Analyzing Cause–Specific Mortality Trends using Compositional Functional Data Analysis

Marco Stefanucci

University of Padova, Italy

E-mail: marco.stefanucci@unipd.it

Stefano Mazzuco

University of Padova, Italy

E-mail: stefano.mazzuco@unipd.it

Summary. We study the dynamics of cause–specific mortality rates among countries by considering them as compositions of functions. We develop a novel framework for such data structure, with particular attention to functional PCA. The application of this method to a subset of the WHO mortality database reveals the main modes of variation of cause–specific rates over years for men and women and enables us to perform clustering in the projected subspace. The results give many insights of the ongoing trends, only partially explained by past literature, that the considered countries are undergoing. We are also able to show the different evolution of cause of death undergone by men and women: for example, we can see that while lung cancer incidence is stabilizing for men, it is still increasing for women.

Keywords: Causes of Death; Compositional Data Analysis; Functional Data Analysis; Mortality.

1. Introduction

Overall mortality trends may be partially explained by cause-specific data. A recent example is provided by Woolf and Schoomaker (2019) who try to shed light on the decreasing trend of US life expectancy inspecting mortality by cause, finding that midlife mortality caused by drug overdoses, alcohol abuse, suicides, and a diverse list of organ system diseases have particularly increased in the latest years. However, analyzing trends of cause specific mortality rates (CSMRs) is not straightforward, as dimensionality of data (rates by cause, year and country) might easily reach a size difficult to manage. One common solution is reducing dimensionality by collapsing one or more components: for instance, Canudas-Romo et al. (2020) use TCAL indicator to summarize evolution over time of cause-specific mortality. TCAL (Truncated Cross-sectional Average Length of life) is an indicator summarizing cohort–specific and period–specific information but in this way it is no longer possible to explain the time trends. An additional issue when dealing with cause-specific mortality is the competing risk setting: a cause-specific mortality rate can decline because there has been a significant improvement in treatment and/or prevention of that disease or just because other causes have risen meanwhile. Therefore, if we want to analyze the time trend of CSMRs we need to take into account this feature. One way to do this is by means of Compositional Data Analysis (CDA) (Aitchinson, 1986; Egozcue and Pawlowsky-Glahn, 2011), which is the study of compositions, i.e. data where quantities are

part of a whole, by their representation as points on a D–dimensional simplex. Cause-specific mortality rates can be seen as compositional data, in the sense that their sum is the overall mortality rate but, if we consider their time trend, we have not only points in a simplicial space, but curves. Therefore we suggest that CDA might be combined with elements of Functional Data Analysis (FDA) (Ramsay and Silverman, 2005). For example, Functional PCA can be applied to reveal what are the main components driving the latest trends of CSMRs in a selected group of countries. Moreover, countries could be grouped with respect to the evolution of their CSMRs. Therefore, by combining CDA and FDA we analyze trends of causes of death in 22 countries, which are eventually clustered according to their CSMRs' evolution. The analysis proposed here is essentially descriptive, however it has the advantage of encompassing all causes of death and, at the same time, allowing to focus on specific causes.
The paper is organized as follows: in the next section we motivate our analysis and describe the data we used; section 3 describes the way in which FDA and CDA are combined together, section 4 shows the results and section 5 illustrates the conclusions.

2. Motivations and Data

Cause-specific mortality rates (CSMRs) on calendar year t are derived from all–causes rates m_x^t and deaths for each cause i, age x and time t, ${}^iD_x^t$. Following Preston et al. (2001), the CSMRs are defined as

$$ {}^im_x^t = m_x^t \cdot \frac{{}^iD_x^t}{D_x^t}, \tag{1} $$

where D_x^t is the number of deaths for all causes occurred at time t and age x, m_x^t is the mortality rate at age x in period t and ${}^im_x^t$ is the cause–specific mortality rate at age x in period t, for cause i. Therefore, considering that $\sum_i {}^iD_x^t = D_x^t$ we have that

$$ \frac{{}^im_x^t}{m_x^t} = \frac{{}^iD_x^t}{D_x^t} \qquad \text{and} \qquad \sum_i \frac{{}^im_x^t}{m_x^t} = 1. \tag{2} $$

Equation (2) shows that CSMRs are compositional data, conditional to age. Interestingly, Oeppen (2008) and Kjærgaard et al. (2019) also use compositional analysis applied to cause-specific mortality, but with forecasting purposes, which lead them to consider multiple–decrement life tables deaths ${}^id_x^t$ rather than rates, as the sum of them over i and x is one (the radix of life table). So they consider ${}^id_x^t$ as a composition over causes i and over ages x, and in this way they only have to make one forecasting step. However, in a descriptive perspective – which is the one we are taking – mixing cause pattern and age pattern, which are varying across time, countries and causes, would make interpretation more difficult, so we condition to a specific age group (40–64) that we consider more important: we excluded younger ages, as the causes involved in mortality are very limited (especially infant mortality). Ages older than 65 are also excluded, considering that the leading causes at old ages are fundamentally different from those in midlife, as also shown by Horiuchi et al. (2003). So we focus on premature mortality cause of deaths pattern and old age can be analyzed separately. It would be possible analyse CSMRs without conditioning to a specific age class. Chen et al. (2017) for instance, propose to marginalize over age the considered quantities, but in this way we would focus on average rates and incidence

Table 1. Countries considered.

Area	Countries
North EU	Denmark, Finland, Norway, Sweden, Iceland
West EU	Austria, Belgium, Switzerland, France, Ireland, Netherlands, UK
South EU	Italy, Spain, Greece
Central EU	Hungary, Poland
Extra-EU	Australia, Canada, Japan, New Zealand, USA

of causes of death varies a lot across ages: the average would be the result of very different prevalence. Thus, we prefer to have a focus on a specific age in order to have results much easier to interpret.

From (2) it follows that analyzing the ratio between cause-specific deaths (${}^iD_x^t$) and all-causes deaths (D_x^t) is equivalent to analyze the ratio between cause-specific rates (${}^im_x^t$) and all-causes rate (m_x^t). This brings about the advantage of using a unique source of data: the World Health Organization (2019) mortality database, which provides data on cause and age specific deaths for all countries. Should we use also all-causes mortality rates, we would need to turn to an additional source (for instance the Human Mortality Database), but given equation (2), ${}^iD_x^t$ and D_x^t is all we need.

2.1. Data

WHO mortality database is an archive of the causes of death information for several countries. The longest time series starts in 1950, however for many countries the information is available only since 1959. We need to consider the same time window for each country, so we focus on years 1959–2015, and we only consider countries for which data are available in this time frame. Moreover, we excluded Luxembourg for which data was too noisy, given the limited population size, so our final analysis restricts to countries listed in Table 1.

An important issue is that classification of causes of death has greatly changed since 1959, passing from ICD 7^{th} to ICD 10^{th} revision. Following Canudas-Romo et al. (2020), we therefore use broad classes of causes that are little affected by changes between different revisions, see Table 2. Deaths due to other causes (classified as "Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified" ICD–10 codes R00–R99) and infant mortality related causes (ICD–10 codes P00–Q99), have been removed. These two sets of causes had a very limited prevalence in our data, also considering the specific age–group we focused on, and not a specific pattern, so including them would only have increased the statistical noise. Note that some adjustments has been made: for example, HIV has been classified among Endocrine diseases in ICD 9^{th} revision and moved to Infectious diseases class in 10^{th} revision, thus we moved it in this class also for 9^{th} revision. In addition, we found that in Austria diabete mellitus with circulatory complications is classified among endocrine diseases, and we moved it to class CIRC (circulatory diseases) to make the classification consistent with what have been done for other countries. Then, we obtained regular curves from noisy data by spline smoothing. This has be done in a transformed space, similarly to the methodology proposed in section 3. A common tuning parameter across countries – with the exception of Iceland – was chosen to balance the bias–variance trade off. Iceland, due to limited population, exhibits a stronger measurement

Table 2. Causes of death considered with related International Classification of Diseases (ICD) codes for each revision.

Cause group	ICD–7	ICD–8	ICD–9	ICD–10
Certain infectious and parasitic diseases (INF)	A001–A043, A104, A132, B001–B017, B043	A001–A044	B01–B007	A00-B99
Endocrine, nutritional and metabolic diseases (END)	A061–A064, B020	A062–A066	B18–B19	E00-E88
Circulatory system diseases (CIRC)	A070, A079–A086, B022, B024–B029	A080–A088	B25–B30	I00–I99
Neoplasms (NEOP)	A044–A060, B018–B019	A045–A050, A052–A061	B08–B09, B100, B109, B11–B17	C00–C32, C34–D48
Lung cancer (LUNG)	A050	A051	B101	C33-C34
Respiratory diseases (RESP)	A087–A097, B030–B032	A089–A096	B31–B32	J00-J98
Digestive system diseases (DIG)	A098–A107, B033–B037	A097–A104	B33–B34	K00–K93
External causes of death (EXT)	A138–A150, B047–B050	A138–A150	B47–B56	V00–Y89

error and we opted for a country–specific tuning parameter. Finally, since in some years data are missing for few countries (namely 2005 for Australia, 1997–1998 for Poland, 2000 for UK and 2015 for New Zealand), we applied the approach suggested by Kraus (2015) to impute missing parts of the curves.

3. Methodology

As pointed out in the previous sections, CSMRs data can be thought as realizations of random functions taking values in the D–dimensional simplex. Petersen and Müller (2016) and Hron et al. (2016) consider the similar problem of analyzing samples of densities, i.e. functional data with the additional constraints $f(x) \geq 0$ and $\int f(x)dx = 1$ that can be though as *continuous compositions*: the density at a specific point x is an infinitesimal part of the density function and the total contribution of the parts is fixed to 1. The main difference with the present work is that in our case we consider *compositional functional data*: discrete compositions where each part is a function $f_d(t)$ and the constraint $\sum_d f_d(t) = 1$ holds for every t on a given interval. As an alternative to the methodology we describe in the next sections, the approach of Dai and Müller (2018) using a different transformation can be considered. As suggested in Scealy et al. (2015), in the applications one should always check which transformation is the most appropriate.

This section aims to furnish the main tools to analyze *compositional functional data* and, to do that, we extend the theory in Aitchinson (1986) about standard compositions to deal with our complex data structure. In particular, computation of the mean function, covariance operator and principal component analysis is presented in sections 3.1-3.2 while section 3.3 discusses details about a clustering procedure we applied to data.

