Mortality Confidence Intervals: A measure of prediction goodness

Federico López-Carrión[(1)], José Luís Gutiérrez de Mesa[(2)] and Luís Felipe Rivera-Galicia[(3)]

Departamento de Estadística, Estructura económica y OEI
Facultad de CC.EE. y Empresariales. University of Alcalá, Spain
[(1)]federico.lopez.carrion@ine.es, [(2)]joseluis.gutierrez@uah.es, [(3)]luisf.rivera@uah.es

Abstract

With prediction, error measurement is a primary goal of data analysis. There is no exception for mortality projections. In this paper we develop a method to determine the confidence interval for mortality rates in small areas, whose population is much more exposed to random risks than the National population as a whole, much more stable in time for each age. The main result is a predictable confidence interval for the province death risk as a function of the National death risk. We specify a model for the death risk of a province as a sum of the National death risk plus an error term normally distributed whose variance depends on the National value and two predictable parameters. We apply the model to the Spanish particular case and observe an exponential relationship between Spain's death risk and the estimated variance. The logarithm of the estimated variance is linear in terms of the logarithm of Spain's death risk.

Key words: Mortality, risk of death by age, confidence interval, life table.

Date of last revision: May 2012

1. Introduction

One of the most interesting problems for a society is to quantify the risk of death. To estimate accurately the probability of death at different ages is very important to Public and private sectors. Certain State institutions need reliable data on probabilities of death, or to calculate population projections and establish appropriate policies for education, health and other measures, or to calculate life expectancies for determining certain benefits such as retirement wages. Moreover, managers need life insurance estimates of the probability of death of insured individuals to set adequate premiums to cover compensation for risk of death.

The probabilities of death by age cannot be calculated for a particular individual, but have to be estimated for a generic individual of a population possessing certain socio-economic characteristics, from the number of deaths and population size. In Spain, the probabilities of death by age are included in the annual life tables published by the Spanish National Statistics Institute (INE) at national and provincial level since 1991. It can be observed that the estimations at the provincial level maintain the trend of the national level; nevertheless, they present large variations, due to estimation sensibility under unexpected factors when population size is not big enough.

Assuming that annual incidence of mortality is equal in all provinces so that the risk of death observed in each province depends on national risk of death, this paper aims to construct a confidence interval to estimate the risk of death by age of any province from national data. Moreover, this confidence interval can be used to ascertain the validity of future estimates of risk of death at the provincial level.

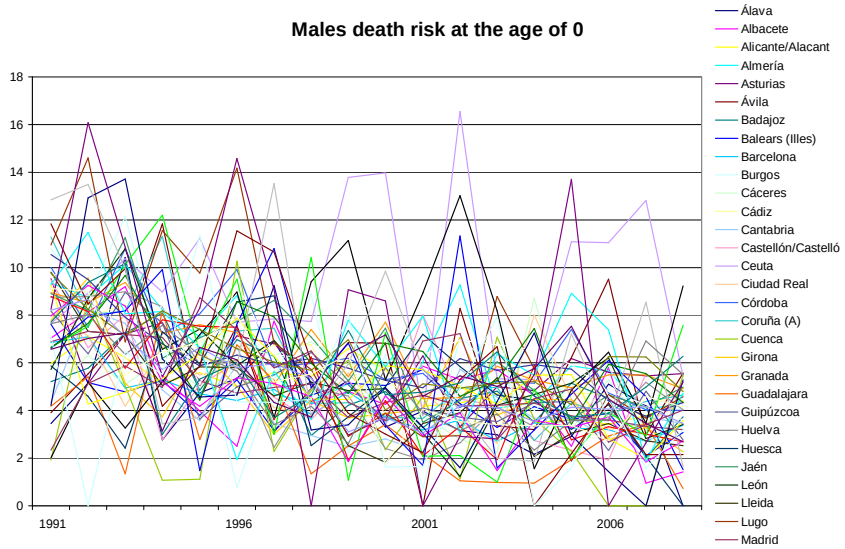
This work is organized as follows: in sections 2 and 3 we present the data and the specification of the model. In section 4 we estimate the variance. In section 5 we test the model assumptions. In section 6, we compare rate and risk mortality in terms of dispersion, and in section 7, the main conclusions of the paper are presented.

