0.

2. The evolution of b_x gives an idea of how fast the ratios decrease in response to changes in k_t for Region 1,

$$\frac{d\text{logit}(q_{xt1})}{dt} = b_x \frac{dk_t}{dt}.$$

3. The evolution of $b_x I_i$ gives an idea of how fast the ratios decrease in response to changes in k_t for Region i,

$$\frac{d\text{logit}(q_{xti})}{dt} = b_x I_i \frac{dk_t}{dt}.$$

4. The values k_t represent the tendency of mortality during the period.

7

2.2.3 A new model

Our proposal, henceforth called the Additive model, seeks a model similar to (7) but simpler, because we add the region as an additive index. This new model can easily be compared with (7) as both share the same number of parameters. Its expression is,

$$\log\left(\frac{q_{xti}}{1-q_{xti}}\right) = a_x + b_x k_t + I_i + \epsilon_{xti}. \tag{9}$$

This model, as in (7), is adjusted by Maximum Likelihood (ML).

The problem of identifiability is solved, as in the above model, by putting $k_{t_0}=0$, $b_0=1$, but $I_1=0$. Then, for t_0 we can write (9) in the form,

$$\text{logit}(q_{xt_01}) = a_x + b_x k_{t_0} + I_1 = a_x, \tag{10}$$

and a_x's are the logit of the probabilities of death for each age in the period t_0 for Region 1 . Consequently,

$$q_{xt_01} = \frac{\exp a_x}{1 + \exp a_x}.$$

Comparing the odds of the probabilities of death in any period t with those of the reference period t_0 for Region 1, we have

$$\log(OR) = \log \frac{q_{xt1}/(1-q_{xt1})}{q_{xt_01}/(1-q_{xt_01})} = a_x + b_x k_t - a_x = b_x k_t. \tag{11}$$

Comparing the odds of the probabilities of death in any region i with those of the reference Region 1, we have

$$\log(OR) = \log \frac{q_{xti}/(1-q_{xti})}{q_{xt1}/(1-q_{xt1})} = a_x + b_x k_t + I_i - (a_x + b_x k_t) = I_i. \tag{12}$$

In addition for $I_1=0$, $\log\left(\frac{q_{xt1}}{1-q_{xt1}}\right) = a_x + b_x k_t$.

The following is the interpretation of the parameters:

1. a_x coefficients describe the shape of the age profile in Region 1 for the period t_0.

2. The evolution of b_x gives an idea of how fast the ratios decrease in response to changes in k_t.
$$\frac{d\text{logit}(q_{xti})}{dt} = b_x \frac{dk_t}{dt},$$

3. The values k_t represent the trend of mortality in Region 1 during the period.

4. The succession of values I_i gives an idea of the differences in the pattern of mortality in any region i with respect to Region 1.

2.3 Comparing model structures

As far as age parameters are concerned, model (6) differs from (7) and (9) in the a_{xi}, which are specific for each region, while a_x is a general term for the whole population. In the three models, b_x is a common parameter for whole population. With regard to time parameters, they all contain a common parameter k_t which reflects the behaviour of the whole population over the time period. The errors ϵ_{xti}, with a zero mean and variance σ_ϵ^2, reflect the historical influences of each specific age that are not captured by the model.

As regards (7) and (9), both models differ essentially in the meaning of their I_i indexes. In (7) the index stands for the specific change shown in each region by the increments, $b_x k_t$, taking place with respect to the general behaviour of the logit of mortality, a_x. In (9), taking the whole population of the country as a reference, and making $I_1 = 0$, the index means the additive change necessary for transforming the $\text{logit}(q_{xt1}) = a_x + b_x k_t$ in Region 1 to that of a region i. In addition, this model assumes that differences in the mortality of specific regions are age and time independent.

The results of the application of all of these models to real data are provided in Section 3.

3 Analysis of mortality data from Spain

All these models will be used to adjust mortality for Autonomous Communities in Spain. The adjustment is carried out through ML using the R Development Core Team (2005).

