

<https://archined.ined.fr>

# **Modelling COVID-19 mortality at the regional level in Italy**

**UgoFilippo Basellini et Carlo Giovanni Camarda**

**Version**

Libre accès

**Licence / License**

CC Attribution 4.0 International (CC BY)

## **POUR CITER CETTE VERSION / TO CITE THIS VERSION**

UgoFilippo Basellini et Carlo Giovanni Camarda, 2020, "Modelling COVID-19 mortality at the regional level in Italy". SocArXiv Papers, Center for Open Science. <https://doi.org/10.31235/osf.io/ykc6w>

**Disponible sur / Available at:**

<http://hdl.handle.net/20.500.12204/AXQ0SuDAkgKZhr-bleCV>

# Modelling COVID-19 mortality at the regional level in Italy

Ugo Filippo Basellini<sup>\*1,2</sup> and Carlo Giovanni Camarda<sup>\*†2</sup>

<sup>1</sup>*Max Planck Institute for Demographic Research (MPIDR), Rostock*

<sup>2</sup>*Institut national d'études démographiques (INED), Aubervilliers*

August 27, 2020

## Abstract

Italy was harshly hit by COVID-19, registering more than 35,000 deaths between February and July, 2020. The virus spread unequally across the country, with northern regions witnessing more cases and deaths than those in the centre and south. We investigate demographic and socio-economic factors that contributed to the diverse regional impact of the virus in Italy. Within a smoothing framework, we divide regions into three well-defined groups of High, Middle and Low mortality by cluster analysis. Extending the Poisson regression model to account for regional clusters, we find that COVID-mortality is positively associated with the share of ICU utilization, GDP per capita, proportion of the older population and the number of COVID-19 positive cases, while it is negatively associated with the delay of region-specific outbreaks and the number of tests performed. Our results have relevant policy implications for potential resurgence of COVID-19 infections in Italy and across the world.

**Keywords:** Mortality modelling · SARS-CoV-2 · Poisson regression · Cluster analysis · Smoothing · Socio-economic determinants · Demographic factors

## 1 Introduction

The novel coronavirus disease (COVID-19) is an infectious disease that has rapidly spread globally since the beginning of 2020. The disease is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and it was firstly identified in the city of Wuhan in December 2019 (Du et al., 2020). Within a matter of weeks, the World Health Organization declared the COVID-19 outbreak a public health emergency of international concern (January 30, 2020), and then a pandemic (March 11, World Health Organization, 2020b).

The global number of reported cases of COVID-19 has been rising at a very fast pace during 2020, increasing from about 10 thousands at the beginning of February to more

---

<sup>\*</sup>The authors contributed equally to this work.

<sup>†</sup>Corresponding author: [carlo-giovanni.camarda@ined.fr](mailto:carlo-giovanni.camarda@ined.fr)

Address: 9 cours des Humanités. 93322 Aubervilliers - France

Phone: +33-1-5606 2166

than 17 millions at the beginning of August. Similarly, the number of deaths attributed to the disease has increased from around 250 to more than 675 thousands during the same period ([Johns Hopkins University CSSE, 2020](#); [World Health Organization, 2020a](#)).

In Italy, the first case of COVID-19 was confirmed on February 20, although the virus was already present in the country since January ([Cereda et al., 2020](#)). Since the identification of “patient one”, the country has been harshly hit by the spread of the virus, with regard to both infections and deaths. The number of reported cases exceeded that of China on March 27, totalling almost 250 thousand cases by the end of July. In addition, Italy registered more COVID-19 deaths than any other country between March 19 and April 7, surpassing 35 thousand deaths at the beginning of August (data from [Johns Hopkins University CSSE, 2020](#)).

In response to the rapid spread of the virus, the country adopted a series of measures to slow down its transmission. On February 23, eleven municipalities in the north of Italy were identified as the main cluster of the epidemic and put under quarantine. Simultaneously, six northern regions implemented different restrictions, ranging from school closures to cancellations of public, religious or sport events. On March 1, the Council of Ministers divided the country in three areas of risk: a red zone, comprising the eleven municipalities in quarantine; a yellow zone, comprising the regions of Lombardia, Veneto and Emilia-Romagna, where public and sport events were suspended and school closed; and the rest of the country, with softer safety and preventive measures. Then, schools were closed nationwide on March 4, the entire country was put in lockdown on March 9 and non-essential activities were suspended on March 23 ([Ministero Della Giustizia, 2020](#)). Unfortunately, all these measures only partially mitigated the diffusion of the virus.

Since the early phases of the Italian outbreak, regional differences emerged in terms of timing and magnitude of the virus’ diffusion. At first, the virus spread in the north of Italy: the regions in the yellow zone (Lombardia, Veneto and Emilia-Romagna) were the first to confirm more than 200 cumulative infections at the start of March. Southern regions, which were initially rather unaffected, witnessed increasing rates of contagions throughout the months of March and April. In terms of COVID-related deaths, similar regional differences occurred, with northern regions generally more affected than southern ones (cf. Figure 4 in the [Results](#) section).

Economic and demographic differences across Italian regions have been the subject of a vast body of literature (see, e.g., [Helliwell and Putnam, 1995](#); [Billari and Ongaro, 1998](#); [Di Giulio and Rosina, 2007](#); [Kertzer et al., 2008](#)). However, little is known about the effects of such differences in shaping the COVID-19 epidemic in Italy. In this paper, we analyse the demographic and socio-economic factors that contributed to the diverse regional impact of the virus in Italy. We focus on COVID-related mortality during the first half of 2020, specifically from the end of February to mid-July, a period during which about 35 thousands COVID-19 deaths were registered. We study the association between the reported number of deaths by COVID-19 and a set of explanatory variables at the regional level. We investigate the main explanatory variables that have been linked to the outbreak of COVID-19 across the world. By employing a Poisson regression model, we determine the demographic and economic variables that had the strongest impact on the number of deceased individuals. Finally, to take into account the different territorial experiences, we stratify our analysis by clustering regions that shared similar trajectories of the epidemic.

This article is organized as follows. Section 2 describes the methodology that we

use for our analysis, the mortality data that we employ and the relevant covariates that we consider in our study. In Section 3, we illustrate the results of our analysis, and we conclude with a discussion of our findings in Section 4.

## 2 Methods

### 2.1 Mortality data and modelling

Since February 24, 2020, the [Dipartimento Della Protezione Civile \(2020\)](#) publishes the cumulative number of reported COVID-19 deaths for each of the 21 NUTS-2 Italian regions on a daily basis. Rather than the cumulative counts, we are interested in the new daily number of deaths, so we start our analysis from February 25. Let  $m$  be the number of days available in the dataset, and  $n$  be the number of regions. Let us denote by  $\mathbf{D} = (d_{t,r})$  the  $m \times n$  matrix containing the observed deaths at time  $t = 1, \dots, m$  in region  $r = 1, \dots, n$ . Time  $t$  is a natural number between February 25 and July 15.

Observed deaths  $d_{t,r}$  are assumed to be realizations from a Poisson distribution ([Brillinger, 1986](#)) with mean  $\check{e}_r \mu_{t,r}$ , where  $\check{e}_r$  denotes the population exposures to the risk of death for each region  $r$  and it is assumed fixed over the period under study. The vector  $\mu_{t,r}$  denotes the force of mortality at time  $t$  for each region  $r$  and its estimation is the object of the proposed model.

As mentioned, we first perform a preliminary step by classifying Italian regions into clusters that shared similar trajectories of the epidemic (see Section 2.3). This analysis will help to better disentangle the effects of explanatory variables as drivers of additional differences between regions. We then model our Poisson death counts via a log-link function in a Generalized Additive Model framework employing both linear and nonlinear effects. Specifically over time  $t$ , for region  $r$  belonging to cluster  $c$ , the logarithm of the force of mortality is given by

$$\ln[\mu_{t,r}] = \eta_{t,r} = \eta_t^0 + \gamma_t^c + \mathbf{Z}_{t,r} \boldsymbol{\beta}, \quad (1)$$

where  $\eta_t^0$  is fixed for each region and it represents the common epidemic dynamic over time. Cluster-specific variations over time with respect to the common time-trend are described by  $\gamma_t^c$ . For each region  $r$ , the design matrix  $\mathbf{Z}_{t,r}$  contains the values of the explanatory variables that could eventually change over time  $t$ . The coefficients vector  $\boldsymbol{\beta}$  is common over regions and it can be interpreted as in a classic regression setting. Note that, without loss of generality, we consider the first cluster as reference, i.e. estimated log-mortality for regions in this group is given by the simple sum of  $\eta_t^0$  and associated  $\mathbf{Z}_{t,r} \boldsymbol{\beta}$ .

