# A Model for Mortality Forecasting Based on Self Organizing Maps

Marina Resta* and Marina Ravera

Department of Economics, University of Genova, via Vivaldi 5, 16126, Genova, Italy
{resta,ravera}@economia.unige.it
http://www.economia.unige.it

**Abstract.** In this paper we introduce a general model framework based on Self Organizing Maps (SOMs) to explore the behavior of populations mortality rates and life expectancy. In particular, we show how to employ SOM clustering capabilities to construct coherent mortality rates, i.e. mortality rates that can be applied unchanged to a wide range of countries. To such purpose, we will employ various countries mortality data downloaded from the Human Mortality Database. Our aim is two–fold. On the one hand, we are going to prove that a data mining approach can be meaningful to build mortality forecasts in a way which is less pretending (in terms of both computing time and parameters to estimate) than traditional techniques. This issue is very important, provided that mortality forecasts are widely employed to develop insurance products. On the other hand, we will show that SOM clustering can be very effective to extract similar mortality patterns from apparently very different countries, thus highlighting non–linear hidden features that are missing for more standard techniques.

**Keywords:** Longevity risk, Self Organizing Maps, Clustering, Mortality forecasting.

## 1 Background

Mortality forecasting is an important topic, as it may considered the basis of social and economic planning, and fundamental to many other forecasting exercises as well. In particular, in this paper we are concerned with the link existing between mortality trends and insurance contracts, namely those contracts providing individuals with annuities, pensions and other benefits paid during their lifetime (the so–called *living benefits*).

The main issue is of financial (and balancing) nature: on the one hand, paying benefits implies that insurance companies must have a proper *reserve*, i.e. a fund from which money can be retrieved; on the other hand, pensions and annuities are usually paid depending on proper amounts of money (premium) the individuals have conveyed throughout their active (i.e.: at work) life. The balance between such different amounts of money is guaranteed if and only if the behavior of

* Corresponding author.

P.A. Estévez et al. (Eds.): *Advances in Self-Organizing Maps*, AISC 198, pp. 335–343.
DOI: 10.1007/978-3-642-35230-0_34

mortality rates is correctly estimated. However, since mortality rates in many countries are persistently decreasing, the systematic misunderstanding of such behavior could lead to serious financial consequences in the longer term, as far as their premiums and reserves are concerned. This focus has led to identify *longevity risk* [9] as a new type of risk affecting the management of annuity and pensions portfolios.

Provided the importance of the issue, a number of methodologies have been proposed to model (and forecast) the dynamics of mortality rates, although it aids to remember that choosing of methodology is not without controversy, since it can lead to very marked difference in forecasts [7], [8]. Actually more popular models are trend–based, and they can be viewed as belonging to the research vein pionereed by the Lee–Carter model –LCM–[4], we will explain in detail in Section 2. In a nutshell, LCM assumes to represent mortality rates as functions of age $x$ and time $t$, identifying a single time index which summarises past trends, which affects mortality at time $t$ at all ages simultaneously, and which can be modelled with a view to extrapolation. Over the past decades several weaknesses of LCM have been highlighted, and various modification of the original model have been suggested (see among others: [3], [1], [6]).

Despite of the wide literary corpus, however, the techniques actually in use are of heavily statistical type, and soft computing approaches are rather unexplored. With this is mind, we are going to introduce a general model framework based on Self Organizing Maps (SOMs) [2], to explore the behavior of populations mortality rates. In particular, we will focus on so–called coherent models, and we will explore mortality data of various countries (downloaded from the Human Mortality Database–HMD) in search of similar mortality experiences. In this way we will be able to show how to employ SOM clustering capabilities to construct coherent mortality rates, i.e. mortality rates that can be applied unchanged to a wide range of countries. Our aim is two–fold. On the one hand, we are going to prove that a data mining approach can be meaningful to build mortality forecasts in a way which is less pretending (in terms of both computing time and parameters to estimate) than traditional techniques. On the other hand, we will show that SOM clustering can be very effective to extract similar mortality patterns from apparently very different countries, thus highlighting non–linear hidden features that are missing for more standard techniques.

The structure of the paper is therefore as follows. In Section 2 we will introduce definitions and notational conventions related to the notion of mortality trend, to move then to the description of the Lee-Carter model. Section 3 will be devoted to the presentation of our simulations and to the discussion of related results. Section 4 will conclude.