3.1 Description of the data

Figures 2 to 9 show the results obtained for male mortality data. Similar patterns are observed in the logit crude mortality rates for age, model parameters and deviance residual plots for female data. The findings for women are similar to those obtained for men, and for the sake of brevity they are not reproduced here. Data used in this analysis come from the INE (www.ine.es). In particular, we have worked with micro-mortality data reporting individual dates of birth and death. The crude estimates of q_{xti}, necessary for the models under study, were obtained with the new methodology recently proposed by the Spanish National Institute of Statistics (INE) (2009), based on Elandt-Johnson and Johnson (1980), pages 128-146, who explain that given complete, continuous-time observations on all births and deaths for all people in a population exposed to the risk of mortality, it is possible to produce direct estimates of the central mortality rates, m_{xt}, by means of

$$\dot{m}_{xt} = \frac{d_{xt}}{1/2 P_{xt} + 1/2 P_{x(t+1)} + \sum_i \delta_{xtj}}, \tag{13}$$

where d_{xt} are deaths in the year t at age x, and P_{xt} and $P_{x(t+1)}$ are the populations that are x years old on December 31st of year t and year $t+1$, respectively.
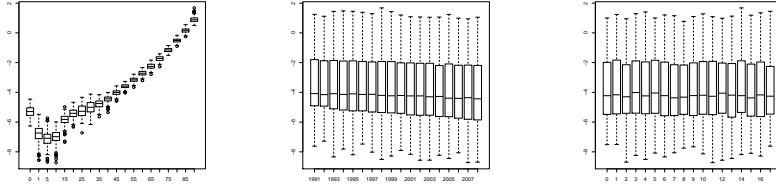
Figure 2: Behaviour of the logit of the crude mortality rates for age, time and region.

Finally, δ_{xtj} is defined as the difference, in years, between the date of death and the birthday in year t, of each individual j who dies in year t at age x. We can obtain $\dot{q}_{xt}$ from (13),

$$\dot{q}_{xt} = \frac{m_{xt}}{1 + (1 - a_{xt})m_{xt}}, \tag{14}$$

where a_{xt} is the average number of years that people dying in year t have lived between the ages x and $x + 1$. This concept was introduced by Chiang (1960, 1968, 1972) in a reformulation of the classic linear survival function model. The survivor functions, l_{xt}, and theorical deaths, d_{xt}, in the table are obtained recursively: $l_{0t} = 100000$, $d_{xt} = l_{xt}\dot{q}_{xt}$ and $l_{(x+1)t} = l_{xt} - d_{xt}$.

The life table for an Autonomous Community measures the incidence of mortality for its resident population. In order to avoid unwanted distortions of the results provided by a complete life table for simple ages, we abridged a life table with 5-year age groups as the Spanish National Institute of Statistics (INE) (2009) do in its publications. This can make interpretations difficult as a direct consequence of the variability of the information that small populations exhibit.

We have included 17 of the 19 Spanish Autonomous Communities in the analysis. Two of them, Ceuta and Melilla, were eliminated due to their very small population size and because the tables obtained for them by the Spanish National Institute of Statistics (INE) (2009) have 90 as the final age and not 95 as for the other regions. Equations (13) and (14) were the applied to each region i to obtain the corresponding q_{xti} grouped into $q_{[x,x+n]ti}$, from the survivor function l_{xt} for ages $x = 0, 1, 5, \ldots, 95$ by means of the expression

$$q_{[x,x+n]ti} = \frac{l_x - l_{x+n}}{l_x},$$

for age groups $0, 1, [1, 5), [5, 9) \ldots, [95, \infty)$. Figure 2 shows the behaviour of the logit of the crude mortality rates according to age x, year t and region i.