2. Data

Data is available in the national and 52 provincial life tables since 1991 to 2008 and ages from 0 to 94 and both sexes. Mortality tables are calculated according to the methodology of the Spanish National Statistical Institute base on the HMD protocol. In this paper we study the mortality rate and risk and both for males and females.

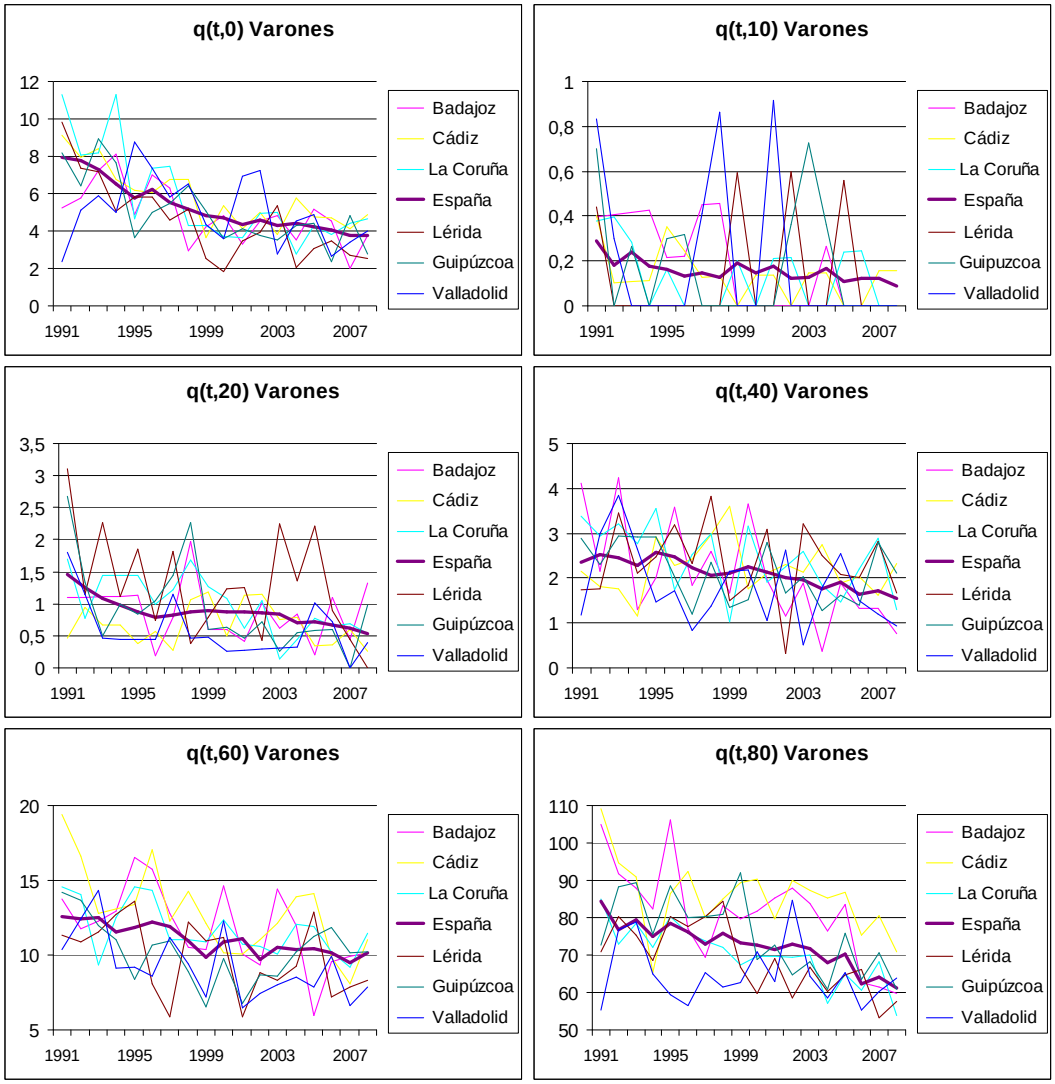
In Figure 1, the male death risk at the age of 0 is represented, from years 1991 to 2008. The cloud observed follows the National tendency. Each province seems to behave with a certain randomness degree inside a quite regular range.