3.1. The functional simplex

A functional composition can be defined as a random function $\boldsymbol{f} : \mathcal{I} \subset \mathbb{R} \to \mathcal{S}^D$ from a subset of the real line to the D–dimensional simplex. With different terms, a functional composition is a multivariate random function, $\boldsymbol{f}(t) = [f_1(t) \dots f_D(t)]$ where each $f_d : \mathcal{I} \subset \mathbb{R} \to \mathbb{R}, d = 1, \dots, D$ is a function from a subset of the real line to $\mathbb{R}$ with the additional constraint that $f_d(t) \geq 0 \ \forall d, \forall t \in \mathcal{I}$ and $\sum_{d=1}^{D} f_d(t) = 1 \ \forall t \in \mathcal{I}$. We refer to realizations of functional compositions as compositional functional data. These data enjoy most of the properties of multivariate functional data but the constraint imposes some important modifications. Aitchinson (1986) shows that compositional data lie on the simplex which has a different geometry with respect to $\mathbb{R}^D$. Similarly, compositional functional data lie in a function space that we call *functional simplex* for which usual operations in $\mathcal{L}_2$ cannot be applied. We define the functional simplex $\mathcal{S}_f^D(\mathcal{I})$ as the collection of all D–variate functions $\boldsymbol{f} : I \subset \mathbb{R} \to \mathcal{S}^D$ that satisfy

$$\sum_{d=1}^{D} \int_{\mathcal{I}} \Big[\log\Big\{\frac{f_d(t)}{\tilde{f}(t)}\Big\}\Big]^2 dt < \infty \tag{3}$$

where $\tilde{f}(t) = \big[\prod_{d=1}^{D} f_d(t)\big]^{1/D}$ is the geometric mean of the components of $\boldsymbol{f}$. The functional simplex is a separable Hilbert space equipped with vector operations as summation and multiplication to a scalar. Specifically, the sum of two elements in this space takes the form of a perturbation

$$\boldsymbol{f} \oplus \boldsymbol{g} = \mathscr{C}[f_1(t) \cdot g_1(t), \dots, f_D(t) \cdot g_D(t)], \tag{4}$$

and multiplication to a scalar becomes powering

$$\alpha \odot \boldsymbol{f} = \mathscr{C}[f_1^{\alpha}(t), \dots, f_D^{\alpha}(t)], \tag{5}$$

where $\mathscr{C}$ denotes the closure operator, i.e. $\mathscr{C}\boldsymbol{x} = \big[\frac{x_1}{\sum x_i}, \dots, \frac{x_D}{\sum x_i}\big]$. These operations are the natural extensions of perturbation and powering on the simplex, and their existence is necessary for the computation of quantities relevant from a statistical point of view. In addition, the geometry of the space can be defined by its inner product that, for any two functional compositions $\boldsymbol{f}$ and $\boldsymbol{g}$, is

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{S}_f^D} = \sum_{d=1}^{D} \int_{\mathcal{I}} \log\Big\{\frac{f_d(t)}{\tilde{f}(t)}\Big\} \log\Big\{\frac{g_d(t)}{\tilde{g}(t)}\Big\} dt. \tag{6}$$

The centered log ratio (clr) transform (Aitchinson, 1986; Egozcue and Pawlowsky-Glahn, 2011) for compositional data is a map between the D–dimensional simplex and $\mathbb{R}^{D-1}$. This transformation is particularly important since not only establishes an isomorphism between the two spaces, but it allows to compute quantities on the simplex through operations outside of it. Similar transformations can be considered as, for example, the additive log ratio (alr) and the isometric log ratio (ilr). The methodology described here can be applied for these choices as well. However, we chose the clr transform for its wider use in the applications and a simpler structure for the presentation. Scealy et al. (2015) and Dai and Müller (2018) highlight that log-ratio transformation might be inappropriate when compositions are equal or very close to zero. This

issue does not apply to data considered here, since marginal causes such as ill-identified ones and infant mortality related ones, have been excluded. Hence, clr is an appropriate transformation, while when dealing with compositions with several values close to 0, the transformations suggested by Dai and Müller (2018) should be considered.

Here we define the clr transform for functional compositions. This extension can be though as the application of the standard clr transform to the composition evaluated at a specific point t, for every $t \in \mathcal{I}$. Let $\mathcal{U}$ be the subset of the cartesian product of D copies of $\mathcal{L}_2$ defined as $\mathcal{U} = \{\boldsymbol{u}(t) \in \mathcal{L}_2^D | \sum_{d=1}^D u_d(t) = 0\}$, then $\mathcal{U}$ is isomorphic to $\mathcal{L}_2^{D-1}$. The centered log ratio transform for functional compositions is a function $\mathtt{clr} : \mathcal{S}_f^D \to \mathcal{U} \subset \mathcal{L}_2^D$ from the functional simplex to $\mathcal{U}$ defined by

$$\boldsymbol{f}^* = \mathtt{clr}\{\boldsymbol{f}\} = \log\left[\frac{f_1(t)}{\tilde{f}(t)} \dots \frac{f_D(t)}{\tilde{f}(t)}\right], \tag{7}$$

where $\tilde{f}(t)$ is the geometric mean of the components of $\boldsymbol{f}$. This function maps elements of the functional simplex into elements of a subspace of $\mathcal{L}_2^D$, allowing us to work outside the functional simplex. To go back to the original space, we use the inverse transform $\mathtt{clr}^{-1} : \mathcal{U} \subset \mathcal{L}_2^D \to \mathcal{S}_f^D$ defined by

$$\boldsymbol{f} = \mathtt{clr}^{-1}\{\boldsymbol{f}^*\} = \mathscr{C}\{\exp[f_1^*(t) \dots f_D^*(t)]\}. \tag{8}$$

As an immediate consequence of these definitions we have that the functional simplex is isomorphic to $\mathcal{L}_2^{D-1}$ and the inner product defined in (6) is the same as the usual $\mathcal{L}_2^D$ inner product between the two clr transformed functions, i.e. $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{S}_f^D} = \langle \boldsymbol{f}^*, \boldsymbol{g}^* \rangle_{\mathcal{L}_2^D}$. The clr transform has the important property that, for any two real constants α and β and any two functional compositions $\boldsymbol{f}$ and $\boldsymbol{g}$,

$$\mathtt{clr}\{(\alpha \odot \boldsymbol{f}) \oplus (\beta \odot \boldsymbol{g})\} = \alpha \cdot \mathtt{clr}\{\boldsymbol{f}\} + \beta \cdot \mathtt{clr}\{\boldsymbol{g}\}. \tag{9}$$

This property is central in the computation of the mean and principal components of compositional functional data, since perturbations and powerings on the functional simplex can be performed by other operations in $\mathcal{L}_2^D$.

3.2. Mean, covariance operator and principal component analysis

We define the expectation of a functional composition as $\boldsymbol{\mu}_f = \mathbb{E}\boldsymbol{f} = \mathtt{clr}^{-1}\{\mathbb{E}\mathtt{clr}\{\boldsymbol{f}\}\}$. Suppose to have a sample of compositional functional data $\{\boldsymbol{f}_1(t), \dots, \boldsymbol{f}_n(t)\}$ on a given time interval $\mathcal{I}$ as, for example, trends of CSMRs. The mean function $\boldsymbol{\mu}_f$ on the functional simplex representing the average trend can be estimated by

$$\widehat{\boldsymbol{\mu}}_f(t) = \bigoplus_{i=1}^n \left[\frac{1}{n} \odot \boldsymbol{f}_i(t)\right]. \tag{10}$$

Equation (10) consists in perturbations and powerings and, thanks to property (9) we can compute these quantities only through operations outside the functional simplex:

$$\widehat{\boldsymbol{\mu}}_f(t) = \mathtt{clr}^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \mathtt{clr}\{\boldsymbol{f}_i(t)\}\right\}. \tag{11}$$

Equation (10) extends the classical definition of an average of a composition to the case of functional compositions. As shown in Aitchinson (1986), the mean in simplicial terms can be obtained by back transforming the arithmetic mean of transformed quantities. This is justified by the isomorphism between the two spaces we consider. In the next section, we will apply estimator (11) to the sample of countries introduced in section 2.1 in order to depict the average trend of CSMRs.

Next we have to define a valid covariance operator for functional compositions. The definition of covariance is not straightforward for compositional data: Aitchinson (1986) proposed three different specifications based on transformations of data. One of them relies on the clr transform and we follow this specification in our work. Basically, for a composition $\boldsymbol{x}$ each entry of the covariance matrix $\mathbf{R}$, r_{jl}, is equal to the covariance between the j–th and l–th transformed part of $\boldsymbol{x}$, i.e. $r_{ij} = \text{cov}(\texttt{clr}\{x_j\}, \texttt{clr}\{x_l\})$. Thus, the covariance operator of a functional composition can be defined as the integral operator $\mathscr{R}\boldsymbol{f}^* = \int r(s,t)\boldsymbol{f}^*(t)dt$ where $r(s,t)$ is a $D \times D$ block kernel such that for block $\{j,l\}$

$$r_{jl}(s,t) = \mathbb{E}\Big\{\texttt{clr}\{f_j(s)\} \cdot \texttt{clr}\{f_l(t)\}\Big\} = \mathbb{E}\left\{\log\left(\frac{f_j(s)}{\tilde{f}(s)}\right) \cdot \log\left(\frac{f_l(t)}{\tilde{f}(t)}\right)\right\}, \tag{12}$$

where for simplicity we assume $\mathbb{E}\boldsymbol{f}^* = 0$. Similarly to the scalar case, the covariance kernel $r(s,t)$ is singular in the sense that satisfies the conditions $\sum_{j=1}^{J} r_{jl}(s,t) = \sum_{l=1}^{L} r_{jl}(s,t) = 0$. However it is still positive semi–definite and admits the decomposition

$$r(s,t) = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\phi}_k^*(s)\boldsymbol{\phi}_k^*(t), \tag{13}$$

where $\{\lambda_1, \lambda_2, \ldots\} \in \mathbb{R}$ are the eigenvalues and $\{\boldsymbol{\phi}_1^*, \boldsymbol{\phi}_2^*, \ldots\} \in \mathcal{U}$ are the eigenfunctions of the covariance operator $\mathscr{R}$. Since the eigenfunctions of the covariance operator $\mathscr{R}$ constitute an orthonormal basis for $\mathcal{U}$, their compositional counterparts $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots\}$ obtained through the function $\texttt{clr}^{-1}$ are an orthonormal basis in the functional simplex. Finally, as a consequence of the Karhunen–Loève theorem, we can represent every functional composition as

$$\boldsymbol{f} \ominus \boldsymbol{\mu}_f = \bigoplus_{k=1}^{\infty} \Big[\xi_k \odot \boldsymbol{\phi}_k(t)\Big], \tag{14}$$

where $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots\} \in \mathcal{S}_f^D$ are transformed eigenfunctions, i.e. $\boldsymbol{\phi}_k = \texttt{clr}^{-1}\{\boldsymbol{\phi}_k^*\}\ \forall k$ and $\{\xi_1, \xi_2, \ldots\} \in \mathbb{R}$ are computed as $\xi_k = \langle \boldsymbol{f}, \boldsymbol{\phi}_k \rangle_{\mathcal{S}_f^D}$. We will call $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots\}$ principal components. Again, for (9) we have that

$$\boldsymbol{f}(t) = \texttt{clr}^{-1}\left\{\texttt{clr}\{\boldsymbol{\mu}_f(t)\} + \sum_{k=1}^{\infty} \xi_k \texttt{clr}\{\boldsymbol{\phi}_k(t)\}\right\}. \tag{15}$$

To summarize, in order to extract principal components we considered the covariance kernel of transformed functions rather than original functions, and then we back transformed the eigenfunctions of the kernel. Other approaches using a different transformation, like in Dai and Müller (2018), or the application of PCA directly to the original data are possible but, as already pointed out, since none of the data in the application is zero o very close to zero, our method

poses no issues. Moreover we want to point out that, although computations have been done through the use of the clr transform, the quantities of interest live in the original compositional space.