We assume smoothness for both common time-trend and cluster-specific distances. Following a  $P$ -spline approach, we model these function as a linear combination of  $B$ -splines and associated coefficients which are penalized by discrete penalties ([Eilers and Marx, 1996](#)). Estimation procedure has been implemented in R ([R Development Core Team, 2020](#)) and the code can be obtained in the following blind repository: [https://osf.io/mu7hy/?view\\_only=7c03829c4c8a4859a37038581e1ce6fd](https://osf.io/mu7hy/?view_only=7c03829c4c8a4859a37038581e1ce6fd). Additional statistical aspects of the model are provided in Appendix A.

## 2.2 Explanatory variables

Here we describe the explanatory variables that we employ in our regression setting, we provide their sources and some descriptive statistics.

We consider two types of covariates in our analysis: (1) “constant” variables that do not change over the time frame analysed; (2) a set of time-varying variables, which are allowed to change on a daily basis throughout the observation period. We start by introducing the first type of covariates.

Exposures data, i.e. the total population for each region at the start of the year 2019 (the last available date), is retrieved from [Istat \(2020\)](#) and they will serve as offset in the regression setting. Moreover, the same source provides data on the share of population aged 65 years or more, the population density, the average size of the households, and the gross domestic product (GDP) by regions. Furthermore, we obtain the regional number of nursing homes from the [Istituto Superiore di Sanità \(2020\)](#). Finally, to account for the fact that regions experienced outbreaks at different time points, we compute a proxy for the delay of the epidemic starting date. This is derived as the number of days (since February 25) until cumulative cases surpassed 0.0001% of the regional population. For example, this regional-specific threshold is 10 in Lombardia, 4 in Piemonte and 2 in Calabria. This allows us to consider the population size in the timing of the epidemic outbreak across regions, and the choice of the threshold (0.0001%) does not modify outcomes of the following analysis.

Furthermore, we consider four time-varying covariates. The first is time, defined as the number of days since the first available date in the dataset. The other covariates are the cumulative number of tests performed, the total amount of current positive COVID-19 cases (including both hospitalised patients and home confinement), and the percentage of intensive care units (ICU) utilization. The first two variables are provided by the [Dipartimento Della Protezione Civile \(2020\)](#). The last variable is computed by dividing the daily number of patients currently in ICU (retrieved from [Dipartimento Della Protezione Civile, 2020](#)) by the total ICU availability in 2019 (the last available date). Regional data on the number of ICU available in 2019 is obtained from the [Ministero della Salute \(2020\)](#). It should be noted that, in some regions, the percentage utilization of ICU exceeds 100%, because the denominator does not take into account the large additions to the stock of ICU in response to the outbreak of COVID-19 during the first months of 2020.

To summarize, Table 1 reports the explanatory variables that we use in our analysis, together with their characterization (constant or time-varying), the transformations that we perform to some variables, and the descriptive statistics of the transformed variables. Figures B.1 and B.2 in Appendix B provide an exploratory analysis of these variables and COVID-related mortality. This analysis allows us also to identify the appropriate transformations of the covariates used in the following.

## 2.3 Regional clustering

The variability displayed by COVID-mortality data at the regional level is very high, and a simple regression analysis with the covariates described in Subsection 2.2 is not powerful enough to provide an adequate fit to the data. We thus opt for a preliminary step in which we group Italian regions that have experienced a similar unfold of the pandemic. This allows us to incorporate in Eq. (1) cluster-specific distances with respect to the general time-trends,  $\gamma_t^c$ ; these distances describe changes over  $t$  devoid of the effects of the explanatory variable in  $\mathbf{Z}_{t,r}$ .

**Table 1.** Explanatory variables considered in this study, together with their characterization (constant or time-varying), their transformation and some descriptive statistics (mean, standard deviation, minimum and maximum values).

Variable	Type	Transformation	Mean	SD	Min	Max
Population (offset)	constant	–	2,874.3	2,490.5	125.7	10,060.6
% population 65y+	constant	–	23.0	2.3	18.5	28.4
Population density	constant	–	177.7	110.8	38.5	424.4
GDP per capita	constant	log	10.2	0.3	9.7	10.8
Nursing homes	constant	–	114.2	175.0	1.0	677.0
Household size	constant	–	2.3	0.2	2.0	2.7
Delay of the epidemic	constant	–	5.0	3.2	0.0	10.0
Time	time-varying	–	71.5	41.0	1.0	142.0
% ICU utilization	time-varying	square root	3.5	3.4	0.0	16.4
Positive cases	time-varying	square root	34.5	35.2	0.0	193.2
Cumulative tests	time-varying	/1000	120.6	188.9	0.0	1,165.5

*Notes:* Figures for the population variable are divided by one thousand in the Table for illustrative purposes only.

*Sources:* [Ministero della Salute \(2020\)](#); [Dipartimento Della Protezione Civile \(2020\)](#); [Istat \(2020\)](#); [Istituto Superiore di Sanità \(2020\)](#)

Our aim is thus to classify regions according to their COVID-mortality time-trend. In practice, we first describe each regional mortality pattern in a non-parametric framework:

$$\ln [\mu_{t,r}] = \eta_{t,r} = \mathbf{B} \boldsymbol{\alpha}_r, \quad (2)$$

where  $\mathbf{B}$  is a matrix of  $k = 31$   $B$ -splines, which are common for all regions. Regional-specific coefficients  $\boldsymbol{\alpha}_r$  are estimated in a  $P$ -spline setting ([Camarda, 2012](#)). This approach reduces both randomness and dimensionality for our classification problem from  $m = 142$ , number of available days, to  $k$ , length of the vector  $\boldsymbol{\alpha}_r$ .

The estimated coefficients  $\mathcal{A} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_r, \dots, \boldsymbol{\alpha}_n]$  contain all relevant features of the observed regional mortality trends, and our aim is to classify them, and consequently classify  $\mu_{t,r}$ . Several options are available when a cluster analysis is performed. In this study we opt for a k-means clustering approach, which partitions the  $n$  vectors in  $\mathcal{A}$  into  $s$  sets so as to minimize the within-cluster sum of squares ([Hartigan and Wong, 1979](#)). This approach allows us to extract the principal components of the correlation structure of  $\mathcal{A}$ . The investigation of these components helps in interpreting the outcomes of the cluster analysis and consequently underpin the choice of  $s$  on explicit conjectures. The R-routine `kmeans()` was used to perform the clustering on our matrix  $\mathcal{A}$  ([R Development Core Team, 2020](#)).

The main problem in any classification approach concerns the choice of  $s$ , the number of clusters. Various criteria are available for optimizing the number of clusters and they have shown to produce diverse outcomes in our analysis. In the following we opt for a choice that balances interpretation of the produced clusters via the associated principal components and robustness of the outcomes from the regression analysis presented in Section 2.1. In other words, we select  $s$  such that the interpretation of both partitions and principal components of  $\mathcal{A}$  is clear, and estimated  $\boldsymbol{\beta}$  in Eq. (1) remain practically unchanged with respect to similar choices of  $s$ . Given this approach, we select  $s = 3$  sets of regions.

Finally the k-means approach allows us to extract “centers” of each cluster providing outcomes which are interesting by themselves. These “centers” could serve as prototypes for identifying regional evolution of the pandemic, and they could be used to identify specific patterns regardless the effects of the explanatory variables included in the regression analysis.

### 3 Results

The first step of our model consists in dividing the 21 Italian regions into three clusters. Figure 1 shows the results of this procedure. The first two principal components (PCs) of the smooth COVID-19 mortality pattern capture 93.1% of the overall variability found in the  $P$ -splines coefficients. Whereas the left panel depicts the partitioning of the regions into three well-separated clusters, the right panels of Figure 1 show the first two principal components.

