scientific reports

Check for updates

OPEN

Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy

Tobia Boschi[1], Jacopo Di Iorio[2], Lorenzo Testa[2], Marzia A. Cremona[1,3,4]✉ & Francesca Chiaromonte[1,2]✉

We investigate patterns of COVID-19 mortality across 20 Italian regions and their association with mobility, positivity, and socio-demographic, infrastructural and environmental covariates. Notwithstanding limitations in accuracy and resolution of the data available from public sources, we pinpoint significant trends exploiting information in curves and shapes with Functional Data Analysis techniques. These depict two starkly different epidemics; an "exponential" one unfolding in Lombardia and the worst hit areas of the north, and a milder, "flat(tened)" one in the rest of the country—including Veneto, where cases appeared concurrently with Lombardia but aggressive testing was implemented early on. We find that mobility and positivity can predict COVID-19 mortality, also when controlling for relevant covariates. Among the latter, primary care appears to mitigate mortality, and contacts in hospitals, schools and workplaces to aggravate it. The techniques we describe could capture additional and potentially sharper signals if applied to richer data.

At the end of January 2020, two Chinese tourists were hospitalized in Rome and tested positive to SARS-CoV-2. At the beginning of February, a group of Italian citizens was repatriated from Wuhan – among them, one tested positive. As the news media reported these headlines, neither the Italian public nor the Italian authorities appeared to perceive an imminent threat, though retrospective analyses now suggest that the virus may have been circulating in the north of the country as far back as December 2019 (e.g., detection of SARS-CoV-2 in the wastewater of Milan and Turin[1]). The first recorded non-travel related COVID-19 case occurred in Codogno (Lombardia)—where a 38 years old male visited the hospital first on February 17, and then again on February 19 with worsening respiratory symptoms; in this date, he was tested and diagnosed. On February 20, two individuals tested positive in Vo' Euganeo (Veneto). Notably, the outbreaks in Lombardia and Veneto took two very different paths, something many observers attributed to the early response and aggressive testing strategy adopted by the regional authorities in Veneto[2,3]. After some initial, much debated inconsistencies (e.g., hesitations in implementing local lock-downs in areas hosting major industrial production hubs, contested decisions to move patients between hospitals and nursing homes and to keep major sports events open to the public in Lombardia), starting in early March, local and central authorities took progressively more stringent measures to limit mobility and social gatherings—culminating with a general nationwide lock-down on March 9 and the suspension of all nonessential production activities on March 23 (starting in early May, activities restarted and mobility and gathering restrictions were gradually loosened).

Lock-down notwithstanding, based on official records, Italy saw a total of ≈35,200 COVID-19 deaths as of the beginning of August 2020. While other countries (e.g., the U.S. and Brazil) reached much higher death counts, Italy's relative death toll remained rather stark at 58.25 per 100,000 inhabitants. This may be partially attributable to the fact that Italy's population is very old (nationally, the median age is almost 46 years and the percentage of individuals over 65 almost 22%[4]), and that age itself correlates with conditions such as type II diabetes, hypertension and chronic respiratory ailments, which substantially worsen illness and increase the likelihood of death for individuals affected by the virus[5]. But perhaps the most striking aspect of the COVID-19 epidemic in Italy has been its heterogeneity. Some parts of Lombardia and of other regions in the industrialized north were hit early and especially hard, yet other demographically and socio-economically similar areas fared better[6,7]. Moreover, most of the central and southern regions of the country experienced a much milder epidemic—notwithstanding waves of relocations from employment-related domiciles in the north back to family homes in the center and

[1]Dept. of Statistics and Huck Institutes of the Life Sciences, Penn State University, University Park, PA 16802, USA. [2]Institute of Economics and EMbeDS, Sant'Anna School of Advanced Studies, 56127 Pisa, Italy. [3]Dept. of Operations and Decision Systems, Université Laval, Quebec G1V 0A6, Canada. [4]CHU de Québec - Université Laval Research Center, Quebec G1V 4G2, Canada. ✉email: marzia.cremona@fsa.ulaval.ca; fxc11@psu.edu

south around the time of the nationwide lock-down. Potential contributors to this heterogeneity discussed by both scientists and the media include human density characteristics; centralized, hospital-based vs distributed, primary health care systems; and pollution levels[8–12].

A broad and extremely sophisticated literature exists on epidemiological models[13], which many research groups are utilizing both to aid policy through forecasts and to dissect what happened, in Italy and around the world. We did not utilize these models. Instead, we applied a mix of statistical tools from the field of Functional Data Analysis (FDA[14,15]), some well-established, and some recently developed by our group—which are still undergoing peer review and have not yet been validated by the community at large. FDA offers very powerful approaches to analyze data sets composed of curves or surfaces, exploiting information in their shapes. These techniques, which have been successfully applied in a variety of scientific domains[16–18], can effectively complement traditional epidemiological analyses and provide useful insights[19]. We used them to characterize patterns of COVID-19 deaths occurring around the country and analyze their statistical association with two *key predictors*; namely, mobility and positivity (the fraction of performed tests returning positive results). We also considered various socio-demographic, infrastructural and environmental *covariates*. We focused on the period from February 16, 2020, right before the first cases were recorded in Codogno and Vo' Euganeo, to April 30, 2020, right before the first lock-down relaxations (restarting of manufacturing and construction activities at the beginning of May). Based on data availability, we performed our analyses at the spatial resolution of regions, which is suboptimal for several reasons. An epidemic is certainly better studied at a much finer resolution (municipalities, urban areas, perhaps the provinces within which Italian regions are further partitioned)—and so are its links to predictors and covariates whose signals may dilute when aggregated at the regional level. Moreover, operating with 20 observational units (the Italian regions) limits the size of the statistical models one can reliably fit on the data. The techniques we employed allowed us to pinpoint significant trends working with what we could retrieve from public data sources. Unquestionably though, access to data at higher resolution would allow more nuanced, in-depth analyses and likely produce sharper results.

Results

Below we describe the salient outcomes of our analyses. After addressing some shortcomings in publicly available COVID-19 deaths records, we characterize two starkly different epidemic patterns and rank regional mortality curves. Next, we relate mortality to mobility and positivity, and to a number of socio-demographic, infrastructural and environmental factors.

Under-counting deaths. Since February 24, 2020, the Italian Civil Protection Agency (Dipartimento della Protezione Civile; DPC) has released daily counts of recorded COVID-19 deaths at the coarse resolution of regions (only the number of recorded cases were released at the finer resolution of provinces). In Italy and elsewhere, official death records have often been criticized as undercounts[20,21]. Alternative data sources do exist, e.g., daily mortality rates—which can be contrasted to those from prior years to gauge differential mortality. In Italy these are provided by the National Statistical Institute (ISTAT) at the resolution of municipalities. We aggregated the data over municipalities belonging to the same region and subtracted averages over the prior 5 years (2015-19, see Methods)[22]. Figure 1(a) shows smoothed DPC and ISTAT differential mortality curves (per 100,000 inhabitants) for some example regions (Lombardia, Veneto, Emilia Romagna and Campania). The under-counting in the official DPC records was dramatic, especially in badly affected areas and in the initial stages of the epidemic. However, ISTAT differential mortality curves have themselves limitations, especially in less affected areas, where they can fluctuate at small levels and even take negative values—idiosyncratically or reflecting other COVID-19 related phenomena (e.g., increases in mortality due to untreated emergencies or reductions in mortality due to fewer accidents during the lock-down). We therefore formed maxima curves (MAX), where the largest between the DPC and the ISTAT datum is taken in each day and for each region, and then smoothed. These are shown in Fig. 1(b) (DPC and ISTAT smoothed curves for all regions are shown in Figs. S1 and S2). We repeated our analyses on all three data sets; given the small number of observational units at our disposal ($n = 20$ regions), this allowed us to borrow strength replicating results across data sets, with their differences and limitations.

Two different epidemics. Italy saw the unfolding of two very different epidemics; a relatively mild one in the majority of the country, and a tragic, seemingly out of control one in its most hard-hit regions. These two epidemics can be effectively characterized with *probKMA*, an FDA technique designed to identify recurrent motifs within a set of curves, and group the curves based on the motifs they comprise[22,23]. Here, the motifs are the temporal patterns of deaths that characterize alternative epidemic unfoldings, which may in fact start at different times in different curves (regions). Thus, the algorithm also produces the shifts required to align regions comprising the same motif to each other. *ProbKMA* is similar to a K-mean algorithm; it requires the user to specify the number of motifs (K) at the outset, and to select a distance—which can be defined on the curve levels, their derivatives, or a combination of both (see Methods).

The solution with $K = 2$ and distance defined on curve levels depicts two starkly different epidemics, shown for the MAX curves in Fig. 2(a). Allowing for shifts, these are represented by 65-day long motifs. Group 1 undergoes a steep ascent (the "exponential" pattern) followed by a slower descent from the peak; it includes many northern regions. Based on the shifts, Lombardia was first, followed by Emilia Romagna, Marche, Liguria, Piemonte, Trento/Bolzano, and last Valle d'Aosta. Lombardia and Valle d'Aosta presented the most extreme peaks—but Valle d'Aosta's descent was steeper (with a second late ascent likely due to data recording imprecisions; Valle d'Aosta is a very small region with only $\approx 125,000$ inhabitants). Group 2 follows a "flat(tened)" pattern; it includes all regions in southern and central Italy and, remarkably, Veneto—where the curve was successfully