Given a finite sample of compositional functional data $\{\boldsymbol{f}_1(t), \ldots, \boldsymbol{f}_n(t)\}$, the estimation of each block of the covariance kernel can be done through

$$\widehat{r}_{jl}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(\frac{f_{ij}(s)}{\tilde{f}_i(s)}\right)\cdot\log\left(\frac{f_{il}(t)}{\tilde{f}_i(t)}\right)\right\}, \tag{16}$$

where $f_{ij}(t)$ and $f_{il}(t)$ are the j–th and l–th part of the i–th observed function minus the sample mean. The eigenfunctions of the covariance operator $\mathscr{R}$ can be computed as the solution of the system of equations $\int r(s,t)\boldsymbol{f}^*(t)dt = \lambda\boldsymbol{f}^*(t)$, as shown in Ramsay and Silverman (2005). From a practical point of view, one has to compute empirical eigenfunctions $\{\widehat{\boldsymbol{\phi}}_1^*, \widehat{\boldsymbol{\phi}}_2^*, \ldots\}$ using the covariance kernel $\widehat{r}(s,t)$ estimated from data, solving the problem

$$\int \widehat{r}(s,t)\boldsymbol{f}^*(t)dt = \lambda\boldsymbol{f}^*(t). \tag{17}$$

This can be done by discretitizing the integral and by using standard arguments from eigenanalysis. Then, transformed eigenfunctions can be obtained as $\widehat{\boldsymbol{\phi}}_k = \texttt{clr}^{-1}\{\widehat{\boldsymbol{\phi}}_k^*\}$. Computation of the empirical eigenfunctions is important because the result provides information about the main modes of variation in the considered sample. For complex data as CSMRs over time, these functions represents a straightforward way to inspect the simultaneous directions of variability of all causes of death. The empirical scores $\{\widehat{\xi}_{i1}, \widehat{\xi}_{i2}, \ldots\}$ for the i–th observation $\boldsymbol{f}_i$ can be computed as

$$\widehat{\xi}_{ik} = \langle \boldsymbol{f}_i^*, \boldsymbol{\phi}_k^*\rangle = \int \texttt{clr}\{\boldsymbol{f}_i\}(t)\boldsymbol{\phi}_k^*(t)dt. \tag{18}$$

Finally, by truncating the infinite series in (15) and replacing population quantities with estimates, one obtains the approximation

$$\boldsymbol{f}(t) \simeq \texttt{clr}^{-1}\left\{\texttt{clr}\{\widehat{\boldsymbol{\mu}}_f(t)\} + \sum_{k=1}^{K}\widehat{\xi}_k\texttt{clr}\{\widehat{\boldsymbol{\phi}}_k(t)\}\right\}. \tag{19}$$

This representation is particularly useful because it gives an approximation of a functional composition using few nonrandom functions shared among observations and few scalar random variables, thus reducing the problem to a finite–dimensional one. The relevance in our application is that CSMRs observations can be summarized by their behaviour with respect to few main modes of variation. The projections on the finite–dimensional space can also serve as a basis for a clustering procedure, as explained in the next subsection.

3.3. The Clustering Step

A common practice in FDA is to reduce the dimensionality of data and use projections on a finite dimensional space as a starting point for other procedures, such as clustering or classification (see for instance Leng and Müller (2006); Sangalli et al. (2009)). In our context, the compositional

functional PCA allows us to drastically reduce the complexity of CSMRs over time by considering only few components, accounting for most of the variability in the sample. Then, in order to detect a clustering structure, one possibility is to apply a procedure for euclidean data to the scores of our original compositional functional data.

Spectral clustering (see von Luxburg (2007) for an overview) is a relatively recent clustering method. It is based on the eigenanalysis of the Laplacian matrix constructed on the similarity graph of the data. Its advantage over alternative procedures is the ability to detect complex and nonlinear structures. The application of this method to our context needs no further adjustments with respect to the original formulation, but we have to choose some parameters. In particular, we set the gaussian similarity kernel defined as $s(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \exp(-\left\|(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\right\|^2/(2\sigma^2))$ with $\sigma = 1$ as kernel to compute the similarity graph among scores, and k–means (Izenman, 2008) as base method. The output of the clustering procedure is given by a set of G centroids $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_G$ and a vector of labels of cluster membership. Since the method is sensible to initialization, we run spectral clustering algorithm for B times and compute the *majority vote* as the most frequent partition of data over B repetitions. For this method, the decay of the spectrum of the Laplacian matrix can be informative about the number of clusters. However, we did not observe a clear signal when we applied the method to CSMRs data. Thus we decided to use the silhouette index for determining the number of clusters, defined as the average of all silhouette values $s(i)$,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{20}$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_k)^2, \qquad a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)^2, \tag{21}$$

where $\boldsymbol{\xi}_i$ is the vector of scores for the i–th observation and $|C_k|$ is the cardinality of the k–th cluster (Izenman, 2008). The silhouette index ranges from -1 to 1 and higher values correspond to better cluster configurations.

4. Results

In this section we present the results of the analysis on the WHO mortality database separately for men and women. We first compute the compositional functional mean as in (11), then we estimate the covariance kernel using (16) and compute principal components and scores with (17) and (18). Finally, we apply spectral clustering to the scores in order to obtain a grouping structure of the countries in the study. The silhouette values for each country are reported in Table 1 of the supplementary material.

4.1. Men

The 22 curves of the men sample for 8 causes of death we consider are given in Figure 1. We depict with a larger colored line the compositional functional mean. The ranges among causes are different, denoting a different contribution on the total mortality rate. Highest ranges are for circulatory diseases (0.20–0.50) and neoplasms (0.10–0.40), representing the two most

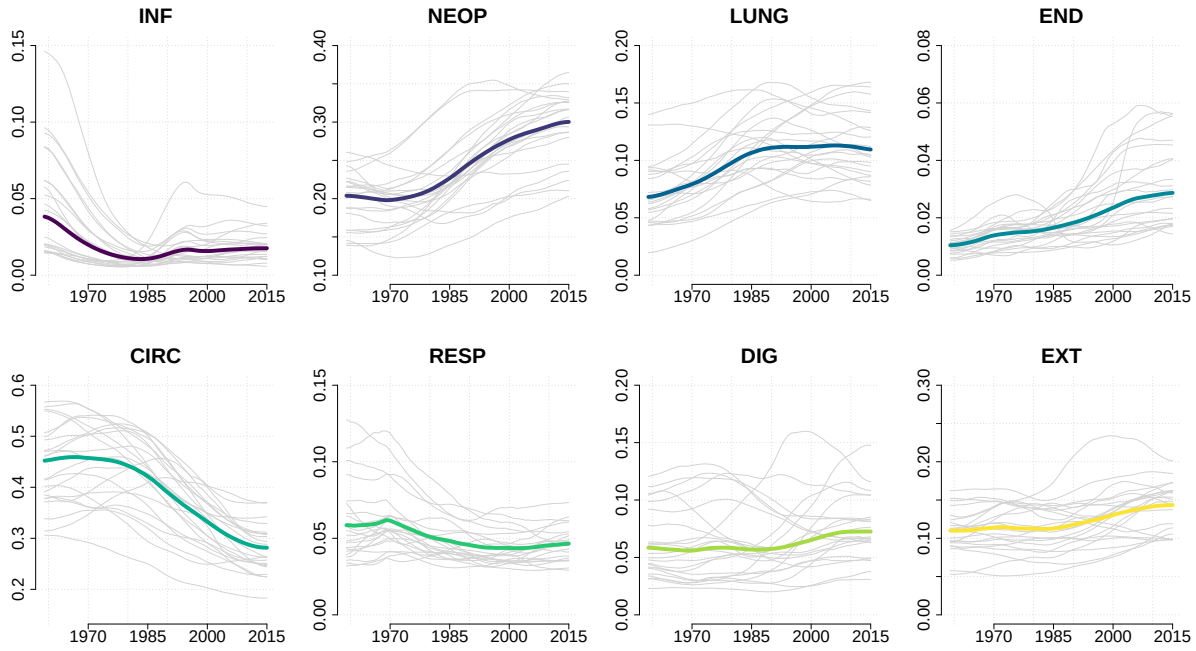

Fig. 1. Composition of mortality rates over years for men population. Curves for 22 countries in each panel are colored in grey. Colored curves represent the compositional functional mean.

important causes of death from 1959 to 2015. However, dynamics are quite different among all causes: neoplasms and lung cancer in particular exhibit an increasing trend over years, while circulatory diseases constantly decrease their contribution. A slightly increasing trend can be noticed also for endocrine, metabolic and nutritional diseases and external causes while respiratory and digestive diseases show heterogeneous patterns: some countries have a declining trend, others an increasing one. Infectious diseases experienced a particular behavior over years: they diminish their contribution at the beginning of the period, then started to increase and finally stabilized around a value of 0.02. These results are in line with the epidemiological transition theory with increasing prevalence of diseases associated with aging (e.g., neoplasms) and of man-made diseases (lung cancer caused by smoking, metabolic and nutritional disorders, caused by obesity and external causes of deaths) and a decline of infectious diseases. Inspection of individual trajectories reveals that some countries have a peculiar pattern: Finland, for example, has the highest level of external causes – see (Saarela and Finnas, 2008) – whereas for Greece decline of circulatory diseases starts much later than the other countries.

