Causes of death patterns and life expectancy: explaining USA gap.

Stefano Campostrini[1], Emanuele Aliverti[2], Daniele Durante[3], and Stefano Mazzuco[4]

[1]Università Ca' Foscari di Venezia, email: stefano.campostrini@unive.it
[2]Università Ca' Foscari di Venezia, email: emanuele.aliverti@unive.it
[3]Università Bocconi di Milano, email: daniele.durante@unibocconi.it
[4]Università degli studi di Padova, email: stefano.mazzuco@unipd.it

May 14, 2021

Abstract

The evolution of longevity across countries is quite diverse and it still remains unclear what determined such different patterns throughout the last decades. In this paper we consider a functional regression for life expectancy gap of United States with other countries with causes of deaths compositions. A preliminary comparison among Italy and United States indicates interesting trends in the evolution of life expectancy, with trajectories that overlap until the early 80s and then diverge substantially. We attempt to justify such differences by studying variations in the causes of mortality across these periods of interest, trying to justify potential driving factors for such divergences.

1 Introduction

In the economically more advanced countries, longevity has steadily increased throughout the last decades. Comparing, for example, life expectancy at birth (e_0) in Italy and USA, we can observe in Figure 1 how these countries show comparable trends until the early 80s, resulting afterwards in substantially diverging trajectories. This result is even more striking if we consider that USA invests relatively almost twice in health of its resources (17% of GDP vs 8.7% of Italy, see OECD 2020) and suggests the need of an improved understanding of the mechanism behind different patterns of longevity evolution, especially in relation to the mortality structure. For example, Bergeron-Boucher

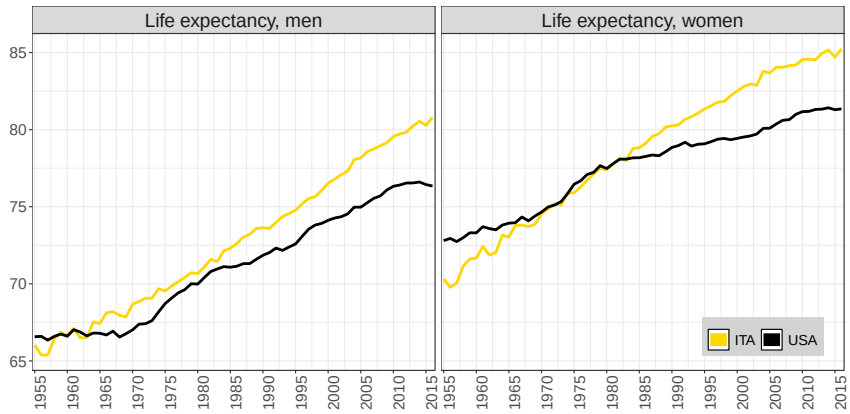

Figure 1: Life expectancy in Italy and USA. Source: Human Mortality Database

et al. (2020) have recently shown that the extension of longevity is usually accompanied by a diversification of the causes of death. More specifically, Woolf and Schoomaker (2019) analyze the trend of causes of death in USA, finding that midlife mortality caused by drug overdoses, alcohol abuses, suicides, and a list of organ system diseases have particularly increased in the last years. However, this finding has been contested. For instance, Mehta et al. (2020) argue that cardiovascular diseases are the main responsible of USA life expectancy stagnation. Such a controversy reflects the issue when dealing with cause-specific mortality, related with a 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 grown meanwhile. Therefore, if we want to analyze the time trend of causes of death we need to take into account this feature. Moreover, in public health perspective a further link need to be created between causes of deaths and risk factors, particularly relevant for non-communicable diseases).

Stefanucci and Mazzuco (2020) propose to combine Functional Data Analysis (FDA — see Ramsay and Silverman (2005)) with Compositional Data Analysis (CDA — see Aitchinson (1986); Egozcue and Pawlowsky-Glahn (2011)). While their contribution provides key initial findings, the focus is mainly on descriptive analyses of causes of death patterns. Here, we consider a model-based approach which allows principled statistical inference.

Countries	AUS	AUT	BEL	CAN	CHE	DNK	ESP	FIN	FRA	GBR
	HUN	IRL	ISL	ITA	JPN	NLD	NOR	PRT	SWE	
Causes	INF	NEO	LUNG	END	CIRC	RESP	DIG	EXT	OTH	RES

Table 1: Countries and causes of death

2 Data and methods

Data are collected from the Human Mortality Database (2020), that ensures high quality data on mortality profiles of different European and non-European countries. Specifically, we focus here on Australia, Austria, Belgium, Canada, Switzerland, Denmark, Spain, Finland, France, Great-Britain, Hungary, Ireland, Island, Italy, Japan, the Netherlands, Norway, Portugal, Sweden and United States of America. We conduct analysis on these $n = 20$ countries, considering sex-specific and age-adjusted rates over a time period of $T = 62$ years ranging from 1955 to 2016. We also consider 8 classes of causes of mortality, namely infections, neoplasms (all cancers with the exception of lung cancer), lung cancer, endocrines diseases, circulatory diseases, respiratory diseases, digestive diseases and external causes, see Table 1.