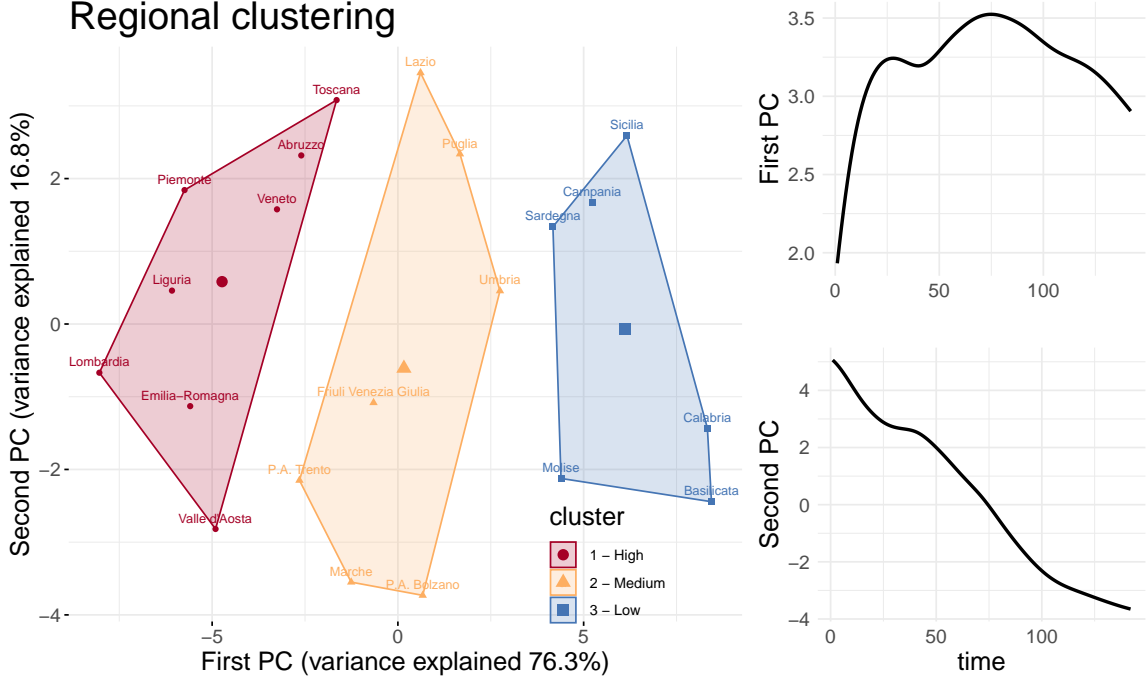
We labeled the three clusters High, Medium and Low mortality since the first PC captures a large part of the overall variability (76.3%), and it describes the average shape of the epidemic over time. Consequently, the values associated to this component describe the regional mortality levels and, with the exception of Toscana, the clustering is based on the different level of mortality. However, the following regression analysis will identify additional features in the cluster-specific mortality patterns.

Moreover, the second principal component (see bottom-right panel in Figure 1) represents how the main pattern deviates over time and, given its essential linearity, it indicates that regions with higher (lower) associated PC value have a relatively higher (lower) mortality in the first period with respect to the second.

The smooth mortality pattern by regions as well as the cluster centers are shown in Figure C.1 of Appendix C.

Figure 2 shows the map of Italy, with regions coloured according to their corresponding cluster. Moreover, colour transparency is proportional to the regional values associated to the first principal component, with higher transparency corresponding to higher values of the component, and hence higher mortality level. As expected, the clustering division broadly follows the territorial positions of the regions, with Northern regions belonging to the High mortality clusters, Centre regions to the Medium one, and Southern regions to the Low one. However, some exceptions are observable, such as P.A. Trento and Bolzano, Friuli-Venezia Giulia, Abruzzo and Puglia. Regions at the border of the transparency-opacity gradient identify boundaries of the clustering divisions in terms of the first principal component. For example, the high transparency of Toscana and Abruzzo indicate that they are closer to the Medium cluster than regions with high opacity such as Lombardia and Liguria.

Next, we run our model on the covariates described in Subsection 2.2. To select the covariates that comprise our final model, we employ a bottom-up, or forward selection approach. Starting from a very parsimonious model, which includes only the cluster effects, we add covariates once at a time, and we retain them only when they are statistically significant (using the Wald test statistic). The rationale behind the bottom-up approach instead of a top-down (backward elimination) one is that some covariates are highly correlated between each other, so that employing a very large model can introduce collinearity in the regression setting. However, different selection procedures have been tested and lead to the same final model.



**Figure 1.** Divisions of the twenty-one regions in Italy into three clusters (High, Medium and Low mortality) according to the first two principal components (PCs). The three clusters are reported in the left panel, while the first two principal components are depicted in the right panels.

*Source* (Figs. 1–4): Authors' own elaborations.

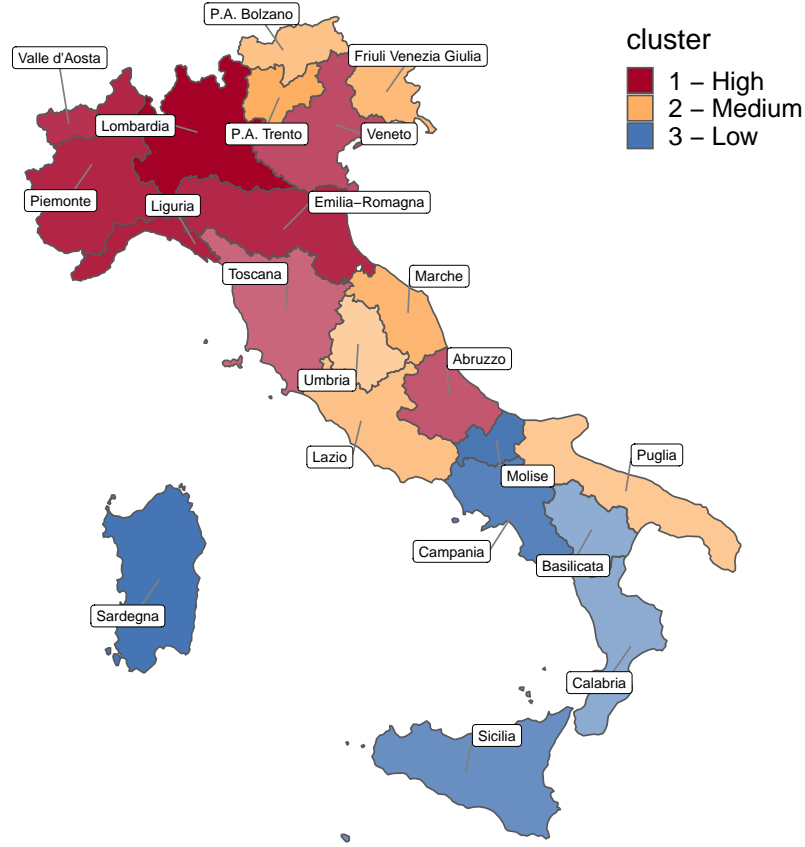
Figure 3 and Table 2 show the results of our model. The figure displays the estimated reference and cluster-specific curves over time with 95% confidence intervals.

The reference curve,  $\eta_t^0$ , describes the average time-profile of the epidemic in the High mortality cluster with a rapid log-mortality increase up to mid-March, followed by a relatively stable pattern and a late downward trend from early June. It is important to also highlight two relatively small mortality increases during April and the end of May. The former increase is particularly puzzling as occurring while the entire country was in lockdown.

In addition, the right panel of Figure 3 presents the cluster-specific terms,  $\gamma_t^c$ , that directly act on the reference curve to modify both its level and curvature. Both curves have inverse-U shapes with more prominent edges in the Low cluster, i.e. whereas all Italian regions have shared a similar log-mortality pattern from the end of March to the end of April, regions in the Medium and Low clusters have lower mortality at the beginning and toward the end of the whole period. Moreover, the Low cluster dramatically approached log-mortality level in the High cluster around the end of March, but those regions simultaneously departed from the highest mortality level immediately after mid-April, and at a faster pace with respect to the regions in the Medium cluster.

Six variables are retained as statistically significant in the model selection: (i) the share of ICU utilization, (ii) the delay in the start of the epidemic, (iii) the cumulative number of tests performed, (iv) the GDP per capita, (v) the share of population aged 65 years or more, and (vi) the number of positive COVID-19 cases.

