Evolution of Sub-National Longevity and Causes of Death Composition Using Data on Italian Provinces

Amerigo Novaro, Davide Benussi, Emanuele Aliverti and Stefano Mazzuco

1 Introduction and Theoretical Framework

Analysing health disparities and their evolution over time within a country is an increasingly key request to data analysts, stimulated by the increasing availability of data at the sub-national level; when sub-national mortality estimates are not reliable enough (e.g., in small populations), specific models are proposed to overcome such difficulties and infer mortality levels and trends (Alexander et al., 2017). A further step can be taken by considering data on causes of death to explain the mechanism driving this evolution. However, as in Alexander et al. (2017), this task requires specific modeling for two reasons: first, when we break down death counts at the province level by age and underlying cause of death we are inevitably faced with many small (or even zero) counts, with consequences on the reliability of estimates. Second, by considering the causes of death, we increase the dimensions of the time dynamics we should inspect for each province, cause, and age group. This makes it more challenging to infer indications from data that can be useful for policymakers. To overcome these challenges, we propose a state space factorial model in this article that allows us to understand the temporal evolution for each cause, the underlying factors that include most of the variability due to main causes of death, and the spatial correlation across provinces.

Our aim is to model death counts as a dynamic process that evolves with respect to the variables included in the analysis. Once the model demonstrates reliability, we can examine its components to understand how mortality patterns change over time and whether these dynamics reveal meaningful geographical differences. We denote as $D_{ijx}(t)$ the number of death in province i at age x for cause j, for $i = 1, \dots, N$, $x = 1, \dots, X$ and $j = 1, \dots, J$. Denoting as E_{ix} the average number of individuals at risk, similarly to Pavone et al. (2024), a state space Poisson model for the death rates can be specified as

$$D_{ijx}(t) \sim \text{Poisson}(\mathrm{E}_{ix}(t) m_{ijx}(t)). \tag{1}$$

When treating mortality count data, it is often observed that the Poisson model fails to capture the large variability exhibited in the data, leading to estimates that are too confident and not so reliable compared to the observed values. Consequently, we also consider a specification based on a Negative-Binomial model, which can be written as

$$D_{ijx}(t) \sim \text{Negative-Binomial}(\mathrm{E}_{ix}(t) m_{ijx}(t), \phi_{ijx}(t)). \tag{2}$$

with $\phi_{ijx}(t)$ being the dispersion parameter, which can vary over the cells indexed by (i, j, x). See, for instance, Fung et al. (2019) and Zhang et al. (2022). It is easy to see that there is an increase in dispersion compared to the Poisson case, where, by definition, the mean and the variance coincide. This can be useful in the case of mortality data.

After choosing the initial distributional assumption for the response variable, the next relevant step will focus on the characterization of the predictor $m_{ijx}(t)$, since the analysis of the coefficients of the included variables will allow us to determine whether mortality shows differences, from both a temporal and a spatial perspective.

2 Data and Methods

The National Institute of Statistics Italy (ISTAT) provides data on death counts by 95[1] provinces, age group (five-year age groups until age 100+), sex, and underlying cause of death from 2004 until 2019.

[1] We remove some provinces since the composition is varying over time with some changes over the last 20 years.

The causes of death are codified with ICD 10 protocol (3 digits). At this stage, we use a classification into 10 main categories: cardiovascular diseases, respiratory diseases, malignant neoplasms, mental diseases, external causes of death, diseases of the nervous system, ill–defined diseases, benign neoplasms or blood disorders, infectious or parasitic diseases and finally endocrine or metabolic diseases. Furthermore, the age component is summarized into four macro-groups, excluding the youngest age and obtaining the following classes: 10-39, 40-64, 65-79, 80+.

As mentioned in the first section, it becomes crucial to understand how to model the component $m_{ijx}(t)$ in order to obtain a model that is both reliable and interpretable, allowing us to answer the research questions. A possible specification assumes a dynamic trend for the annual effect, with its distribution defined as

$$\mu(t) \sim \mathcal{N}\big(\mu(t-1),\, \sigma_\mu^2\big),$$

leading to the general form of the predictor as

$$\log m_{ijx}(t) \;=\; \mu(t) + f_t(\alpha_i, \zeta_j, \delta_x).$$

We decide to consider the spatial effect separately from the effects related to cause of death and age group, obtaining the general specification

$$\log m_{ijx}(t) = \mu(t) + \alpha_i(t) + \zeta_j + \delta_x + B_{jx}(t),$$

with the spatial component modeled dynamically, and the cause and age effects summarized into a general dynamic term $B_{jx}(t)$, providing the opportunity to reduce dimensionality or to select specific combinations of these factors. We also add two baselines, one for the province, to separate the static effects from the dynamic ones, and one for the cause of death, to incorporate information about the different mean impacts of the different causes.

