Cause-specific senescence: classifying causes of death according to the rate of aging

Carlo G. Camarda* Markéta Pechholdová† France Meslé‡

Short paper submitted to the
Population Association of America Conference 2015, San Diego, CA
September 2014

This paper was written within the following projects:
- *"AXA project on Mortality Divergence and Causes of Death"*
- *"Project ANR-12-FRAL-0003-01 DIMOCHA"*

Abstract

With an intention of producing a statistically-based cause-of-death shortlist, we reclassified the causes of death based on their age profile, expressed here in terms of the rate of aging. We are using individual-level mortality data from the Czech Republic for period 1998-2011 and connect them to the HMD exposures. The analysis is limited to males aged 30 and over. The cause-specific mortality curves were smoothed using non-parametric procedures and an optimized cluster analysis was applied to the smoothed data. We ended up with three dissimilar age patterns, which we entitled as "premature", "Gompertzian", and "degenerative". The findings suggest that a similar but refined analysis can produce a meaningful disease grouping with a potential of application in future mortality research.

KEYWORDS: Classification; Cluster analysis, Mortality; Cause of death; Smoothing; Rate-of-Aging.

*Corresponding author: Institut National d'Études Démographiques. 133, Bd Davout, 75980 Paris Cédex 20, France; Tel. +33(0)1 5606 2155, email: carlo-giovanni.camarda@ined.fr

†Department of Demography, University of Economics, Prague; email: Marketa.Pechholdova@seznam.cz

‡Institut National d'Études Démographiques, Paris; email: mesle@ined.fr

1 Introduction

The current revision of International classification of diseases (ICD) contains around 12,000 items (causes of death, diseases, health-related events). As it is impossible to take into account all of the ICD detail in demographic analyses, researchers seek for a convenient cause-of-death shortlist, which would be small in size but explicative in its design. Typically, a simple list of main ICD chapters is used, although its inconvenience is obvious (contains a heterogeneous mixture of diseases of diverse origin and its explanatory power is therefore quite limited).

Evidence-based COD shortlists, as well as analyses of cause-specific senescence, remain rare attempts. An elaborated reclassification of diseases based on their etiology was published in the 1990s (Meslé, 1999). An analytical shortlist derived from a thorough study of comparability was published recently in 2012 (Pechholdová, 2012). A few other studies also focused at the relationship between aging and cause of death (Horiuchi et al., 2003) or aimed at regional similarities of age-cause-specific profile (Brouard and Lopez, 1985; Meslé and Vallin, 2002).

In our paper we propose a statistically-oriented reclassification of causes of death based on their intrinsic characteristic: the age profile (the rate of aging, more precisely). We reduce the element of subjectivity to minimum by using non-parametric mortality models and by performing a cluster analysis with optimized number of clusters. We then analyse the content of each cluster and describe, if possible, its main etiological and risk factors. The analysis of the outcomes can shed new lights on the debate about the controversial hypothesis on rate of aging as a fundamental human invariant: working on the rate-of-aging by cause of death, one can better isolate cause- specific features which are confounded in the overall mortality age-pattern.

2 Data and methods

We are using individual-level data on deaths in the Czech Republic covering the period 1998-2011. We have information on the age, sex, date of death, underlying COD and six contributing CODs (a multiple-cause of death dataset). Based on these data, we are able to compute death counts by single year of age at the ICD10 3-digit level. Correspondent exposure population are taken from the Human Mortality Database (2014).

2.1 Estimating cause-specific rate-of-aging

In this paper we present the analysis for males only and starting from age 30. Moreover, in order to produce reliable outcomes for cause-specific age-patterns, we select only causes in which we have got at least 10 available data-points over ages. The final dataset contains

$m = 71$ ages for $n = 531$ CODs and it can be described as a set of two matrices:

$$\boldsymbol{D} = (d_{i,c}) \qquad \text{and} \qquad \boldsymbol{E} = (e_{i,c}) = \boldsymbol{e}\,\mathbf{1}_{1,n}\,, \tag{1}$$

where $d_{i,c}$ are the death counts for age i and COD c, $\boldsymbol{e}$ is the series of exposure over ages and $\mathbf{1}_{1,n}$ is a $1 \times n$ matrix of ones. From these matrix we can compute the matrix of cause-specific death rates $\boldsymbol{M} = (m_{i,c}) = (\frac{d_{i,c}}{e_{i,c}})$, which can also be seen as a fully non-parametric representation of the underlying cause-specific force of mortality over ages, $\mu^c(x)$.

