

# MODELING TRENDS AND INEQUALITY OF LONGEVITY IN THE EUROPEAN UNION COUNTRIES.

A. Debón<sup>1</sup> S. Haberman<sup>2</sup>

<sup>1</sup>Centro de gestión de la Calidad y del Cambio  
Universitat Politècnica de València (Spain)  
**e-mail: [andean@eio.upv.es](mailto:andean@eio.upv.es)**

<sup>2</sup>Cass Business School  
City University London (UK)

14 July 2014

## Main objective

We present a method for **clustering information about differential survival by country**. Therefore, this approach is used to group mortality surfaces for European Union countries.

## Motivation

Interest in health inequalities between European Union (EU) countries and their regions as well as the various social clusters in the EU population is growing. This is driven by the fact that European and national epidemiological studies highlight a widening gap between northern and southern Countries and regions of the EU as well as within countries and regions and between socio-economic groups Spinakis et al. (2011).

# Software

## R-packages

- demography Hyndman and with contributions from Heather Booth, Leonie Tickle and John Maindonald (2014).
- missMDA Husson and Josse (2013)
- e1071 Meyer et al. (2012)

## R project



# Plan of the Talk

- 1 Introduction
- 2 Cluster methodology
  - Principal Component Analysis (PCA)
  - Cluster validity indices
- 3 Characterization of groups
  - Classification by Random Forest
  - Mortality Indicators
- 4 Analysis of mortality data from Human mortality database
  - data
  - PCA with missing data for the countries of EU
  - Clustering of the countries of EU using Principal Components
  - Random forest for mortality indicators in 2009
- 5 Conclusions

# Plan of Talk

- 1 Introduction
- 2 Cluster methodology
  - Principal Component Analysis (PCA)
  - Cluster validity indices
- 3 Characterization of groups
  - Classification by Random Forest
  - Mortality Indicators
- 4 Analysis of mortality data from Human mortality database
  - data
  - PCA with missing data for the countries of EU
  - Clustering of the countries of EU using Principal Components
  - Random forest for mortality indicators in 2009
- 5 Conclusions

# Introduction

The statistical methodology that this article proposes was developed with the aim of establishing an operating procedure which permits the comparison of mortality life tables for countries, in particular the European Union countries. For these reasons, the paper has three steps:

- 1 To **cluster** the behavior of European Union (EU) mortality during the period 1990-2009 using fuzzy cluster algorithm.
- 2 To select the most **important mortality indicators**.
- 3 **Characterization of the groups** in the EU.

# Plan of Talk

- 1 Introduction
- 2 Cluster methodology
  - Principal Component Analysis (PCA)
  - Cluster validity indices
- 3 Characterization of groups
  - Classification by Random Forest
  - Mortality Indicators
- 4 Analysis of mortality data from Human mortality database
  - data
  - PCA with missing data for the countries of EU
  - Clustering of the countries of EU using Principal Components
  - Random forest for mortality indicators in 2009
- 5 Conclusions

## Some notation

- *Dynamic life tables* can be considered as mortality data array ( $q_{xt}$ ), where  $x$  denotes age (row) and  $t$  denotes calendar time (column).
- Each column in this array represents the constituents of the period life table for year  $t$ .

### Cluster

The purpose of the present section is to show how to **measure distances between mortality surfaces and then clustering them in homogeneous groups**. In **fuzzy clustering** data elements can belong to more than one cluster each with an associated membership level (Hatzopoulos and Haberman, 2013). This indicates the strength of the association between the data element and a particular cluster.



# Principal Component Analysis (PCA)

- Cluster analysis is a reasonable approach to separate the countries into clusters with similar mortality dynamics.
- To avoid the curse of dimensionality, the most common solution is to reduce information about countries mortality surfaces to most significant features by means of PCA.
- The idea of using PCA mortality is not new, many papers since Lee and Carter (1992) have tried to improve the classical model Renshaw and Haberman (2006); Debón et al. (2010); De Jong and Tickle (2006).
- None of them deal with missing data nor choose the number of PC with an objective criterion. In this paper, we build upon the idea of PC modeling, using the R package missMDA to choose the number of dimensions by cross-validation.

## Cluster validity indices

- Additionally, cluster validity indices are introduced, which assess the average compactness and **separation of fuzzy partitions generated by the fuzzy c-means algorithm**.
- There are a number of cluster validation indices available, in this study, **the validity measures used are: xie.beni, fukuyama.sugeno, partition.coefficient and partition.entropy**, a short description can be found in Ramze Rezaee et al. (1998).

# Plan of Talk

- 1 Introduction
- 2 Cluster methodology
  - Principal Component Analysis (PCA)
  - Cluster validity indices
- 3 **Characterization of groups**
  - Classification by Random Forest
  - Mortality Indicators
- 4 Analysis of mortality data from Human mortality database
  - data
  - PCA with missing data for the countries of EU
  - Clustering of the countries of EU using Principal Components
  - Random forest for mortality indicators in 2009
- 5 Conclusions

# Random forest

- The purpose of the analyses via tree-building algorithms (CART) is to determine a set of **if-then logical (split) conditions that permit accurate classification of cases**.
- **random forests**: each tree using a different **bootstrap sample** of the data and each node is split using the best among a **subset of predictors** randomly chosen at that node.
- The randomForest package optionally produces two additional pieces of information: a measure of the **importance of the predictor variables**, and a measure of the **internal structure of the data** (the proximity of different data points to one another).