## 2 Mortality Trends and Related Issues

### 2.1 Understanding Actuarial Notations

Modelling the dynamics of mortality rates over time implies to understand the data we are dealing with. Assume the random variable $D_{x,t}$ to denote the number of deaths in a population at age $x$ and time $t$. Corresponding realizations are generally denoted by $d_{x,t}$, and represent the observed number of deaths, while $e_{x,t}$ generally refers to the matching exposure (in person-years) to the risk of death. The probability of death at age $x$ for a given time $t$ is then given by: $q_{x,t} = \frac{d_{x,t}}{d_{x-1,t}}$. Finally, empirical mortality rates are given by: $m_{x,t} = \frac{d_{x,t}}{e_{x,t}}$ whose stochastic counterpart is the hazard rate (or force of mortality) for age $x$ and time $t$: $\mu_{x,t}$. In order to provide a cross classification, one can fix a calendar year $t$ in the range $[t_1, t_n]$, and an age $x$ in the interval $[x_1, x_k]$, either grouped into $k$ ordered categories, or by individual year (range $k$). The main issue an actuary must face is how to model $\mu_{x,t}$ for every $t \in [t_1, t_n]$ and $x \in [x_1, x_k]$.

### 2.2 The Lee–Carter Model

As said in Section 1, Lee and Carter [4] suggested a framework to model the force of mortality $\mu_{x,t}$ for age $x$ and time $t$:

$$\ln \mu_{x,t} = \alpha_x + \beta_x \, \kappa_t + \epsilon_{x,t}, \tag{1}$$

subject to the constraints:

$$\sum_{t=t_1}^{t_n} \kappa_t = 0, \text{ and: } \sum_{x=x_1}^{x_k} \beta_x = 1 \tag{2}$$

Here $\alpha_x$ is a fixed parameter exploiting the age profile; by Eqs.(1)–(2) it is possible to prove [4] that the least squares estimator of $\alpha_x$ is given by:

$$\hat{\alpha}_x = \ln \prod_{t=t_1}^{t_n} \mu_{x,t}^{1/h}, \; h = t_n - t_1 + 1. \tag{3}$$

In this way $\alpha_x$ expresses the fixed general shape of the logarithmic transformation of the age–specific mortality rates. For what it concerns remaining parameters, $\kappa_t$ describes the underlying time trend, while (constant) $\beta_x$ is the sensitivity of $\ln \mu_{x,t}$ at age $x$ to the time trend represented by $\kappa_t$. Finally, $\epsilon_{x,t}$ renders age and time specific effects not captured by the model, and it is assumed to be an independent, identically distributed random variable.

In order to fit the model, [4] proposed a three–steps procedure detailed on following.

**Step 1.** Estimate $\alpha_x$ as from Eq.(3) above.

**Step 2.** Compute the matrix of statistics $[Z_{x,t}] = [\ln m_{x,t} - \hat{\alpha}_{x,t}]$ and then estimate $\kappa_t$ and $\beta_x$ as, respectively, first right and first left singular vectors in the Singular Value Decomposition (SVD) [10] of the matrix $[Z_{x,t}]$ subject to the above constraints.

**Step 3.** Adjust the estimated $\kappa_t$ such that, for each $t$:

$$\sum_{x=x_1}^{x_k} d_{x,t} = \sum_{x=x_1}^{x_k} e_{x,t} exp\left(\hat{\alpha}_x + \hat{\beta}_x \hat{\kappa}_t\right), \text{ for all } t \tag{4}$$

By running the procedure one can get proper estimates for $\mu_{x,t}$, and hence it will be able to derive any other related actuarial variable.

## 3 Simulation and Results

### 3.1 Experimental Settings

We build a framework aimed to develop coherent mortality forecasts. This choice may be easily justified: over the past two decades the populations of the world have become more closely linked by communication, transportation, trade, technology, and disease [5]. It is then reasonable and perfectly straightforward to forecast mortality for a pool of countries (and hence populations), taking advantage of commonalities in their historical experience and age patterns. Obviously populations that are sufficiently similar to be grouped together might have somewhat different mortality histories; however, such past differences should not lead to continuing long-run divergence in the future.

With this in mind we employed data extracted from the Human Mortality Database (HMD)[1], that contains original calculations of death rates and life tables for national populations (countries or areas), as well as the input data (death counts from vital statistics, census counts, birth counts, and population estimates from various sources) used in constructing those tables. Six data types are available from the HMD: births, deaths, population size (annual estimates), exposure to risk of death, death rates, and life tables. At present the database contains detailed data for 37 countries: Table 1 lists the countries as well as the acronym we employed to refer to them in our simulations.