The signs of the estimated coefficients are in line with our expectations. The share of ICU utilization, the level of GDP per capita, the share of older population and the

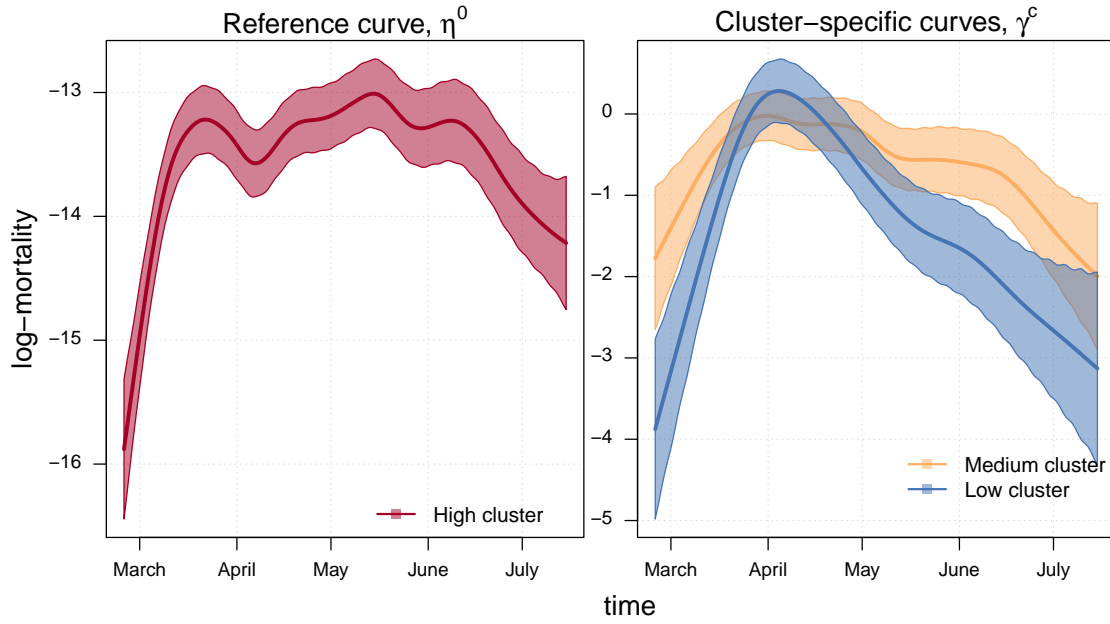


**Figure 2.** Map of the twenty-one Italian regions associated with their respective clusters (High, Medium and Low mortality). Colour transparency is proportional to regional value associated to the first principal component.

**Table 2.** Covariates included in our Poisson regression model, with estimated coefficients and associated 95% confidence intervals. Variables are ordered in terms of absolute coefficient magnitude.

Variable	Estimate	95% CI
% ICU utilization (square-root)	0.84	(0.81; 0.88)
Delay of the epidemic	-0.37	(-0.40; -0.33)
Cumulative tests (/1000)	-0.19	(-0.22; -0.16)
GDP per capita (log)	0.13	(0.09; 0.18)
% population 65y+	0.09	(0.07; 0.11)
Positive cases (square-root)	0.08	(0.06; 0.10)
Deviance	8140	
Effective Dimension	36	
BIC	8431	

number of positive cases are positively associated with COVID-19 mortality. Conversely, the delay in the start of the epidemic and the number of cumulative tests performed are negatively associated with mortality. To aid the interpretation and comparison of the

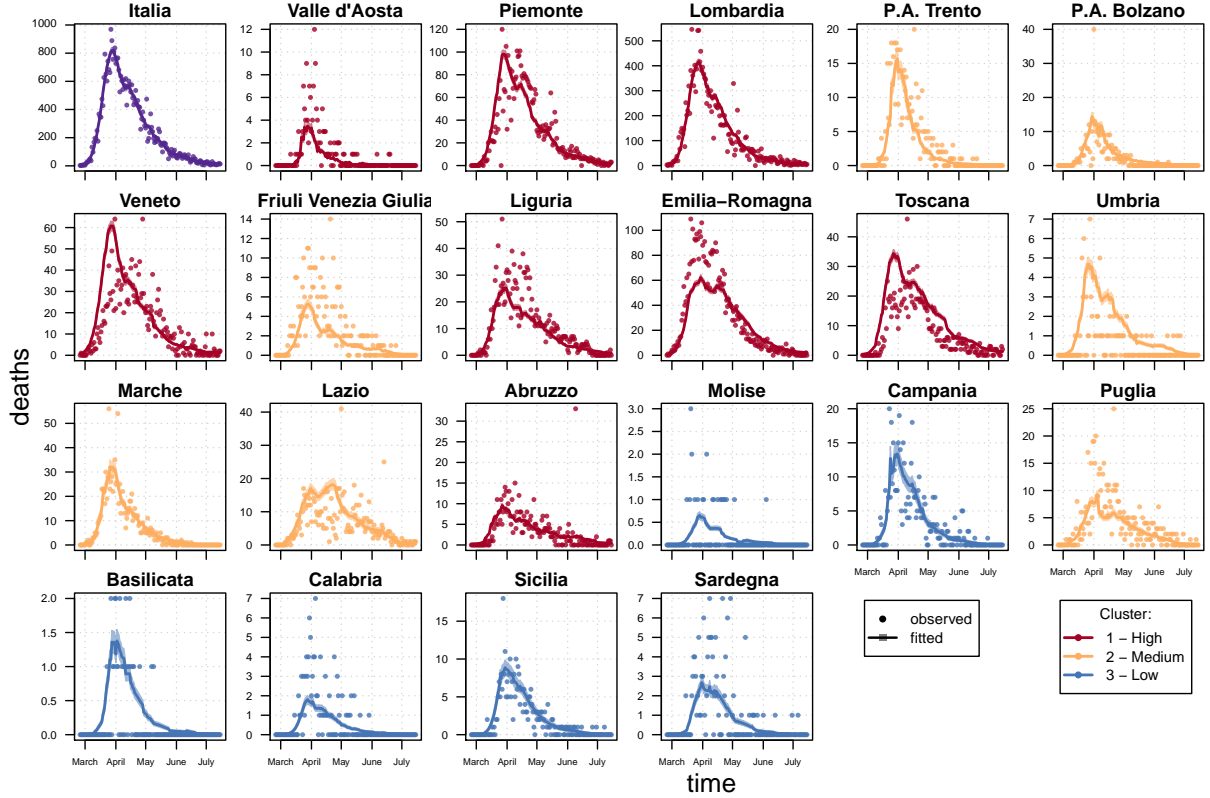


**Figure 3.** Estimated smooth reference and cluster-specific curves with 95% confidence intervals from the Poisson regression model on COVID-related mortality and the six explanatory variables in Table 2.

different coefficients, we normalize each variable by mean-centering and scaling to one standard deviation (1SD). As such, the estimated coefficients can be considered as the change in the log-mortality corresponding to a 1SD increase in the variable. The strongest effect on mortality is found for the (squared root of the) share of ICU utilization, with a 1SD (3.4 percentage points, cf. Table 1) increase associated with a 0.92 increase in the log-mortality. The lowest effects are observed for the share of the older population and the number of positive cases, with coefficients equal to 0.09 and 0.08, respectively.

It is worth noticing that the remaining covariates introduced in Subsection 2.2, namely population density, household size and the number of nursing homes, were not retained in our model. For population density and the number of nursing homes, the estimated coefficients were not statistically significant, leading to an increase in the BIC as compared to our selected model. For the average size of the household, issues of collinearity arose in the model (altering the signs of other estimated coefficients). This effect was mainly due to the high correlation between household size and two other variables in our model (GDP per capita and share of the older population, cf. Fig. B.2 in Appendix B).

The estimated model allows us to compare the observed and fitted regional evolution of the pandemic. Figures 4 show the observed and fitted COVID-19 death counts with 95% confidence intervals in each region as well as for the overall country (top-left panel). The model captures the data well, although fitted values tend to underestimate the data in Emilia-Romagna and can only partially capture the relatively different shape of the curve observed in Veneto. For completeness, in the Appendix C, we report a similar plot for log-mortality as well as the model’s deviance residuals for a model diagnostic.



**Figure 4.** Observed and fitted (with 95% confidence intervals) number of COVID-positive deceased individuals in Italy (top-right) and across its 21 regions from February 25 to July 15, 2020, stratified by three clusters of High, Medium and Low mortality.

## 4 Discussion

The research community has been very responsive to analyse the spread of COVID-19 across the globe. For the purpose of this article, we focus on the relevant literature for the Italian context.

Early efforts have been directed towards monitoring the virus’s spread at the national or subnational level using Poisson models. [Chiogna and Gaetan \(2020\)](#) proposed a dynamic generalized linear model for the Poisson distribution of new and total cases at the national, regional and provincial level. The analysis of the time-varying slope of the local linear trend allows them to detect changes in the underlying process in terms of acceleration, deceleration or stabilization of the diffusion of the disease. Moreover, [Bonetti and Basellini \(2020\)](#) introduced a tool for visualizing the spread of COVID-19 in Italian provinces and regions by modelling the total number of cases with Poisson regression and parametric hazard functions. Furthermore, [Agosto et al. \(2020\)](#) proposed a Poisson autoregression on the daily number of cases, and compared the Italian context with China and other European countries.

