FUNCTIONAL REGRESSION OF MORTALITY COMPOSITIONS: AN APPLICATION TO ITALIAN PROVINCES

A PREPRINT

Stefano Mazzuco
University of Padova

Marco Stefanucci
University of Rome "Tor Vergata"

Gaia Bertarelli
University of Venice "Ca' Foscari"

October 30, 2025

ABSTRACT

There is growing literature considering causes death as compositional data for forecasting purposes or in regression models. As for the latter, compositions of mortality by cause has been used as covariates of overall mortality. In this work, we propose to consider the compositions of death by cause as response variable of functional regression model, in order to identify what are the determinants of specific compositional patterns of mortality, particularly focusing on mortality at age 40–64 in Italian provinces between 2003 and 2021. The analysis first involves a functional PCA of compositions, which are eventually regressed with a set of province characteristics. Preliminary reuslts show that causes of death structure of compositions is mostly associated with economic and working conditions of the population.

***K*eywords** functional data analysis · compositional data analysis · causes of death · Italy

1 Introduction

A growing literature [Kjærgaard et al., 2019, Stefanucci and Mazzuco, 2022, Feraldi and Zarrulli, 2022, Paoli et al., 2024] is considering to analyse causes of death as compositions. The main advantage of such approach is that it moves beyond analyzing isolated causes and instead provides a mathematically sound and interpretable framework for understanding how the entire system of causes evolves and changes in relation to itself. This can lead to more robust, reliable, and insightful conclusions about population health trends. In a recent work, [Paoli et al., 2024] implement a functional regression model to inspect how compositions of causes of death at specific age groups affect the overall life expectancy. In this paper, we consider compositions of causes of death as a response variable, which is regressed with several covariates related to economic condition, health services supply and other characteristics of the region. More specifically we are considering data on number of deaths by age, sex, and cause of Italian provinces from 2003 up to 2021, focusing on age group 40–64. The evolution of compositions of deaths of every province is cosnidered as a functional compositional datum that is modelled in function of several covariates

2 Modelling strategy

We consider a functional compositional datum as a multivariate function $f : \mathcal{T} \subset \mathbb{R} \to \Delta^D$ mapping from a subset of the real line to the D-dimensional simplex Δ^D. Formally, the space of such functions is defined as:

$$\Delta_f^D = \left\{ f \in [L_2(\mathcal{T})]^D : f_1(t) + \ldots + f_D(t) = 1,\ f_i(t) \geq 0\ \forall i,\ \forall t \in \mathcal{T} \right\}.$$

To relate these compositions to external covariates, we propose a class of linear models. Let $\boldsymbol{z}_i(t) = \text{clr}\{\boldsymbol{y}_i(t)/\tilde{y}(t)\}$ be the centered log-ratio (CLR) transformation of the functional composition for the i-th subject. Our base model is specified as:

$$\boldsymbol{z}_i(t) = x_i(t)\boldsymbol{f}(t) + \boldsymbol{\varepsilon}_i(t), \tag{1}$$

where $x_i(t)$ is a functional covariate and $\boldsymbol{f}(t)$ is the unknown regression coefficient function to be estimated, which itself is constrained to be a CLR-transformed composition. A special case of this model, regression on a scalar variable x_i, is obtained when the covariate is constant over t.

Our modeling strategy employs a spline basis expansion for $\boldsymbol{f}(t)$. Let $\boldsymbol{\Phi}(t)$ be a B-spline basis; we then express the coefficient function as:

$$\boldsymbol{f}(t) = \boldsymbol{B}\boldsymbol{\Phi}(t), \tag{2}$$

where $\boldsymbol{B}$ is a matrix of coefficients. The zero-sum constraint of the CLR transform is enforced by requiring $\mathbf{1}_p^\top \boldsymbol{B} = \mathbf{0}$ for all t. This leads to the following constrained optimization problem for estimation:

$$\arg\min_{\boldsymbol{B}} \frac{1}{2} \sum_{i=1}^{n} \int \|\boldsymbol{z}_i(t) - x_i(t)\boldsymbol{B}\boldsymbol{\Phi}(t)\|_2^2 dt \quad \text{subject to} \quad \mathbf{1}_p^\top \boldsymbol{B} = \mathbf{0}. \tag{3}$$