We are interested in evaluating the effect of the causes of death on life-expectancy. In particular, we focus on the differences among the selected countries with respect to United States of America (USA); see the first part of Table 1 for a list of the countries under investigation. For a fixed year t, we denote as $y_i(t)$ the difference among the life expectancy of country i and USA, with $i = 1, \dots, 19$ countries and focusing on the years $t = 1955, \dots, 2016$. Similarly, we focus on compositions of causes-specific mortality rates for ages below 70 (in order to remove the effect of varying age-structure across countries), and denote as $x_{ij}(t)$ the difference in the j-th cause of death between country i and USA. In this work we consider each $y_i(t)$ and $x_{ij}(t)$ as functional data (with respect to t), and develop functional linear models to characterize the effect of the causes of deaths on the evolution of life expectancy.

2.1 Concurrent model

A concurrent functional regression model (e.g. Ramsay, 2004, Chapter 14) can be interpreted as a varying-coefficient model (Hastie and Tibshirani, 1993) where both the response and the covariates are evaluated with respect to the same argument t; therefore, the effect of the covariates on the response is simultaneous, and x_{ij} influences y_i through

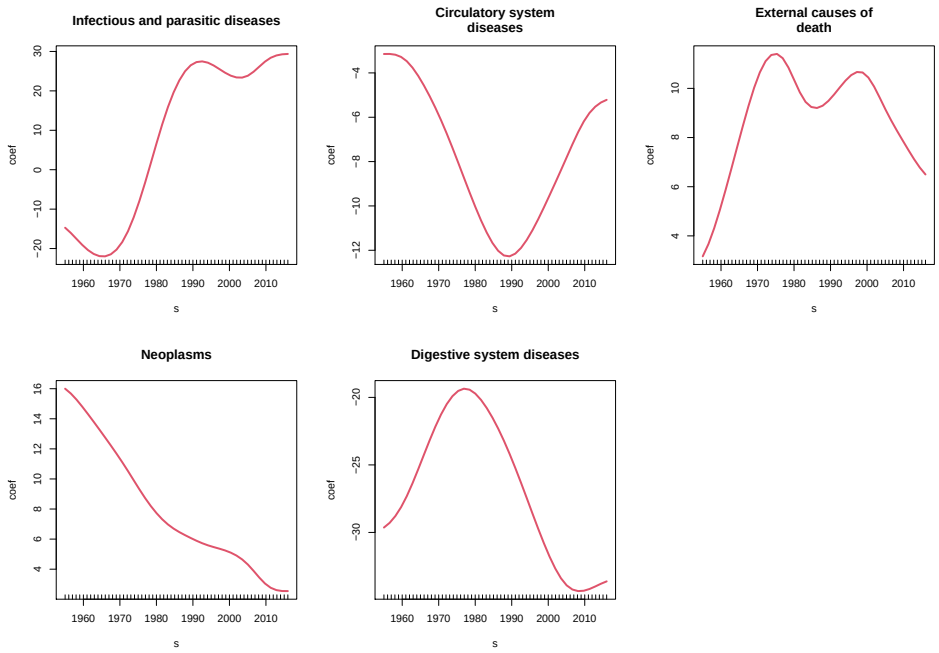

Figure 2: Female populations, results from a concurrent functional regression model. Graph reports the estimates for the coefficients $\beta_j(t)$ for the selected causes.

its value at time t. More formally, we let

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{p} x_{ij}(t)\beta_j(t) + \varepsilon_i(t) \quad i = 1, \ldots, n, \tag{1}$$

where $\beta_0(t)$ denotes a time-varying intercept, $\varepsilon_i(t)$ an error term and $\beta_j(t)$ the effect of the covariate j on the response. We model each $\beta_j(t)$ smoothly via P-splines, leveraging the package `FSboost` (Brockhaus et al., 2020) to perform estimation via component-wise gradient boosting with early stopping to induce variable selection of the covariates; see Brockhaus et al. (2017) for details. Figures 2 and 3 illustrate estimates for the smooth coefficients $\hat{\beta}_j(t)$, focusing on the variables selected by the gradient boosting algorithm. (note: functional R^2 approximately 75% and 70% for the female and male populations, respectively).