**Table 1.** Countries included in the Human Mortality Database and related abbreviations

| Country & ID | Country & ID | Country & ID |
|---|---|---|
| Australia (AUS) | Germany (GER) | Norway (NOR) |
| Austria (AUT) | Hungary (HUN) | Poland (POL) |
| Belarus (BIE) | Iceland (ICE) | Portugal (POR) |
| Belgium (BEL) | Ireland (EIRE) | Russia (RUS) |
| Bulgaria (BUL) | Israel (ISR) | Slovakia (SLK) |
| Canada (CAN) | Italy (ITA) | Slovenia (SLO) |
| Chile (CHI) | Japan (JAP) | Spain (SP) |
| Czech Rep. (CR) | Latvia (LAT) | Sweden (SWE) |
| Denmark (DEN) | Lithuania (LIT) | Switzerland (SWI) |
| Estonia (EST) | Luxembourg (LUX) | Taiwan (TW) |
| Finland (FIN) | Netherlands (NL) | United Kingdom (UK) |
| France (FRA) | New Zealand (NZ) | U.S.A. (USA) |
| | | Ukraine (UKR) |

[1] `http:\www.mortality.org`

In our simulations we employed life tables: we can think to them as matrices whose components are time ($t$), age ($x$), observed number of deaths ($d_{x,t}$), exposure to risk of death ($e_{x,t}$), probability of death ($q_{x,t}$), and empirical mortality rates ($m_{x,t}$): while generally it is $x \in [0, 110]$, since all ages from birth ($x = 0$) to extremal age (i.e. the highest age at which someone in the population is still living, e.g.: $x = 110$) are represented, $t$ depends on the year from which the country's demographic bureau began to collect data. In the case of Sweden, for instance, data began to be collected since 1751, so that the available life table has more than $28,000$ entries (obtained as $111 \times 258$, i.e. 111 years for each collection time $t = 1751, \ldots, 2009$). Moving to Russia and Ukraine, on the other hand, the dataset is sensitively smaller (approximately $6,000$ rows), because data began to be collected after 1953. In order to make meaningful comparisons, we use as starting time $t = 1960$, thus having for each country an input matrix of 5439 rows. Moreover, although it is possibile to access and examine separated life tables for both male and female populations, we considered global life tables, giving statistics for the population as whole.

We then implemented a three steps procedure running as follows.

**Step 1.** For each country's lifetable we run a separate SOM, with rectangular topology, initialization at random, and logarithmic transformation of all input variables (with the exception of time and age that have been used to label the data and hence have not been processed).

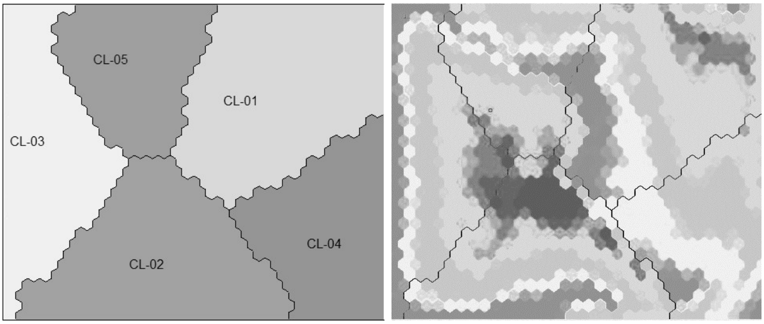
**Step 2.** We then examined the similarity among maps obtained in the previous step, thus getting a $37 \times 37$ symmetric scores table $SCT$, whose generic $i, j$ entry represents the degree of similarity between the i–th and j–th map. Using SCT values we were then able to group countries hence defining the number of populations sharing common mortality features.

**Step 3.** For each group defined in Step 2. we have then built mortality forecasts, according to formulas already provided in Eqs. (1)–(4).

### 3.2 Discussion

As said in previous rows, SOMs operate in two stages over three of the implemented procedure. For what is concerning **Step 1.**, Figure 1, representing Australian life tables, offers some insights about the kind of information SOMs can provide.