## Mortality indicators

- The most usual mortality indicators such as life expectancy, modal age at death and Gini Index can be calculated from a dynamic period life table, details in Debón et al. (2012).
- Life expectancy calculated refers to **life expectancy at birth and at 65**,  $e_{0t}$  and  $e_{65t}$ , respectively.
- The **modal age at death** is the age associated with the maximum frequency of death.
- The above two indicators do not provide any information about whether the improvement in mortality rates applies equally to different age groups. The **Gini index** is the most common statistical index used in demography for measuring the contribution of different ages to mortality over time inequality or diversity.

# Plan of Talk

- 1 Introduction
- 2 Cluster methodology
  - Principal Component Analysis (PCA)
  - Cluster validity indices
- 3 Characterization of groups
  - Classification by Random Forest
  - Mortality Indicators
- 4 Analysis of mortality data from Human mortality database
  - data
  - PCA with missing data for the countries of EU
  - Clustering of the countries of EU using Principal Components
  - Random forest for mortality indicators in 2009
- 5 Conclusions

## Crude estimates

- The data are downloaded from the Human Mortality Database (2013) and the corresponding life tables have been obtained through the commands `hmd.mx` and `lifetable` from the demography R-package.
- There are no complete data for all the EU countries and available countries have a common time range 1990-2009 for the age ranges from 0 to 99 for men.
- The data are completed for early ages with techniques to deal with missing values.
- In our method, for each country each age group for each year is considered as an separate variable only for men, so each country is represented by  $100 \times 20 = 2000$  variables.

## PCA with missing data for the countries of EU

- The dimensionality is reduced by using only the first few principal components (PCs) losing as little information as possible. The optimal number of components can be defined as the minimum number of components which accounts for **the maximum possible variance**.
- One common criteria is to include all those PCs up to a **predetermined total percentage of explained variance**, such as 90 %.
- Recently a class of “objective” cross-validation methods have been developed to determine this quantity.

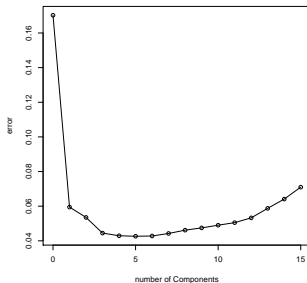


## PCA with missing data for the countries of EU

To perform PCA on an incomplete dataset using missMDA R-package.

- 1 the first step consists of estimating the number of dimensions that will be used in the regularized iterative PCA algorithm using the *estim\_ncpPCA* command.
- 2 the second is to represent the prediction error for different numbers of dimensions calculated by cross-validation. The error for the model without components corresponds to reconstruction of the data.
- 3 Cross-validation criterion have a well-marked minimum for 5 components with our data.

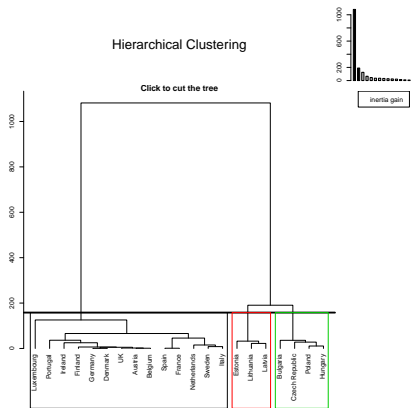
Cross-validation (CV) of PCA applied to mortality surfaces (the logit transformation of the period lifetables) for the time range 1990-2009 and for ages from 0 to 99 for all countries.



## Clustering of the countries of EU using Principal Components

- We want to gather the 21 countries of the dataset into a number of clusters which would correspond to **different mortality profiles**.
- Mortality surfaces have been summarized on the corresponding **5 PCs** obtained previously.
- Firstly, we are going to perform a **hierarchical classification** on the principal components of a factorial analysis.
- In addition, **fuzzy clustering** generalizes partition clustering methods (such as k-means and medoid) by allowing a country to be **partially classified into more than one cluster**.

# Dendrogram of the hierarchical clustering of the countries based on the resulting 5 PCs for men.



## Results of the validation indices by using data of the EU countries.

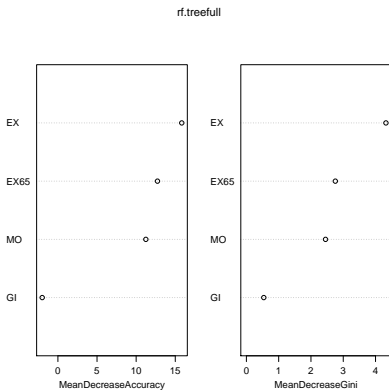
The optimal number of clusters for all the indices as it corresponds to maximum partition coefficient (pc) and minimum entropy (pe), xie.beni (xb) and fukuyama.sugeno (fs).

index	number of clusters		
	2	3	4
xb	0.01	<b>0.01</b>	0.03
fs	-26720.47	<b>-26720.92</b>	-18588.22
pc	0.73	<b>0.73</b>	0.61
pe	0.50	<b>0.50</b>	0.73

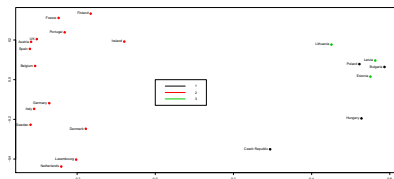
## The membership values of the EU countries to the clusters.