Figure 1. Male death risk at the age of 0 for the 52 Spanish provinces from 1991 to 2008.



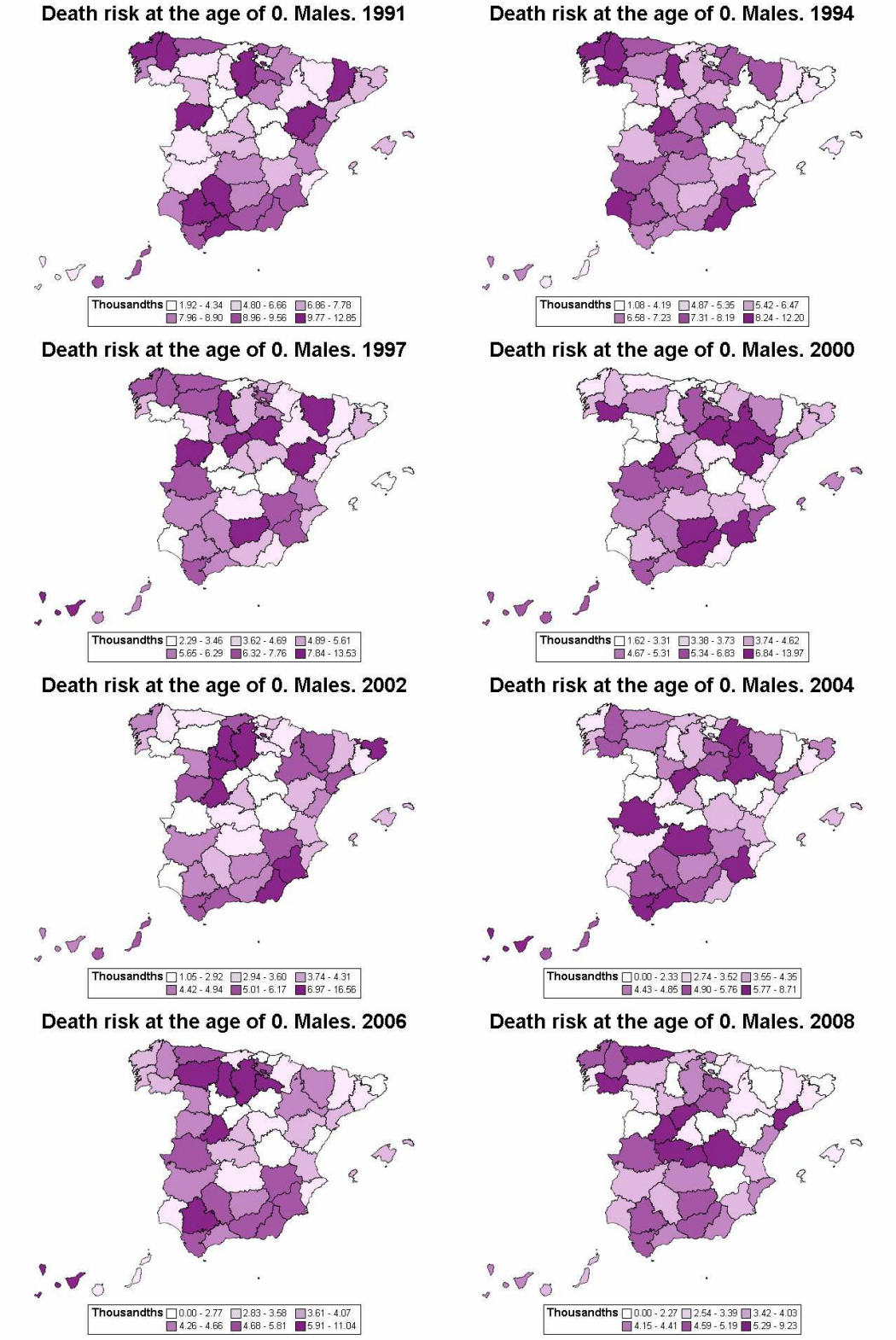
In addition, to clarify the analysis we select six representative middle sized provinces and plot their death risk series for ages 0, 10, 20, 40, 60 and 80. We observe the range evolution with age and check how variation grows not with age but with the mean of the value of the National risk of death. Figure 2 shows how range grows. First, the age of 10 with a range of 1 and a mean around 0,2; second, age of 20 with a range of about 2 and mean around 1; third, age 40 with a range of 3 and mean around 2; forth, age 0 with a range of 4 and mean around 5; fifth, age 60 with a range of 10 and mean around 10 and sixth, age 80 with a range of 30 and mean around 70.