The first compositional functional principal component for men population is depicted in Figure 2. This function is computed as described in section 3.2 and represents the main mode of variation of the men sample. The eigenvalue associated with this eigenfunction is $\lambda_1 = 17.06$ and the Fraction of Explained Variance (FEV) is $\lambda_1 / \sum \lambda_k = 0.337$, thus this component account for about one third of the total variability. This component is related to all causes but respiratory diseases. A variability connected with the *level* of the cause – values always above or below the mean – is present in digestive diseases. For infectious and endocrine diseases, as we already noticed, the main variability is related only to specific years – 1959-1985 for the former and

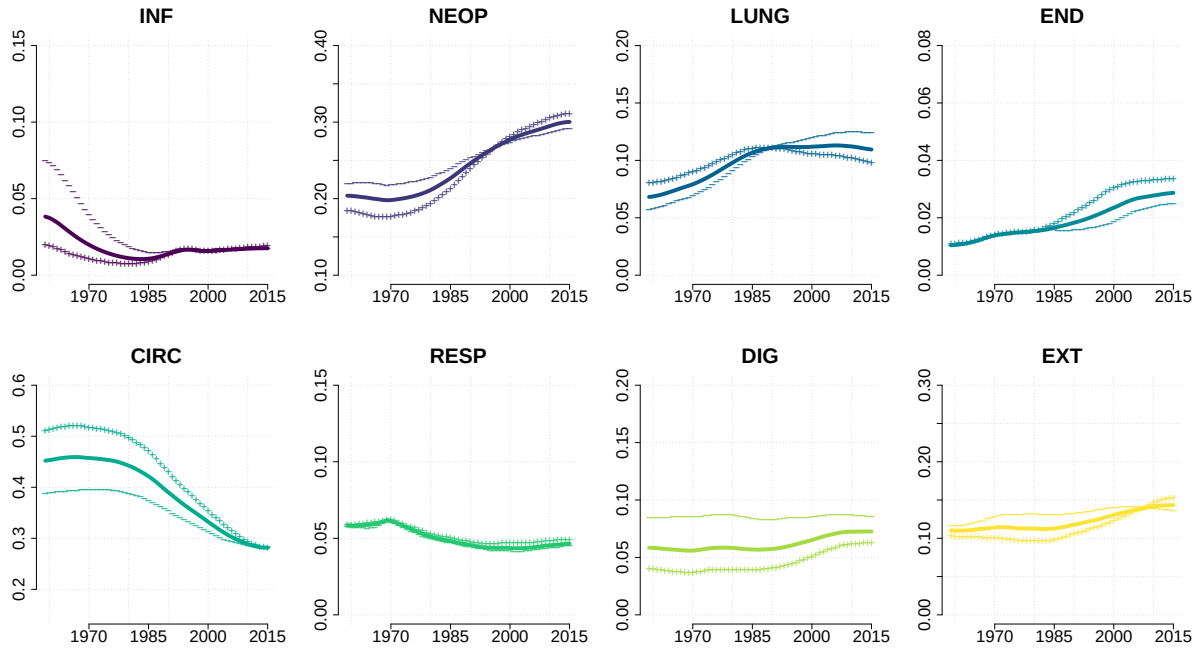

Fig. 2. First compositional functional principal component for men population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

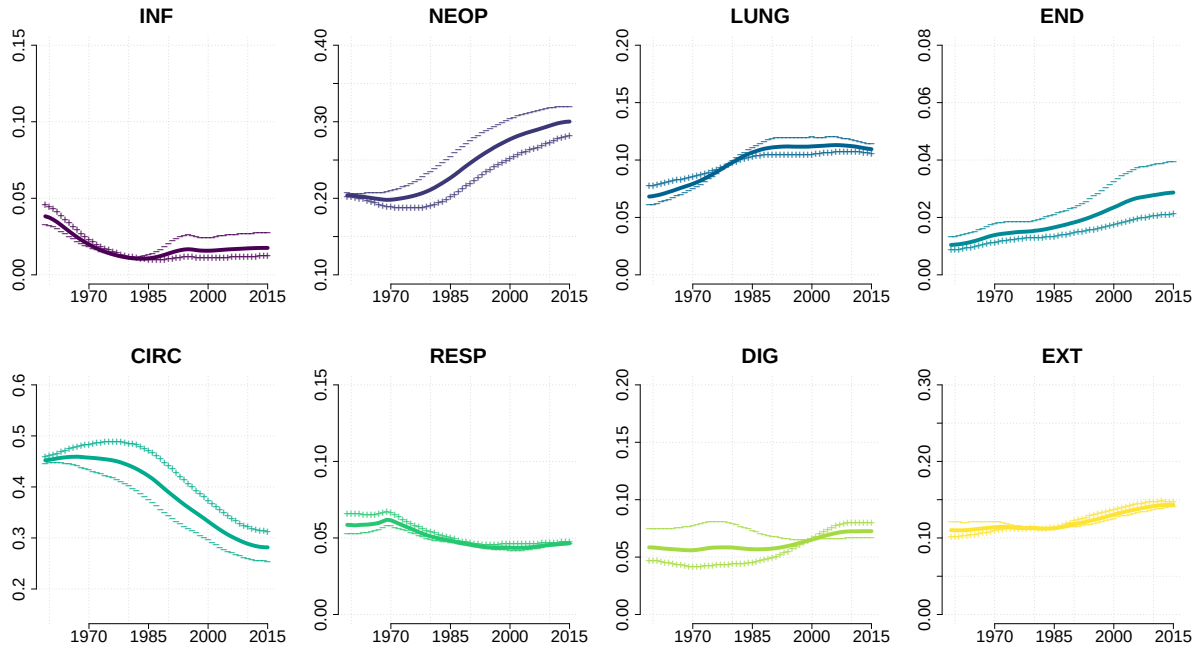

Fig. 3. Second compositional functional principal component for men population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

1985-2015 for the latter – and we call this aspect *local* level variability. This behaviour can be partially observed for circulatory diseases and external causes too. Lastly, a more structured

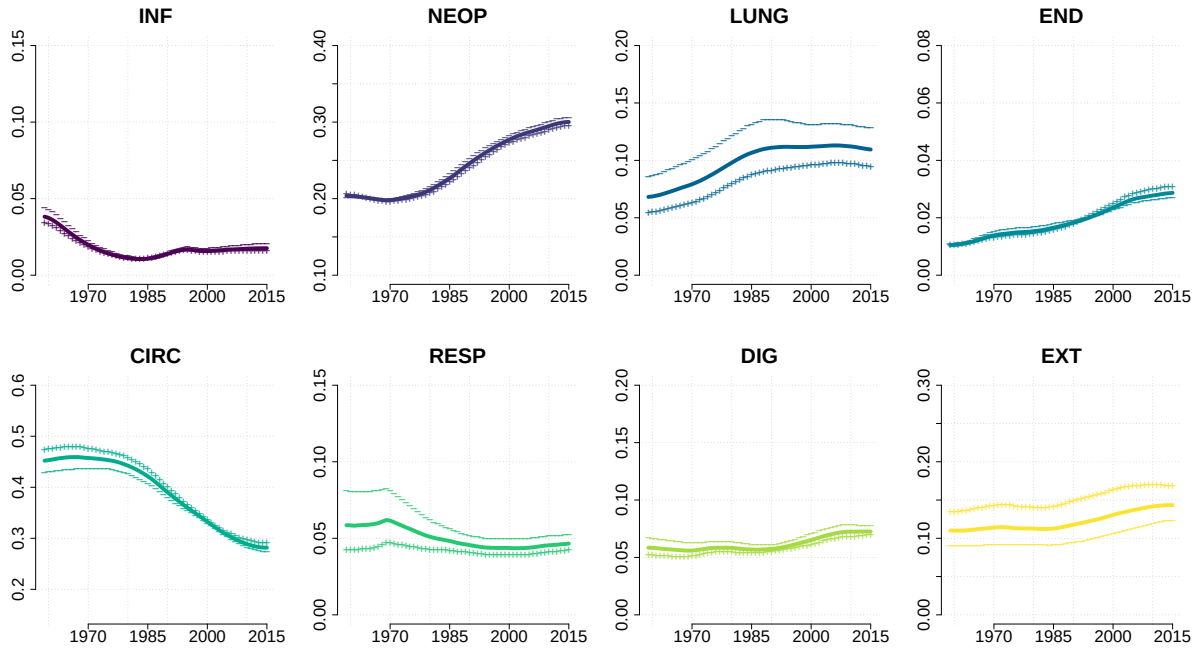

Fig. 4. Third compositional functional principal component for men population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

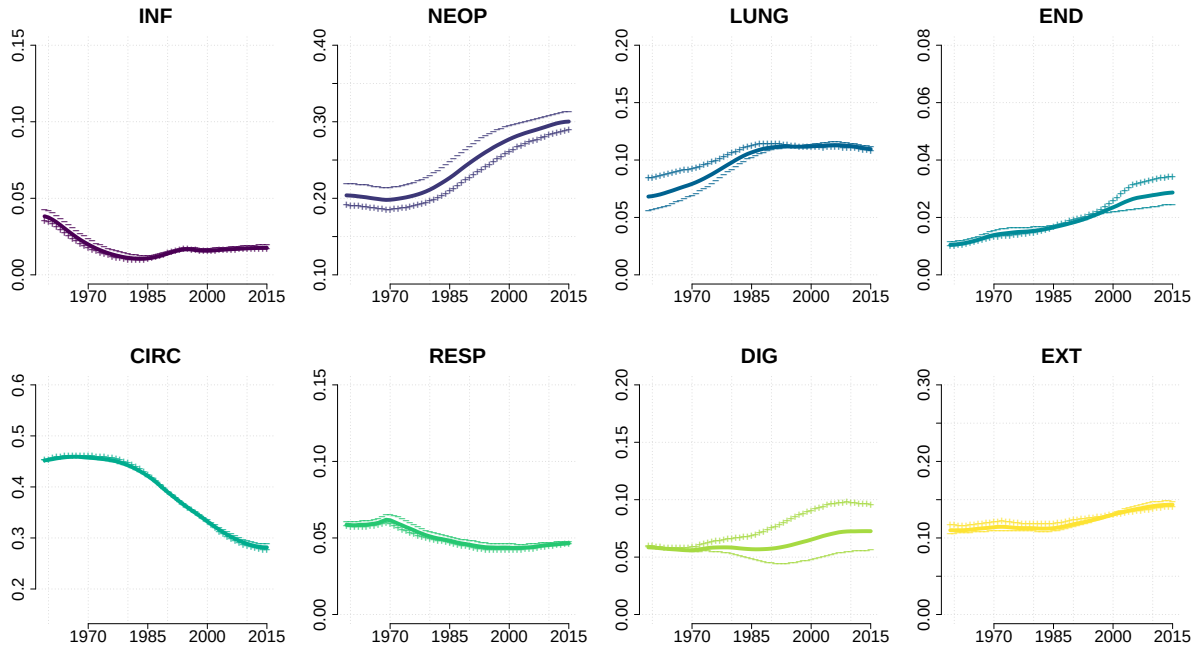

Fig. 5. Fourth compositional functional principal component for men population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

pattern can be seen for neoplasms and lung cancer: for these causes the first compositional functional principal component crosses the mean at some point. This reflects the fact that in

Table 3. Scores of the 22 countries in the study. Men sample is reported on the left and women sample on the right.