The four models described in Section 2 have been used to adjust mortality data in the 17 Autonomous Communities for the period 1991-2008 and a range of ages from 0 to 90. The adjustments have been made separately for women and men.
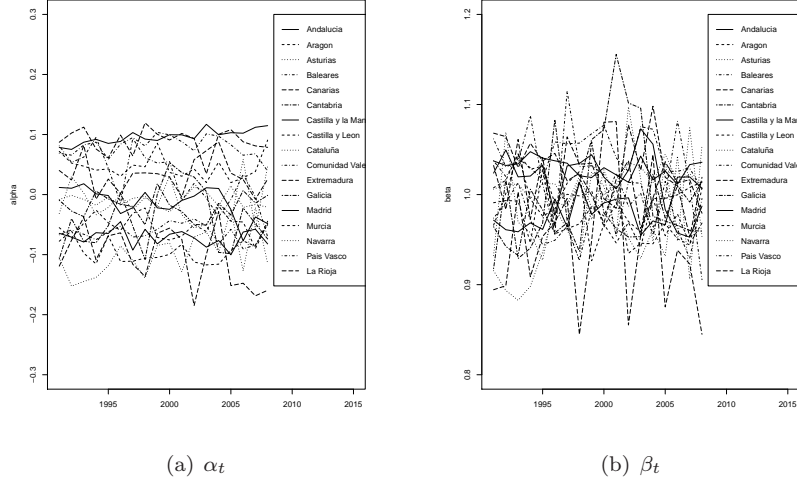
10

(a) α_t (b) β_t

Figure 3: Estimated values for logit Brass model.

3.2 Model fitting

We have included data for Spain as a region. In the case of models (1) and (9) it is used as a reference, and in models (5) and (7) to obtain more robust parameters that permit comparison with the other two.

Figures 3 and 4 show the estimations of the parameters obtained with the 4 models for the 17 regions. Firstly it is necessary to evaluate to what extent the proposed models are able to obtain mortality tables for the different regions that do not differ too much from each other, eliminate the irregularities in those with smaller populations and finally, respect their peculiarities. The first two conditions are analysed with comparative graphs of the different regions and the last by means of graphs and measures of goodness-of-fit.

Renshaw and Haberman (2006) suggest carrying out diagnostic checks on the fitted model by plotting residuals. These are done in Figures 5 to 8 with Deviance residuals,

$$r_{dev_{xti}} = sign(d_{xti} - \hat{d}_{xti})\sqrt{2\left[d_{xti}\log\left(\frac{d_{xti}}{\hat{d_{xti}}}\right) + (E_{xti} - d_{xti})\log\left(\frac{E_{xti} - d_{xti}}{E_{xti} - \hat{d}_{xti}}\right)\right]}. \tag{15}$$
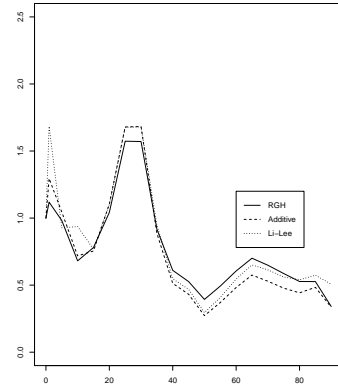
The model's performance is evaluated with the Deviance, a measure of the distance between observed $\dot{q}_{xti}$ and adjusted values $\hat{q}_{xti}$, whose expression is

$$D(\hat{q}) = 2\log L(\dot{q}_{xti}) - 2\log L(\hat{q}_{xti}), \tag{16}$$
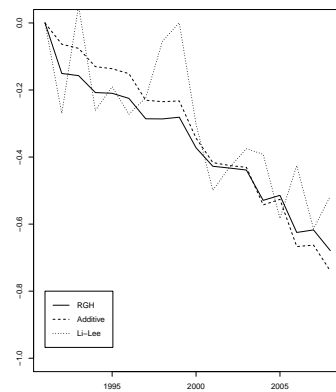
where $\log L()$ is the Binomial loglikelihood function, as we have assumed that the number of deaths is distributed as a Binomial. Table 1 shows the Deviance
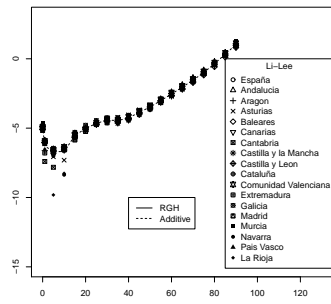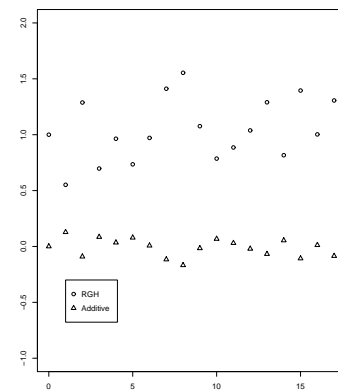
(a) a_x