We estimate the model using an **Augmented Lagrangian** approach, which iteratively updates paramet

ers to accommodate the linear constraint. The $(t+1)$-th iteration proceeds as follows:

$$\boldsymbol{b}^{(t+1)} = (\boldsymbol{K} + \rho \boldsymbol{M})^{-1} \left(\boldsymbol{J} - \tilde{\boldsymbol{L}}^\top \boldsymbol{u}^{(t)} \right), \tag{4}$$

$$\boldsymbol{u}^{(t+1)} = \boldsymbol{u}^{(t)} + \rho \tilde{\boldsymbol{L}} \boldsymbol{b}^{(t+1)}, \tag{5}$$

where $\boldsymbol{b} = \text{vec}(\boldsymbol{B})$, and the key components are:

$$\boldsymbol{K} = \sum_{i=1}^{n} \int \tilde{\boldsymbol{\Phi}}_i(t)^\top \tilde{\boldsymbol{\Phi}}_i(t) dt, \qquad \tilde{\boldsymbol{\Phi}}_i(t) = (\boldsymbol{I}_p \otimes x_i(t))\boldsymbol{\Phi}(t),$$

$$\boldsymbol{M} = \tilde{\boldsymbol{L}}^\top \tilde{\boldsymbol{L}}, \qquad \tilde{\boldsymbol{L}} = \mathbf{1}_p \otimes \boldsymbol{I}_k,$$

$$\boldsymbol{J} = \sum_{i=1}^{n} \int \tilde{\boldsymbol{\Phi}}_i(t)^\top \tilde{\boldsymbol{z}}_i(t) dt, \qquad \tilde{\boldsymbol{z}}_i(t) = \text{vec}(\boldsymbol{z}_i(t)^\top).$$

This framework naturally extends to a multiple regression setting with q functional covariates:

$$\boldsymbol{z}_i(t) = \sum_{j=1}^{q} x_{ij}(t) \boldsymbol{f}_j(t) + \boldsymbol{\varepsilon}_i(t). \tag{6}$$

For this more general model, estimates are obtained via a **backfitting algorithm**. This procedure updates the estimate for each coefficient function $\boldsymbol{f}_j(t)$ in turn by fitting a model to the partial residuals with respect to all other predictors:

$$\boldsymbol{z}_i^j(t) = \boldsymbol{z}_i(t) - \sum_{r \neq j} x_{ir}(t) \hat{\boldsymbol{f}}_r(t). \tag{7}$$

Following [Paoli et al., 2024], the model can be enriched with sparsity-inducing penalties to perform variable selection, either on specific parts of the composition or on entire predictors. This is achieved by introducing a ℓ_1/ℓ_2 (Group Lasso) penalty. The resulting optimization problem becomes:

$$\arg\min_{\boldsymbol{B}} \frac{1}{2} \sum_{i=1}^{n} \int \|\boldsymbol{z}_i(t) - x_i(t)\boldsymbol{B}\boldsymbol{\Phi}(t)\|_2^2 dt + \lambda \sum_{d=1}^{D} \|\boldsymbol{B}_d\|_2$$
$$\text{subject to} \quad \mathbf{1}_p^\top \boldsymbol{B} = \mathbf{0},$$

where $\boldsymbol{B}_d$ denotes the d-th row of the coefficient matrix. While this penalty enhances interpretability, the trade-off is a more complex estimation procedure.

3 Data

The analysis is based on mortality data for Italy, obtained from official statistics released by the Italian National Institute of Statistics (Istat). The dataset spans the years 2003 to 2021, providing a nearly two-decade-long panel for investigation.

To focus on premature mortality, which is a key indicator of population health and socioeconomic conditions [Stefanucci and Mazzuco, 2022], we restrict our analysis to the age group of *40 to 64 years*. The geographical detail of the data is at the *NUTS-3 level*, corresponding to Italian provinces, allowing for a fine-grained spatial analysis.