Countries	PC1	PC2	PC3	PC4
AUS	4.38	-2.08	0.32	0.28
AUT	-3.83	-1.03	0.95	3.90
BEL	0.44	1.43	-3.02	-0.21
CAN	3.26	-3.80	0.80	0.75
DNK	3.41	-1.51	0.23	3.42
FIN	-0.63	6.34	3.59	3.36
FRA	-5.20	-3.66	-0.10	0.18
GRE	-2.90	1.77	-2.14	-2.28
HUN	-5.80	4.11	0.09	3.32
ICE	3.01	3.13	4.77	-3.32
IRL	2.15	4.72	-2.67	-2.63
ITA	-2.49	-3.81	-2.90	0.64
JPN	-6.76	-2.10	2.89	-2.98
NL	5.55	-1.00	-2.77	0.72
NZL	6.49	-0.69	0.82	-2.21
NOR	2.47	0.72	2.93	-1.17
POL	-5.27	3.93	0.19	-1.58
SPA	-5.76	-2.51	-3.61	-2.02
SWE	1.37	-1.45	4.52	0.43
SWI	-0.65	-2.19	0.55	0.06
UK	4.28	4.84	-5.20	1.05
USA	2.50	-5.16	-0.24	0.27

Countries	PC1	PC2	PC3	PC4
AUS	3.08	1.87	0.86	-0.42
AUT	-2.76	1.02	-1.31	-3.53
BEL	-1.67	2.06	-0.51	-0.01
CAN	2.74	3.76	0.17	-0.37
DNK	4.12	1.92	-2.26	-3.47
FIN	-3.01	-2.79	-3.92	0.06
FRA	-6.77	2.70	-2.35	0.90
GRE	-0.52	-2.37	2.50	0.72
HUN	-2.96	-3.25	-1.79	-5.40
ICE	5.40	0.21	-4.05	4.43
IRL	4.00	-5.82	1.25	1.01
ITA	-3.72	1.75	4.19	-1.47
JPN	-4.21	-1.69	-0.99	4.99
NL	1.98	2.59	0.75	-1.22
NZL	5.10	-0.62	3.00	0.59
NOR	2.37	-0.71	-1.49	1.85
POL	-3.63	-4.13	1.14	-0.74
SPA	-5.71	-0.81	4.00	1.87
SWE	1.02	1.40	-2.02	0.17
SWI	-2.54	2.31	-1.56	0.81
UK	5.97	-3.21	0.61	-2.05
USA	1.73	3.82	3.76	1.29

our sample some countries had a faster increase of this causes with respect to the mean while others had a slower increase than the mean. Overall, a positive score on this component would lead to a high level of circulatory diseases and endocrine diseases (only after 1985), a low level of digestive diseases and infectious diseases (only before 1985), a fast increase of neoplasms and a slow increase of lung cancer. Negative scores characterize an opposite behaviour. The scores obtained by each country on the first 4 PCs for men sample are reported in Table 3.

Figure 3 shows the second compositional functional principal component for men population. The eigenvalue associated with this eigenfunction is $\lambda_2 = 10.92$ and the FEV is 0.216. This component represents a *level* variability for neoplasms, circulatory and endocrine diseases, especially in the last 30 years. *Local* variability is observed for infectious diseases – note that the time interval connected with this component is 1985-2015 – and for respiratory diseases, but to a lesser extent. For lung cancer, digestive diseases and external causes this component shows a cross–mean behaviour, reflecting different velocities among countries in the evolution of these contributions. Overall, a positive score on this component would lead to a high level of circulatory diseases, a low level of neoplasms and endocrine diseases, a fast decay of infectious diseases, a slow increase of lung cancer and an increase of digestive diseases after 2000.

The third compositional functional principal component for men population is illustrated in Figure 4. The eigenvalue associated with this eigenfunction is $\lambda_3 = 7.27$ and the FEV is 0.144. This component is mostly connected with four causes of death. It represents a *level* variability for lung cancer, respiratory diseases and external causes while *local* variability is observed for circulatory diseases, in relation to the period 1959-1990. Overall, a positive score on this com-

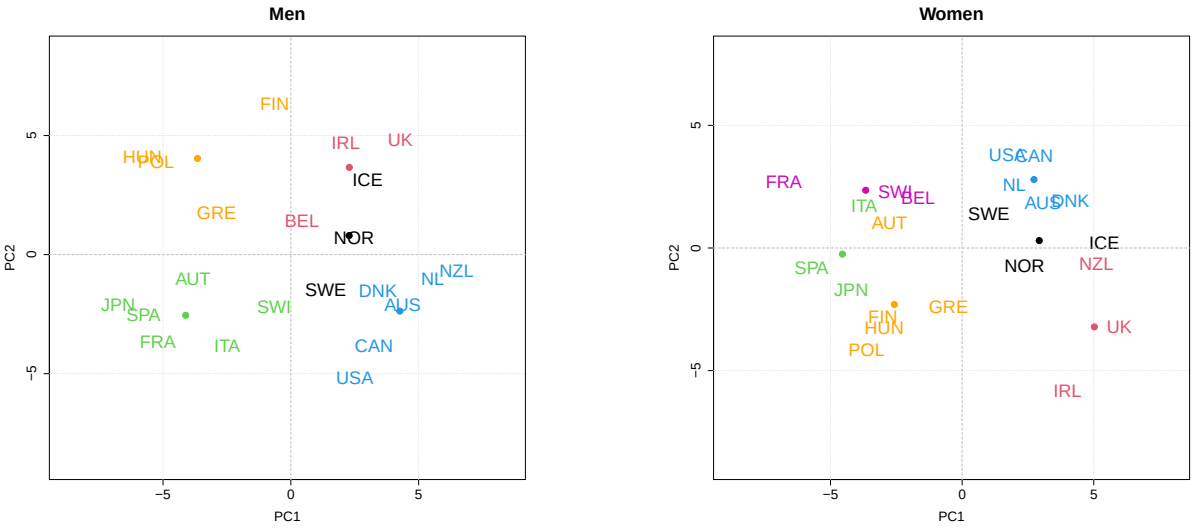

Fig. 6. Empirical scores $\{\widehat{\xi}_{i1}, \widehat{\xi}_{i2}\}_{i=1}^{n}$ related to the first 2 principal components computed for all countries in the sample for men and women. Different colors represent different clusters.

ponent would lead to a low level of lung cancer and respiratory diseases and to a high level of external causes and circuatory diseases – especially at the beginning of the considered period. Figure 5 depicts the fourth compositional functional principal component for men population. The eigenvalue associated with this eigenfunction is $\lambda_4 = 4.62$ and the FEV is 0.091. Also this component refers to only four causes of death. It represents a *level* variability for neoplasms and a *local* variaility for lung cancer (1959-1995), endocrine diseases (after 2000) and digestive diseases (from 1970). Overall, a positive score on this component would lead to a low level of neoplasms and a high level of lung cancer, endocrine diseases and digestive diseases in the aforementioned periods.

A clustering procedure has been applied to the men sample as described in section 3.3. We considered the first 4 PCs, accounting for 78.7% of the total variability. For different values of G, we ran the spectral clustering algorithm for $B = 1000$ times and save the result of the *majority vote* . The spectrum of the Laplacian matrix shows no evidence for a particular number of clusters, while the silhouette index is maximized for $G = 5$ clusters, which are reported in Table 4. A visual representation of the scores of the first two PCs along with the clustering structure can be found in Figure 6. Using (19) with $K = 4$ and $\gamma_{1g}, \gamma_{2g}, \gamma_{3g}, \gamma_{4g}$ as the centroids of the spectral clustering output, we reconstructed the compositional functional centroids $\boldsymbol{f}_g, g = 1, \ldots, G$, depicted in Figure 7. These trajectories summarize the behavior of each cluster with respect to the eight causes of death.

We can draw some comments. The first group (including USA, Canada, Australia, New Zealand, Netherlands and Denmark) is characterized by relatively low circulatory related deaths (but at the beginning of the observation period it was high) a decreasing trend of lung cancer and an almost plateauing trend of neoplasms. Notably the level of endocrine and metabolic related deaths is much higher than other clusters, with a particularly high slope between 1980 and 2000. This trend is close to that of obesity prevalence in USA, as shown, for example by Fryar et al. (2014), and New Zealand and Australia show similar figures, so it is likely that this is

Table 4. Clustering output for men sample.

Clusters	Countries
Cluster 1	USA, Canada, Australia, New Zealand, Denmark, Netherlands
Cluster 2	Italy, France, Spain, Austria, Switzerland, Japan
Cluster 3	Hungary, Poland, Finland, Greece
Cluster 4	UK, Ireland, Belgium
Cluster 5	Norway, Sweden, Iceland

Table 5. Clustering output for women sample.

Clusters	Countries
Cluster 1	USA, Canada, Australia, Denmark, Netherlands
Cluster 2	Italy, Spain, Japan
Cluster 3	France, Switzerland, Belgium
Cluster 4	Hungary, Poland, Austria, Finland, Greece
Cluster 5	UK, Ireland, New Zealand
Cluster 6	Norway, Sweden, Iceland

the leading cause. It should also be noted that external causes of death are rapidly increasing in the last years, in line with what is shown by Woolf and Schoomaker (2019). The second group consists of Italy, Spain, France, Japan, Austria and Switzerland, and is characterized by high lung cancer mortality – in all countries smoking prevalence is high among men, see for instance, Brennan and Bray (2002) – and low circulatory-related deaths. At the beginning of the observation period up to 1980, this group had particularly high level of digestive system diseases and infectious related deaths, as for most of them the epidemiological transition started later than in other countries. Third group is made up of Hungary, Poland, Finland and Greece. The main feature of this group is the increasing relevance of digestive system diseases related deaths. This might be driven by rising alcohol–attributable diseases, considering that such health problems are a growing concern in these countries. It must be noted that this is the least homogeneous group, as can be seen from Table 1 of the supplementary material. Finland, for example, has a peculiar pattern with an extremely high incidence of external causes of deaths that is more consistent with the group including Sweden, Norway and Iceland but, at the same time, a rapidly increasing prevalence of digestive diseases that matches more with Eastern Europe group. Greece, instead is in between cluster 3 and 2 having a particularly high incidence of lung cancer (as in cluster 2) but also a high level – and very slowly declining – of circulatory diseases related deaths, as in cluster 3. The fourth group is made up of UK, Ireland and Belgium, characterized by high – albeit descending – levels of lung cancer, respiratory diseases and circulatory diseases and increasing relevance of digestive system diseases. This might be brought about by risk factors like smoking, alcohol consumption and poor diet. The last cluster includes Sweden, Norway and Iceland, characterized by the lowest level of lung cancer and respiratory diseases and the highest level of external causes.