From left to right, the first picture in Figure 1 represents age–time clusters for the Australian population in the period: 1960–2009. Note that five cluster emerged: data were at least equally distributed among them. Independently from the reference time $t$, Cluster 1 (CL01) collects data for population aged in the interval $[75 - 97]$, Cluster 2 (CL02) gathers individuals whose age is in the range $[98, 111]$, Cluster 3 (CL03) refers to ages $x \in [31, 60]$, Cluster 4 (CL04) to ages $x \in [0, 30]$, and Cluster 5 (CL05) considers $x \in [61, 74]$. Moving to the second picture, it offers a view into the map organization by time, that is how life tables data referring to different years are spread on the SOM: various gray tones



(a) (b)

**Fig. 1.** From left to right: age–time clusters, and map organization by time in a sample country (AUS). Various gray tones represent different years.

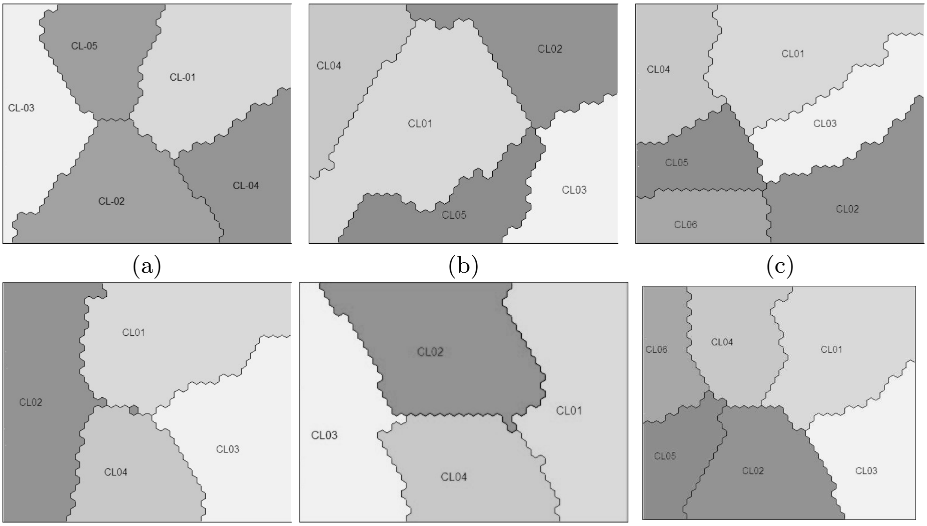
(from white to black) represent different years (in the interval 1960–2009), so that one can easily view that latest years statistics are mainly concentrated on the left hand side of the map, years around later 20th century and earlier 21th century are essentially represented in the internal part of the SOM, while in the center of the map we find data referring to initial years of the sample.

In the second step, we turn to evaluate the similarity among the various maps. This was done by looking at the following factors: (*i*) number of clusters; (*ii*) representativeness of each cluster; (*iii*) ages collected in each cluster. In this way we were able to find out six homogeneous groups (given in Table 2), for which it is then possible to move to **Step 3**, and hence to coherent mortality forecastings.

**Table 2.** Groups identified by SOM for coherent mortality forecasts. The underlined country is the group central country.

| Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|
| AUS | DEN | BIE | AUT | CHI | CR |
| CAN | FIN | BUL | BEL | ICE | HUN |
| EIRE | NOR | EST | FRA | ISR | POL |
| NZ | SWE | LAT | GER | POR | RUS |
| UK |  | LIT | ITA | TW | SLO |
| USA |  | UKR | JAP |  | SLK |
|  |  |  | LUX |  |  |
|  |  |  | NL |  |  |
|  |  |  | SP |  |  |
|  |  |  | SWI |  |  |

The groups evidence strong coherence among anglo–saxon countries (Group 1), Northern Europe countries (Group 2), Baltic countries (Group 3), (mainly) Western Europe countries (Group 4) and Eastern Europe countries (Group 6). Group 5 appears of residual nature. In order to stress the difference among



(a) (b) (c)

(d) (e) (f)

**Fig. 2.** From top to bottom and from left to right: SOM organization corresponding to Groups 1 to 6 central countries identified by our procedure. In the top row, moving in clockwise sense, the picture labelled by $(a)$ is associated to Group 1 central country SOM, the picture labelled by $(b)$ corresponds to Group 2 central country SOM, and so on. In the second row, once again in clockwise sense, the picture labelled by $(d)$ is associated to Group 4 central country SOM, and son on up to the picture labelled by $(f)$ which represents Group 6 central country SOM.

countries in the groups, Figure 2 shows the SOM appearance for the central country of each group.