The causes of death have been aggregated following the methodology of [Paoli et al., 2024] into *17 distinct categories*. This aggregation ensures a meaningful and manageable classification for compositional analysis. A notable feature of the dataset is the introduction of a specific "Covid-19" category for the years 2020 and 2021, which accurately captures the significant impact of the pandemic on mortality patterns during this period.

As for covariates, we include the following:

1. Health for All project by Istat.
    - Demographic (Resident foreigners, Fertility rate, Birth rate, ...),
    - Economic (Activity rate, Employment rate, Total employed in agriculture, manufacturing, or services, ...)
    - Healthcare (Ordinary hospital bed rate, Average length of stay, ...) variables.
2. Educational (Graduates by residence) variables (MUR)
3. Environmental (Air quality indicators) variables (ISPRA)

4 Preliminary Results

We begin by examining the dominant modes of variability in the cause-of-death compositions over time using Functional Principal Component Analysis (FPCA). The FPCA characterizes the primary patterns of relative increases and decreases in cause-specific mortality shares with respect to the mean composition on the simplex.

The estimated eigenvalue spectrum decays slowly, indicating a complex, high-dimensional structure in the data. The first four principal components collectively explain 37% of the total variability for males and 29% for females, suggesting that the remaining variability is distributed across many minor modes of variation.

The temporal patterns of the mean composition reveal several key trends. The share of deaths from neoplasms remains largely stable over the study period for both sexes. In contrast, circulatory diseases show a pronounced declining trend. Concurrently, we observe a slight but consistent increase in the shares attributable to respiratory and mental diseases. Notably, lung cancer exhibits divergent sex-specific patterns: a decreasing trend among men contrasted with an increasing trend among women.

The spatial distribution of the FPCA scores, illustrated in Figure 1 for the male population, reveals clear geographical patterns. The first four components all exhibit significant spatial structure, with the second component, for instance, showing a strong North-South gradient. The spatial patterns of the scores for females (not shown) were qualitatively similar to those observed for males, indicating shared underlying geographic determinants of mortality structure.

Initial regression models, which link the compositional mortality patterns to socio-economic covariates, suggest that the structure of causes of death is most strongly associated with the economic and working conditions of the population.

An illustrative example is provided by the relationship between labour market indicators and the share of lung cancer mortality. Figure 2 displays the estimated functional coefficients for activity rate and employment rate. For both men and women, higher rates are associated with a greater share of deaths from lung cancer. However, the strength of this association is dynamic, declining at a faster rate in recent years for men than for women.

These initial findings highlight the potential of the functional-compositional framework to uncover complex, time-varying relationships. Further refinement of the model specification, including covariate selection and cause-of-death aggregation, is underway to confirm and elaborate upon these results.

Acknowledgments

This research is supported by the MUR–PRIN 2022 project "CARONTE" (Prot. 2022KBTEBN).