4.2. Women

The 22 curves of the women sample for 8 causes of death we consider are represented in Figure 8. Similarly to men population, the ranges among causes are different. Highest ranges are observed

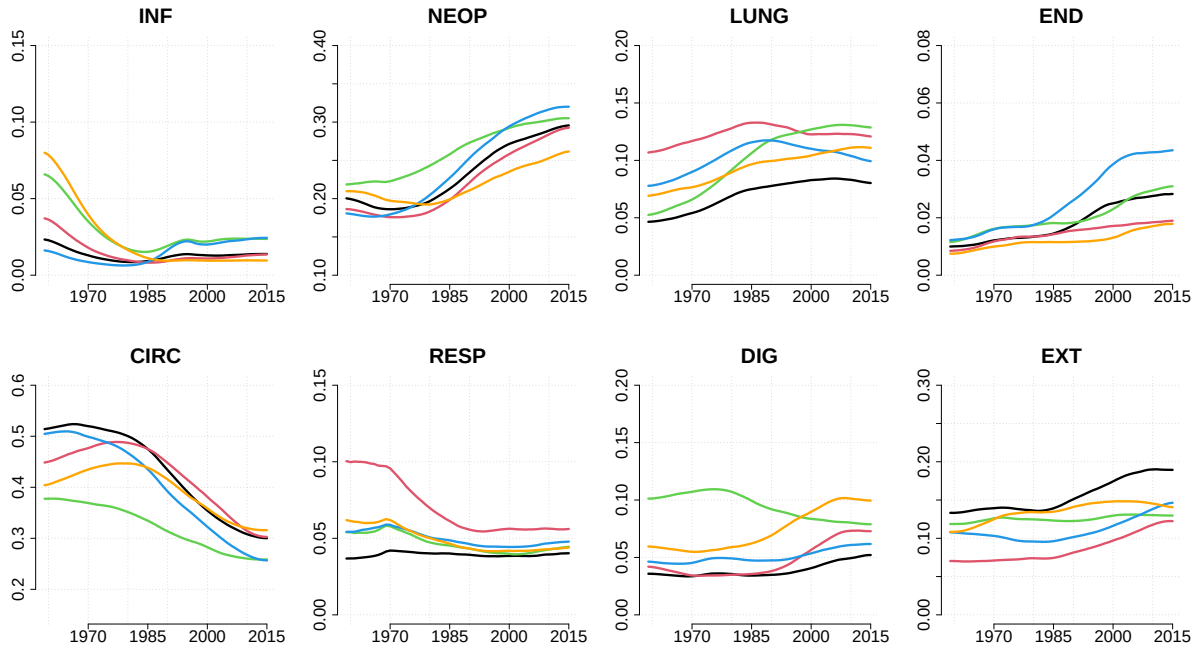

Fig. 7. Functional centroids of the spectral clustering for men sample. Legend: Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5.

for circulatory diseases (0.1–0.4), neoplasms (0.3–0.6) and lung cancer (0.02–0.2). With respect to men, women have a higher contribution of neoplasms and a lower contribution of circulatory diseases, though trends over years are similar: the former moved from 0.38 to 0.48 and the latter reduced from 0.4 to below 0.2. Lung cancer had almost negligible contribution at the beginning of the considered period but, thanks to a strong and constant increase, it is nowadays as important as in men (slight above 0.1). On the other side, the positive trend for other neoplasms seems to have reached a plateau since the last decade of the twentieth century. An opposite behaviour can be found in men, where lung cancer had a stable contribution since 1985 while other neoplasms are still increasing. Another difference can be found for endocrine diseases, where women do not show an increasing trend. Lastly, infectious diseases experienced a similar evolution for both men and women. Inspection of individual trajectories reveals that Austria has a peculiar pattern characterized by a strong raise of endocrine diseases around year 2000. The first compositional functional principal component for women population is shown in Figure 9. It represents the main mode of variation of the women sample. The eigenvalue associated with this eigenfunction is $\lambda_1 = 14.87$ and the Fraction of Explained Variance (FEV) is $\lambda_1 / \sum \lambda_k = 0.326$, thus, similarly to men, this component accounts for about one third of total variability. Surprisingly, this component is related to all causes but circulatory diseases, one of the main causes of death. A *level* variability can be found for lung cancer, respiratory and digestive diseases. Similarly for men, *local* variability is present in infectious and endocrine diseases, linked to time periods 1959-1985 for the former and 1959-1990 for the latter, but for the women sample we observe *local* variability also for external causes, in the interval 1980-2005. Lastly, neoplasms exhibit a variability pattern related to the velocity of the evolution of this

cause. Overall, a positive score on this component would lead to a high level of lung cancer and

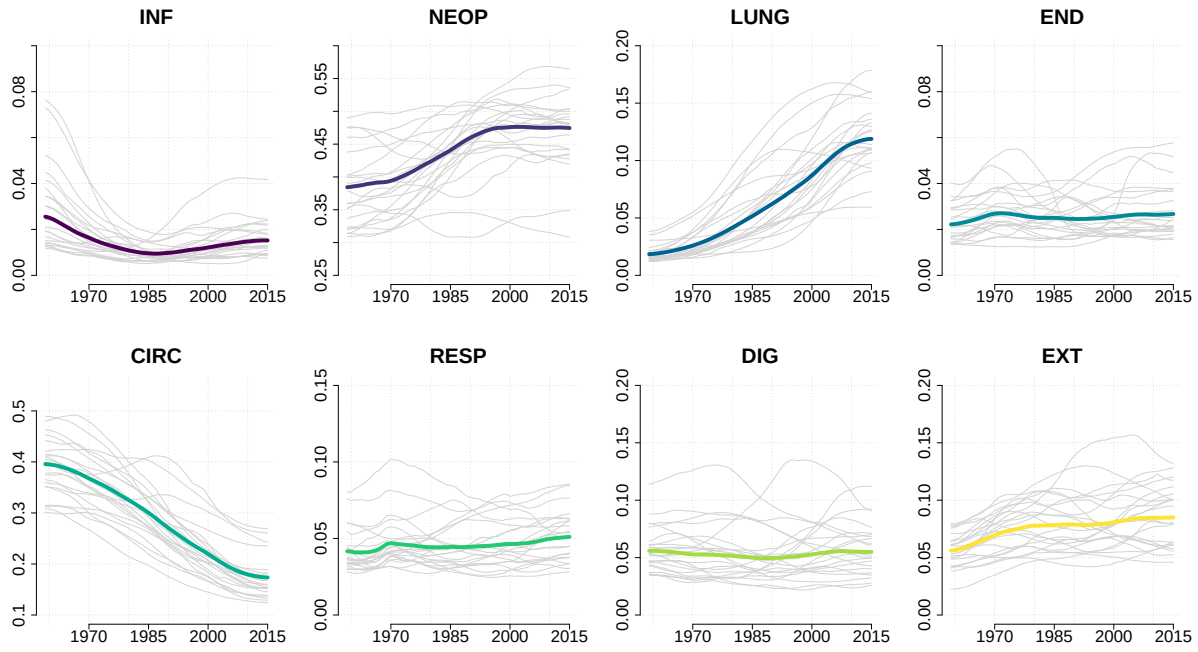

Fig. 8. Composition of mortality rates over years for women population. Curves for 22 countries in each panel are colored in grey. Colored curves represent the compositional functional mean.

respiratory diseases, a low level of digestive diseases, infectious diseases (especially before 1985), endocrine diseases (only before 1990) and external causes (between 1980 and 2005) and slow increase of neoplasms. Negative scores characterize an opposite behaviour. The scores obtained by each country on each component for women sample are reported in Table 3. The second compositional functional principal component for women population is depicted in Figure 10. The eigenvalue associated with this eigenfunction is $\lambda_2 = 7.38$ and the FEV is 0.162. This second component is related to all causes of death. Circulatory and endocrine diseases and external causes show a *level* variability. A *local* variability can be observed for neoplasms (1959-2000), lung cancer (1980-2015) and respiratory diseases (1959-1995). Lastly, a crossing-mean pattern can be found in infectious and digestive diseases. Overall, a positive score on this component would lead to a low level of circulatory diseases and respiratory diseases (before 1995), a high level of endocrine diseases, external causes, lung cancer (since 1980) and neoplasms (until 2000), and a increase of infectious diseases after 1985.

Figure 11 shows the third compositional functional principal component for women population. The eigenvalue associated with this eigenfunction is $\lambda_3 = 5.94$ and the FEV is 0.13. This component represents a *level* variability for endocrine, circulatory and respiratory diseases and for external causes and a *local* variability for neoplasms (before 1990) and lung cancer (after 1990). Overall, a positive score on this component would lead to a high level of endocrine, circulatory and respiratory diseases, a low level of external causes, a low level of neoplasms before 1990 and a low level of lung cancer after 1990.