Figure 2. From left to right, from top to bottom, graphs for males' death risk for six provinces and National value. At the age of 0, 10, 20, 40, 60 and 80.



According to maps in Figure 3 for these provinces, relative sample position of each provincial value change from year to year and can go from the lower values to the upper values with no apparent pattern. There are certain factors that condition the value such as the size of the population of the province that stabilises the tendency as it grows. Anyway there is no straight rule for the sign of the bias for each provincial value against the national.

Figure 3. From left to right, from top to bottom, graphs for males' death risk at the age of 0 for years 1991, 1994, 1997, 2000, 2002, 2004, 2006 and 2008.



In spite of all conditional factors against the randomness of provincial values empirical results lead us to conclude that it is possible to assume we are dealing with a pseudo random sample. Next, we propose a model to explain provincial behaviour from National values.

3. Model specification and confidence intervals

Let's propose the following model:

$$q(t,x,h) = q(t,x) + \varepsilon(t,x,h)\text{, for } h = 1 \text{ to } H\text{,} \tag{1}$$

where *t* is the year, *x* the age and *h* the province (in the Spanish case, *H* is 52, since there are 52 provinces)

This model is ruled by the following hypothesis:

I. Additivity. It is the sum of a deterministic part $q(t,x)$, National death risk and $\varepsilon(t,x,h)$ a random part or residual.
II. $\varepsilon(t,x,h)$ follows a normal distribution with 0 mean and variance $\sigma^2(t,x)$.

III. We also have:

$$\sigma(t,x) = a(t)\cdot q(t,x)^{b(t)} \tag{2}$$

Dividing (1) by $\sigma(t,x)$ we get the standardized model:

$$q^*(t,x,h) = q^*(t,x) + \varepsilon^*(t,x,h) \tag{3}$$

with

$$q^*(t,x,h) = \frac{q(t,x,h)}{a(t)\cdot q(t,x)^{b(t)}} \tag{4}$$

and

$$q^*(t,x) = \frac{q(t,x)}{a(t)\cdot q(t,x)^{b(t)}}\,. \tag{5}$$

In this case, the standardized residual, $\varepsilon^*(t,x,h) = \dfrac{\varepsilon(t,x,h)}{a(t)\cdot q(t,x)^{b(t)}}$, follows a normal distribution with cero mean and variance equal to one. According to the hypothesis (III) of model we get:

$$P\left(\left|q^*(t,x,h) - q^*(t,x)\right| \le z_{1-\alpha/2}\right) = 1-\alpha/2\,, \tag{6}$$

and the following confidence interval is found:

$$q(t,x,h) \in \left[\, q(t,x) \pm z_{1-\alpha/2}\cdot a(t)\cdot q(t,x)^{b(t)} \,\right],$$