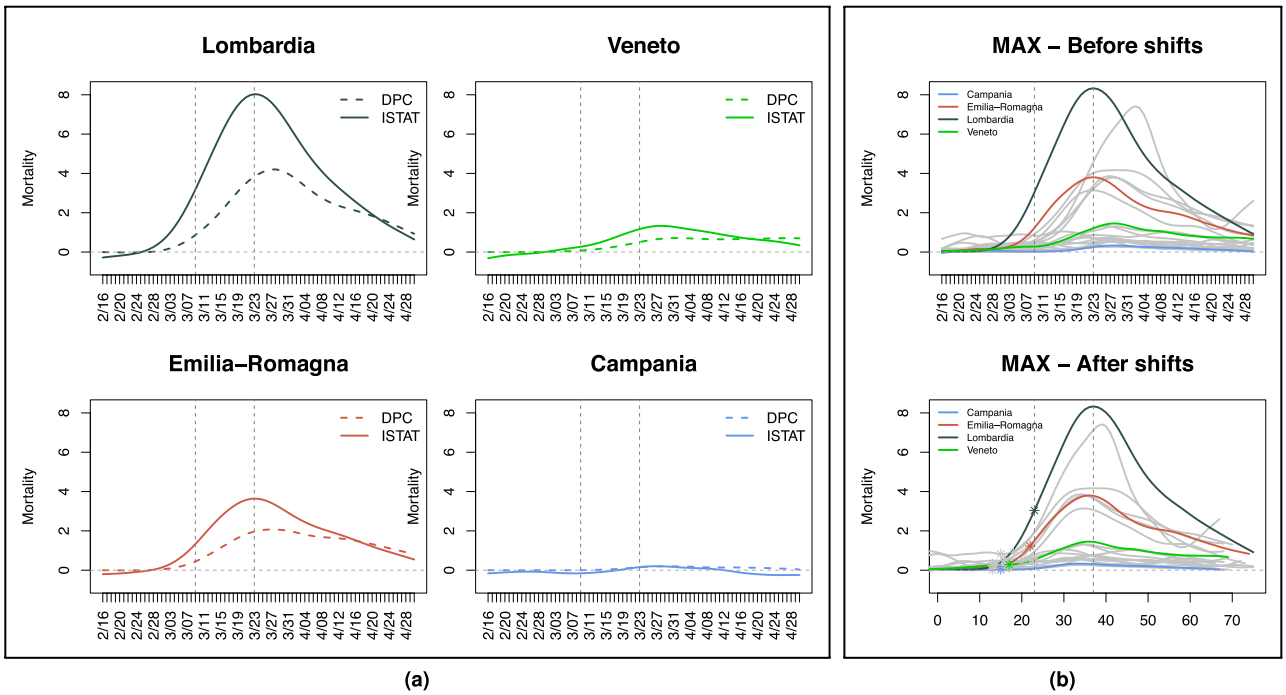


Figure 1. Mortality curves. (**a**) DPC (dashed) and ISTAT (solid) differential mortality curves (per 100,000 inhabitants) in four example regions; Lombardia, Veneto, Emilia Romagna and Campania. Curves are smoothed with splines, with degree of smoothing selected by generalized cross-validation (see Methods). ISTAT curves "take off" earlier and in some regions are as much as twice as high at their peak—possibly due to many COVID-19 deaths happening at home and/or not being recorded as such in hospitals, especially in the early stages of the epidemic. (**b**) MAX mortality curves (per 100,000 inhabitants) in the 20 Italian regions, before (top) and after (bottom) the shifts produced by *probKMA* run with $K = 2$. In the bottom panel, time is marked as a day number (as opposed to a date); this represents the region-specific time of the epidemic unfolding, and corresponds to actual time (starting on February 16 and ending on April 30) only for regions with no shifts, e.g., Lombardia). Curves are again smoothed with splines, with degree of smoothing selected by generalized cross-validation. Lombardia, Veneto, Emilia Romagna and Campania, also shown in (**a**), are highlighted in color. In all panels, vertical lines mark the dates of the national lock-down (March 9) and of the suspension of all nonessential production activities (March 23). In the bottom panel of (**b**) vertical lines still show these dates without shifts; stars on the curves mark the lock-down after the region specific shifts.

curbed. The shifts produced for this group are less stable and less meaningful in terms of interpretation, as flatter profiles leave more leeway in aligning curves against each other. All results (except for the shifts in Group 2) are rather consistent when using DPC and ISTAT curves (see Fig. S3a and Fig. S3c), and when using distances defined on derivatives instead of curve levels. The solution with $K = 3$ places Lombardia (ISTAT curves) or Lombardia and Valle d'Aosta (MAX and DPC curves) in a cluster of their own (see Fig. S4). We also validated our results using a modification of *funBI*[24], a functional biclustering technique, and *IWTomics*[25], a functional testing technique which contrasts two sets of aligned curves pinpointing the locations and scales at which they differ (see Methods). Figure 2(b) shows how, starting a little over two weeks from the beginning of their motif (wherever that was in each curve), Group 1 and Group 2 differ significantly at all temporal scales (see also Fig. S3b and Fig. S3d, and Table S1).

Why the two epidemics? The pattern of deaths characterizing Group 1 may be due, in large part, to the fact that the virus had circulated silently in the north of Italy for a long period of time before any kind of behavioral changes by the general public, medical protocols, or mitigation policies by local and central authorities were put in place. Mounting evidence suggests that a large share of COVID-19 cases are asymptomatic and yet contagious[3,26]; their numbers may have increased until a pent-up reservoir of virus found its way to vulnerable individuals (some researchers also hypothesize Antibody-Dependent-Enhancement of SARS-CoV-2[27], and thus a role for re-infections). But a variety of additional factors may have contributed to shaping the two epidemics; we explore some below.

Ranking mortality curves. Non-parametric FDA methods can be used to rank curves based on the notion of depth—from the innermost to the most extreme, and to identify outliers[28,29]. Figure 3 shows a functional box plot of the MAX mortality curves and a depth ranking of the curves in the DPC, ISTAT and MAX data sets—shifted based on *probKMA* run with $K = 2$ and restricted to their aligned 65-day portions. The ranking is directional; we attributed signs to the depth measurements, so that curves far over or under the median curve are at the top or bottom of the ranking, respectively (see Methods). The top portion of the ranking comprises regions with "exponential" epidemics (Group 1) and is rather stable across data sets; Lombardia and Valle d'Aosta are

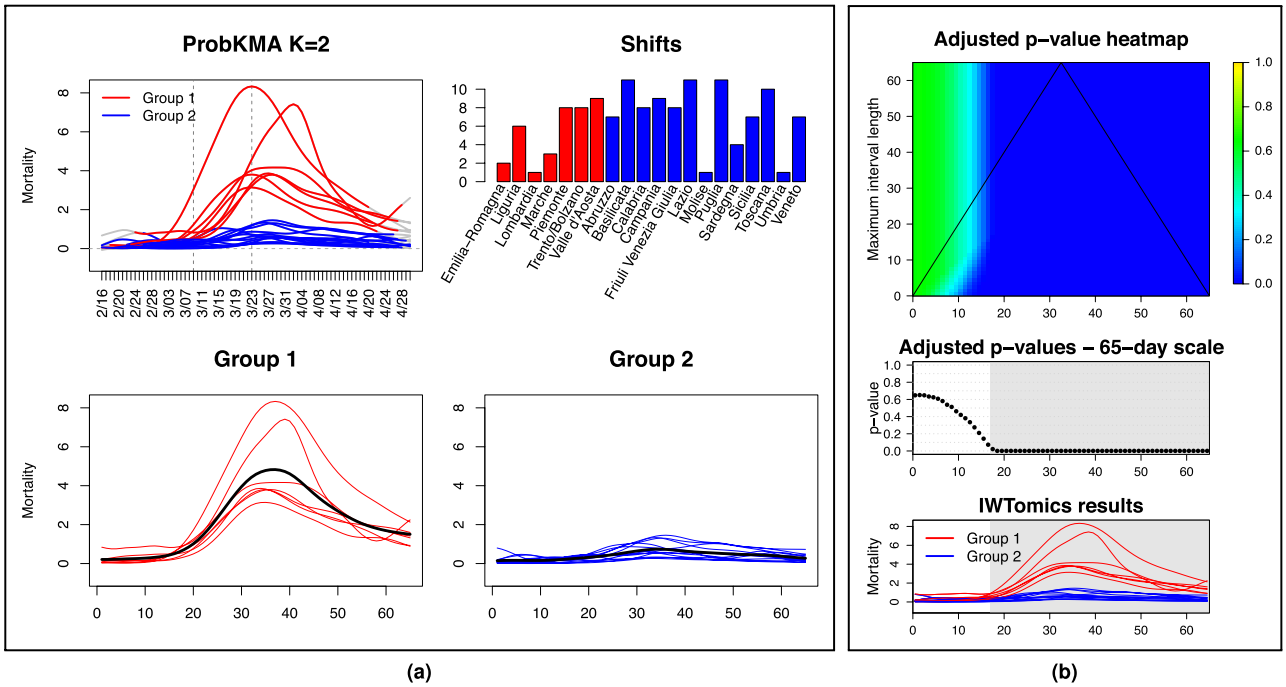


Figure 2. Characterizing two epidemics. (**a**) MAX mortality curves are shown in the top left panel with 65-day portions identified by *probKMA* with $K = 2$ in red (Group 1; "exponential" pattern) and blue (Group 2; "flat(tened)" pattern). The curve portions are shown again, this time aligned with each other and separated by group, in the bottom panels. Black lines indicate group averages. The shifts produced by *probKMA* are shown in the top right panel (motifs, groups and shifts for Group 1 are stable across data sets; shifts for Group 2 are less stable and less interpretable—see Fig. S3). (**b**) Shifted Group 1 and Group 2 MAX mortality curves are tested against each other with *IWTomics*. The heatmap at the top shows *p*-values adjusted at all possible scales (from 1 to 65 days). The middle panel shows in detail the top-most row of the heatmap; i.e. the *p*-values adjusted across the whole 65-day interval. The bottom panel shows again the shifted curves. Gray areas in the middle and bottom panels mark days when the difference between the two groups has an adjusted *p*-value $\leq 5\%$ (see Table S1). Starting a little over two weeks from the beginning of their epidemic, curves in the two groups differ at all temporal scales with adjusted *p*-values$\leq 5\%$.

consistently among the most extreme curves (they are also identified as outliers in the MAX and DPC data sets). The mid- and bottom portions of the ranking comprise regions with "flat(tened)" epidemics (Group 2) and are less stable across data sets, as the flatter profiles can more easily switch in their depth ranks. However, Toscana (which is the median in the MAX and ISTAT data sets) and Veneto are consistently among the deepest, most central curves. This analysis highlights again the tragic epidemic unfolding in Lombardia, and, by contrast, confirms how Veneto managed to "flatten" its curve back into the bulk.

Local mobility and positivity as statistical predictors of mortality. Next, we focus on two key variables. The first is one of the most discussed policy-actionable variables, mobility, which has been curtailed to various degrees through lock-down measures in most of the countries affected by COVID-19. The second is one of the most discussed sentinel indicators, positivity, i.e. the fraction of performed tests returning positive results. For both these variables daily values for the period February 16–April 30, 2020, were obtained from data in the public domain at regional resolution.