Our aim is to classify CODs according to their rate-of-aging would could be seen as the first relative derivative of the force of mortality with respect to age:

$$r(x) = \frac{\frac{\partial \mu^c(x)}{\partial x}}{\mu^c(x)} = \frac{\partial \ln(\mu^c(x))}{\partial x}\,. \tag{2}$$

Computing derivatives directly on the raw death rates implies the usage of difference operator over ages and it is common knowledge that such computation is highly affected by noise in the data. A possible solution would be to impose a parametric structure to each cause-specific mortality pattern, but this leads to two main issues: (1) it is hard to select a parametric which well suits each cause-specific age pattern; (2) a given parametric model will impose a rigid outcome in terms of derivative, e.g. a Gompertz model imposes a constant rate-of-aging over age.

In this paper we decide to follow a non-parametric approach for modeling raw death rates. This allows us to identify instantaneous rate-of-aging for each cause-specific age-pattern which would be free from any predetermined structure. Specifically we smooth our data using a P-spline approach (Camarda, 2012; Eilers and Marx, 1996). For our aims, the advantage of this approach lies in the straightforward computation of relative derivatives of mortality age-pattern with respect to age.

For a given COD and within a P-spline approach, estimated cause-specific force of mortality can be described, in a logarithm scale, as a linear combination of B-splines spanning over ages and coefficients which are penalized to achieve a smooth behavior of the age-patterns. In formula:

$$\ln(\hat{\boldsymbol{\mu}}^c) = \boldsymbol{B}^q\,\hat{\boldsymbol{\beta}}^c\,.$$

where $\boldsymbol{B}^q \in \mathbb{R}^{m \times r}$ is a matrix of r B-splines of degree q with knots equally spaced by a distance h. They are common for all CODs patterns. This formulation reduces also the dimensionality of our problem from m, number of ages, to r, length of the vector $\hat{\boldsymbol{\beta}}^c$. In the following we choose $r = 17$.

This representation naturally allows for incorporating the relative derivative of the force of mortality. A linear combination of B-splines can be derived with respect to x and equivalently, operating directly on the coefficients $\hat{\boldsymbol{\beta}}^c$, we can write the relative derivative

of the force of mortality as follows:

$$r^c(x) = \boldsymbol{C}\,\hat{\boldsymbol{\beta}}^c = \frac{1}{h}\boldsymbol{B}^{q-1}\,\hat{\boldsymbol{\alpha}}^c \tag{3}$$

where $\boldsymbol{C}$ is a matrix incorporating both original B-splines and computing first order difference of the estimated coefficients $\hat{\boldsymbol{\beta}}^c$ (Camarda, 2008). The vector $\hat{\boldsymbol{\alpha}}^c$ denote the first difference of the estimated coefficients for a given COD and its length is equal to $r-1$.

In other words we can either use the original estimated coefficients or apply difference operator on them to obtain identical $r^c(x)$. This result allows us to apply cluster analysis directly of the first difference of estimated coefficients $\hat{\boldsymbol{\alpha}}^c$ without loosing smoothness behavior of the age-patterns as well as keeping the relationship between instantaneous rate-of-aging for each COD and the associated age profiles.

2.2 Clustering cause-specific rate-of-aging

We aim to classify all n COD by their instantaneous rate-of-aging which we reduced to a vector of lagged-coefficients for each COD c. Hence our set of observations is $[\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\alpha}}^2, \ldots, \hat{\boldsymbol{\alpha}}^n]$, where each observation is a $r-1$ dimensional real vector.

Several options are available when a cluster analysis is performed. In the following we followed a k-means clustering which allows us to cluster our observations as well as to extract a center for each cluster. These "centers" could served as prototypes for the different rates-of-aging present in human mortality, although they may not necessarily be a member of the data set themselves.

In few words, we aim to partition the n observations into k sets $\boldsymbol{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares. In formula our objective would be:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\boldsymbol{\alpha} \in S_i} \|\boldsymbol{\alpha} - \boldsymbol{\nu}_i\|^2 \tag{4}$$

where ν_i is the mean of points in S_i, i.e. the center for cluster i. The algorithm proposed by Hartigan and Wong (1979) and implemented in the `R`-routine `kmeans()` was used for this paper.