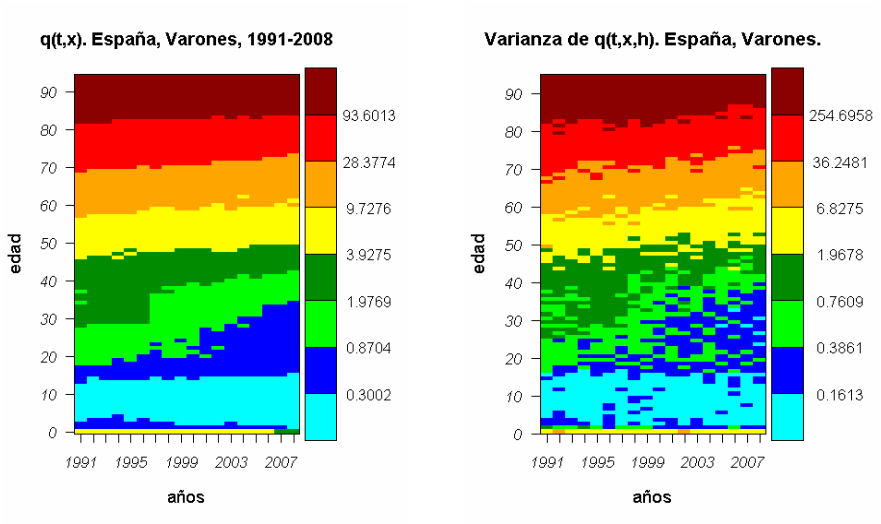
where $z_{1-\alpha/2}$ is such as $P\left[\left|X\right| < z_{1-\alpha/2}\right] = 1-\alpha$ for *X* a standardized normal random variable.

4. Variance estimation

According to the model proposed in (1), $\varepsilon(t,x,h) = q(t,x,h) - q(t,x)$ for each province. Error mean and variance are estimated, respectively, by $\bar{\varepsilon}(t,x) = \dfrac{\sum_{h=1}^{H} \varepsilon(t,x,h)}{H}$, and $\hat{\sigma}^2(t,x) = \dfrac{1}{H-1} \cdot \sum_{h=1}^{H-1} (\varepsilon(t,x,h) - \bar{\varepsilon}(t,x))^2$.

As a generalization of the study in section 2, Figure 4 shows that there is a growing parallelism between both graphs that can be explained through the exponential relation (III).

Figure 4. From left to right, observed National death risk and estimated residual variance. Death risk measured in thousandths.