All these works consider the spread of the virus in different territories *in isolation*, i.e. each region or province is analysed independently without taking into account their correlations. However, regional contexts have played a relevant role in the unfolding of the Italian pandemic. As such, in this paper we have introduced a methodology that considers the country as the sum of different regional experiences. Our regression model allows us to identify the most significant variables that have contributed to the

greater or lower burden of deaths across regional units from the end of February until mid July 2020. In particular, we find that the percentage of ICU utilization, the GDP per capita, the share of the older population and the number of positive COVID-19 cases are positively associated with COVID-19 mortality. Conversely, the delay in the start of the epidemic and the cumulative number of tests performed are negatively associated with mortality.

Few attempts have been made to take into account the cross-regional dependence in the spread of COVID-19 in Italy. [Maltagliati \(2020\)](#) was among the firsts to suggest analysing the Italian epidemic as the sum of region-specific outbreaks. The author describes the cumulative number of deaths using a logistic model that considers the delay in the start of the regional outbreaks. The proposed model produces large differences in the regional-specific asymptotes, and the author argues that the regional perspective is fundamental to understand the evolution of COVID-19 in Italy. Furthermore, [Boschi et al. \(2020\)](#) employ Functional Data Analysis techniques to investigate the association of COVID-19 mortality with mobility, positivity and other covariates at the regional level from mid-February to April. The authors document the outbreak of two starkly different epidemic types: an exponential one in the worst hit areas in the north of the country, and a flat one in the remaining regions.

Our analysis shares commonalities with these two studies, but they substantially differ in a number of ways. First, the parametric model proposed by [Maltagliati \(2020\)](#) does not consider the role of explanatory variables in shaping the effects of the COVID-19 outbreak across the regions. Second, while the semi-parametric analysis of [Boschi et al. \(2020\)](#) is conceptually closer to our approach, the focus of the two studies is rather different. In their work, [Boschi et al. \(2020\)](#) concentrate the analysis on the role of mobility and positivity as predictors of COVID-19 mortality, and other covariates are only considered – one at a time – as control variables in the regression model. In our work, we take a more comprehensive view and assess the competing effect of several factors on mortality during the pandemic. Nonetheless, we acknowledge the contributions of these two studies, and our work is meant to complement and provide additional insights on the dynamic of the Italian epidemic.

Our findings are generally in line with those reported by recent research on the COVID-19 pandemic. The share of ICU utilization has the strongest association with mortality, after controlling for the (cluster-specific) epidemic time trend and additional factors. Regions where ICU needs exceeded the available stock experienced the highest levels of COVID-19 mortality. The saturation of ICU beds and ventilator availability has indeed been an important factor to explain the higher number of COVID-related deaths in Lombardia and Italy ([Favero, 2020](#); [Volpato et al., 2020](#)). Furthermore, the negative association between mortality and (i) the delay of the outbreak and (ii) the cumulative number of tests performed, are reasonable: regions that experienced later outbreaks had relatively more time to prepare for it, and higher number of tests could have allowed for a better tracing of infected individuals – and therefore a lower number of new cases and deaths. For example, an empirical study in the municipality of Vo’ ([Lavezzo et al., 2020](#)) has shown that viral transmission of the disease can be suppressed when these factors interact (early isolation of infected people combined with community lockdown).

Finally, the positive, albeit comparatively weaker, associations of mortality with the remaining covariates (GDP per capita, share of older persons and number of positive COVID-19 cases) are also in line with previous studies. The COVID-19 mortality gradient in Italian regions closely matches the economic gradient, with Northern regions having

higher GDP per capita than Southern ones. Furthermore, the role of population age structures has been suggested to explain the higher number of COVID-deaths in older versus younger populations (Dowd et al., 2020).

Interestingly, some of our findings are not aligned with recent research on the pandemic. Rocklöv and Sjödin (2020) suggested that population densities are an important catalyst for the spread of COVID-19. Moreover, nursing homes have been identified as hotspots of COVID-related deaths in Italy (Trabucchi and Leo, 2020; di Giacomo et al., 2020). In our analysis, these two variables were not significantly associated with COVID-19 mortality, after controlling for other factors. Similarly, the dimension of households has been proposed as a key factor (together with the population age structure) to determine the vulnerability of countries to outbreaks of COVID-19 (Esteve et al., 2020). Conversely, the dimension of households is negatively correlated with COVID-19 mortality in Italian regions during the period analysed, and a similar negative correlation is observable across clusters for the number of nursing homes (see Figure B.2 in Appendix B). It is likely that these discrepancies are related to the contexts of the analysis, as our subnational setting differs from those considered in these studies. As such, our findings highlight the importance of context-specific analysis, providing a warning to the generalizability of COVID-related hypothesis and results.

In conclusion, our study shed light on the most significant factors that have contributed to the spread of COVID-19 in Italian regions. In addition to their scientific value, our findings provide important insights for policymakers in the event of future pandemics or a potential second wave of COVID-19. Finally, the methodology that we propose in this article is a novel contribution to the analysis of mortality during epidemics, which can be fully replicated and applied to other countries and frameworks (even outside epidemic research) using the codes provided along with our article.

## Acknowledgements

We would like to thank Marília Nepomuceno as well as the members of the Laboratory of Digital and Computational Demography at the Max Planck Institute for Demographic Research for providing useful comments on a previous version of this manuscript.

## Authors' contributions

Both authors designed the study, retrieved data, performed analysis and wrote the manuscript.

## References

- Agosto, A., Campmas, A., Giudici, P., and Renda, A. (2020). Monitoring Covid-19 contagion growth in Europe. CEPS working paper. Available at: <https://www.ceps.eu/ceps-publications/monitoring-covid-19-contagion-growth-in-europe/>.
- Billari, F. and Ongaro, F. (1998). The transition to adulthood in Italy. Evidence from cross-sectional surveys / Le passage à l'âge adulte en Italie. *Espace, populations, sociétés*, 16(2):165–179.

- Bonetti, M. and Basellini, U. (2020). Epilocal: a real-time tool for local epidemic monitoring. arXiv preprint. Available at: <https://arxiv.org/abs/2003.07928>.
- Boschi, T., Iorio, J. D., Testa, L., Cremona, M. A., and Chiaromonte, F. (2020). The shapes of an epidemic: using Functional Data Analysis to characterize COVID-19 in Italy. arXiv preprint. Available at: <https://arxiv.org/abs/2008.04700>.
- Brillinger, D. R. (1986). A biometrics invited paper with discussion: The natural variability of vital rates and associated statistics. *Biometrics*, 42(4):693–734.
- Camarda, C. G. (2012). MortalitySmooth: An R Package for Smoothing Poisson Counts with  $P$ -Splines. *Journal of Statistical Software*, 50:1–24. Available on [www.jstatsoft.org/v50/i01](http://www.jstatsoft.org/v50/i01).
- Cereda, D., Tirani, M., Roviola, F., Demicheli, V., and al. (2020). The early phase of the COVID-19 outbreak in Lombardy, Italy. arXiv preprint. Available at: <https://arxiv.org/abs/2003.09320>.
- Chiogna, M. and Gaetan, C. (2020). COVID-19 in Italy. Available at: <https://github.com/cgaetan/COVID-19>. Accessed on August 24, 2020.
- di Giacomo, E., Bellelli, G., Pesci, G., Scarpetta, S., Colmegna, F., de Girolamo, G., and Clerici, M. (2020). Management of older people during the COVID19 outbreak: Recommendations from an Italian experience. *International Journal of Geriatric Psychiatry*, 35(7):803–805.
- Di Giulio, P. and Rosina, A. (2007). Intergenerational family ties and the diffusion of cohabitation in Italy. *Demographic Research*, 16:441–468.
- Dipartimento Della Protezione Civile (2020). Dataset of COVID-19 infected cases in Italy. Available at: <https://github.com/pcm-dpc/COVID-19/tree/master/>. Accessed on August 27, 2020.
- Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y., and Mills, M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences*, 117(18):9696–9698.
- Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B. J., and Meyers, L. A. (2020). Risk for transportation of coronavirus disease from wuhan to other cities in china. *Emerging Infectious Diseases*, 26(5):1049–1052.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with  $B$ -splines and Penalties (with discussion). *Statistical Science*, 11:89–102.
- Esteve, A., Permanyer, I., Boertien, D., and Vaupel, J. W. (2020). National age and coresidence patterns shape COVID-19 vulnerability. *Proceedings of the National Academy of Sciences*, 117(28):16118–16120.
- Favero, C. A. (2020). Why is COVID-19 Mortality in Lombardy so High? Evidence from the Simulation of a SEIHCRC Model. *Covid Economics. Vetted and Real-Time Papers*. Accessed on August 20, 2020. Available at <https://iris.unibocconi.it/retrieve/handle/11565/4026151/122470/CovidEconomics4.pdf>.

- Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28:100–108.
- Helliwell, J. F. and Putnam, R. D. (1995). Economic Growth and Social Capital in Italy. *Eastern Economic Journal*, 21(3):295–307.
- Istat (2020). Resident population and demographic indicators for year 2019. <http://demo.istat.it/pop2019/index1.html>. Accessed on April 6, 2020.
- Istituto Superiore di Sanità (2020). Survey nazionale sul contagio covid-19 nelle strutture residenziali e sociosanitarie. Aggiornamento nazionale: 6 aprile 2020. Available at: <https://www.epicentro.iss.it/coronavirus/sars-cov-2-survey-rsa>.
- Johns Hopkins University CSSE (2020). Novel Coronavirus (COVID-19) Cases. <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>. Accessed on August 17, 2020.
- Kertzer, D. I., White, M. J., Bernardi, L., and Gabrielli, G. (2008). Italy’s Path to Very Low Fertility: The Adequacy of Economic and Second Demographic Transition Theories. *European Journal of Population / Revue européenne de Démographie*, 25(1):89–115.
- Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Vecchio, C. D., Rossi, L., Manganelli, R., Loregian, A., Navarin, N., Abate, D., Sciro, M., Merigliano, S., Canale, E. D., Vanuzzo, M. C., Besutti, V., Saluzzo, F., Onelia, F., Pacenti, M., Parisi, S. G., Carretta, G., Donato, D., Flor, L., Cocchio, S., Masi, G., Sperduti, A., Cattarino, L., Salvador, R., Nicoletti, M., Caldart, F., Castelli, G., Nieddu, E., Labella, B., Fava, L., Drigo, M., Gaythorpe, K. A. M., Brazzale, A. R., Toppo, S., Trevisan, M., Baldo, V., Donnelly, C. A., Ferguson, N. M., Dorigatti, I., and Crisanti, A. (2020). Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo’. *Nature*, 584(7821):425–429.
- Maltagliati, M. (2020). COVID-19 in Italia: una o tante epidemie? Neodemos. Available at: [https://www.neodemos.info/articoli/covid\\_19-in-italia-una-o-tante-epidemie/](https://www.neodemos.info/articoli/covid_19-in-italia-una-o-tante-epidemie/). Accessed on April 3, 2020.
- Ministero Della Giustizia (2020). Gazzetta Ufficiale della Repubblica Italiana. volume 161. Ufficio Pubblicazione Leggi e Decreti. All COVID-related decrees are available at: <https://www.gazzettaufficiale.it/dettaglioArea/12>.
- Ministero della Salute (2020). Dataset of beds per hospital facility. Available at: <http://www.dati.salute.gov.it/dati/dettaglioDataset.jsp?menu=dati&idPag=18>. Accessed on March 30, 2020.
- R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rocklöv, J. and Sjödin, H. (2020). High population densities catalyse the spread of COVID-19. *Journal of Travel Medicine*.
- Trabucchi, M. and Leo, D. D. (2020). Nursing homes or besieged castles: COVID-19 in northern Italy. *The Lancet Psychiatry*, 7(5):387–388.
- Volpato, S., Landi, F., and Incalzi, Raffaele Antonelli, on behalf of the Italian Society of

Gerontology and Geriatrics. (2020). A Frail Health Care System for an Old Population: Lesson form the COVID-19 Outbreak in Italy. *The Journals of Gerontology: Series A*. Available at <https://doi.org/10.1093/gerona/glaa087>.

World Health Organization (2020a). Coronavirus disease 2019 (COVID-19): Situation Report – 194. [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200801-covid-19-sitrep-194.pdf?sfvrsn=401287f3\\_2](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200801-covid-19-sitrep-194.pdf?sfvrsn=401287f3_2). Accessed on August 10, 2020.

World Health Organization (2020b). Rolling updates on coronavirus disease (COVID-19). Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen>.

## A The regression model: statistical details

In this Appendix, we provide additional information about the estimation procedure we performed for the model introduced in Section 2.1. Beside the explanatory variables collected in the regional-specific matrices  $\mathbf{Z}_{t,r}$ , the proposed model requires two datasets as input data: the matrix of deaths  $\mathbf{D} = (d_{t,r})$  over time  $t$  for region  $r$ , and the vector  $\check{\mathbf{e}} = (\check{e}_r)$  that collects the population exposures to the risk of death for each region  $r$ . For the purpose of regression, we arrange death counts as column vectors, i.e.  $\mathbf{d} = \text{vec}(\mathbf{D})$ . Likewise we arrange the matrix of exposures  $\mathbf{e} = \text{vec}(\mathbf{E})$ , where  $\mathbf{E} = (e_{t,r}) = \mathbf{1}_m \check{\mathbf{e}}'$ , with  $\mathbf{1}_m$  a  $m \times 1$  matrix of ones.

As mentioned in Section 2.1, we assume deaths to be realizations from a Poisson distribution with mean  $\check{e}_r \mu_{t,r} = e_{t,r} \mu_{t,r}$ . We then model our Poisson death counts for all Italian regions via a log-link function:

$$\ln [\mathbb{E}(\mathbf{d})] = \ln(\mathbf{e}) + \ln(\boldsymbol{\mu}) = \ln(\mathbf{e}) + \boldsymbol{\eta} = \ln(\mathbf{e}) + \mathbf{X} \boldsymbol{\theta}, \quad (3)$$

where  $\boldsymbol{\eta} = \ln(\boldsymbol{\mu})$  is the linear predictor. The design matrix  $\mathbf{X}$  contains all features described in (1) and the coefficients vector  $\boldsymbol{\theta}$  is a vector of the model's coefficients which needs to be estimated.

In order to present all model components and without loss of generality, we sort  $n = 9$  regions according to their cluster membership and we consider these regions equally distributed in  $s = 3$  clusters. Design matrix and coefficients vector are then given by

$$\mathbf{X} = \begin{bmatrix} \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_1 \\ \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_2 \\ \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 \\ \mathbf{B} & \mathbf{B} & \mathbf{0} & \mathbf{Z}_4 \\ \mathbf{B} & \mathbf{B} & \mathbf{0} & \mathbf{Z}_5 \\ \mathbf{B} & \mathbf{B} & \mathbf{0} & \mathbf{Z}_6 \\ \mathbf{B} & \mathbf{0} & \mathbf{B} & \mathbf{Z}_7 \\ \mathbf{B} & \mathbf{0} & \mathbf{B} & \mathbf{Z}_8 \\ \mathbf{B} & \mathbf{0} & \mathbf{B} & \mathbf{Z}_9 \end{bmatrix}, \quad \boldsymbol{\theta} = [\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^2, \boldsymbol{\alpha}^3, \boldsymbol{\beta}], \quad (4)$$

where  $\mathbf{B}$  are  $k$  equally spaced  $B$ -spline bases and  $\mathbf{0}$  are  $m \times k$  matrices of zeros. The last columns in  $\mathbf{X}$  are made of all region-specific  $m \times p$  matrices with the values of the explanatory variables that could eventually change over  $t$ .