We considered differential mobility curves provided by Google for the category "Grocery & pharmacy". These express the fractional reduction with respect to January 2020 levels, and refer to mobility linked to first necessities—such as buying food, medicine, etc. For Italy, they were provided at the resolution of regions. Even though individuals were allowed to leave their homes for these necessities also during the most restrictive phase of the lock-down, the reduction captured by Google's "Grocery & pharmacy" was substantial. Mobility in weekdays fell by roughly 0.30, i.e. 30%, in the week after the lock-down (March 9), and further decreased in following weeks—reaching the lowest levels (between approximately -0.60 and -0.40 depending on the region) in the week after the suspension of nonessential production activities (March 23). It then slowly increased, getting back in a range between approximately -0.40 and -0.20 at the very end of April (see Fig. S5). In Lombardia, the peak MAX mortality was between March 20 and 25—i.e., roughly, simultaneous to the lowest mobility and two weeks after its first substantial drop. Notably, in most Italian regions mobility during lock-down weekends reached -1.00, i.e. -100%. For comparison, in the state of New York, which had among the strongest restriction measures in the U.S., Google's "Grocery & pharmacy" never fell below -0.40. We refer to Google's "Grocery & pharmacy" curves as *local* mobility because they measure how much individuals move around where they live, as opposed to how much individuals move from place to place—e.g., to go from Wuhan to Milan, or from Milan to Palermo,

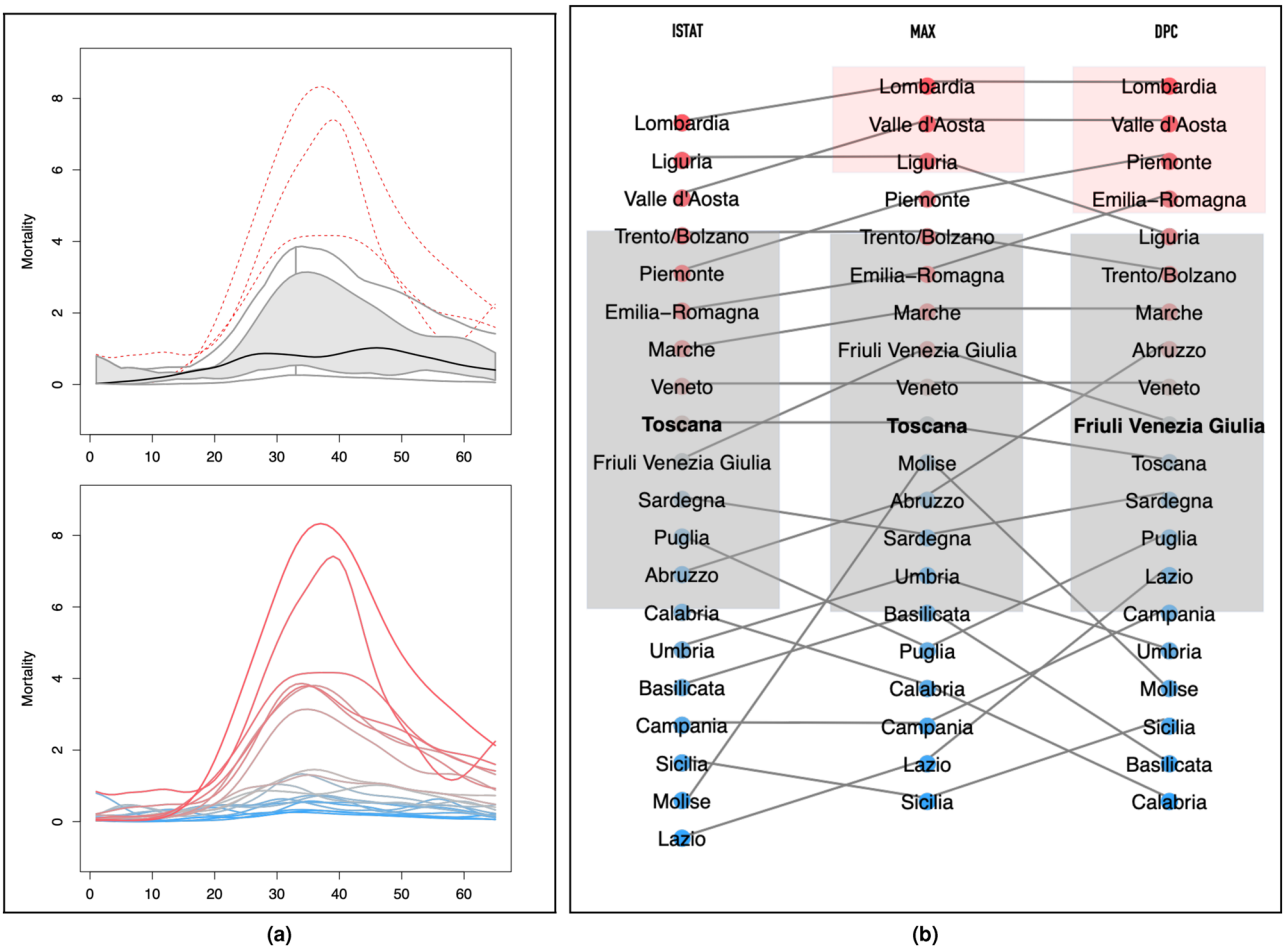


Figure 3. Functional boxplot and ranking. (**a**) Functional boxplot of the MAX data set (top) and MAX mortality curves (bottom) color-coded according to their ranking, as shown in the MAX column of (**b**). In the boxplot, Toscana is the median (black continuous line); Lombardia, Valle d'Aosta and Liguria are identified as outliers (red dashed lines); and the 50% innermost "box" (grey area) include the curves for Trento/Bolzano, Emilia-Romagna, Marche, Friuli Venezia Giulia, Veneto, Toscana, Molise, Abruzzo, Sardegna, Umbria, and Basilicata. Note that the "box" is skewed upwardly. (**b**) Rankings of the ISTAT (left), MAX (center) and DPC (right) mortality curves. The median regions are in bold, gray rectangles mark the 50% innermost boxes, and pale red rectangles mark outliers (no region is labeled as an outlier in the ISTAT data set; see Methods). The dots representing each region are color-coded (from intense red, through gray, to intense blue) according to their signed depth values (see Methods). In all three data sets, Lombardia's curve is the most extreme at the very top of the ranking and, in contrast, Veneto's curve is deep in the bulk close to the median (Toscana for ISTAT and MAX, Friuli Venezia Giulia for DPC). Segments joining the regions across the three rankings show how the top portion remains rather stable, while the mid- and bottom portions contain several crossings. Regions at the top are those characterized by "exponential" epidemics (Group 1), while regions in the middle and at the bottom are those with "flat(tened)" epidemics (Group2), whose curves can more easily switch in their depth ranks.

or New York City. Obviously both types of mobility are relevant for the spread of a virus, and definitions depend on scale/resolution, but the first one is the one we analyzed.

To construct positivity curves, we combined daily public records on number of tests performed and number of new cases, which are also provided by the Italian Civil Protection Agency. Taking daily ratios of new cases on tests performed is clearly imperfect, because of (variable and unreported) delays in test results. But regularizing and smoothing these ratios (see Methods) produced a reasonable proxy. Smoothed positivity surpassed 0.1, i.e. 10%, as early as February 20 in some hard hit regions, peaked in a staggered fashion throughout March, and fell below 0.10 for all regions by around April 22 (see Fig. S5). Lombardia surpassed 0.10 around February 22 and peaked around March 15-18; that is, roughly, about a month and about a week prior to the peak of MAX mortality, respectively. Though we cannot draw exact parallels (our positivity curves are approximate and smoothed), this is consistent with what was observed, e.g., in New York City—where positivity was above 0.10 approximately from March 6-7 to May 12-13 and peaked at about 0.70 around March 28, with deaths peaking between April 5 and 13.

To anchor local mobility and positivity curves to the epidemic unfolding in each region, we shifted them congruently with the mortality curves. Figure 4(a) displays shifted curves based on *probKMA* run on MAX data

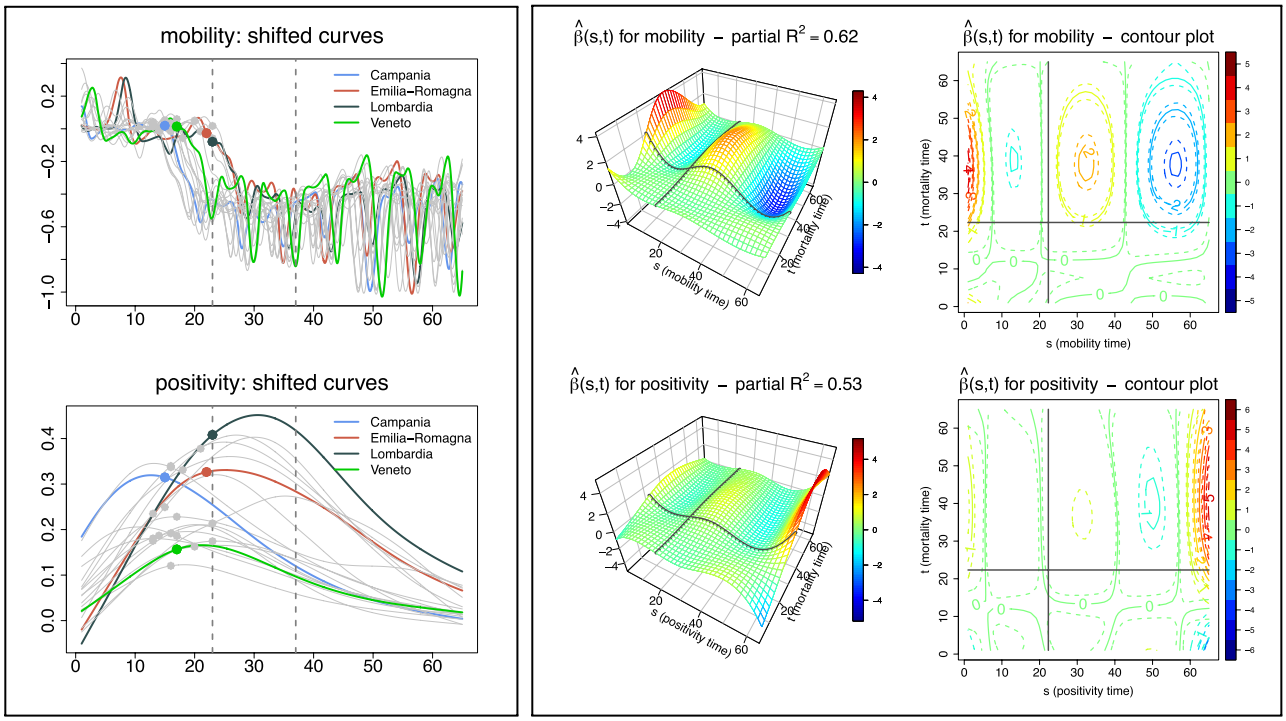


Figure 4. Associating mortality to local mobility and positivity. (**a**) Local mobility curves (Google's "Groceries & pharmacy") and positivity curves (regularized ratios of new cases to number of tests performed) in the 20 Italian regions. Curves are smoothed with splines, with degree of smoothing selected by generalized cross-validation, and shifted based on *probKMA* run on the MAX mortality curves with $K = 2$; time is marked as a day number representing the region-specific time of the epidemic unfolding, and corresponds to actual time (starting on February 16 and ending on April 30) only for regions with no shifts, e.g., Lombardia. Vertical lines show the days corresponding to the nationwide lock-down (March 9) and the suspension of all nonessential production activities (March 23) without shifts, stars on the curves mark the lock-down after the region specific shifts. The example regions of Fig. 1(a) are highlighted in color. (**b**) Estimated effect surfaces from the joint function-on-function regression of MAX mortality on local mobility and positivity shown in 3D and as contour plots (March 9, without shift, is again marked on both). Early and mid-period local mobility levels are strong positive predictors of mortality at its peak. Positivity has similar but much weaker predictive signals, likely because the effects are subsumed by mobility. Late local mobility has a negative association with mortality at its peak (mobility resumed faster in regions with milder epidemics), and late positivity a strong positive one (positivity remained elevated in regions with worse epidemics). The regression captures a large share of the variability in mortality curves (in-sample $R^2 = 0.90$, LOO-CV $R^2 = 0.52$), with substantial and comparable contributions of the two predictors (partial R^2s $= 0.62, 0.53$).

with $K = 2$ (Fig. S6 displays shifted curves based on *probKMA* run on DPC and ISTAT data). The horizontal axis now indicates again days in the region-specific epidemic unfolding, restricted to the 65-day portions where mortality curves align forming the two *probKMA* motifs.