(b) b_x

(c) k_t

(d) I_i

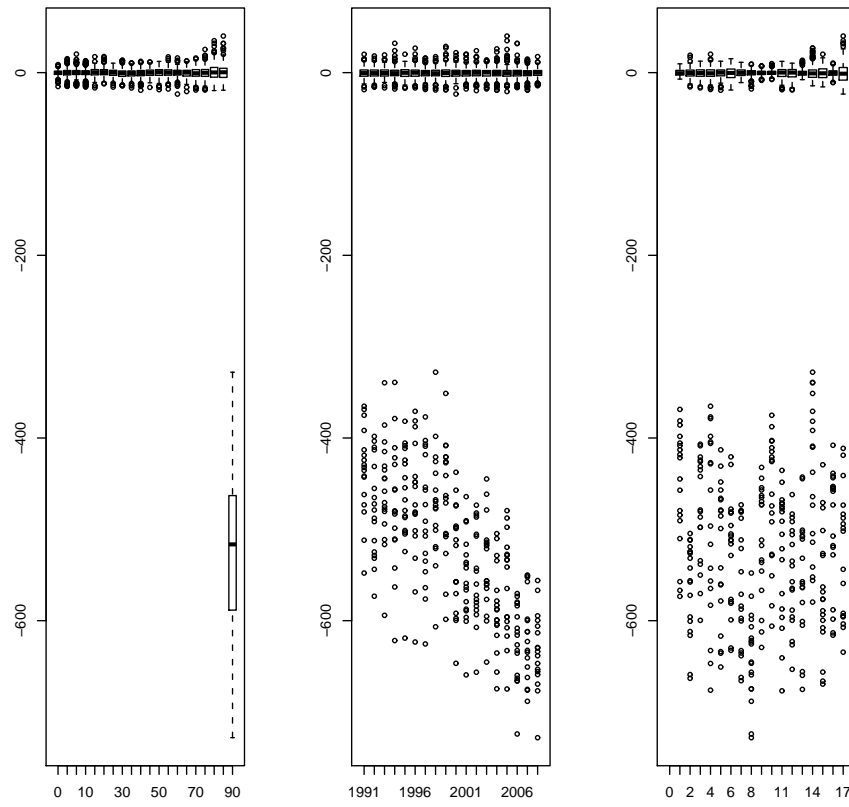Figure 4: Estimated values for Li and Lee, RGH and Additive models.

12

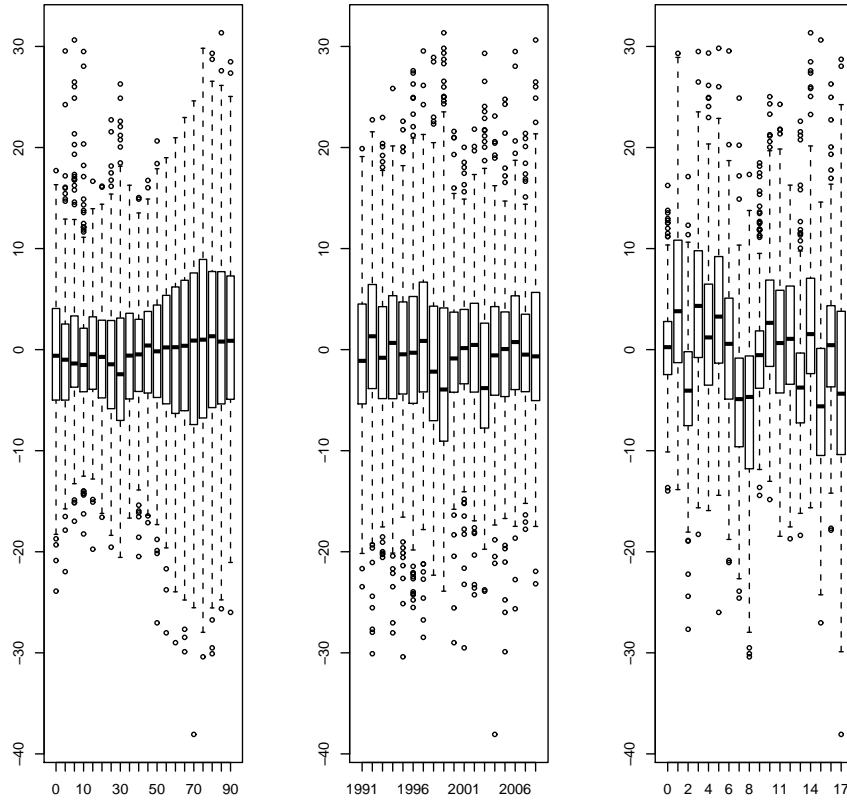Figure 5: Deviance residuals for the logit Brass model for men.