The vector  $\boldsymbol{\theta}$  concatenates all associated coefficients and consequently common time-trend and clustering deviation from it can be written as follows:

$$\boldsymbol{\eta}^0 = \mathbf{B} \boldsymbol{\alpha}^0 \quad \boldsymbol{\gamma}^2 = \mathbf{B} \boldsymbol{\alpha}^2 \quad \boldsymbol{\gamma}^3 = \mathbf{B} \boldsymbol{\alpha}^3. \quad (5)$$

Following a  $P$ -splines approach, we use a generous number of  $B$ -spline and enforce smoothness of  $\boldsymbol{\eta}^0$ ,  $\boldsymbol{\gamma}^2$  and  $\boldsymbol{\gamma}^3$  by penalizing the associated coefficients. As result, we can adapt the iteratively re-weighted least-squares algorithm for estimating GLMs by adding a penalty term:

$$(\mathbf{X}'\mathbf{W}\mathbf{X} + \mathbf{P})\boldsymbol{\theta} = \mathbf{X}'\mathbf{W}\mathbf{z}, \quad (6)$$

where  $\mathbf{z} = (\mathbf{d} - \mathbf{e} * \boldsymbol{\mu}) / \mathbf{e} * \boldsymbol{\mu} + \boldsymbol{\eta}$  is the working dependent variable and  $\mathbf{W}$  is a diagonal matrix of weights,  $\mathbf{W} = \text{diag}(\mathbf{e} * \boldsymbol{\mu})$ . The penalty term  $\mathbf{P}$  must work only on  $\boldsymbol{\alpha}^0$ ,  $\boldsymbol{\alpha}^2$  and  $\boldsymbol{\alpha}^3$  and it takes the following diagonal structure:

$$\mathbf{P} = \begin{bmatrix} \lambda_0 \boldsymbol{\Delta}'\boldsymbol{\Delta} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 \otimes \lambda_\gamma \boldsymbol{\Delta}'\boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad (7)$$

where  $\boldsymbol{\Delta}$  is a difference matrix over  $t$ . The smoothing parameters  $\lambda_0$  and  $\lambda_\gamma$  control the trade-off between smoothness of  $(\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^2, \boldsymbol{\alpha}^3)$  and model fidelity. We assume second order differences for  $\boldsymbol{\Delta}$  and the same smoothing parameter for all cluster-specific deviations. Finally the Bayesian Information Criterion was employed to optimize both  $\lambda_0$  and  $\lambda_\gamma$ .

## B Exploratory analysis

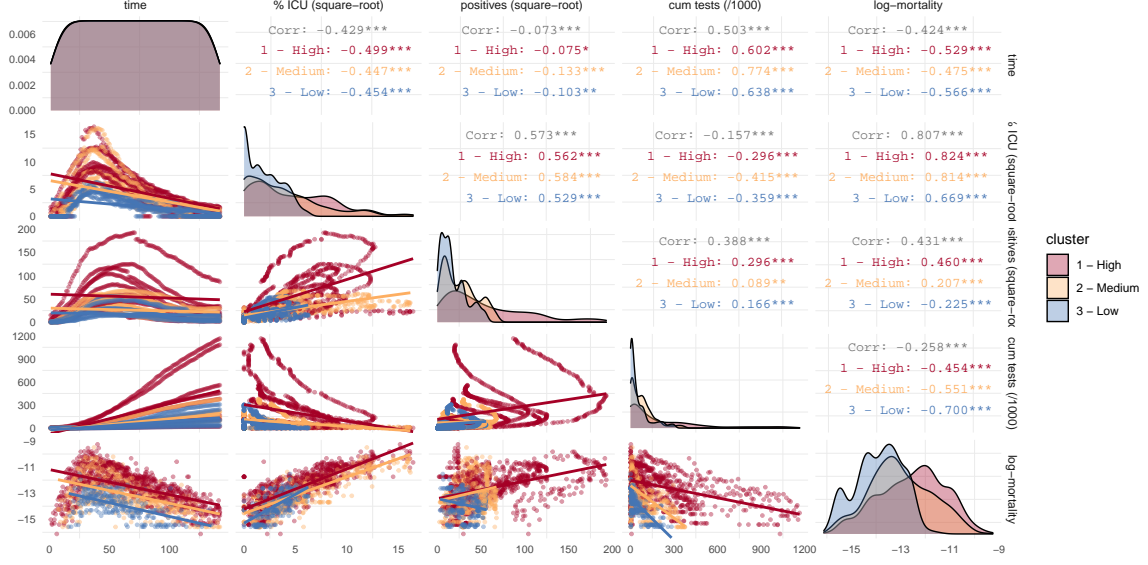
In this Appendix, we perform some exploratory analysis of the dataset that we employ in our study.

Figure B.1 shows the distributions, correlations and scatter plots of COVID-mortality (in log-scale) and the time-varying covariates described in Subsection 2.2, stratified by the three clusters of High, Medium and Low mortality. The Figure provides several information. The scatter plots in the last row show the relationship between mortality and the variables, and their correlations can be read from the last column of the graph. The non-linear relationship between time and mortality clearly emerges, as well as the strong linear relationship between mortality and ICU utilization (after a square root transformation). The negative correlation between cumulative tests and mortality is also evident, and the mortality density plot (last panel in the last row) show the ordering of the clusters in terms of different mortality levels (i.e. higher clusters are characterized by higher mortality).

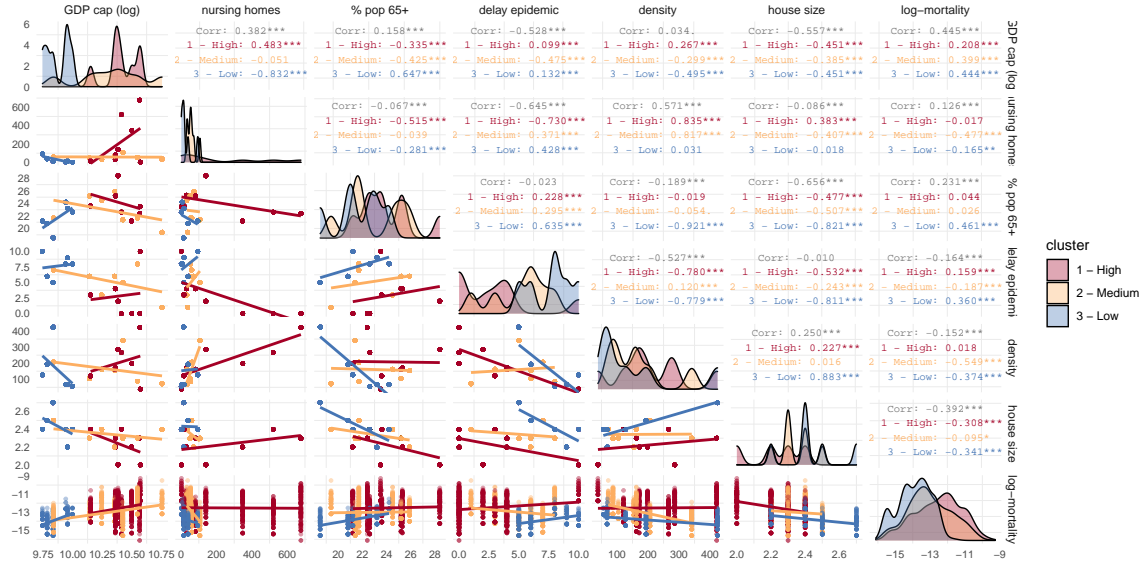
A similar plot for the time-constant variables is provided in Figure B.2. On one hand, the six constant variables have a rather low correlation to mortality; on the other, some variables are highly correlated between each other.

## C Additional results

Here we present additional results of our analysis. Figure C.1 shows the smooth COVID-19 mortality patterns over time for the 21 regions in Italy, as well as the overall patterns for the centers of the three clusters. The figure clearly shows how the centers distinguish



**Figure B.1.** Exploratory analysis of the four time-varying variables described in Subsection 2.2 and COVID-related mortality, stratified by the three clusters employed in this study. *Source* (Figs. B.1-B.2): Authors' elaborations on data from Dipartimento Della Protezione Civile (2020); Ministero della Salute (2020); Istat (2020); Istituto Superiore di Sanità (2020).