Using data from central countries, we then performed the final stage of our procedure, i.e. mortality forecasting. The main gain deriving from our technique is primarily in the fact that instead of needing to provide different estimations for 37 countries, we are now asked to give six estimations, at each age $x$, and for every time $t$ in a proper time range. This means obviously a gain in terms of both time and computational efforts.

Figure 3 shows thirty-year life expectancy forecasts $(e_{x,t})$ obtained in the final stage of our procedure for each group central country.

## 4 Final Remarks

In this paper we introduced a SOM–based framework to model and forecast mortality rates dynamics.

The importance of the topic is related to the emergence of longevity risk, as a new type of risk affecting the management of annuity and pensions portfolios, due to misundertandings in the behaviour of mortality.

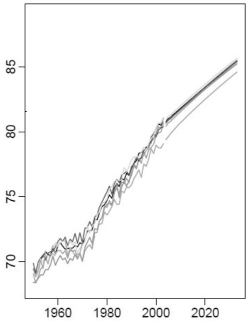


**Fig. 3.** Coherent life expectancy forecasts for each group central country

The main issue faced by existing methods relies in the fact that in order to provide forecasts at a given time $t$ in future and every age $x \in [0, 110]$, they need a very big amount of information going back in time as much as possible. Moreover, according to the traditional approach, each country must be considered as a unique experience, so that generally forecasts for a population cannot be *tout–court* applied to people in a different geographical area.

Our contribution moves in the research vein of coherent mortality forecasts, assuming that if countries share proper common features (e.g. geographic, politic or economic ones) then they are coherent and hence they can also share mortality statistics and forecasts. We then introduced a three–stages procedure which offers a way to create coherent groups. SOM operate in two of three steps, since in the first phase they are employed to get a representation of countries lifetables, while in the second step the clusters originated by SOMs (in particular: their number, as well as their stastistical representativeness) are used to build coherent groups. Data of central country groups are then employed to provide mortality forecasts.

We tested our approach on 37 countries dataset, as resulting from the Human Mortality Database (HMD). The procedure lets us to identify six meaningful groups, whose composition seems to mirror mainly geopolitic differences: we have groups gathering Anglo–Saxon countries (Group 1), Northern and Eastern Europe countries respectively (Groups 2 and 6), Baltic countries (Group 3), and Western Europe lands (Group 4). Group 5, on the other hand, appears of residual nature, collecting areas with apparently no immediate connections.

The results we have obtained prove the effectiveness of a data mining approach to build mortality forecasts. Besides in this way the estimation procedure is less pretending (in terms of both computing time and parameters to estimate) than traditional techniques. This issue is very important, provided that mortality forecasts are widely employed to develop insurance products. Finally we have shown that SOM clustering can be effective to extract similar mortality patterns from apparently very different countries, thus highlighting non–linear hidden features that are missing for more standard techniques.

## References

1. Haberman, S., Renshaw, A.: A cohort–based extension to the LeeCarter model for mortaility reduction factors. Insur. Math. Econ. 38, 556–570 (2006)
2. Kohonen, T.: Self–Organizing Maps. Springer, Berlin (2002)
3. Koissi, M.C., Shapiro, A., Hognas, G.: Evaluating and extending the LeeCarter model for mortality forecasting: bootstrap confidence interval. Insur. Math. Econ. 38, 1–20 (2006)
4. Lee, R., Carter, L.: Modelling and forecasting US mortality. J. Am. Stat. Assoc. 87, 659–671 (1992)
5. Li, N., Lee, R.: Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. Demography 42(3), 575–594 (2005)
6. Li, S.H., Chan, W.S.: The Lee-Carter Model for Forecasting Mortality Revisited. Presented at the Living to 100 and Beyond Symposium Sponsored by the Society of Actuaries, Orlando, Fla., January 12-14 (2005)
7. Oeppen, J., Vaupel, J.: Broken limits to life expectancy. Sc. 296, 1029–1031 (2002)
8. Olshansky, J., Passaro, D., Hershaw, R., Layden, J., Carnes, B., Brody, J., Hayflick, L., Butler, R., Allison, D., Ladwig, R.: A potential decline in life expectancy in the United States in the 21st Century. New Engl. J. Med. 352, 1138–1145 (2005)
9. Pitacco, E.: Longevity risks in living benefits. In: Fornero, E., Luciano, E. (eds.) Developing Annuity Market in Europe, pp. 132–167. Edward Elgar, Cheltenham (2004)
10. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Singular Value Decomposition. In: Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd edn., pp. 51–63. Cambridge University Press, Cambridge (1992)