We then used function-on-function regressions[14,30] to model the statistical dependence of mortality on local mobility and positivity; in symbols, we fit the joint model $y(t) = \alpha(t) + \int \beta_{mob}(s,t)x_{mob}(s)ds + \int \beta_{pos}(s,t)x_{pos}(s)ds + \varepsilon(t)$, where $y(t)$ is the response curve, i.e. mortality, $\alpha(t)$ is the intercept, $\varepsilon(t)$ is the model error, and $x_{mob}(s)$ and $x_{pos}(s)$ are the predictor curves —mobility and positivity, respectively. These predictors are integrated over time, with "effects" represented by *surfaces*; $\beta_{mob}(t,s)$ is the association of mortality at time t with local mobility at time s, and similarly $\beta_{pos}(t,s)$ for positivity (see Methods).

Figure 4(b) shows the effect surfaces for local mobility and positivity estimated using the MAX curves as response. $\hat{\beta}_{mob}(t,s)$ suggests that local mobility levels early on and mid-way through the epidemic (e.g., around the March 9 lock-down date for Lombardia) are strong positive predictors of mortality at its peak, with the early predictive signal stronger than the mid-way one. In contrast, the local mobility level late in the epidemic has a negative association with mortality at its peak, likely reflecting a faster resumption of mobility in regions with milder epidemics. $\hat{\beta}_{pos}(t,s)$ suggests that positivity levels early on and mid-way through the epidemic are also positive predictors of mortality at its peak—though the predictive signals are substantially weaker than those of mobility, likely because they are confounded with the latter. However, the positivity level late in the epidemic has a marked positive association with mortality at its peak. Here the signal is "detangled" from that of mobility, and one finds a sort of retrospective signature; regions which fared worse still had hightened positivity in the late stages of their epidemics. The data at our disposal does not allow an accurate evaluation of the lags that might occur between mobility, positivity and mortality. However, we performed some additional analyses to

Covariate	Description [comment in legend]	Year and Source
% Over 65	Aging of the population [1]	2018, ISTAT
% Diabetics	Prevalence of relevant pre-existing conditions [2]	2018, ISTAT
% Allergic	Another potentially relevant pre-existing condition	2018, ISTAT
Adults per family doctor	Quality of distributed, primary health care	2017, Ministry of Health
ICU beds per 100K inhabitants	Quality of centralized, hospital-based health care [3]	2018, Ministry of Health
Ave. beds per hospital (whole)	Ability of hospitals to act as contagion hubs	2018, Ministry of Health
Ave. beds per nursing home (ward)	Ability of nursing homes to act as contagion hubs	2018, Ministry of Health
Ave. students per classroom	Ability of schools to act as contagion hubs	2018, Ministry of Education
Ave. employees per firm	Ability of work places to act as contagion hubs	2017, ISTAT
Ave. members per household	Ability of households to act as contagion hubs [4]	2017, ASR Lombardia
Public transport rides per capita	Ability of public transport to act as contagion hub	2017, ISTAT
PM10	Pollution levels (particulates)	2018, ISTAT

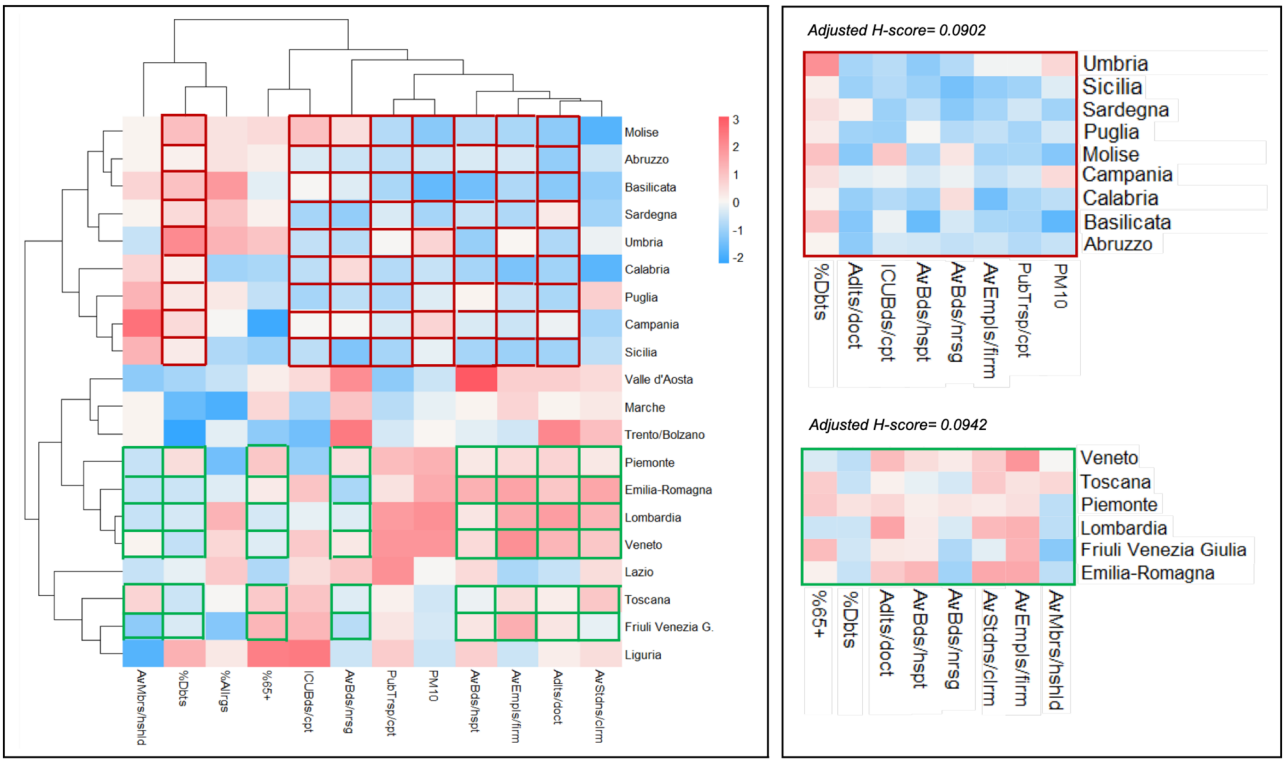
Table 1. Scalar covariates potentially affecting COVID-19 mortality. [1] The percentages of over 65, 70, 75 and 80 are highly correlated at the resolution of regions; we took over 65 as representative. [2] The prevalence of diabetes, hypertension and chronic bronchitis are highly correlated at the resolution of regions; we took diabetes as representative (allergies are not as highly collinear and were retained as a separate covariate). [3] Availability of ICU beds is also directly relevant for withstanding the impact of COVID-19 surges. [4] Average members per household is *not* a direct proxy of inter-generational contacts, but it may capture some of its effects.

investigate this. We further denoised the curves projecting them on their first functional principal components[14], and measured the distances between the peaks of such projections. On average, there were $\approx$ 20 days between the peak of mobility and the peak of positivity, and $\approx$ 10 days between the peak of positivity and the peak of mortality (see Fig. S7). Back to the estimated effect surfaces, we found them to be remarkably similar across the three data sets (MAX, DPC and ISTAT). The joint models all have in-sample R^2s above 90% and leave-one-out cross-validated (LOO-CV) R^2s above 50% (see Table S2), with strong and comparable contributions of local mobility and positivity (e.g., for the MAX curves, the partial R^2s are 62% and 53%, respectively). Also, while this is not the case for all regions, residuals are rather consistent across data sets for Veneto, whose mortality is well predicted, and for Lombardia, whose mortality is always and sizably underestimated (see Fig. S8a and Fig. S9).

In order to further assess the roles of local mobility and positivity, we also considered *marginal* function-on-function regressions for mortality on each, separately; in symbols, $y(t) = \alpha(t) + \int \beta_{mob}(s,t)x_{mob}(s)ds + \varepsilon(t)$ and $y(t) = \alpha(t) + \int \beta_{pos}(s,t)x_{pos}(s)ds + \varepsilon(t)$. Effect surface estimates for local mobility are very similar to those in the joint models for all three data sets (see Fig. S10). Those for positivity confirm a strong association with mortality at its peak, but are less defined in terms of time profile (see Fig. S11). In summary, we find substantial evidence that local mobility and positivity are associated with COVID-19 mortality, and can predict it with some lag-time. Though the data at our disposal does not allow us to pinpoint lag lengths with accuracy, our analysis does support their roles as policy-actionable and monitoring variables, respectively. We also find that, even when considered jointly, these variables are not enough to fully account for the massive numbers of COVID-19 deaths recorded in Lombardia, the worst hit region in the country.

The role of socio-demographic, infrastructural and environmental factors. We considered 68 scalar (non-longitudinal) covariates retrieved from public sources, proxying for socio-demographic, infrastructural and environmental factors debated by scientists and policy-makers during the epidemic (see Table S3). Many of these are suboptimal proxies; they refer to the closest times we could find data for (in some cases 2016 or earlier) and are, too, at the coarse resolution of regions. We performed an initial screen among these covariates to guarantee reasonable data quality (eliminating older and less complete data sets), facilitate interpretations and control collinearity. Fig. S12 shows a histogram of the pair-wise correlations, about a quarter of which exceeds 0.5 in absolute value, and Fig. S13 shows a dendrogram where the covariates agglomerate in distinct groups. We thus selected 12 covariates which were relatively recent (2017 or 2018) and well spread across the dendrogram groups. These capture aging of the population; prevalence of pre-existing conditions believed to affect disease severity; quality of distributed primary health care vs. centralized hospital-based health care; the potential of hospitals and nursing homes, but also schools, workplaces, households and public transport to act as contagion hubs; and pollution levels (see Table 1; Fig. S14 provides marginal densities, pair-wise scatter plots and correlations for the 12 selected covariates).