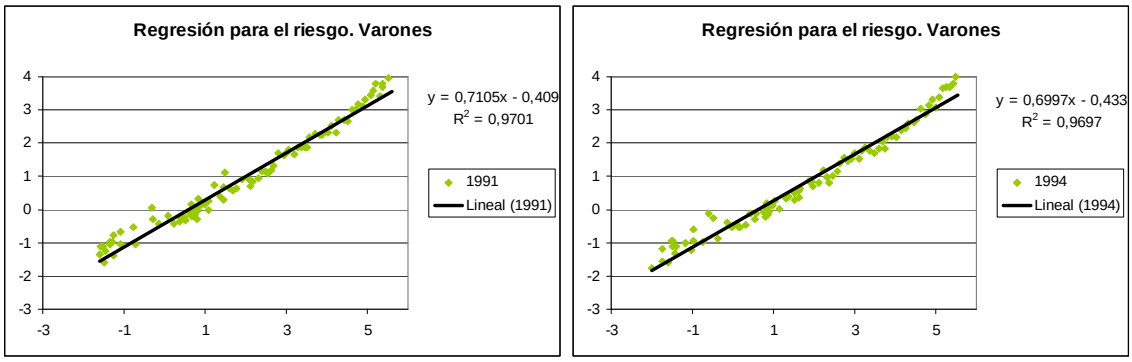


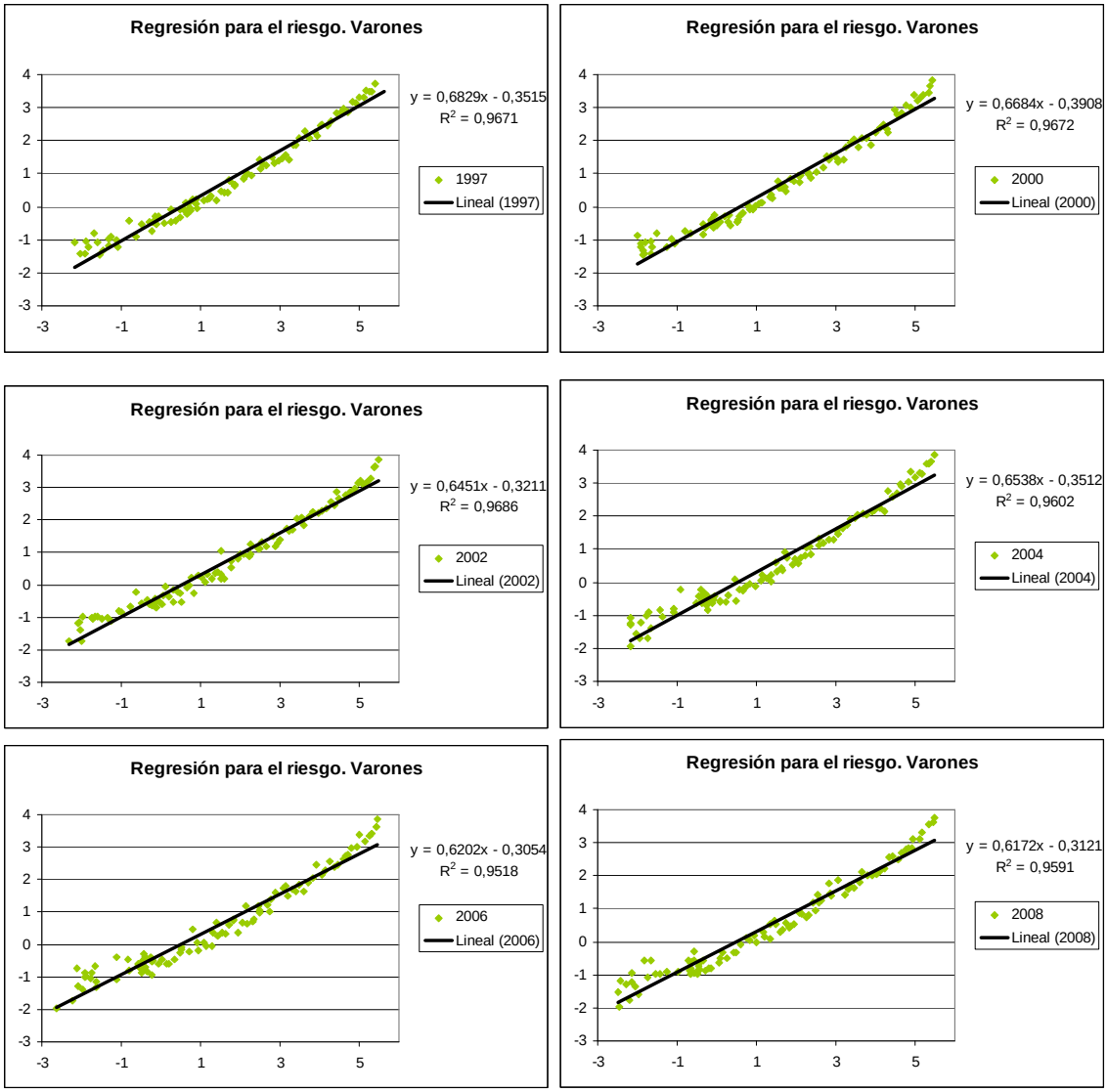
Applying logarithms in equation (2):

$$\ln \sigma(t,x) = \alpha(t) + b(t) \cdot \ln q(t,x) \text{ where } \alpha(t) = \ln a(t) \qquad (7)$$

We adjust a linear model to (7) by OLS using the observed values $q(t,x)$ and the estimated variance $\hat{\sigma}^2(t,x)$.

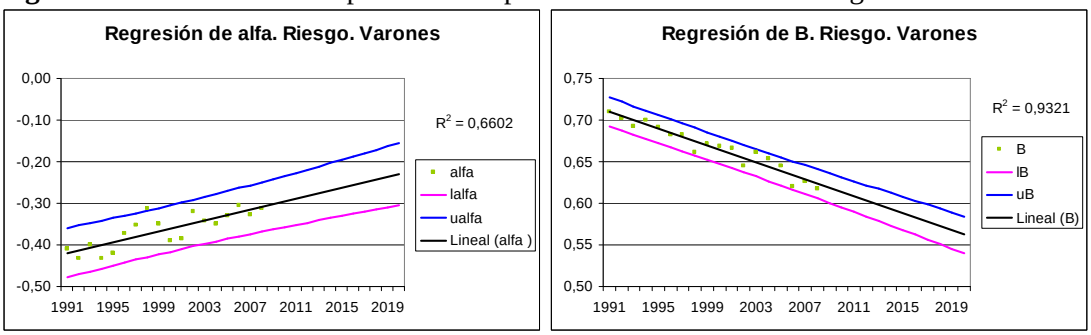
Figure 5. Linear regression for equation (7). Years 1991, 1994, 1997, 2000, 2002, 2004, 2006 and 2008.