Figure 6: Deviance residuals for the Li and Lee model for men.

values for all models.

Deviance (16) is a global measure of the fit of the model; its evaluation for each age, period and region is carried out using the graphs of the residuals (15) which are shown in Figures 5 to 8.

The predictions beyond the last time period are carried out by the projection of time series previously adjusted to the time parameters. The corresponding $ARIMA$ models are obtained using the functions $auto.ARIMA$ and $forecast$ from the R package forecast (Hyndman, 2008).

Fitted time series were studied using the function $tsdiag$, which produces a diagnostic output containing a plot of the residuals, the autocorrelation of the residuals and the p-values of the Ljung-Box statistic for the first 10 lags. As an example, Figure 9 shows the diagnoses of the series of the k_t for RGH model
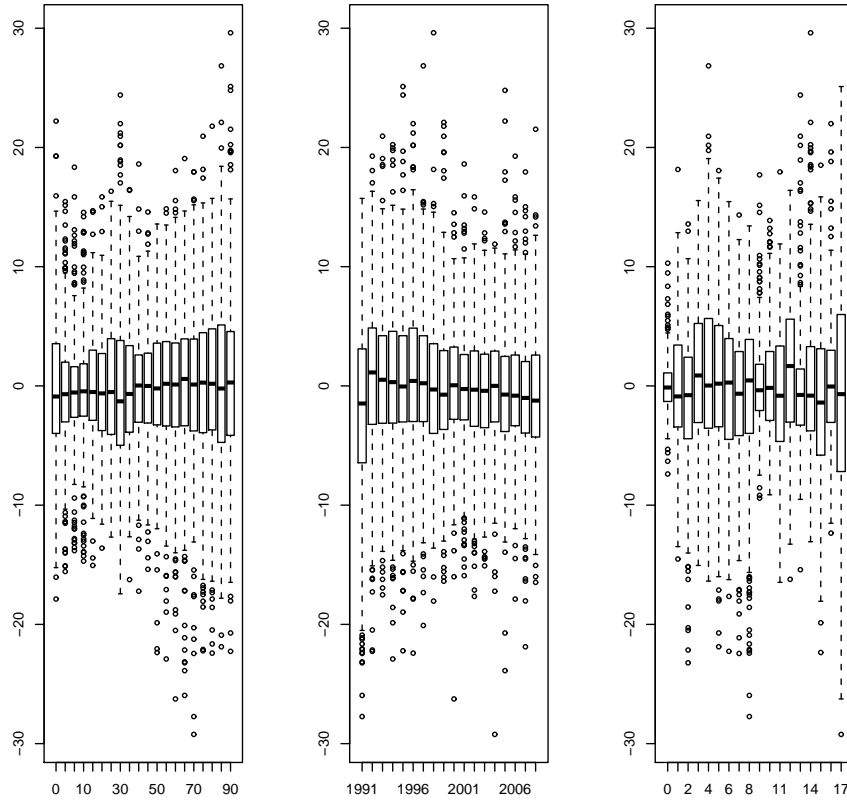
14

Figure 7: Deviance residuals for the RGH model for men.

applied to men. It should be highlighted that Brass's model (1) has $34 = 2 \times 17$ time parameters. A similar problem occurs with the Li and Lee model (5), but as we said previously, for the purposes of comparison we used expression (6) which contains one k_t and avoids the problem.

Figures 10 and 11 show the geographical distribution of the indexes corresponding to each of the regions. The results seem to be contradictory, as small index values of the Additive model correspond to large index values in the RGH model. This is because a negative index in the Additive model indicates that the region has a mortality rate below the reference one, and therefore a greater reduction in mortality which corresponds to an increased parameter b_x, giving an index greater than the one in the RGH model.