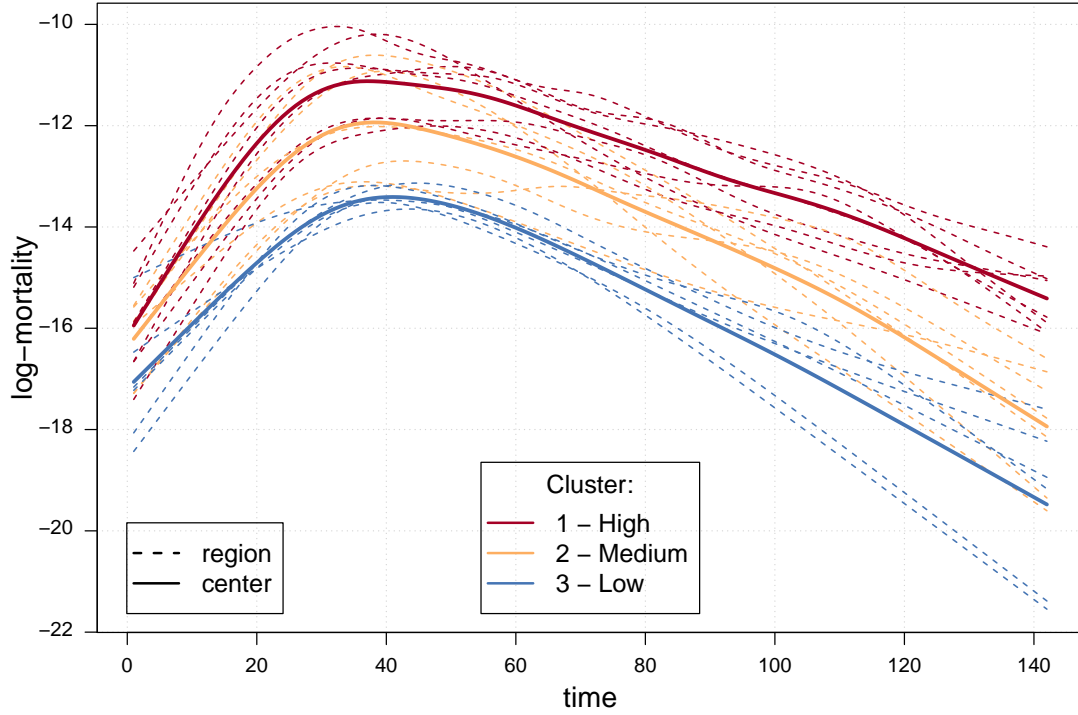


**Figure B.2.** Exploratory analysis of the six constant variables described in Subsection 2.2 and COVID-related mortality, stratified by the three clusters employed in our study.

three patterns of the pandemic, with mortality level and shapes well stratified over time. The three clusters are clearly ordered into High, Medium and Low mortality.

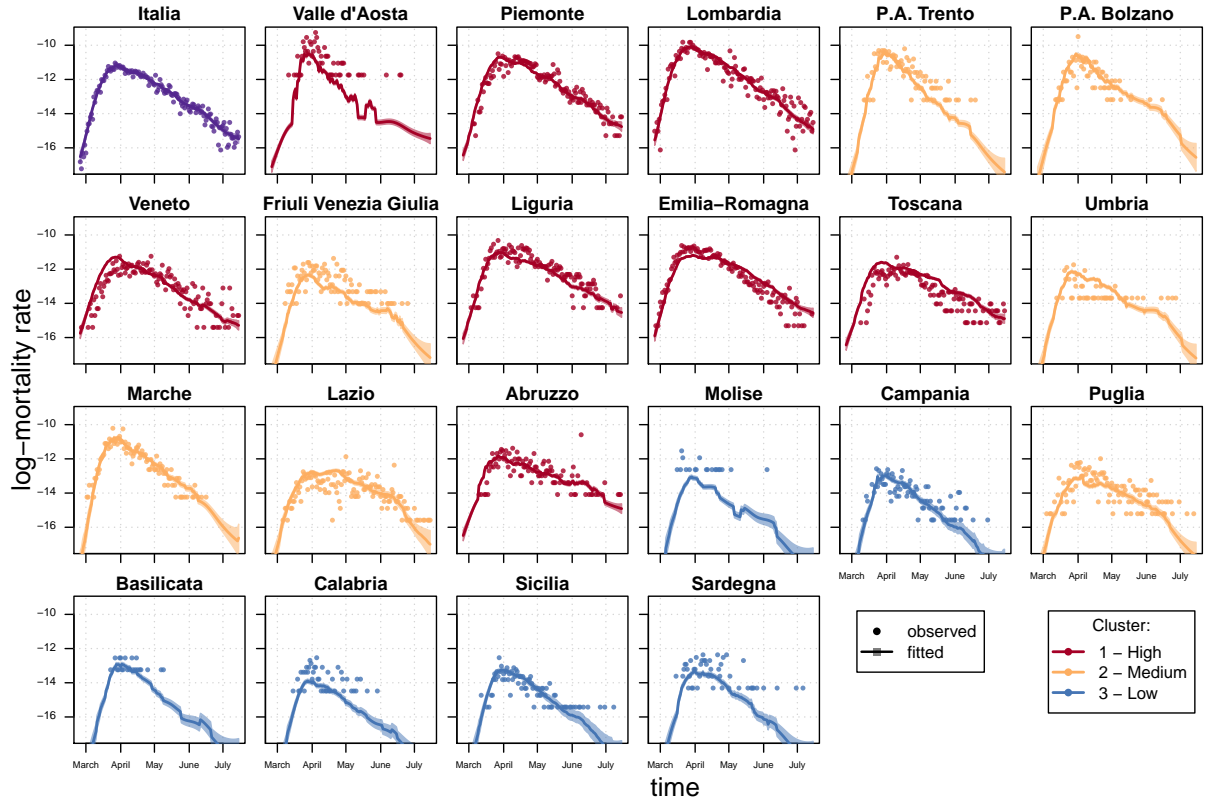
Figures C.2 show the observed and fitted log-mortality rates of COVID-positive deceased individuals with 95% confidence intervals in each region as well as for the overall country (top-left panel).

Finally, Figure C.3 presents the Poisson deviance residuals of our model. As shown in Figures C.2 and 4 of the manuscript, the model captures the observed data well, although

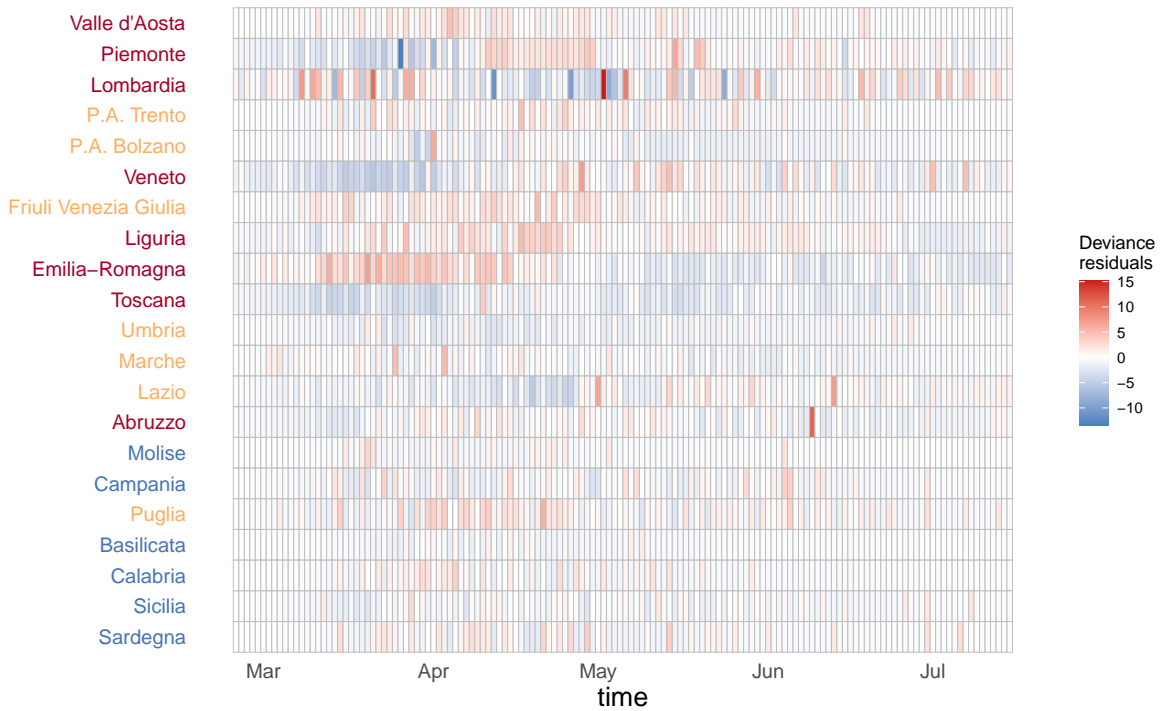


**Figure C.1.** Smooth COVID-19 mortality patterns over time for the 21 Italian regions (dashed lines) and cluster centers (solid lines), with colors corresponding to the three clusters.

fitted values underestimate observed data in Emilia-Romagna and in Piemonte since mid-April. An overestimation is also observable in the first part of the dataset for Piemonte and Veneto.



**Figure C.2.** Observed and fitted (with 95% confidence intervals) log-mortality rates of COVID-positive deceased individuals in Italy and across its 21 regions from February 25 to July 15, 2020, stratified by three clusters of High, Medium and Low mortality.



**Figure C.3.** Deviance residuals of the Poisson regression model for 21 regions in Italy from February 25 to July 15, 2020, stratified by three clusters of High, Medium and Low mortality.