We calculate estimators $\tilde{\alpha}(t)$ y $\tilde{b}(t)$ for respective parameters through OLS for each year from 1991 until 2008. We adjust a linear trend to the data. In Figure 6, results appear to be statistically significant. In addition, according to the parameters, we find that variance decreasing with time.

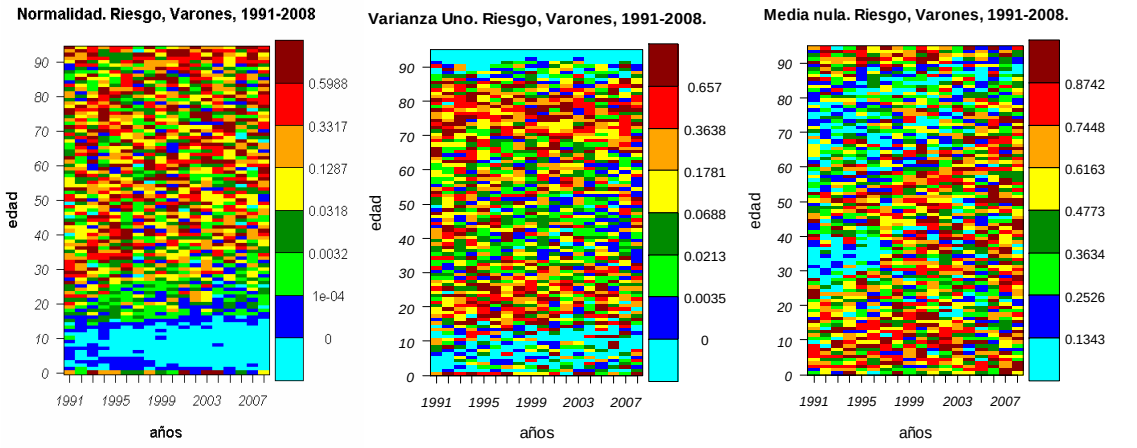
Figure 6. Time series for the parameters alpha on the left and b on the right.



5. Residual analysis

Under model hypothesis, the residual $\hat{\varepsilon}^*(t,x,h) = \dfrac{\varepsilon(t,x,h)}{\hat{\sigma}^*(t,x)}$ follows a normal distribution with cero mean and variance equal to one where $\hat{\sigma}^*(t,x) = \hat{a}(t) \cdot q(t,x)^{\hat{b}(t)}$. This estimation is different from $\hat{\sigma}^2(t,x)$.

Figure 7. From left to right. p-value for the Normality Shapiro-Wilks test, ji-square variance test and T-student test for males death risk.

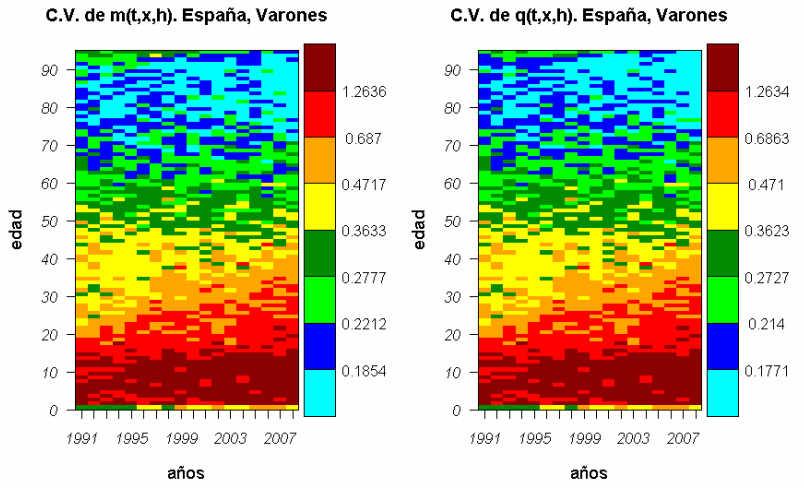


According to Figure 7, in the region with national death risk over a thousandth not enough evidence is found to reject the hypothesis of normality, variance equal to one and cero mean for the residuals. As a result the three hypothesis are accepted in that region.