As one can easily see, the procedure highly depends upon the number of selected clusters k and an inappropriate choice of k may yield poor results. Various criteria are available for optimize the number of clusters. Instead of subjectively pick a given number of clusters, or a specific selection criterion, we run 24 different indices for determining the number of clusters. To easily obtain all these outcomes, we modified the main routine available in the `R`-package `NbClust` (Charrad et al., 2014). Each index will have thus its

associated optimal k. We select the modal values of the final distributions.

Figure 1 presents the frequency distribution for 24 indices according to the associated optimal number of cluster. Though there is a large spectrum of possible optimal k, it seems that having 3 clusters is preferable for the relative majority of the proposed indices. We thus run our k-means in (4) with $k = 3$.

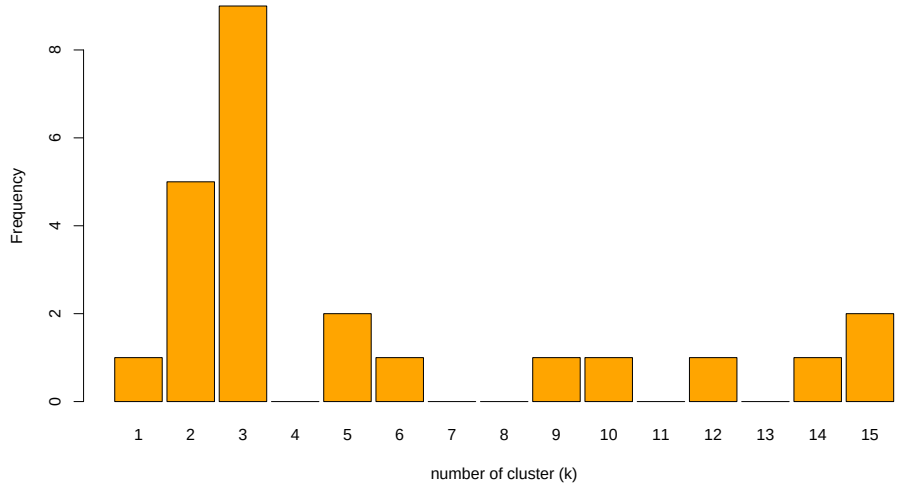

Figure 1: Distribution of optimal number of clusters for cause-specific rate-of-aging for 24 different methods. Czech Republic, Males, 1998-2011, ages 30-100.

Once we obtain the means for each cluster $\boldsymbol{\nu}_i$, we can assign each observation to the cluster with the nearest mean. The center of each cluster can then be used to evaluate the centers of the 3 different instantaneous rate-of-aging

$$\boldsymbol{\gamma}_i = \frac{1}{h}\boldsymbol{B}^{q-1}\,\boldsymbol{\nu}_i \qquad \text{for} \quad i = 1, 2, 3$$

and the associated age-profiles.

3 Results

Figure 2 presents the outcomes in terms of instantaneous rate-of-aging and age-patterns for the 3 selected clusters. They can be seen as the representative profiles for the 531 cause-specific age-patterns. To feel the leverage of each cluster in terms of overall mortality, we employed line widths proportional to the number of deaths within each cluster.

The first cluster (red solid lines in Figure 2) is typified by decreasing rate of mortality change and by an unusual age pattern, reaching maximum around age 50 and declining since then. About 6% of all deaths are concentrated within this class.

Based on the CODs belonging to this cluster, it can be entitled as ***Premature***. The first cluster contains a mixture of genetically-conditioned diseases (such as rarer types of cancer, epilepsy, lupus, ulcerative collitis), accidental deaths (traffic accidents, suicide, homicide, drowning), but also alcohol related mortality.