The fourth compositional functional principal component for women population is illustrated in

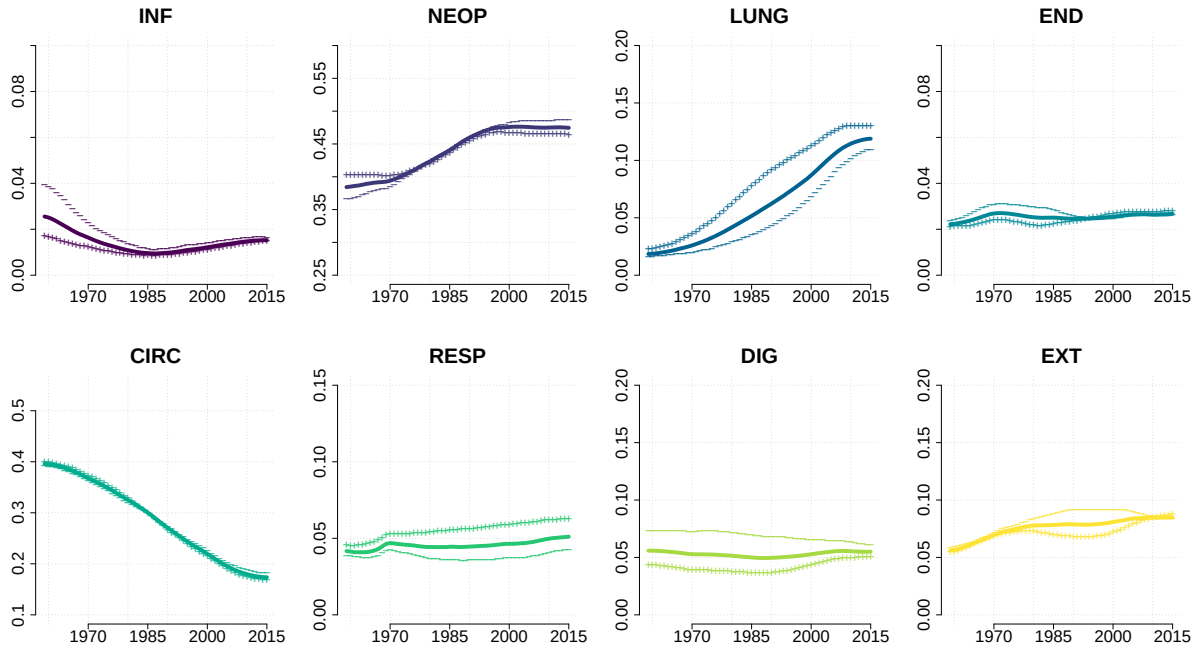

Fig. 9. First compositional functional principal component for women population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

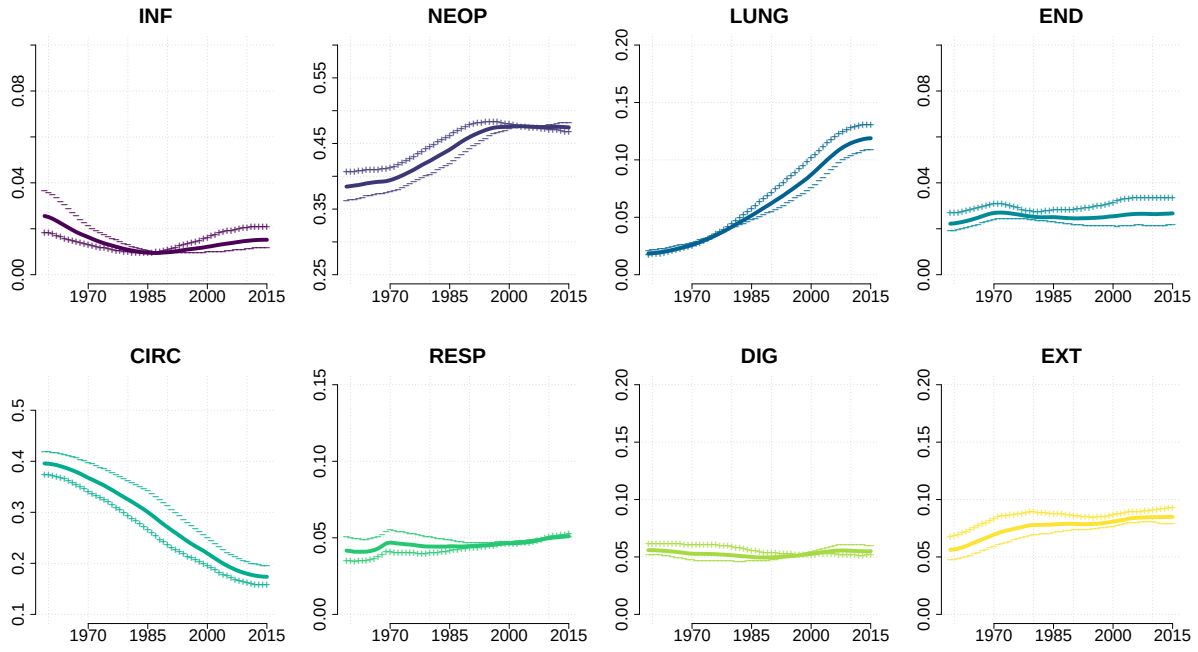

Fig. 10. Second compositional functional principal component for women population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

Figure 12. The eigenvalue associated with this eigenfunction is $\lambda_4 = 5.66$ and the FEV is 0.124.

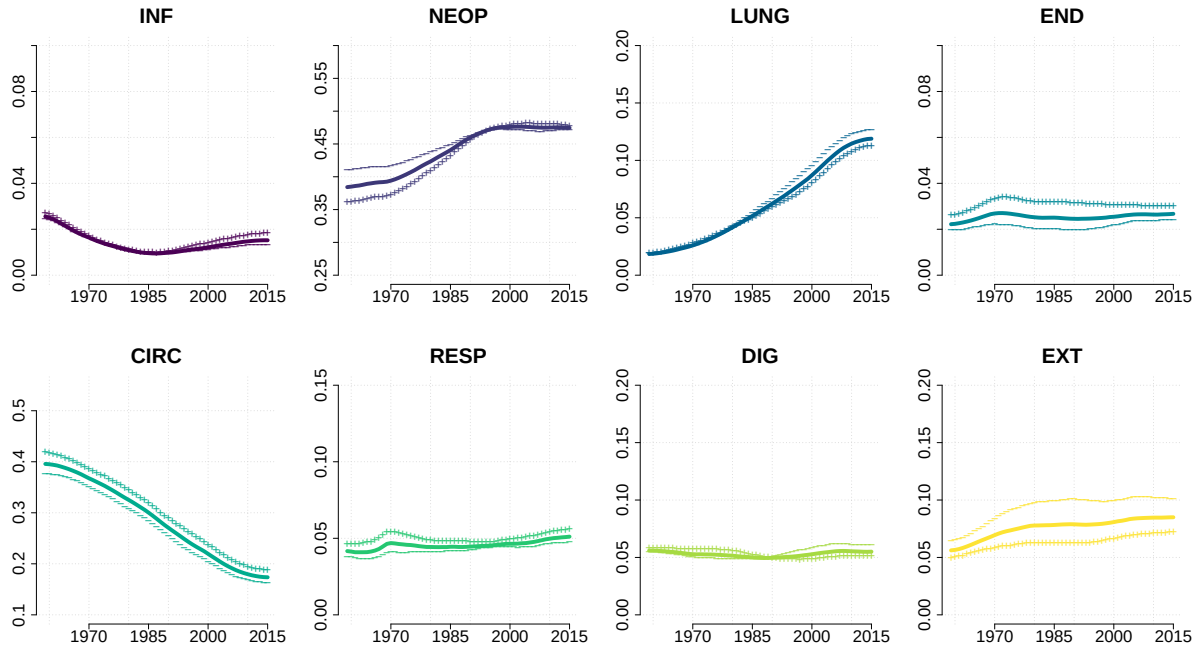

Fig. 11. Third compositional functional principal component for women population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

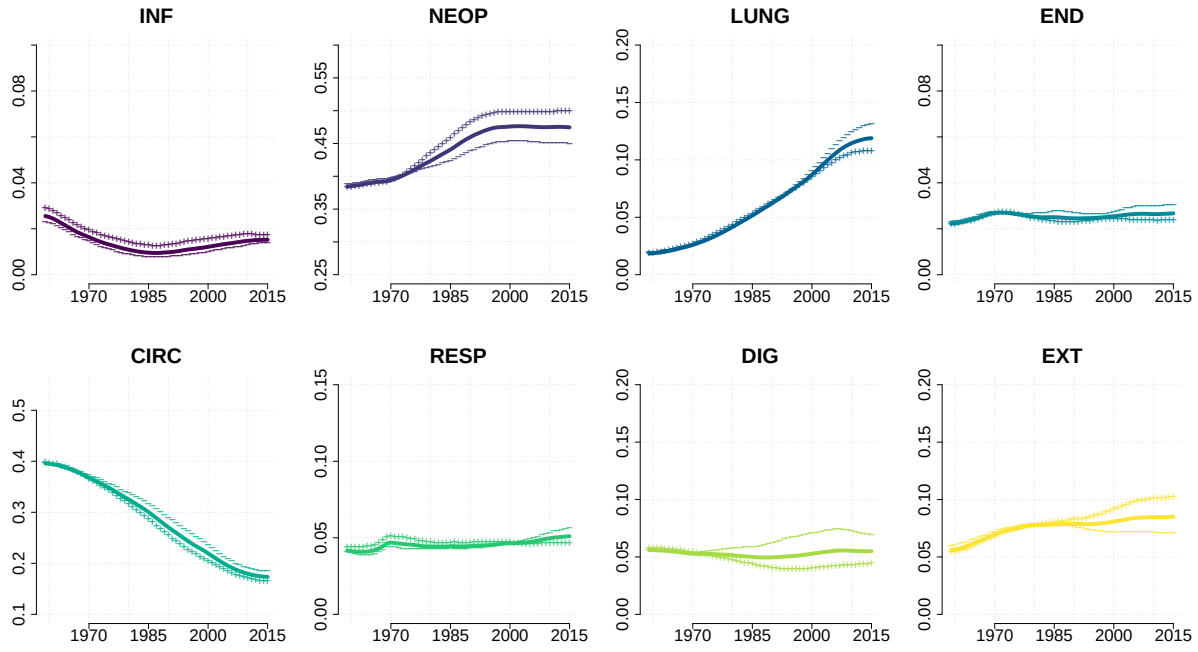

Fig. 12. Fourth compositional functional principal component for women population. The continuous line represents the mean function while outer lines represent the mean function +/- the component.