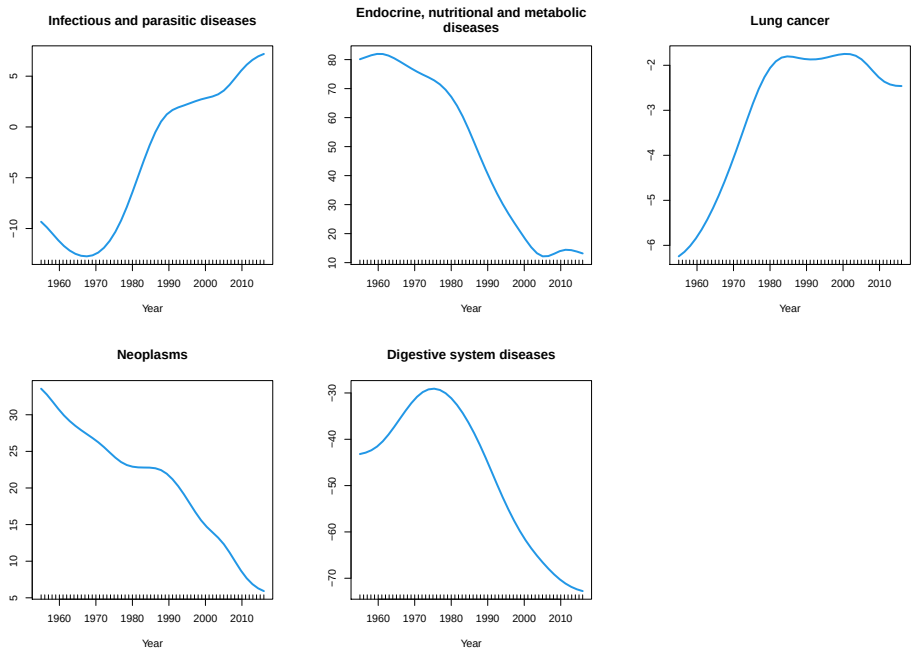

Figure 3: Male populations, results from a concurrent functional regression model. Graph reports the estimates for the coefficients $\beta_j(t)$ for the selected causes.

2.2 Functional linear model

We also consider a functional linear model, where we fully characterize the functional nature of the response and the covariates, allowing their effect to influence the response at different time points. In particular, we extend Equation (1) letting

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{p} \int_{1955}^{t} x_{ij}(s)\beta_j(s,t)\mathrm{d}s + \varepsilon_i(t), \quad i = 1, \ldots, n, \quad (2)$$

where $\beta_j(s,t)$ denotes the bivariate surface effect charactering the effect placed by the covariate j at year s on the life-expectancy in year t; we also force $s \leq t$ to induce a model with "historical" effects, where the covariates cannot influence the response retroactively; see also Ramsay (2004, Chapter 16) for further arguments. We model $\beta_j(s,t)$ smoothly via tensor-product P-splines relying again on the package `FSboost`. Results are reported in Figures 4 and 5. (note: functional R^2 approximately 63% and 57% for the female and male populations, respectively).

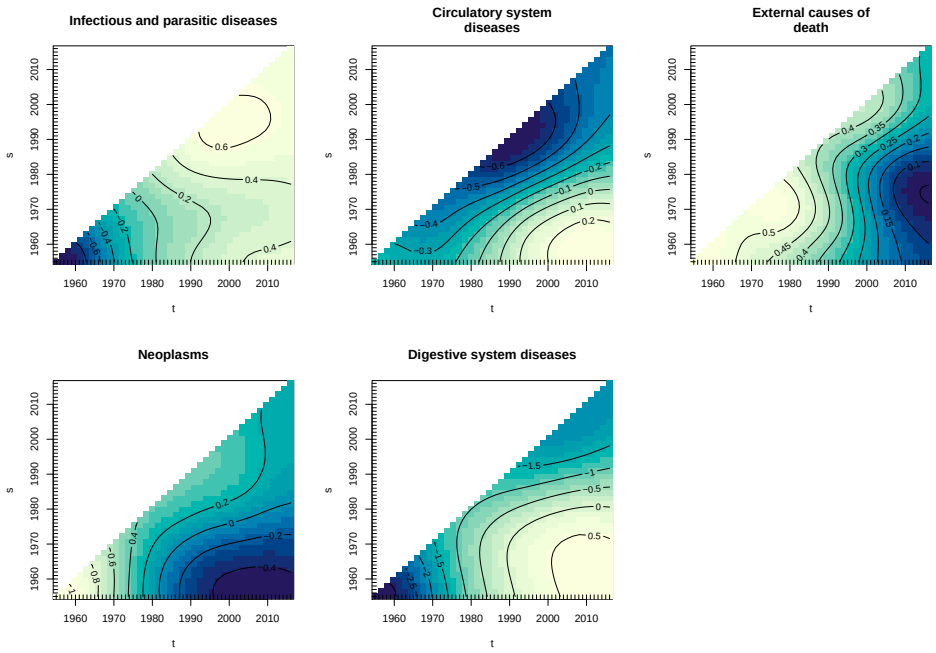

Figure 4: Female populations, results from a functional regression model. Graph reports the estimates for the coefficients $\beta_j(s,t)$ for the selected causes.