Regarding the spatial component, each dynamic trajectory is summarized using an orthogonal polynomial, so as to preserve centering and avoid competition with the baseline. We impose a multivariate normal prior on the polynomial coefficients, thereby inducing a geographically motivated correlation structure. In order to take into account the spatial dependence, we consider the following specification of the covariance matrix across provinces

$$[\Sigma_\alpha]_{i,j} \propto \exp\left(-\gamma \cdot d_{ij}\right)$$

which has an exponential decay with a scale parameter γ to be calibrated.

As anticipated, the component $B_{jx}(t)$ aims to capture the joint effect without adding excessive complexity to the model. For this reason, a simplification is introduced through a latent structure, which specifically corresponds to a bilinear factor model

$$B_{jx}(t) \;=\; u_j^\top w_x(t),$$

where, assuming K latent factors, $u_j \in \mathbb{R}^K$ is the loading vector for cause j, while $w_x(t) \in \mathbb{R}^K$ is the time–varying score vector for age group x. We can also rewrite this quantity with matrices, obtaining

$$\mathbf{B}(t) \;=\; \mathbf{U}\,\mathbf{W}(t),$$

where $\mathbf{U}$ is $J \times K$, $\mathbf{W}(t)$ is $K \times X$, so $\mathbf{B}(t)$ is $J \times X$. Clearly the number K directly implies the scale of the reduction of the dimensionality, capturing with a larger number coarser subpopulations in Italy. It should be noted that the factor $B(t)$ is dynamic, as the component $\mathbf{W}(t)$ evolves over time, with a first order Random Walk prior imposed

$$\mathrm{vec}\big(\mathbf{W}(t)\big) \sim \mathcal{N}\big(\,\mathrm{vec}\big(\mathbf{W}(t-1)\big),\; \Sigma_W\big),$$

which allows the age–group loadings to evolve smoothly over time and with an adequate structure of dependence Σ_W.

An extension can be considered by including a joint effect between provinces and causes. Accordingly, we propose the following alternative specification

$$\log m_{ijx}(t) = \mu(t) + \zeta_j + \delta_x + B_{jx}(t) + \nu_{ij}(t)$$

in which the previous spatial term $\alpha_i(t)$ is replaced by $\nu_{ij}(t)$ and so with the global spatial effect divided across the J causes. We again consider the dynamic structure based on orthogonal polynomials. This extended model introduces greater complexity; therefore, we incorporate regularization to prevent potential overfitting. In particular, the additional parameters are modeled using a global horseshoe prior. Across all model specifications, certain parameters may capture overlapping information; hence, constraints on the joint trajectories are necessary to ensure identifiability and preserve interpretability. Model estimation is conducted within a Bayesian framework using Markov Chain Monte Carlo (MCMC) methods, with prior specifications chosen to be consistent with the expected range of the observed data.

3 Expected Findings

Our analysis primarily focuses on the study of the dynamic components associated with the spatial trajectories included in the models. In the model without interactions, the main objective is to understand if, over the considered period, there are provinces that show a significant evolution, trying to visualize geographical meaning, which is enforced by the spatial correlation imposed on the coefficients of the orthogonal polynomials. In the first model, it is also possible to analyze how the dynamic evolution of the latent factors leads to certain causes showing changes over time in their composition with respect to age groups. We expect to identify marginal improvements particularly in the north-eastern provinces, with opposite patterns mainly in the South and in some western areas. For example, we can analyze the linear coefficients of the spatial polynomials to identify the provinces where mortality is linearly increasing or decreasing. Figure 1 shows these patterns on the Italian map with two specific examples of provinces showing opposite trends.

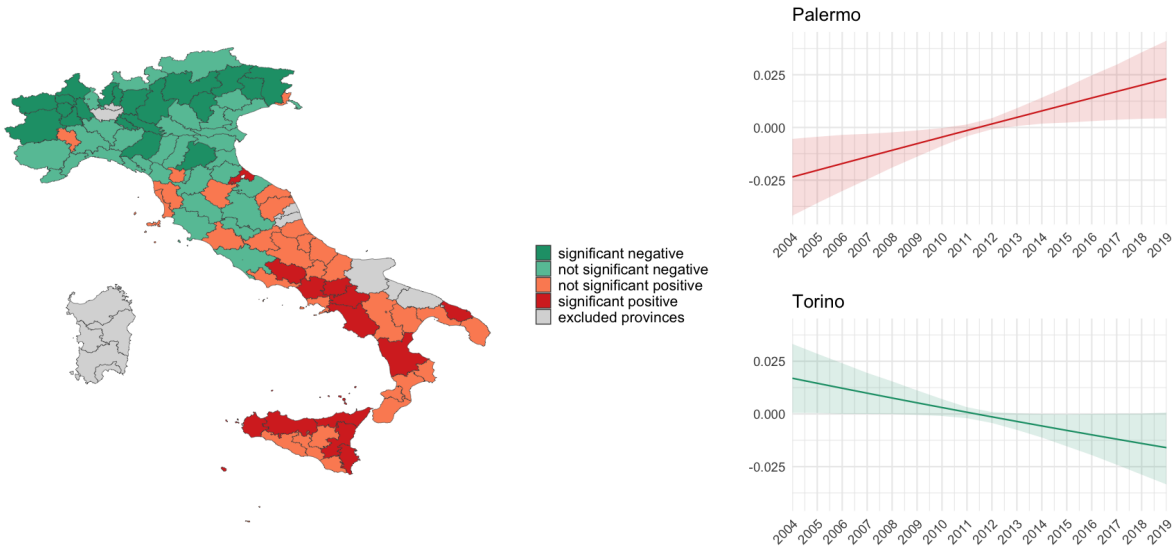

Figure 1: Provincial linear effects from province-specific polynomial models, with example trajectories for Palermo and Torino.