Even this restricted set of 12 covariates presents a distinct interdependence structure (see covariates dendrogram in Fig. 5(a) and Variance Inflation Factors in Table S4). For instance, our contagion hubs proxies for hospitals, schools and work places, and our (inverse) proxy for quality of distributed, primary health care (number of adults per family doctor), tend to vary closely together across regions. Also, our contagion hub proxy for public transport and pollution levels tend to vary together (this is not counter-intuitive, as both increase in more industrialized regions with large metropolitan areas), as do the percentages of individuals affected by diabetes and allergies, and our proxy for quality of centralized, hospital-based health care (ICU beds per 100,000 inhabitants) and the percentage of individuals over 65.



(a) (b)

Figure 5. Interdependencies among scalar covariates and regions. (**a**) Heatmap of the 20 (regions) × 12 (covariates) data matrix, with dendrograms from separate hierarchical clustering (correlation distance, complete linkage) of the regions (left) and the covariates (top). Color coding within cells represents values of the standardized covariates (centered and scaled to mean 0 and standard deviation 1). Color coding of some cell borders identifies the biclusters in (**b**). The dendrograms capture a distinct interdependence structure. For instance, there are marked similarities among Lombardia, Veneto, Emilia Romagna and Piemonte, as well as among some groups of southern regions (Sicilia, Campania, Puglia and Calabria; Basilicata, Abruzzo and Molise). There are also marked associations among groups of covariates. The contagion hubs proxies for hospitals, schools and work places, and number of adults per family doctor, vary closely together. So do the contagion hub proxy for public transport and pollution levels; the percentages of individuals affected by diabetes and allergies; and ICU beds and the percentage of individuals over 65. (**b**) Restricted heat-maps further illustrating interdependencies through two biclusters of regions and covariates. Color-coding within cells corresponds to that in (**a**), and each bicluster is identified by a border color and its adjusted H-score (an inverse measure of bicluster strength; see Methods). The first bicluster (adjusted H-score = 0.0902) comprises central and southern regions with "flat(tened)" epidemics (Group 2). The second bicluster (adjusted H-score = 0.0942) comprises northern regions with "exponential" epidemics (Group 1) but also northern and central regions from Group 2.

Conversely, some regions show similar profiles across covariates (see regions dendrogram in Fig. 5(a)). For instance, Lombardia, Veneto, Emilia Romagna and Piemonte have strong similarities, as do groups of southern regions (e.g., Sicilia, Campania, Puglia and Calabria; Basilicata, Abruzzo and Molise). An interesting characterization is produced using the *Cheng and Church's biclustering algorithm*[31], which we implement with an adjusted mean squared residue, or H-score[32]. A bicluster is a subset of regions which exhibit similar behavior across a subset of covariates. Figure 5(b) shows two biclusters with similar adjusted H-score values, obtained through the same run of the algorithm. The first bicluster comprises central and southern regions, all with "flat(tened)" epidemics (Group 2). Its regions have low ratios of adults to family doctors, limited concentrations in hospitals, nursing homes, work places and public transport, and low pollution levels. They also have high percentages of diabetic individuals and limited availability of ICU beds. The second bicluster comprises northern regions with "exponential" epidemics (Group 1), such as Lombardia, Emilia-Romagna and Piemonte, but also northern and central regions with "flat(tened)" epidemics (Group 2), such as Veneto, Friuli Venezia Giulia and Toscana. Its regions have high ratios of adults to family doctors, high concentrations in hospitals, work places and classrooms, and tend to have large percentages of individuals over 65. They also have low percentages of diabetic individuals and medium or small-sized households.

Next, we used functional regressions with a two-fold aim: pursue a more direct, systematic assessment of the associations between the scalar covariates and COVID-19 mortality; and use the scalar covariates as controls in models comprising mobility and positivity to re-assess these key predictors. We stress again that the coarse resolution of the data poses serious limitations for these analyses, because it may dilute some predictive signals

and because it bounds us to a small sample size. With only $n = 20$ observational units (the regions), fitting functional regression models comprising many terms (e.g., several scalar covariates and possibly their interactions; mobility and positivity curves along with more than one scalar covariate) produces unstable, overfit outcomes. Thus, we evaluate only the marginal effects of the scalar predictors, and the effects of mobility and positivity with one scalar control at a time. The marginal function-on-scalar regressions of mortality curves on each of the 12 covariates have in-sample R^2s ranging between $\approx$ 20 and 65%. Here the "effects" are *curves*; $\beta_x(t)$ represents the association of mortality at time t with the covariate x. For 8 of the covariates the $\hat{\beta}_x(t)$'s show the expected signs throughout the peak period of the epidemic. In particular, the (inverse) proxy for quality of distributed, primary health care is the strongest marginal predictor; adults per family doctor shows a very large positive association with mortality. Also hospital, school and work place contagion hub proxies show strong positive associations with mortality. Nursing homes and public transport contagion hub proxies, pollution and the percentage of individuals over 65 are positive but comparatively weaker marginal predictors. For 4 of the covariates the $\hat{\beta}_x(t)$'s show unexpected signs. The percentages of diabetics and allergic individuals show negative associations with mortality, likely due to the fact that their prevalence is high(er) in areas which were spared the brunt of the epidemic. In fact, estimated effect curves become positive when a differential intercept is included in the model to account for different overall mortality levels in Group 1 and Group 2 regions (see Fig. S15). Also the average number of members per household shows a negative association with mortality. Its small range of variation across regions ($\approx$ 2.0−2.8, mean 2.3, s.d. 0.16) may not allow it to properly proxy the effect of household contagions. At the same time, a strong negative correlation with the percentage of individuals over 65 may not allow it to properly proxy inter-generational contacts; regions with more elderly people are in fact those with smaller households. The negative association of average number of members per household with mortality, which persists even when including a differential intercept for Group 1 and Group 2 in the model (see Fig. S15), may simply be a "shadow" of its negative correlation with the percentage of individuals over 65. Finally, ICU beds per 100,000 inhabitants shows a positive association with mortality which, too, persists when including a differential intercept for Group 1 and Group 2 in the model (see Fig. S15), and may be in part a "shadow" of positive correlations with percentage of individuals over 65 and average number of beds per hospital. However, this proxy for quality of centralized, hospital-based health care, so prominent to the public debate during the epidemic, is *not* a negative predictor of mortality in our analysis.

In conclusion, better proxies and finer resolution may reveal stronger aggravating roles for age, nursing homes, public transport and pollution[8,9] and better dissect the roles of chronic conditions, households and inter-generational contacts, and ICU availability[33,34]. But our analysis, notwithstanding limitations in the data, suggests important roles of primary care in mitigating mortality, and of contacts in hospitals, schools and work places in aggravating it.

The results of our marginal function-on-scalar regressions, which are summarized in Fig. 6(a) for MAX mortality curves, are also consistent across data sets (see Fig. S15)—which lends them support, at least at the resolution of regions. To further validate their stability we ran a functional generalization of *SsNAL-EN*[35]—an Elastic Net-like algorithm that performs feature selection for regressions with many predictors, producing reasonably stable outcomes even with small sample sizes and collinear features. Reassuringly, the output of *SsNAL-EN* is consistent with the marginal analysis, and again consistent across data sets (see Table S5): the top feature is always adults per family doctor, and the top 5 always include, in addition to it, average beds per hospital, average students per classroom, average employees per firm, and average members per household.

Finally, we ran again the function-on-function regression of mortality on local mobility and positivity, and re-evaluated the effects of these predictors introducing in the model one of the top 5 scalar covariates at a time (see Fig. S16 for results on DPC, ISTAT and MAX data), as well as their first principal component, which explains $\approx$ 68% of their variability and can act as a "summary" control (see Fig. 6(b) for MAX curves and Fig. S17 for DPC and ISTAT curves). Remarkably, while the control covariate "subsumes" some of the predictive power in each model, the estimated effect surfaces of local mobility and positivity retain the same shapes, and they remain very strong and comparable contributors (e.g., for MAX curves in Fig. 6(b), the overall in-sample R^2 reaches 94%, the LOO-CV R^2 is 70%, and the partial R^2s are 66%, 61% and 39%, respectively, for local mobility, positivity and the first principal component; see also Table S2). Thus, with all the limitations of the data at our disposal, controlling for relevant covariates does not modify how the epidemic unfolding is associated to local mobility and positivity over time. Introducing socio-demographic, infrastructural and environmental factors in the modeling also does not change what we observed concerning residuals: mortality in Veneto is well predicted, and mortality in Lombardia remains sizably underestimated (see Fig. S8b) for MAX and Fig. S17 for DPC and ISTAT).

Discussion

Notwithstanding the limitations of the data employed in this study, using FDA techniques we were able to characterize heterogeneous and staggered epidemics in different areas of Italy—recapitulating and quantitating what scientists, policy makers and the public saw unfolding during the months of February, March and April 2020. In addition, we were able to document strong associations of COVID-19 mortality with local mobility and positivity, which persist in models that control for other relevant covariates. Investigating local mobility and positivity as, respectively, an actionable effector and a sentinel indicator of epidemic strength and progression, possibly to be used to adapt mitigation and containment efforts in real time, will require more and better data. In particular, accurate data on cases and hospitalizations in addition to deaths, and at a resolution much finer than that of Italian regions. Such data would allow a more systematic evaluation of the lags between the temporal patterns of mobility, contagions, illnesses and casualties—an important avenue for future studies, which could again utilize FDA tools (e.g., registration and dimension reduction techniques[36]). Such data would also be critical to better capture predictive signals in a number of covariates—which may weaken and/or become confounded