This component represents mainly *local* and most recent variability. In particular it is related

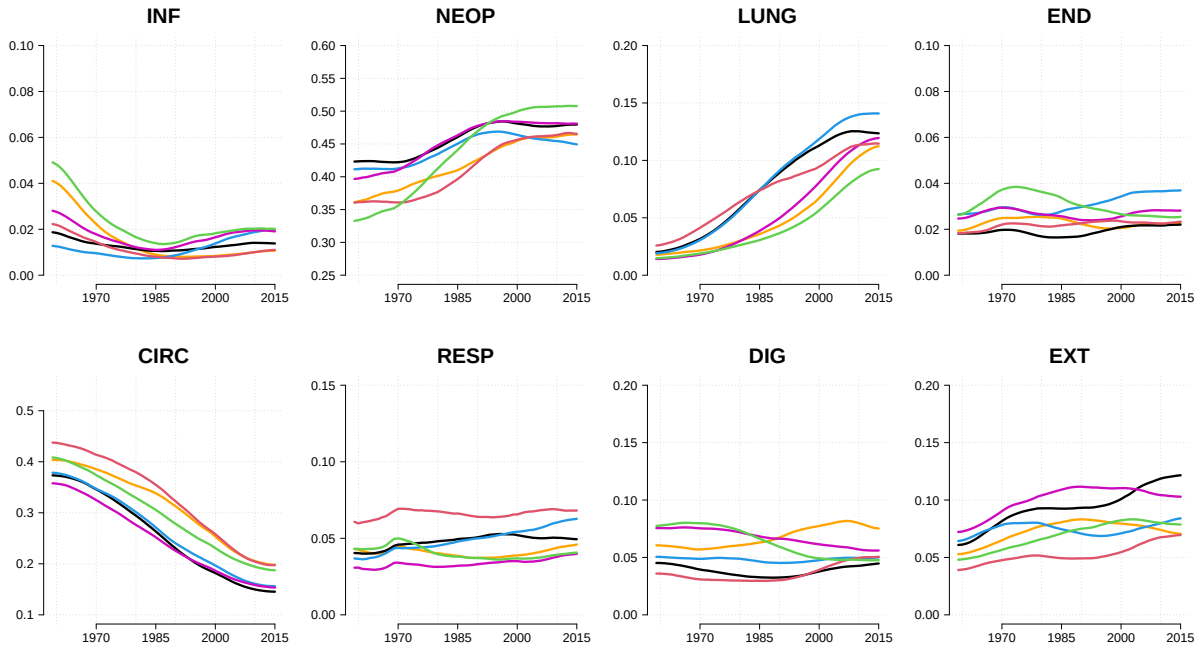

Fig. 13. Functional centroids of the spectral clustering for women sample. Legend: Cluster 1, Cluster 2, Cluster 3 Cluster 4, Cluster 5, Cluster 6.

to neoplasms after 1980, lung cancer and endocrine diseases after 2000, digestive diseases after 1970 and external causes after 1990. Overall, a positive score on this component would lead to a high level of neoplasms and external causes and a low level of lung cancer and digestive diseases, all in the aforementioned periods. As for men, we applied spectral clustering to the women curves projected on the first 4 PCs. The percentage of total explained variability is 74.2%. The algorithm ran $B = 1000$ times and we considered different values for G. Again, the spectrum of the Laplacian matrix does not support any specific number of clusters, while the silhouette index suggests a value between 6 and 8. For parsimony and interpretability reasons, we choose to use $G = 6$ clusters that we present in Table 5 (see also Figure 6). Using (19) with $K = 4$ and $\gamma_{1g}, \gamma_{2g}, \gamma_{3g}, \gamma_{4g}$ as the centroids of the spectral clustering output, we reconstructed the compositional functional centroids $\boldsymbol{f}_g, g = 1, \ldots, G$, depicted in Figure 13. The first group of countries is similar to that of men: it is mainly made up of extra-European countries together with Denmark and Netherlands, with a rising trend of lung cancer and respiratory diseases and endocrine and metabolic diseases, related to increasing prevalence of smoking and obesity among women. The fact that Denmark and Netherlands have been clustered in this group is not surprising, as it is well-known they have undergone a stagnation of life expectancy improvement, and smoking is one of the leading causes of this, as shown also by Lindahl-Jacobsen et al. (2016). The second group is made up by Japan, Italy and Spain, with the lowest level of lung cancer, in contrast to what has been seen for men in the same countries. Indeed, smoking prevalence of women in these countries is much lower than that of men, see Brennan and Bray (2002). The third group includes France, Belgium and Switzerland; the main feature of this group is the high level of external causes deaths, only recently getting lower than that of Scandinavian countries.

The fourth group (which for men includes Hungary, Poland, Greece, and Finland) now includes also Austria and, as for men, it is characterised by high level of digestive system diseases related deaths. It should be noted that also for women this is the least homogeneous group. Austria, in particular, has a peculiar pattern with a recently increasing trend of endocrine and metabolic diseases and a steady level of external causes related deaths. The group with UK and Ireland now includes also New Zealand, instead of Belgium, but the characteristics are similar to the analogous group for men: high respiratory and circulatory diseases and increasing digestive system diseases related deaths. Finally, the Scandinavian group: as for men, the main feature is the high mortality for external causes but, in contrast with men, this group is also characterized by a high and increasing relevance of lung cancer related mortality.

5. Conclusions

We have proposed to combine compositional data analysis (CDA) and functional data analysis (FDA) to describe the global trend of cause–specific mortality of several countries. CDA allows us to analyze cause–specific mortality rates (CSMRs) taking properly into account their competing—risk nature and FDA is then applied to compositional data so that their trends can be decomposed by means of Functional principal components analysis and countries can be clustered by means of PCs we have retrieved. In this way we are able to make a descriptive but comprehensive analysis of trends of CSMRs. Results give us many insights of the ongoing trend that the considered populations are undergoing in terms of composition of causes of deaths and, while many of them come as no surprise, in some cases we find some evidence that has not been highlighted by past literature. We should bear in mind that our analysis is focused on ages 40–64, so all considerations should be applied to mid-life mortality only.

The first finding is that clusters for men and women are very similar, albeit some differences can be found: the main one is that for women we have one cluster more, which is basically the result of splitting the cluster 2 into two. However, even though the clustering results are quite similar for men and women, we can easily notice that evolution of mortality composition is quite different: in all countries, prevalence of lung cancer among men has stopped increasing between 1990 and 2000, while for women it kept on rising until the very recent years, when first signs of plateauing can be seen. Moreover for some countries we can see a high disparity between men and women (e.g., Italy, Japan and Spain are in the cluster with lowest level of lung cancer for women and in that with the highest for men, while for cluster including Nordic countries is the other way round) while for others (e.g., East European countries) women catch up with men. Relevance of digestive system and metabolic, endocrine and nutritional diseases related deaths is increasing especially for men, suggesting that poor dieting and alcohol consumption are increasingly impacting on men's health. For women a rising concern is lung cancer and respiratory diseases, especially in Nordic countries. Interestingly some countries have a peculiar pattern that is difficult to group with others. Finland, in particular, with extremely high external causes and digestive system diseases related deaths, shows a composition of causes of death that is difficult to classify. Such peculiarity is confirmed by the relatively low life expectancy with respect to other Scandinavian countries.

Although FDA and CDA approaches are helpful in explaining cause–specific mortality trends

in a comprehensive way, both of them come with a specific limitation. The limitation of FDA is that it allows a descriptive analysis through functional principal components and cluster analysis, but it can not be used in a straightforward way to forecast the future trends. A forecast application that takes into account the compositional aspect but not the functional one has been implemented by Kjærgaard et al. (2019). The limitation of CDA is that it focus on cause–specific rates, but the overall trend is not considered. However, the clusters we identified largely capture differences in overall mortality rate trends of countries.

On the other hand, such a combination of CDA and FDA can be helpful in other applications: the same analysis can be applied to older ages (65+), where, following Horiuchi et al. (2003), we can expect to find higher prevalence of infectious diseases, mental disorders and cerebrovascular diseases. Another possible application, remaining in the demographic field but turning to a different aspect, could be to describe trends of parity–specific fertility rates, which can be seen as a composition of the overall fertility, although it could be hard to find a sufficient number of countries with parity–specific fertility data for a long enough time window.

Acknowledgements

We acknowledge the support from MIUR–PRIN 2017 project number 20177BR-JXS.

References

Aitchinson, J. (1986) *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Brennan, P. and Bray, I. (2002) Recent trends and future directions for lung cancer mortality in europe. *British Journal of Cancer*, **87**, 43–48.

Canudas-Romo, V., Aldair, T. and Mazzuco, S. (2020) Cause of death decomposition of cohort survival comparisons. *International Journal of Epidemiology*. URL: `https://doi.org/10.1093/ije/dyz276`. Dyz276.

Chen, K., Delicado, P. and Müller, H.-G. (2017) Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 177–196.

Dai, X. and Müller, H.-G. (2018) Principal component analysis for functional data on Riemannian manifolds and spheres. *The Annals of Statistics*, **46**, 3334 – 3361.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2011) *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd.

Fryar, C. D., Carroll, M. D. and Ogden, C. L. (2014) Prevalence of overweight, obesity, and extreme obesity among adults: United states, 1960–1962 through 2011–2012. *Tech. rep.*, National Center for Health Statistics.

Horiuchi, S., Finch, C. E., Meslé, F. and Vallin, J. (2003) Differential Patterns of Age-Related Mortality Increase in Middle Age and Old Age. *The Journals of Gerontology: Series A*, **58**, B495–B507.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K. and Filzmoser, P. (2016) Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis*, **94**, 330 – 350.

Izenman, A. J. (2008) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

Kjærgaard, S., Ergemen, Y. E., Kallestrup-Lamb, M., Oeppen, J. and Lindahl-Jacobsen, R. (2019) Forecasting causes of death using compositional data analysis: the case of cancer deaths. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**, 1351–1370.

Kraus, D. (2015) Components and completion of partially observed functional data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **77**, 777–801.

Leng, X. and Müller, H.-G. (2006) Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**, 68–76.

Lindahl-Jacobsen, R., Rau, R., Jeune, B., Canudas-Romo, V., Lenart, A., Christensen, K. and Vaupel, J. W. (2016) Rise, stagnation, and rise of danish women's life expectancy. *Proceedings of the National Academy of Sciences*, **113**, 4015–4020. URL: `https://www.pnas.org/content/113/15/4015`.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics & Computing*, **17**, 395 – 416.

Oeppen, J. (2008) Coherent forecasting of multiple-decrement life tables: A test using japanese cause of death data. *Tech. rep.*, Catalonia, Spain: Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona.

Petersen, A. and Müller, H.-G. (2016) Functional data analysis for density functions by transformation to a hilbert space. *Ann. Statist.*, **44**, 183–218.

Preston, S. H., Heuveline, P. and Guillot, M. (2001) *Demography. Measuring and Modeling Population Processes*. Blackwell Publishing.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*, vol. 2nd Ed. New York: Springer.

Saarela, J. and Finnas, F. (2008) Cause-specific mortality at young ages: Lessons from finland. *Health & Place*, **14**, 265–274.

Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009) A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, **104**, 37–48.

Scealy, J. L., de Caritat, P., Grunsky, E. C., Tsagris, M. T. and Welsh, A. H. (2015) Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association*, **110**, 136–148.

Woolf, S. H. and Schoomaker, H. (2019) Life Expectancy and Mortality Rates in the United States, 1959-2017. *Journal of the American Medical Association*, **322**, 1996–2016.

World Health Organization (2019) Mortality database. available at `https://apps.who.int/healthinfo/statistics/mortality/causeofdeath_query/start.php`.