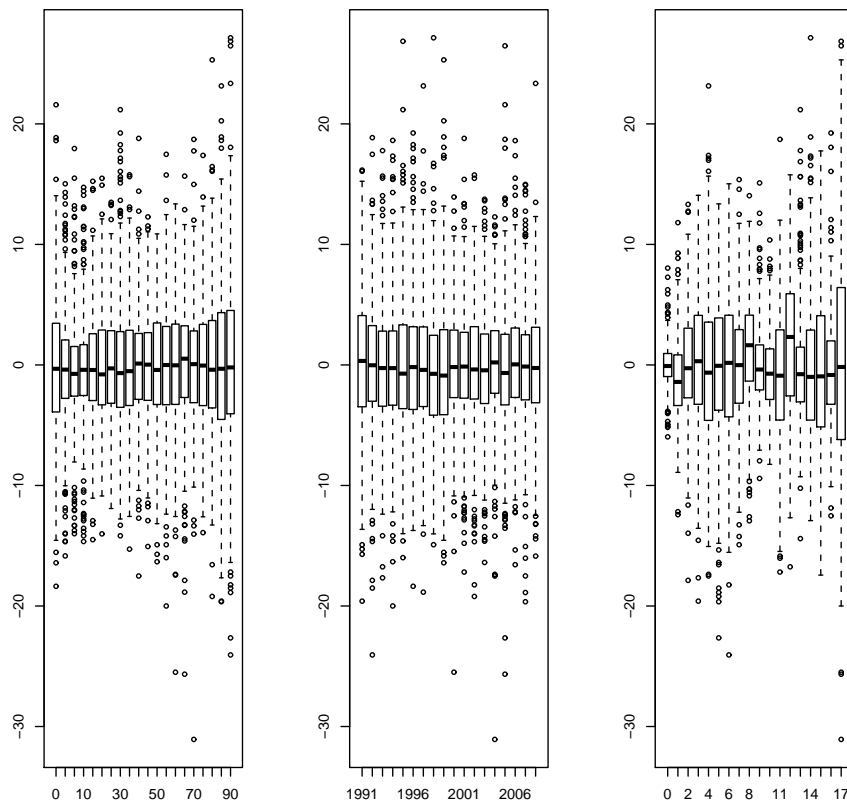The maps in Figures 10 and 11 show a north/south trend, mortality being

15

Figure 8: Deviance residuals for the Additive model for men.

		Deviance	
Model	number of parameters	Men	Women
Li-Lee	17*20+18+20	426151	481672
RGH	20+18+20	234820	209829
Additive	20+18+20	187270	171077
Brass	18*17+18*17	212659504	85733193

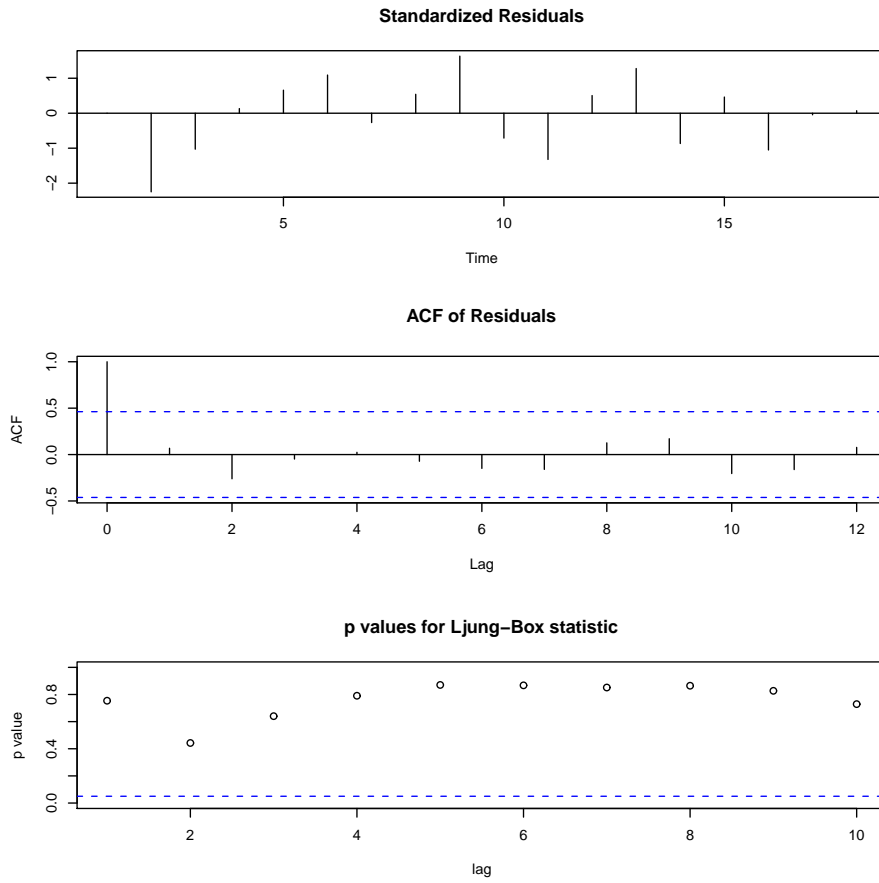Table 1: Goodness-of-fit for the different models