It is easy to see that mortality improves in northern Italy, with several provinces, such as Verona and Torino, showing a positive linear coefficient that is significantly greater than zero. In the south, such as Palermo and Cosenza, we observe the opposite pattern.

Considering the second model, the analysis shifts from an overall spatial context to a more precise focus on the causes. For instance, while in the first model one might note an evolution occurring in

a specific area of Italy, the second model allows for a better understanding of whether there is mainly a cause driving this improvement or worsening, and whether this pattern is shared among neighboring provinces with a similar overall effect. In Figure 2 we show, as an example, the classification of the linear-component patterns into the same four categories used previously, but focusing on mortality from nervous system disorders. Some provinces change color, for instance, Verona exhibits a significant worsening for this cause. For this reason, we also display in the same figure the opposite behavior of two causes for the province of Verona, the trajectories related to nervous system diseases and one related to neoplasms.

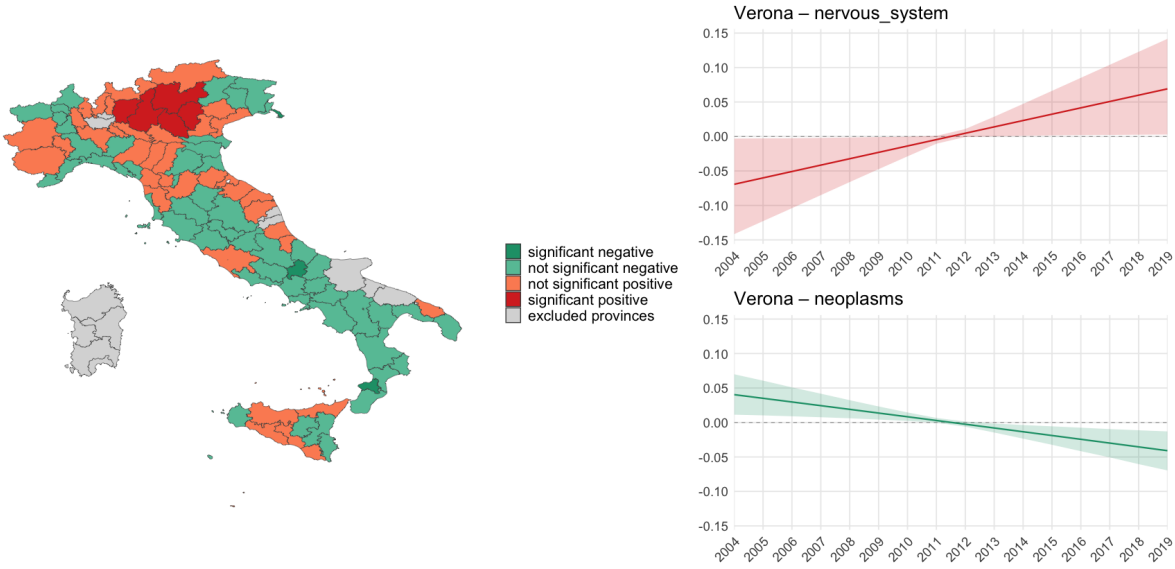

Figure 2: Behaviors of Italian provinces based on the linear component of province-specific polynomials for nervous system disorders (left), and example effects for Verona only considering two causes (right).

Moreover, it becomes possible to gain a deeper understanding of dynamics within provinces that, at a global level, appeared static. In fact, it may occur that a province does not show an overall change in its effect over the 16 years considered, even if two or more causes are in fact evolving, balancing their effects over time. With these analyses, we expect, for instance, to observe how certain causes, including cardiovascular diseases, benign tumors, and some ill-defined causes, play a significant role in influencing the global trend of some provinces.

References

Alexander, M., Zagheni E., and Barbieri M. (2017). A flexible bayesian model for estimating subnational mortality. *Demography 54*, 2025-2041.

Fung, M. C., Peters, G. W., and Shevchenko, P. V. (2019). Cohort effects in mortality modelling: a Bayesian state-space approach. *Annals of Actuarial Science*(13), 109–144.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*(73), 423–498.

Pavone, F., Legramanti S. and Durante D. (2024). Learning and forecasting of age-specific period mortality via B-spline processes with locally-adaptive dynamic coefficients. *The Annals of Applied Statistics 18*(3), 1965-1987.