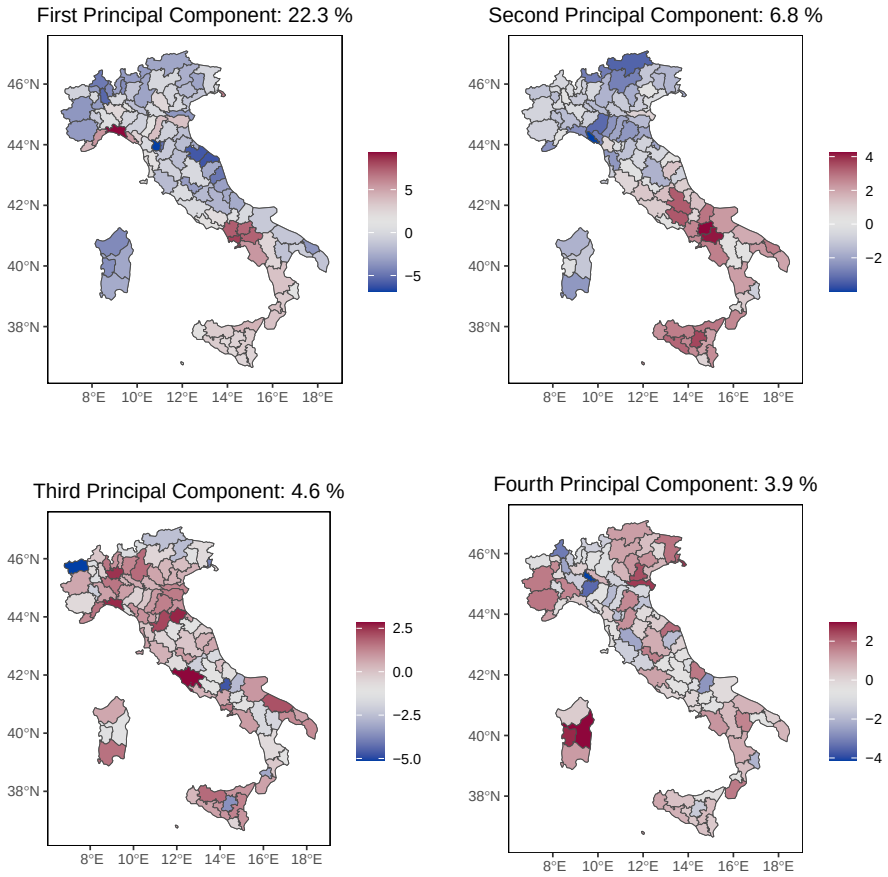

Figure 1: First four Principal components of compositions of causes of death in Italian provinces, 20023–2021, Men, age 40–64.

References

[Feraldi and Zarrulli, 2022] Feraldi, A. and Zarrulli, V. (2022). Patterns in age and cause of death contribution to the sex gap in life expectancy: a comparison among ten countries. *Genus*, 78(23).

[Kjærgaard et al., 2019] Kjærgaard, S., Ergemen, Y. E., Kallestrup-Lamb, M., Oeppen, J., and Lindahl-Jacobsen, R. (2019). Forecasting causes of death by using compositional data analysis: the case of cancer deaths. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5):1351–1370.

[Paoli et al., 2024] Paoli, E. D., Stefanucci, M., and Mazzuco, S. (2024). Functional concurrent regression with compositional covariates and its application to the time-varying effect of causes of death on human longevity. *Annals of Applied Statistics*, 2:1668–1685.

[Stefanucci and Mazzuco, 2022] Stefanucci, M. and Mazzuco, S. (2022). Analysing cause-specific mortality trends using compositional functional data analysis. *Journal of the Royal Statistical Society – Series A*, 185:61–83.

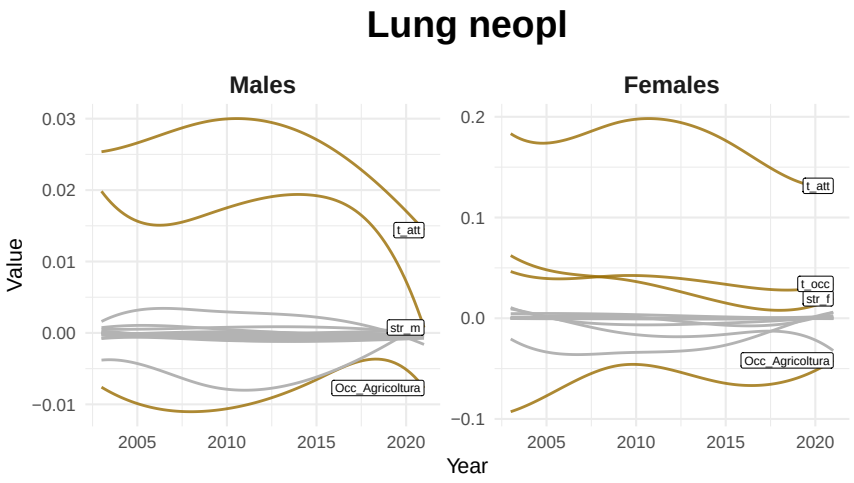

Figure 2: Estimated time-varying regression coefficients for the effect of activity and employment rates on the share of lung cancer mortality. Results are shown separately for males and females.