16

Figure 9: Output from tsdiag for RGH model mortality index for men.

lower in the regions in the north and centre than in the south of Spain.

4 Conclusions

Table 1 shows the goodness-of-fit values for different models. A first conclusion, common to all models, is that fitting performs better for women than for men. This may be due to the fact that male mortality fluctuations for the ages in the accident hump are difficult to capture over the period of time under consideration.

In general, the inclusion of the regional effect improves the models, as can be seen by comparing the residual graphs in Figure 5 with the graphs in Figures 6 to 8. On the other hand, the Additive model shows the best global result for
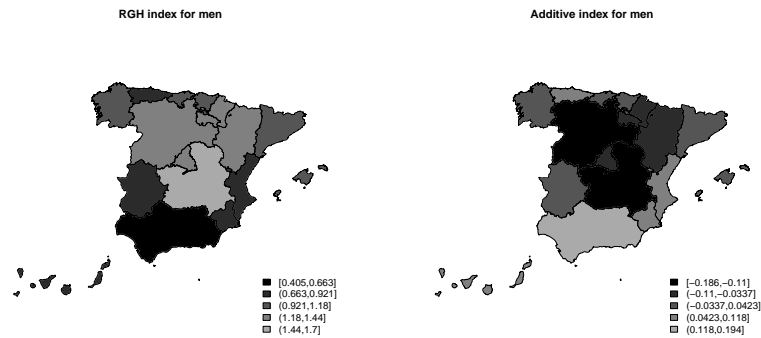
Figure 10: Geographical index for RGH (left) and Additive model (right) for men.
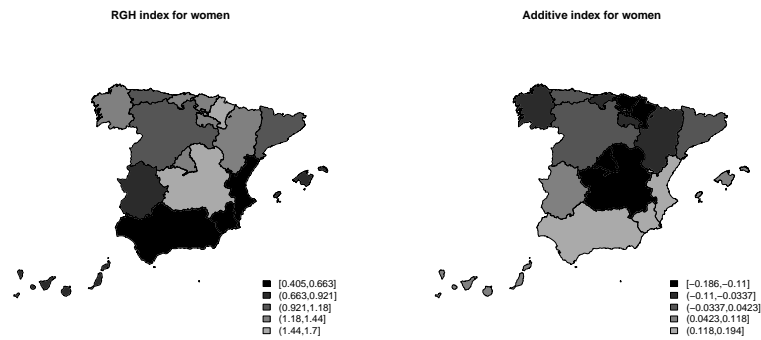

Figure 11: Geographical index for RGH (left) and Additive model (right) for women.

both sexes and for the goodness-of-fit measures. The explanation for this can be found in the introduction of the regional effect as an additive term, which better adapts the model for the regions involved in the study and for advanced and intermediate ages. Note that the Additive model has the simplest structure because it only considers the main effects assuming, as mentioned in Section 2.3, that differences in the mortality of specific regions are age and time independent.

The greatest residuals are observed in the advanced ages. This fact is observed in Figures 5 to 8, which specify the magnitude of residuals for all ages and models. These figures also confirm the point in the above paragraph.