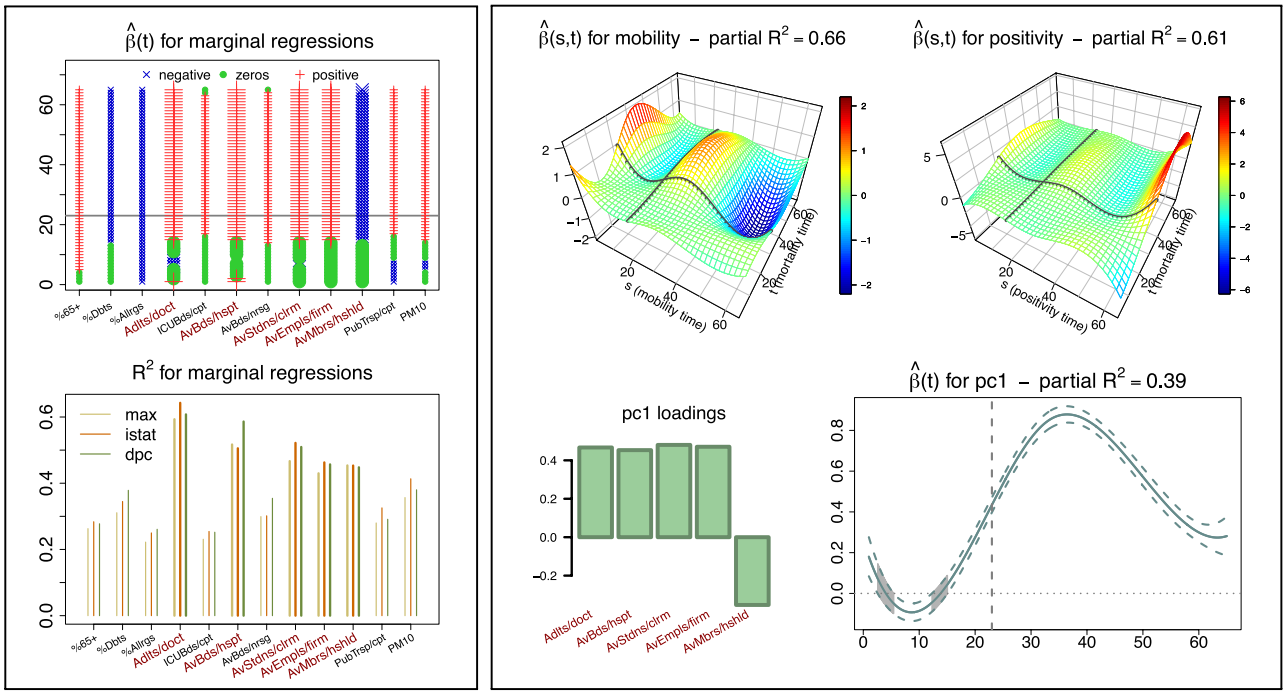


Figure 6. Associating mortality to socio-demographic, infrastructural and environmental factors. **(a)** Results from marginal function-on-scalar regressions. Mortality curves are regressed against each of the scalar covariates in Table 1. The top plot displays the signs of the effect curves estimated on the MAX data. Time, marked as the 65 days of the region specific epidemic unfoldings, is on the vertical axis (the nationwide lock down on March 9, without shift, is marked by a horizontal line. Red, blue and green indicate, respectively, positive, negative and non-significant portions (i.e., where 95% confidence bands around the estimated effect curve are entirely above, entirely below, or contain 0; see Methods). The bottom plot displays in-sample R^2s for the regressions fitted on MAX, ISTAT and DPC data; these are remarkably consistent. The names in red on the horizontal axes indicate the top 5 covariates selected by SnNAL-EN on all three data sets (see Methods); these are also the ones with the largest R^2s. (**b**) Results from the joint function-on-function regression of MAX mortality on local mobility, positivity, and the first principal component (pc1) of the top 5 covariates, used as a "summary" control. This control does not modify the shapes of the estimated effect surfaces for mobility and positivity (shown on top)—which are very similar to the ones in Fig. 4(b). The estimated effect curve for pc1 shows a positive and significant association with mortality at its peak (bottom right; 95% confidence band in dashes, gray corresponds to non significant portions, vertical dashed line corresponds to March 9, without shift). The sign of this effect is consistent with marginal findings, based on the loadings of the first principal component (bottom left; positive for adults per family doctor, average beds per hospital, average students per classroom and average employees per firm, and negative for average members per household). With the addition of pc1, the regression reaches an in-sample $R^2 = 0.94$ and a LOO-CV $R^2 = 0.7$. The contributions of local mobility and positivity remain high (partial $R^2 = 0.66$ and 0.61, respectively). That of our "summary" covariate is also substantial (partial $R^2 = 0.39$).

when aggregating data over broad, internally heterogeneous areas. Clearly, the limited data at our disposal for this study prevent us from drawing causal implications with confidence, but our results, along with those of other recent studies[37,38], do support a role for mobility as a key modulator of COVID-19 spread and for positivity as a monitoring variable. Moreover, they support a role for distributed, primary health care in mitigating mortality, and for hospitals, schools and work places as contagion hubs that may aggravate the epidemic. If confirmed and fine-tuned on higher resolution data, also these findings could inform decision making—e.g., on short- and medium-term investments to boost distributed health care, or "pod" patients, students or employees. Finally, an extension of the temporal span of the data would also be of great interest to properly characterize different phases of the Italian epidemic—including its evolution after the gradual weakening of lock-down measures in May 2020. We believe that our work demonstrates the potential of FDA techniques for analyzing epidemiological data and we note that, while some of the techniques we used are well-established, others are very recent. These novel tools appear to offer original and useful insights, but it is important to point out their limited validation to date. Of course our pipelines and the mix of FDA tools used in this study could be applied to COVID-19 data from other parts of the world.

Methods

Data retrieval and pre-processing. *Functional variables.* Daily cumulative COVID-19 death counts per region were retrieved from the Italian Civil Protection Agency (Dipartimento della Protezione Civile; DPC[39]). *DPC mortality curves* from February 24 to April 30, 2020, were computed for each region as the daily increments in COVID-19 death counts, divided by the population of the region as of January 1, 2019 (as recorded by ISTAT[40]). DPC mortality curves were set to zero for the period February 16-23, 2020, before the Civil Protection Agency started releasing data. Daily death counts from all causes in 7270 Italian municipalities (about 93.5% of the Italian population) for the years 2015-20 were downloaded from the Italian National Institute of Statistics (ISTAT[41]) on June 4, 2020. Data were aggregated by region, and *ISTAT differential mortality curves* from February 16 to April 30, 2020, were computed for each region as the daily difference between 2020 deaths and the average daily deaths in 2015-19, divided by the total population of the municipalities included in the death counts as of January 1, 2019[42]. *MAX mortality curves* were created taking, for each region and each day, the maximum between DPC mortality and ISTAT differential mortality. Daily measurements concerning "Grocery & pharmacy" mobility from February 16 to April 30, 2020, were downloaded for each region from the Google Mobility Report[43] (*local mobility curves*). These measurements express percent changes with respect to the corresponding daily mobility levels in the first five weeks of 2020 (January 3 to February 6). *Positivity curves* were constructed using raw data from the Italian Civil Protection Agency[39]. For each day from February 24 to April 30, 2020, and each region, we took the ratio between the number of new positive cases and the number of new tests performed. The ratios were truncated at 0 and 1 to account for irregularities in the row data (e.g., positive cases $= -1$, or positive cases exceeding tests performed, presumably due to delays in test results). Like DPC mortality, positivity curves were set to zero for the period before the Civil Protection Agency started releasing data (February 16-23, 2020). We point out that positivity curves must be used with caution due the fact that positive cases have been tallied with different rules and approaches in different areas and at different times[44]. For all functional data sets, the two self-governing provinces of Trento and Bolzano were considered together as the Trento/Bolzano region, since not all data were available for both provinces separately. The 20 curves in each functional data set were smoothed using cubic *smoothing B-splines* with knots at each day and roughness penalty on the curve second derivative[14]. For each functional data set the smoothing parameter was selected minimizing the average generalized cross-validation error (GCV[45]) across the 20 curves. All computations were performed using the statistical software `R`[46], and specifically the `R package fda`[47].

Scalar covariates. We considered a large number of scalar covariates of potential interest (see Table S3), and focused on the 12 listed in Table 1 and below. In retrieving and computing various measurements, as was done for the functional variables, the provinces of Trento and Bolzano were aggregated into the Trento/Bolzano region. *% Over 65* was retrieved from ISTAT[40] at the regional level for the year 2018. *% Diabetics* and *% Allergics* were retrieved from ISTAT[48] at the regional level for the year 2018. *Adults per family doctor* was retrieved from the Ministry of Health[49] at the regional level for the year 2017. To compute *ICU beds per 100,000 inhabitants*, we collected the total number of ICU beds in each region in 2018 from the Ministry of Health[50], multiplied by 100,000 and divided by the population of the region as of January 1, 2019[40]. To compute *Ave. beds per hospital (whole)* we used data from the Ministry of Health[51], which provides the number of beds per ward in each hospital in 2018. We first aggregated them over wards belonging to the same hospital, and then averaged over hospitals in each region. *Ave. beds per nursing home (ward)* was also obtained based on data for the year 2018 from the Ministry of Health[52]—here we considered regional averages at the level of wards, without aggregating over wards inside the same nursing home (the ward-level covariate had a slightly higher association with mortality outcomes). To compute *Ave. students per classroom* we used data from the Ministry of Education[53], which provides the number of students in each classroom of each school in the country (public or private, at every level of education), for the year 2018. We averaged them over schools in each region. Data for Trento/Bolzano and Valle d'Aosta were missing, and were imputed through random forest imputation[54] using the `R package missForest`[55] with default parameters `maxiter=10` (maximum number of iterations) and `ntree=100` (number of trees)—the latter is very large compared to the small number of missing values to be imputed, and our runs always converged well before the 10th iteration. To compute *Ave. employees per firm* we used data from ISTAT [56], which provides number of employees per firm at the level of municipalities. We averaged them over firms in each region. Data for Valle d'Aosta were missing, and were again imputed through random forest imputation with default parameters. *Ave. members per household* was retrieved from ASR Lombardia[57] at the regional level for the year 2017. To compute *Public transport rides per capita* we used data from ISTAT[58], which provides the number of rides per capita for each Italian province in 2017. We multiplied these by the provinces' population as of January 1, 2019[40], summed up over provinces in the same region, and divided by the region population as of January 1, 2019[40]. To compute *PM10* we used data from ISTAT[58], which provides the average annual concentrations of PM10 (in $\mu g/m^3$) detected by air quality meters distributed over the Italian territory. We averaged them over meters located in each region.

Multivariate analysis tools. We used a number of standard multivariate techniques to analyze first the entire set of scalar covariates and then the 12 we focused on—including the extraction of *Principal Components*[59], the calculation of *Variance Inflation Factors*[60] to evaluate multicollinearities, and clustering based on hierarchical agglomeration[59]. The latter was used both to agglomerate covariates with similar behavior across regions and to agglomerate regions with similar behavior across covariates. *Agglomerative hierarchical clustering* groups elements in a set with a bottom-up procedure that results in a dendrogram. Each element starts in its own cluster, and pairs of clusters are merged iteratively with a chosen distance for elements and linkage criterion for clusters. We employed the correlation distance, defined as $d(x_1, x_2) = 1 - |corr(x_1, x_2)|$ for two generic elements

x_1 and x_2, and the complete linkage, defined as $D(X_1, X_2) = max_{x_1 \in X_1, x_2 \in X_2} d(x_1, x_2)$ for two generic clusters X_1 and X_2 (thus, the distance between two clusters is defined as the furthest distance between their elements).