6. Dispersion comparison between death risk and rate

We have done a similar research for death rate, and we have compared the results with those obtained with death risk. According to Figure 8 death risk performs better than death rate in terms of variability measured as the quotient of the standard deviation and the national value.

Figure 8. From left to right, quotient of the standard deviation and the national value for males death rate and risk.



7. Discussion

Some of our findings from data analysis are:

(a) Sample dispersion measured with de variance depends not on age but on the variable value. The magnitude of the variation depends on the size population but will not determine the sign of the bias.
(b) There is a functional relationship between the sample variance and the national value. More precisely, variance increase exponentially with the national value.
(c) Variance decrease with time. This result is independent of the age and sex. We empirically observe that variance decrease with time. Although parameter a(t) increases with time the strength of the potential part $q(t,x)^{b(t)}$, since both, q(t,x) and b(t) decrease, leads variance to decrease. In addition, we observe that for each two functions defined by (7) for two years that ordinates in the second year tend to lie under the first year in the region where ln q(t,x) is over a thousand. So variance will always be smaller for the second year in the region.
(d) Variance coefficient for death risk is smaller than the rates variance coefficient for both sexes. This result is independent of the size of the variable. There for it is better to use death risk as the projection variable.
(e) For the hypothesis acceptance region it is possible to obtain a provincial prediction confidence interval projecting the death risk. In addition this interval decreases with time.
(f) Normality of the distribution is rejected for regions with low national risk below a thousandth. This rejection is due to the fact that there is a group of provinces, that increases with time, where death risk for young ages is cero. There for there is a repeated value in the empirical distribution so it is not possible to assume normality.

There is still much more research insight left to do. At the present time, we are focussed on dealing with the rejection in the region with low risk. Our alternative and open research line is to use the sample to adjust a theoretical mixed distribution base on a beta distribution with an accumulative point in cero.

8. References

Alho, J. M.; Spencer, B. D. (2005). *Statistical Demography and Forecasting*. Springer Series in Statistics.

Buse, A. (1973). Goodness of Fit in Generalized Least Squares Estimation. The American Statisticians 27.

Camarda, C. G.; Durban, M. (2008). Goodness of fit in models for mortality data. Madrid: Universidad Carlos III de Madrid, Departamento de Estadística. Mimeo.

Currie, I. D., M. Durban, and P. H. C. Eilers (2004). Smoothing and Forecasting Mortality Rates.

Haberman, S. H. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

INE (2007). *Tablas de mortalidad de la Población Española 1992-2005*.

INE (2009). *Tablas de mortalidad. Metodología*. Electronic text. Available at http://www.ine.es/daco/daco42/mortalidad/metodo_9107.pdf

Gutiérrez de Mesa, J.L. (2002). *Revisión crítica de los actuales modelos de proyección de la población*. Mimeo.

J.R. Wilmoth(2007). *Human Mortality Database Protocol*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at. http://www.mortality.org/Public/Docs/MethodsProtocol.pdf

McCullagh, P.; Nelder, J. A. (1989). *Generalized Linear Model (2nd ed.)*. Monographs on Statistics Applied Probability. Chapman & Hall.

Myers, R.H. (2010). *Generalized Linear Models: with Applications in Engineering and the Sciences*. Wiley Series in Probability and Statistics.
Peña, D.; Tiao, G.C.; Tsay, R.S. (Eds.) (2001). *A Course in Time Series Analysis*. John Wiley, New York.
Peña, Daniel. (1998) Estadística. Modelos y métodos. Alianza Universidad Textos.
Preston, S. H., P. Heuveline, and M. Guillot (2001). Demography. Measuring and Modeling Population Processes. Blackwell.
Windmeijer, F. A. G. (1995). Goodness-of-Fit Measures in Binary Choice Models. Econometric
Sachia, R. M. (1992). The Box-Cox transformation technique: a review. The Statistician