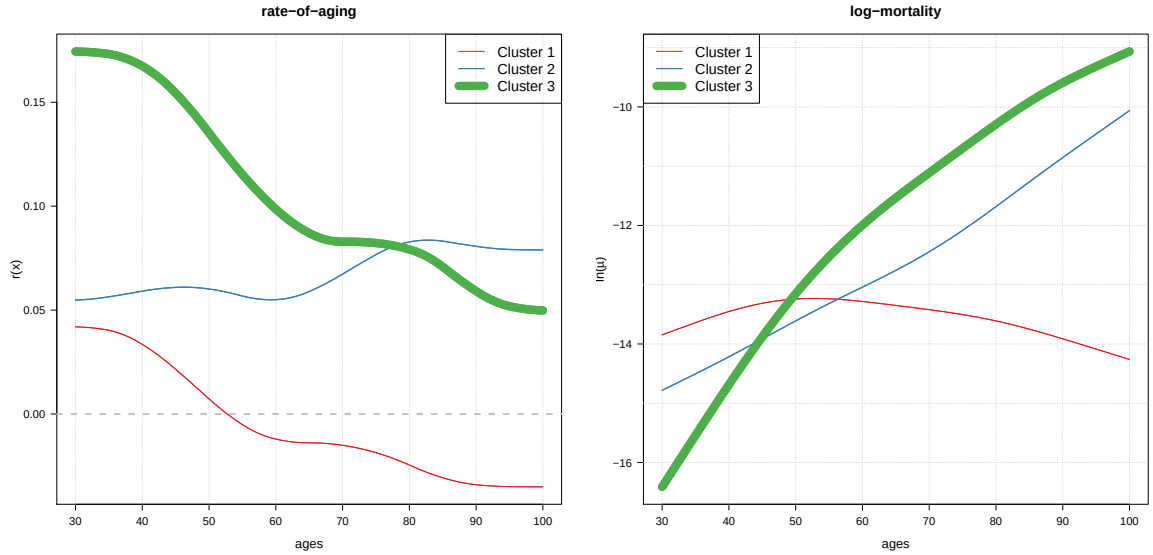

Figure 2: Instantaneous rate-of-aging (left) and log-mortality (right) for the centers of the 3 clusters formed by k-means algorithm. The width of the lines is proportional to the number of deaths within each cluster. Czech Republic, Males, 1998-2011, ages 30-100.

The characteristic pattern of the second cluster is the closest to the Gompertz idea of senescence: a quasi-constant rate of mortality change / linear log of the force of mortality. The cluster contains some bacterial infections, malignant melanoma of skin, acute respiratory diseases, most of digestive and genitourinary diseases, and accidental falls. Surprisingly, the ***Gompertzian*** cluster only contains a small fraction of deaths (about 12%).

As Figure 2 suggests, by far the most deaths are grouped in the third cluster (ca. 82%), which shows an overall decrease of the rate of aging and where death rates rise rapidly until age 60 to increase at decreasing rate until the oldest ages groups. This group has the lowest mortality at the starting age and the highest mortality as of age 50. The group contains tuberculosis, colorectal cancer, stomach cancer, and most of the smoking-related mortality (notably cancer of trachea, bronchus, lung, and upper airways, and other cancer highly suspected to be related to smoking: cancer of bladder and kidney). The cluster also contains smoking-related conditions from the chapter of respiratory diseases (emphysema and other chronic obstructive pulmonary diseases, chronic bronchitis). Other man-made (occupational) respiratory infections are classified here as well (pneumoconioses). Next important CODs in this cluster are diabetes, dementias and Alzheimer disease. We also find majority of circulatory diseases (and thus the majority of all deaths) in the same

group (ischemic heart disease, stroke, hypertension). Interestingly, the group contains virtually no accidental deaths. Therefore, the cluster could be entitled as ***Degenerative***, or ***Man-made*** - it is probably the man-made etiology which creates the deceleration-like age pattern of this cluster, via a lifetime selection of individuals with healthier life styles, implying that an important part of these deaths are preventable.

Finally Figure 3 illustrates the results of the clustering with three optimal clusters highlighted. We can see the clear dissimilarity of the main three groups. Beyond the optimal number of clusters, several sub-clusters can be identified at lower similarity distance. Exploring the content of these sub-clusters is the agenda of the future work.