	1	2	3
Austria	0.01	<b>0.99</b>	0.00
Belgium	0.03	<b>0.97</b>	0.01
Bulgaria	<b>0.66</b>	0.09	0.25
Czech Republic	<b>0.80</b>	0.12	0.08
Denmark	0.06	<b>0.92</b>	0.02
Estonia	0.15	0.05	<b>0.81</b>
Finland	0.11	<b>0.86</b>	0.03
France	0.11	<b>0.84</b>	0.05
Germany	0.03	<b>0.96</b>	0.01
Hungary	<b>0.73</b>	0.08	0.19
Ireland	0.21	<b>0.72</b>	0.07
Italy	0.05	<b>0.93</b>	0.02
Latvia	0.08	0.02	<b>0.90</b>
Lithuania	0.12	0.05	<b>0.83</b>
Luxembourg	0.34	<b>0.49</b>	0.17
Netherlands	0.10	<b>0.87</b>	0.03
Poland	<b>0.90</b>	0.03	0.07
Portugal	0.35	<b>0.52</b>	0.13
Spain	0.09	<b>0.87</b>	0.04
Sweden	0.10	<b>0.86</b>	0.04
UK	0.01	<b>0.99</b>	0.00

# Mortality index importance.



# The metric multi-dimensional scaling representation for the proximity matrix of the EU countries





# Plan of Talk

- 1 Introduction
- 2 Cluster methodology
  - Principal Component Analysis (PCA)
  - Cluster validity indices
- 3 Characterization of groups
  - Classification by Random Forest
  - Mortality Indicators
- 4 Analysis of mortality data from Human mortality database
  - data
  - PCA with missing data for the countries of EU
  - Clustering of the countries of EU using Principal Components
  - Random forest for mortality indicators in 2009
- 5 Conclusions

# Conclusions

- 1 21 EU countries were classified by fuzzy c-means cluster analysis, for the time period 1991-2009.
- 2 one cluster is formed by the western European countries, a second by the Baltic States and a third by Eastern European countries.
- 3 Next, random forest was used to implement coherent classification of the countries based on mortality indicators in 2009.
- 4 For the Baltic States (Estonia, Lithuania and Latvia), the mortality dynamics are distinguishable from the rest of the East-cluster countries, and their life expectancies in 2009 are the lowest ones, less than 70.

## Conclusions (cont.)

- 1 The main aim of this work was to cluster the mortality behavior of the EU countries and we have objectified,
  - the assessment of dimensionality reduction using cross-validation, and
  - the assessment of number of clusters using cluster validity indices.

*Thanks for your attention*

Any question?

Ana Debón

Centro de gestión de la Calidad y del Cambio

Universitat Politècnica de València (Spain)


e-mail: [andeau@eio.upv.es](mailto:andeau@eio.upv.es)

De Jong, P. and Tickle, L. (2006). Extending lee–carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18.

Debón, A., Martínez-Ruiz, F., and Montes, F. (2010). A geostatistical approach for dynamic life tables: The effect of mortality on remaining lifetime and annuities. *Insurance: Mathematics and Economics*, 47(3):327 – 336.

Debón, A., Martínez-Ruiz, F., and Montes, F. (2012). Temporal evolution of some mortality indicators. application to Spanish data. *North American Actuarial Journal*, 16(3):364–377.

Hatzopoulos, P. and Haberman, S. (2013). Common mortality modeling and coherent forecasts. an empirical analysis of worldwide mortality data. *Insurance: Mathematics and Economics*, 52(2):320–337.

Human Mortality Database (2013). *University of California*, 

*Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).*

Husson, F. and Josse, J. (2013). *missMDA: Handling missing values with/in multivariate data analysis (principal component methods)*. R package version 1.7.2.

Hyndman, R. J. and with contributions from Heather Booth, Leonie Tickle and John Maindonald (2014). *demography: Forecasting mortality, fertility, migration and population data*. R package version 1.17.

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

Lemon, J. (2006). Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12.

- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2012). *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien. R package version 1.6-1.
- Ramze Rezaee, M., Lelieveldt, B. P., and Reiber, J. H. (1998). A new cluster validity index for the fuzzy  $c$ -mean. *Pattern recognition letters*, 19(3):237–246.
- 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.
- Spinakis, A., Anastasiou, G., Panousis, V., Spiliopoulos, K., Palaiologou, S., and Yfantopoulos, J. (2011). Expert review and proposals for measurement of health inequalities in the european union. Technical report, European Commission Directorate General for Health and Consumers, Luxembourg. ISBN 978-92-79-18529-8.