3 Preliminary results

Results from the concurrent model and the functional linear one show how the causes of death impacted on difference of USA life expectancy and those of other countries. We should keep in mind that the response variable considered is the difference between countries' life expectancy with USA one, while the independent variables are the differences in the percentage of death due to cause i between countries and USA.
Figures 2 and 3 are easier to interpret: infective diseases effect for both females and males has turned from negative to positive, meaning that while in the 1960 the difference in prevalence of infectious diseases was in favour of US life expectancy, after 1980s it has become a factor reducing the USA life expectancy, in comparison with the other countries, probably mainly due to HIV epidemic. Neoplasm relevance has steadily decreased in the time window considered. Interestingly, for female population we see a decreasing effect of external causes in the latest years and an increasing one for diseases of circulatory system. This could confirm Mehta et al. (2020): latest stagnation of US life ex-

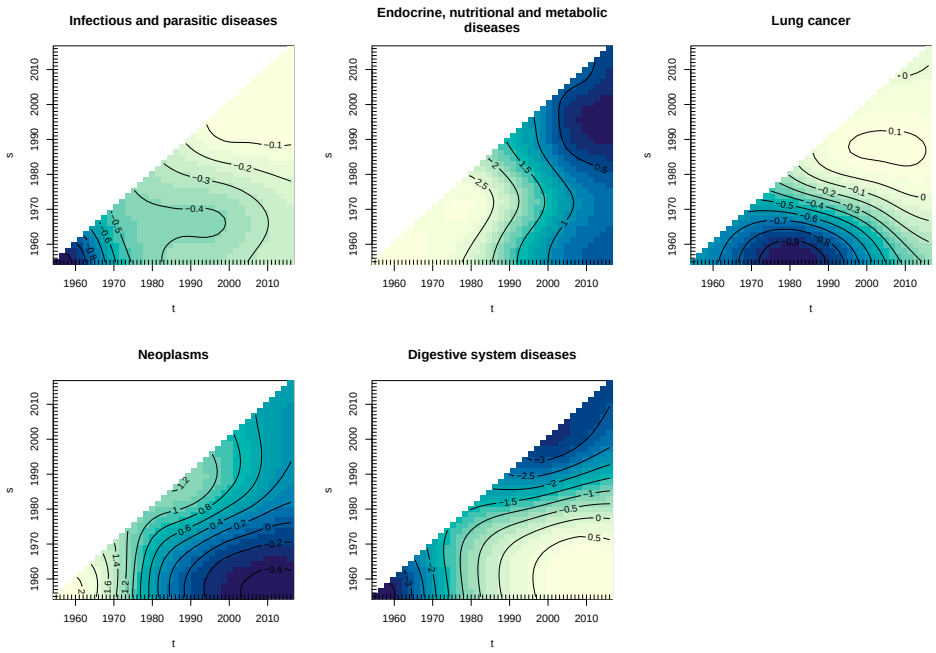

Figure 5: Male populations, results from a functional regression model. Graph reports the estimates for the coefficients $\beta_j(s,t)$ for the selected causes.

pectancy is more due to cardiovascular diseases than external causes, at least for women.
However these results should be considered still preliminary, models 1 and 2 will be further refined, including also other possibile covariates such as trends in risk factors, and some other patterns can emerge.

References

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

Bergeron-Boucher, M.-P., Aburto, J., and van Raalte, A. (2020). Diversification in causes of death in low-mortality countries:emerging patterns and implications. *BMJ Global Health2020*, e002414.

Brockhaus, S., Melcher, M., Leisch, F., and Greven, S. (2017). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing*, 27(4):913–926.

Brockhaus, S., Rügamer, D., and Greven, S. (2020). Boosting functional regression models with fdboost. *Journal of Statistical Software, Articles*, 94(10):1–50.

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

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Human Mortality Database (2020). University of california, berkeley (USA), and max planck institute for demographic research (germany).

Mehta, N. K., Abrams, L. R., and Myrskyla, M. (2020). Us life expectancy stalls due to cardiovascular disease, not drug deaths. *Proceedings National Academy of Sciences* USA, 117(13):6998–7000.

OECD (2020). Health statistics. available at `http://www.oecd.org/els/health-systems/health-data.htm`.

Ramsay, J. O. (2004). Functional data analysis. *Encyclopedia of Statistical Sciences*, 4.

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

Stefanucci, M. and Mazzuco, S. (2020). Analyzing cause-specific mortality trends using compositional functional data analysis. *Arxiv*:2007.15896.

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.