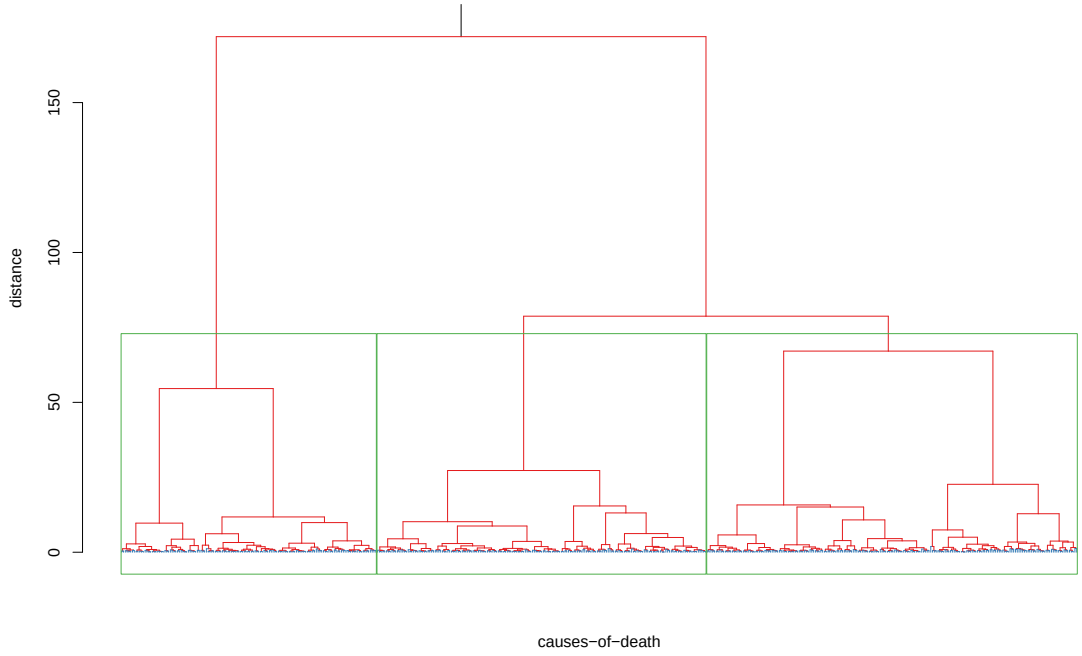

Figure 3: Dendogram of the cluster analysis carried on the cause-specific instantaneous rates-of-aging. Czech Republic, Males, 1998-2011, ages 30-100.

4 Summary

Using optimized clustering methods, we identified three predominant age patterns of human diseases. The first type is virtually age-independent, and is present in individuals with either congenital predisposition or at high risk behaviour (such as severe alcohol consumption). The mortality in second group of diseases rises progressively with age without signs of deceleration in the most advanced ages. Most important in the terms of death counts, however, is the group with decelerating mortality, which we entitle as

degenerative due to a strong man-made component.

In the future work, the analysis design can be further refined. For example, we could pre-group the diseases entering into the analysis based on common medical definitions rather than work on the ICD 3-digit level. We could also explore other levels of clustering, to define more clusters within the main three groups. Finally, to validate the proposed groupings, the method should be tested on other countries as well.

References

Brouard, N. and A. D. Lopez (1985). Cause of death patterns in low mortality countries: a classification analysis. In *International Population Conference, Florence, 1985, June 5-12. Congres International de la Population. Volume 2*, pp. 385–406. Liege, Belgium, International Union for the Scientific Study of Population.

Camarda, C. G. (2008). *Smoothing Methods for the Analysis of Mortality Development*. Ph. D. thesis, Programa de Doctorado en Ingeniería Matemática. Universidad Carlos III, Departamento de Estadística, Madrid.

Camarda, C. G. (2012). MortalitySmooth: An `R` Package for Smoothing Poisson Counts with *P*-Splines. *Journal of Statistical Software 50*, 1–24. Available on `www.jstatsoft.org/v50/i01`.

Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs (2014). *`NbClust`: `NbClust`package for determining the best number of clusters*. R package version 2.0.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with *B*-splines and Penalties (with discussion). *Statistical Science 11*, 89–102.

Hartigan, J. A. and M. A. Wong (1979). A k-means clustering algorithm. *Applied Statistics 28*, 100–108.

Horiuchi, S., C. E. Finch, F. Meslé, and J. Vallin (2003). Differential Patterns of Age-Related Mortality Increase in Middle Age and Old Age. *Journal of Gerontology 58*, 495–507.

Human Mortality Database (2014). *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*. Available at `www.mortality.org`. (Data downloaded on February 2014).

Meslé, F. (1999). Classifying Causes of Death According to an Aetiological Axis. *Population Investigation Committee 53*, 97–105.

Meslé, F. and J. Vallin (2002). *Diversity of mortality structures among western industrialized countries.* Paper presented at PAA 2002 annual Meeting, Atlanta, May 9-11.

Pechholdová, M. (2012). *Causes of Death: Collection, Classification, Continuity and Comparability. Evidence from the Czech Republic, West Germany and France.* Slaný, Melandrium.