We also used biclustering on the 20 (regions) by 12 (covariates) data matrix, to identify subsets of regions exhibiting similar behaviors across subsets of covariates. Following standard literature, we sought sub-matrices of the data whose entries are consistent with the "ideal" additive model $x_{i,j} = \mu + \alpha_i + \tau_j$, where μ is the typical value within the bicluster, and α_i and τ_j are additive adjustments for row i and column j, but we set all α_is to 0 in order to find constant column biclusters, i.e., sub-matrices with constant columns (covariates). We employed the *Cheng and Church Biclustering Algorithm*[31], a greedy algorithm which finds the largest sub-matrices whose departure from the additive model is below a user-defined threshold. The departure is computed using the H-score (or mean squared residue score); in symbols, $H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \left(x_{i,j} - x_{I,j}\right)^2$, where I and J index the sets of rows and columns composing the bicluster, $x_{i,j}$ is a generic cell in the bicluster and $x_{I,j}$ is the mean of column j (note the algorithm thus estimates the typical values in the additive model using means). We implemented this algorithm with a recently proposed *adjustment to the H-score*[32] that corrects a bias towards smaller biclusters in the original formulation. The adjusted H-score is defined as $H_{adj}(I,J) = (\prod_{r=2}^{I-1} \frac{r^2}{r^2-1} \prod_{q=2}^{J-1} \frac{q^2}{q^2-1})^{-1} H(I,J)$, where r and q indicate the number of rows and columns, respectively.

Functional data analysis tools. *Local clustering of curves and functional motif discovery.* We performed local clustering of smoothed mortality curves (DPC, ISTAT and MAX, separately) using *probabilistic K-mean with local alignment* (*probKMA*[22]). *ProbKMA* is a *K*-mean-like algorithm for functional data that finds K groups in a set of curves based on local similarity among portions of the curves themselves. This allows the discovery of functional motifs, i.e. of typical local shapes that recur within and across the curves. In symbols, the algorithm finds K motifs $v_1, \ldots, v_K$, membership probabilities $p_{k,i}$ and shifts $s_{k,i}$ (i.e. the starting points of the motif instances) for each cluster-curve pair that minimize the generalized least-squares functional $J(v_1, \ldots, v_K, p_{k,i}, s_{k,i}) = \sum_{i=1}^{N} \sum_{k=1}^{K} p_{k,i}^2 d^2(\tilde{x}_i, v_k)$, where $\tilde{x}_i$ is the portion of the curve i corresponding to the shift $s_{k,i}$, and d is the distance used to capture local similarity. For each data set, we considered $K = 2$ and $K = 3$ (using larger K values did not improve results and did not produce robust clusters across the three data sets considered). *ProbKMA* is probabilistic; it returns as output a membership probability $p_{k,i}$ for each cluster-curve pair. However, such an output can be turned into a hard partition by assigning each curve to the group with highest membership probability—which is what we did here. Notably, for $K = 2$, membership probabilities showed that Lombardia's and Valle d'Aosta's extreme mortality patterns were not well accommodated even in the "exponential" group[22]. The algorithm can employ different definitions of similarity d and thus capture different aspects of curve shapes. We used Euclidean (L^2) distance between curve levels for our main analysis—in symbols, $d^2 = \frac{1}{c} \int_0^c (x(t) - v(t))^2 dt$ for two generic curves x and v—though using Euclidean distance between curve derivatives produced similar results (not shown). *ProbKMA* allows the length of the motifs to be extended endogenously starting from a minimal one fixed in input. However, to identify epidemic patterns we ran it with a fixed motif length of 65 days—hence allowing a maximum shift of 10 days between curves (the mortality curves are 75 days long). The same clusters and very similar shifts were obtained with a fixed motif length of 50 days, which allows a maximum shift of 25 days (results not shown). The shifts produced by *probKMA* with $K = 2$ on the three mortality data sets (DPC, ISTAT and MAX) were employed to align, in addition to the mortality curves themselves, local mobility and positivity curves. All subsequent analyses employing shifted curves (tests contrasting groups of curves, functional boxplots and depth analyses, and functional regression models) were therefore restricted to the 65-day portions where mortality curves aligned following the two *probKMA* motifs. We also validated the groups produced by *probKMA* with a modified version of *funBI*[24], an algorithm tipically used for finding functional biclusters. We used the modified *funBI* to identify groups of curves characterized by group-specific fixed length motifs, considering all possible sub-curves of a fixed length and clustering them with a divisive hierarchical algorithm (results not shown).

Testing for differences between groups of curves. We employed an *Interval-Wise Testing* algorithm developed for omics data (*IWTomics*[25]) to test for differences between the two groups of shifted mortality curves produced by *probKMA* with $K = 2$ (again, separately for DPC, ISTAT and MAX). *IWTomics* is a non-parametric, permutation-based functional hypothesis test. It contrasts two sets of curves aligned on a common domain to detect locations where their distributions differ significantly, and scales at which such significant differences are displayed (scales correspond to varying degrees of adjustment for multiple testing on intervals of varying lengths). Here locations are represented by the 65 days where the shifted mortality curves are defined, while scales vary from 1 day to the whole 65 days. The test was performed with the `R package IWTomics`[25,61]. The package allows the user to select among various possible test statistics. Since our tests contrasted groups of curves produced by *probKMA* with a Euclidean (L^2) distance, so that cluster centers are in effect the functional means of the aligned curves in each cluster, we employed the mean as test statistic in *IWTomics*. The number of permutations was set to 1000 (default value).

Functional boxplots and depth analyses. The functional boxplot[28] is an exploratory tool used to visualize functional data. It is constructed after ordering a set of curves based on a depth measure, such as the modified band depth[29]. The statistics employed to construct a functional boxplots are: the 50% central region envelope, the median curve, and the maximum non-outlying envelope. The 50% central region envelope corresponds to the box in a classical boxplot; it contains the 50% deepest, most centrally located curves. The median, i.e. the deepest curve, is inside this box and represents a robust "center" of the functional data set. The maximum non-outlying

envelope is obtained by inflating the 50% central region envelope by 1.5 times its range. All curves extending outside of this envelope are flagged as outliers (the fact that the ISTAT data set in Fig. 3(b) lacks outlying curves based on this definition is due to the width of its 50% central region envelope). We ranked the curves based on their depth measurements, after attributing a sign to such measurements with an ad hoc procedure. We subtract the median from each curve, and consider the share of the domain on which the difference is positive. If this is larger than 50%, we attribute a positive sign to the curve's depth—otherwise, we attribute a negative sign. Curves can thus be ranked from the most outlying above the median (labeled as positive), down to those close to the median, down to the most outlying below the median (labeled as negative)—see Fig. 3(b). While this is not a fully general procedure, it works well on the DPC, ISTAT and MAX mortality curves we considered, which are rather unambiguously above/below the median (the share of the domain where the difference from the median is positive is ≥ 70 or $\leq 30\%$ for all curves in all three data sets). Note also that the median curve of a data set, defined as the deepest, does not necessarily have half of the curves above it and half of the curves below it in the signed ranking we created (e.g., Toscana is the median curve in both ISTAT and MAX data sets, but the number of curves above/below it differs).

Functional regression models. We consider models where a functional response variable is regressed against functional predictors and/or scalar covariates[14,15]. All are special cases of the general equation[30]

$$y_i(t) = \alpha(t) + \sum_{\ell=1}^{L} \int \beta_\ell(s,t) x_{i,\ell}(s) ds + \sum_{j=1}^{J} \beta_j(t) x_{i,j} + \varepsilon_i(t) \quad i = 1, \ldots n.$$

n is the number of observations, in our case $n = 20$ regions. $y_i(t), i = 1, \ldots n$ are the aligned mortality curves (DPC, ISTAT or MAX, modeled separately), $\alpha(t)$ is a functional intercept and $\varepsilon_i(t), i = 1, \ldots n$ are i.i.d. Gaussian model errors. L is the number of functional predictors. $x_{i,\ell}(s), i = 1, \ldots n, \ell = 1, \ldots L$, are such predictors, measured on the n observations. The regression coefficient of each functional predictor, $\beta_\ell(s,t)$, is a surface. J is the number of scalar covariates. $x_{i,j}(s), i = 1, \ldots n, j = 1, \ldots J$, are such covariates, measured on the n observations. The regression coefficients of each scalar covariate, $\beta_j(t)$, is a curve. For the marginal regressions of mortality on local mobility and mortality on positivity, we have $L = 1$ and $J = 0$. For the joint regression of mortality on local mobility and positivity, we have $L = 2$ and $J = 0$. For the marginal regressions of mortality on individual scalar covariates, we have $L = 0$ and $J = 1$. In Fig. S15 we fit marginal regressions of this type allowing the estimation of two different intercepts: $\alpha_1(t)$ for curves in Group 1 and $\alpha_2(t)$ for curves in Group 2. Finally, for the joint regression of mortality on local mobility, positivity and one scalar control variable, we have $L = 2$ and $J = 1$. To fit all these functional regressions we used the `R package refund`[62], which estimates the functional coefficients as well as their standard errors. We used these standard errors to construct confidence bands around the estimated functional coefficients. To gauge the explanatory power of each model, we computed the in-sample R^2 as well as the Leave-One-Out Cross-Validation (LOO-CV) R^2. The former is a functional generalization of the classical *coefficient of determination* defined as $SS_{reg}/(SS_{reg} + SS_{res})$, where SS_{reg} and SS_{res} are the regression and the residual sum of squares, respectively. To compute the latter, for each observation i, one replaces the fitted response curve $\hat{y}_i(t)$ (from the model fitted on all observations) with the predicted response curve $\hat{y}_{pred,i}(t)$ obtained for i from the model fitted withholding i itself. Finally, for models with multiple terms (predictors and/or covariates), the partial R^2 of each term is computed as $(R^2 - R^2_{red})/(1 - R^2_{red})$, where R^2 is the coefficient of determination of the complete model, and R^2_{red} that of the model comprising all terms but the one being evaluated.

SsNAL-EN for feature selection. SsNAL-EN[35] is an algorithm to perform Elastic Net[63] feature selection in a standard regression framework (i.e. when both response and features are scalars) which has been designed to provide computational efficiency. The Elastic Net is a hybrid between LASSO and Ridge, which penalizes both the L^1 and the L^2 (Euclidean) norm of the regression coefficients. The L^1 penalty induces sparsity selecting only the most predictive among the features. The L^2 penalty regularizes coefficient estimates mitigating variance inflation due to collinearity. To perform feature selection in the functional regression setting, we applied a generalization of SsNAL-EN which incorporates a group structure in the Elastic Net objective function and uses the Functional Principal Components basis expansion to represent a functional response. In particular, we performed feature selection for the regression of mortality against all 12 scalar covariates in Table 1. Notably, we selected the same top 5 features across all three data sets (DPC, ISTAT and MAX) (see Table S5)—lending strong support to their association with mortality.