One comment must be made with regard to the Additive model. In addition

18

to obtaining the best results, it has the advantage that its parameters are easily interpretable in as much as they describe the evolution of mortality over age, period and region, its computational cost is very low as it only needs an ARIMA model for forecasting, the comparison between regions is reduced to a unique index and it is a robust model considering the outliers. For these reasons, we think that this type of model must be borne in mind for future development.

We have focused our work on fitting and predicting mortality ratios instead of using mortality indicators for life expectancy. The reason for this is that to gain an understanding of forecasting error, the evaluation of error in log death rates is essential (Booth et al., 2006).

In line with the Goerlich and Pinilla (2009) results, we can confirm the marked north-south pattern in mortality which they observed more clearly for 2006 than for 1975 at the beginning of the period studied. In our case it is seen for the whole period 1991-2008.

Acknowledgments

References

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22(3):547–582.

Booth, H., Hyndman, R., Tickle, L., and de Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, 15(9):289–310.

Brass, W. (1971). *Biological aspects of Demography*, chapter On the scale of mortality, pages 69–110. Taylor & Francis, London.

Brouhns, N., Denuit, M., and Vermunt, J. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics & Economics*, 31(3):373–393.

Chiang, C. (1960). A stochastic study of the life table and its applications, ii: Sample variance of the observed expectations of life and other biometric functions. *Human Biology*, 32.

Chiang, C. (1968). *Introduction to Stochastic Processes in Biostatistics*. Wiley, New York.

Chiang, C. (1972). On constructing current life tables. *Journal of the American Statistical Association*, 67.

Cossette, H., Delwarde, A., Denuit, M., Guillot, F., and Étienne Marceau (2007). Pension plan valuation and mortality projection: A case study with mortality data. *North American Actuarial Journal*, 11(2):1–34.

Currie, I., Kirkby, J., Durban, M., and Eilers, P. (2004). Smooth Lee-Carter models and beyond. In *Workshop on Lee-Carter Methods*, www.ma.hw.ac.uk iain/workshop/workshop.html (4th March 2005).

Debón, A., Montes, F., and Puig, F. (2008). Modelling and forecasting mortality in Spain. *European Journal of Operation Research*, 189(3):624–637.

Elandt-Johnson, R. and Johnson, N. (1980). *Survival Models and Data Analysis*. Wiley, New York.

Goerlich, F. and Pinilla, R. (2009). Tablas de mortalidad para españa y sus regiones: 1975-2006 y esperanzas de vida libres de discapacidad por sexo y comunidad autónoma: 2004-2006. base de datos y principales resultados. Technical report, Instituto Valenciano de Investigaciones científicas (Ivie).

Haberman, S. and Renshaw, A. (2008). *2008 Living at 100 Monograph*, chapter On Simulation-Based Approaches to Risk Measurement in Mortality with Specific Reference to Binomial Lee-Carter Modelling. Society of Actuaries.

Hyndman, R. J. (2008). *forecast: Forecasting functions for time series*. R package version 1.11.

Lee, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, 4(1):80–91.

Lee, R. and Carter, L. (1992). Modelling and forecasting U. S. mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Li, N. and Lee, R. (2005). Coherent mortality forecast for a group of populations: an extension of the Lee-Carter method. *Demography*, 42(3):575–593.

Murray, C., Ferguson, B., Lopez, A., Guillot, M., Salomon, J., and Ahmad, O. (2003). Modified logit life table system: principles, empirical, validation, and application. *Population Studies*, 57(2):165–182.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Renshaw, A. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics & Economics*, (3):556–570.

Russolillo, M., Giordano, G., and Haberman, S. (2010). Extending the lee carter model: a three-way decomposition. *Scandinavian Actuarial Journal*, Forthcoming.

Spanish National Institute of Statistics (INE) (2009). Tablas de mortalidad. metodología. Technical report, Instituto Nacional de Estadística.

Turner, H. and Firth, D. (2006). *Generalized nonlinear models in R: An overview of the gnm package.* R package version 0.9-1.

Wilmoth, J., Andreev, K., Jdanov, D., and Glei, D. (2007). Methods protocol for the human mortality database, versión 5. Technical report, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).