Received: 19 August 2020; Accepted: 27 July 2021
Published online: 30 August 2021

References

1. La Rosa, G. *et al.* SARS-CoV-2 has been circulating in northern Italy since December 2019: Evidence from environmental monitoring. *Sci. Total Environ.* **750**, 141711 (2021).
2. Mugnai, G. & Bilato, C. COVID-19 in Italy: Lesson from the Veneto region. *Eur. J. Internal Med.* **77**, 161–162 (2020).
3. Lavezzo, E. *et al.* Suppression of COVID-19 outbreak in the municipality of Vo', Italy. *Nature* **584**, 425–429 (2020)
4. ISTAT. Demographic indicators. http://dati.istat.it/Index.aspx?DataSetCode=DCIS_INDDEMOG1&Lang=en.
5. Lim, S., Bae, J. H., Kwon, H.-S. & Nauck, M. A. Covid-19 and diabetes mellitus: From pathophysiology to clinical management. *Nat. Rev. Endocrinol.* **17**, 11–30 (2021).

6. Pluchino, A. *et al.* A novel methodology for epidemic risk assessment of covid-19 outbreak. *Sci. Rep.* **11**, 1–20 (2021).
7. Rovetta, A. & Castaldo, L. Relationships between demographic, geographic, and environmental statistics and the spread of novel coronavirus disease (covid-19) in Italy. *Cureus* **12**, e11397 (2020).
8. Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. & Dominici, F. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Sci. Adv.* **6**, eabd4049 (2020).
9. Coccia, M. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. *Sci. Total Environ.* **729**, 138474 (2020).
10. Binkin, N., Salmaso, S., Michieletto, F. & Russo, F. Protecting our health care workers while protecting our communities during the COVID-19 pandemic: A comparison of approaches and early outcomes in two Italian regions, 2020 (2020). Preprint at https://www.medrxiv.org/content/10.1101/2020.04.10.20060707v2.
11. Frumento, P. & Sylos Labini, M. Mortalità da coronavirus: quanto vale l'effetto Lombardia. *LaVoce.info* https://www.lavoce.info/archives/65752/mortalita-da-coronavirus-quanto-vale-leffetto-lombardia (2020).
12. Cortés, M. E. Enfermedad por coronavirus 2019 (covid-19): Importancia de la comunicación científica y de la enseñanza actualizada de las zoonosis. *Revista peruana de investigación en salud* **4**, 87–88 (2020).
13. James, L. P., Salomon, J. A., Buckee, C. O. & Menzies, N. A. The use and misuse of mathematical modeling for infectious disease policymaking: Lessons for the covid-19 pandemic. *Med. Decis. Making* **41**, 379–385 (2021).
14. Ramsay, J. O. & Silverman, B. W. *Functional data analysis*, 2nd edn (Springer, 2005).
15. Kokoszka, P. & Reimherr, M. *Introduction to Functional Data Analysis* (CRC Press, 2017).
16. Ramsay, J. O. & Silverman, B. W. *Applied Functional Data Analysis: Methods and Case Studies* (Springer, 2007).
17. Ullah, S. & Finch, C. F. Applications of functional data analysis: A systematic review. *BMC Med. Res. Methodol.* **13**, 43 (2013).
18. Cremona, M. A. *et al.* Functional data analysis for computational biology. *Bioinformatics* **35**, 3211–3213 (2019).
19. Carroll, C. *et al.* Time dynamics of COVID-19. *Sci. Rep.* **10**, 21040 (2020).
20. Ciminelli, G. & Garcia-Mandicó, S. Covid-19 in Italy: An analysis of death registry data. *VOXEU, Centre for Economic Policy Research, London* https://voxeu.org/article/covid-19-italy-analysis-death-registry-data (2020).
21. Modi, C., Böhm, V., Ferraro, S., Stein, G. & Seljak, U. Estimating covid-19 mortality in Italy early in the covid-19 pandemic. *Nat. Commun.* **12**, 1–9 (2021).
22. Cremona, M. A. & Chiaromonte, F. Probabilistic K-mean with local alignment for clustering and motif discovery in functional data (2020). Preprint at arXiv:1808.04773.
23. probKMA. https://github.com/marziacremona/ProbKMA-FMD.
24. Di Iorio, J. & Vantini, S. funbi: A biclustering algorithm for functional datas. *MOX-Report***46/2019** (2019).
25. Cremona, M. A. *et al.* IWTomics: Testing high-resolution sequence-based "Omics" data at multiple locations and scales. *Bioinformatics* **34**, 2289–2291 (2018).
26. Ra, S. H. *et al.* Upper respiratory viral load in asymptomatic individuals and mildly symptomatic patients with sars-cov-2 infection. *Thorax* **76**, 61–63 (2021).
27. Cegolon, L. *et al.* Hypothesis to explain the severe form of COVID-19 in northern Italy. *BMJ Glob. Health* **5**, e002564 (2020).
28. Sun, Y. & Genton, M. G. Functional boxplots. *J. Comput. Graph. Stat.* **20**, 316–334 (2011).
29. López-Pintado, S. & Romo, J. On the concept of depth for functional data. *J. Am. Stat. Assoc.* **104**, 718–734 (2009).
30. Horváth, L. & Kokoszka, P. *Inference for functional data with applications*, vol. 200 (Springer, 2012).
31. Cheng, Y. & Church, G. M. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA*, pp. 93–103 (2000).
32. Di Iorio, J., Chiaromonte, F. & Cremona, M. A. On the bias of h-scores for comparing biclusters, and how to correct it. *Bioinformatics* **36**, 2955–2957 (2020).
33. Dowd, J. B. *et al.* Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc. Natl. Acad. Sci.* **117**, 9696–9698 (2020).
34. Nepomuceno, M. R. *et al.* Besides population age structure, health and other demographic factors can contribute to understanding the COVID-19 burden. *Proc. Natl. Acad. Sci.* **117**, 13881–13883 (2020).
35. Boschi, T., Reimherr, M. & Chiaromonte, F. An efficient semi-smooth newton augmented lagrangian method for elastic net (2020). Preprint at arXiv:2006.03970.
36. Boschi, T., Chiaromonte, F., Secchi, P. & Li, B. Covariance based low-dimensional registration for function-on-function regression. *MOX-Report* (2018).
37. Cintia, P. *et al.* The relationship between human mobility and viral transmissibility during the COVID-19 epidemics in Italy (2020). Preprint at arXiv:2006.03141.
38. Martellucci, C. A. *et al.* Changes in the spatial distribution of covid-19 incidence in Italy using gis-based maps. *Ann. Clin. Microbiol. Antimicrob.* **19**, 1–4 (2020).
39. DPC. Covid19 dati regioni. https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni.
40. ISTAT. Atlante statistico territoriale delle infrastrutture. http://asti.istat.it/asti.
41. ISTAT. Decessi e cause di morte: cosa produce l'istat. https://www.istat.it/it/files/2020/03/Dataset-decessi-comunali-giornalieri-e-tracciato-record-30giugno.zip.
42. ISTAT. Popolazione residente al 1° gennaio. http://dati.istat.it/Index.aspx.
43. Google. Community mobility reports. https://www.google.com/covid19/mobility/.
44. Barone, N. & Bartoloni, M. La giravolta comunicativa sul coronavirus, menotamponi e contare solo i casi gravi. *Il sole 24 ore* https://www.ilsole24ore.com/art/la-giravolta-comunicativa-coronavirus-meno-tamponi-e-contare-solo-casi-gravi-ACQYXQMB (2020).
45. Craven, P. & Wahba, G. Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403 (1978).
46. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2021). Software version 4.1.0.
47. Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. *fda: Functional Data Analysis* (2011). R package version 2.2-6.
48. ISTAT. Aspetti della vita quotidiana. http://dati.istat.it/Index.aspx?QueryId=15448.
49. Ministry of Health. Assistenza primaria. http://www.salute.gov.it/imgs/C_17_pubblicazioni_1203_ulteriorallegati_ulteriorallegato_8_alleg.pdf.
50. Ministry of Health. http://www.dati.salute.gov.it/dati/dettaglioDataset.jsp?menu=dati&idPag=96.
51. Ministry of Health. http://www.salute.gov.it/imgs/C_17_bancheDati_6_0_1_file.xls.
52. Ministry of Health. http://www.salute.gov.it/imgs/C_17_bancheDati_6_0_0_file.xls.
53. Ministry of Education. https://dati.istruzione.it/opendata/opendata/catalogo/elements1/leaf/?area=Studenti&datasetId=DS0030ALUCORSOINDCLASTA,DS0030ALUCORSOINDCLAPAR, DS1114INFANZIACLASTA,DS1115INFANZIACLAPAR.
54. Stekhoven, D. J. & Bühlmann, P. Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
55. Stekhoven, D. J. *missForest* (2012). R package version 1.4.
56. ISTAT. Atlante statistico dei comuni. http://asc.istat.it/ASC/.
57. ASR Lombardia. Numero di famiglie, convivenze e numero medio di componenti per famiglia. https://www.asr-lombardia.it/asrlomb/it/13740numero-di-famiglie-convivenze-e-numero-medio-di-componenti-famiglia-regionale.
58. ISTAT. Ambiente urbano. https://www.istat.it/it/archivio/236912.

59. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pp. 534–541 (Springer, 2009).
60. Allison, P. D. *Multiple Regression: A Primer* 140–145 (Pine Forge Press, 1999).
61. Cremona, M. A. *IWTomics* (2018). R package version 1.16.0. https://bioconductor.org/packages/release/bioc/html/IWTomics.html.
62. Goldsmith, J. *et al.Refund: Regression with functional data* (2016). R package version 0.1.16.
63. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67**, 301–320 (2005).

Acknowledgements

M.A. Cremona acknowledges support from the NSERC. F. Chiaromonte and T. Boschi acknowledge support from the Huck Institutes of the Life Sciences (Penn State University). F. Chiaromonte, J. Di Iorio and L. Testa acknowledge support from the Sant'Anna School of Advanced Studies. We are grateful to Paola Cesari, Christian Esposito, Giovanni Felici, Daniele Licari, Andrea Mina and Flavia Petruso for useful feedback.

Author contributions

All authors conceived ideas and analysis approaches. T.B., J.Di I., L.T. and M.A.C. retrieved and processed data from multiple public sources, implemented pipelines and performed statistical analyses. All authors interpreted findings and participated to the writing of the manuscript. M.A.C. and F.C. supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-95866-y.

Correspondence and requests for materials should be addressed to M.A.C. or F.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.