Epilocal: A real-time tool for local epidemic monitoring

Marco Bonetti, Ugofilippo Basellini

DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 44, ARTICLE 12, PAGES 307–332
PUBLISHED 10 FEBRUARY 2021
http://www.demographic-research.org/Volumes/Vol44/12/
DOI:10.4054/DemRes.2021.44.12

Research Material

Epilocal: A real-time tool for local epidemic monitoring

Marco Bonetti

Ugofilippo Basellini

Contents

Demographic Research: Volume 44, Article 12
Research Material

Epilocal: A real-time tool for local epidemic monitoring

Marco Bonetti[1]

Ugofilippo Basellini[2]

Abstract

BACKGROUND
The novel coronavirus (SARS-CoV-2) emerged as a global threat at the beginning of 2020, spreading around the globe at different times and rates. Within a country, such differences provide the opportunity for strategic allocations of health care resources.

OBJECTIVE
We aim to provide a tool to estimate and visualize differences in the spread of the pandemic at the subnational level. Specifically, we focus on the case of Italy, a country that has been harshly hit by the virus.

METHODS
We model the number of SARS-CoV-2 reported cases and deaths as well as the number of hospital admissions at the Italian subnational level with Poisson regression. We employ parametric and nonparametric functional forms for the hazard function. In the parametric approach, model selection is performed using an automatic criterion based on the statistical significance of the estimated parameters and on goodness-of-fit assessment. In the nonparametric approach, we employ out-of-sample forecasting error minimization.

RESULTS
For each province and region, fitted models are plotted against observed data, demonstrating the appropriateness of the modeling approach. Moreover, estimated counts and rates of change for each outcome variable are plotted on maps of the country. This provides a direct visual assessment of the geographic distribution of risk areas as well as insights on the evolution of the pandemic over time.

CONTRIBUTION
The proposed `Epilocal` software provides researchers and policymakers with an open-

[1] Carlo F. Dondena Research Center, Bocconi Institute for Data Science and Analytics, and Covid Crisis Lab, Bocconi University, Milan, Italy.
[2] The Max Planck Institute for Demographic Research (MPIDR), Rostock, and the Institut national d'études démographiques (INED), Aubervilliers, France. Email: basellini@demogr.mpg.de.

access real-time tool to monitor the most recent trends of the COVID-19 pandemic in Italian regions and provinces with informative graphical outputs. The software is freely available and can be easily modified to fit other countries as well as future pandemics.

1. Introduction

The novel coronavirus (SARS-CoV-2) emerged between the end of 2019 and the beginning of 2020 as a global threat. The disease caused by the virus has been designated COVID-19. The virus has spread very quickly and – at the time of writing – can be found in most of the world's countries, with differential timing both across and within countries (see, e.g., WHO 2020). Such differential spreading of the epidemic over different regions provides the opportunity for optimizing the use of health care resources across a country, thus potentially reducing mortality caused by a lack of sufficient care.

We have developed `Epilocal`, a software tool that allows researchers and policymakers to monitor province- and region-level COVID-19 data in Italy. The software is made available to the community in two distinct ways: (i) as open-access `R` codes, which allow users to produce real-time results on their computers, and (ii) as an online shinyapp (Chang et al. 2020), which provides users with an interface for the direct visualization of different results by selecting the desired inputs.

This article is organized as follows. In Section 2 we describe the methods, and in particular the data sources, the parametric model specifications, the automatic model selection process, and the output provided by the software, to describe cumulative counts. We also describe a nonparametric P-splines approach to model noncumulative counts. In Section 3 we show a few sample plots for some Italian provinces and regions, as well as the graphical representation for the whole country. We close with some discussion in Section 4.

2. Methods

2.1 Data sources and modeling

Daily data on the reported numbers of infected individuals with the SARS-CoV-2 virus are available at the region and province level in Italy from the Department of Civil Protection (Dipartimento della Protezione Civile 2020). The data report the daily new and cumulative counts for each of the 21 regions and 107 provinces (NUTS 2 and 3 levels, respectively) starting from February 24, 2020, together with representative coordinates (longitude and latitude) for each geographical unit. Similar daily data are available, only

at the regional level, for the new and cumulative number of COVID-positive deceased individuals and for the number of individuals who are currently present in intensive care unit (ICU) and non-ICU departments. The resident population of each province and region on January 1, 2019, can be obtained from the Italian Statistical Institute (Istat 2020). In the description below we focus on the cumulative number of reported cases for a given province, but the software produces similar analyses for the other variables at the regional level.

2.2 Parametric models

We use parametric functions to model cumulative counts from a given starting time point. Given the resident population N for a generic province, we let Y_t be the number of individuals reported infected at time t in that province. We model Y_t via a Poisson generalized linear model with hazard rate $h(t)$ as function of time, the natural logarithm link function, and the exposed population N as an offset (Brillinger 1986; McCullagh and Nelder 1989). In what follows, we assume that the exposure N does not change over time. Although the population size is influenced by demographic components (fertility, mortality, and migration), this assumption is reasonable given that we consider a limited time frame. For the hazard rate, we consider some parametric functions of time, specifically first-, second-, or third-degree log-polynomials for all responses, and a logistic parameterization only for cumulative counts. The index t is a natural number that indicates the number of days since the total reported cases in the province reached five.

Let us focus on the third-degree log-polynomial hazard model, i.e., $\eta(t) = \log(h(t)) = \beta_0+\beta_1 t+\beta_2 t^2+\beta_3 t^3$, where $\eta(t)$ denotes the linear predictor. Let λ_t denote the expected value of the Poisson process at time t. We thus assume $Y_t \sim \text{Poisson}(\lambda_t)$ with

$$\begin{aligned} \log(\lambda_t) &= \log(N) + \eta(t) \\ &= \log(N) + \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3. \end{aligned} \tag{1}$$

Note the inclusion of the offset term $\log(N)$ to explicitly extract the time-constant exposure term from the constant parameter of the regression.

The vector parameter $\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \beta_3]$ is estimated by maximum likelihood in `R` using the `glm` command of the `stats` package (R Core Team 2020). We do not fit the model for provinces that have experienced a maximum reported infected rate smaller than a threshold of 5×10^{-5}, as the few data points do not allow for a robust model estimation. In such instances, we simply output a message saying "Prevalence too low."

From Equation (1), the model-based fitted number of reported infected cases at time

t is thus

$$\widehat{Y_t} = \widehat{\lambda}_t = N\,\widehat{h(t)} = N \exp\left[\widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2 + \widehat{\beta}_3 t^3\right]. \tag{2}$$

Here and in what follows, the hat symbol "$\widehat{\ }$" denotes an estimated term.

The estimated rate of growth at time t can be computed from the (relative) first derivative of the rate $h(t)$ by plugging in the maximum likelihood estimator (mle) $\widehat{\boldsymbol{\beta}}$ into the expression

$$\frac{\widehat{\frac{\partial}{\partial t}\,h(t)}}{h(t)} = \frac{\partial}{\partial t}\,\widehat{\log(h(t))} = \widehat{\frac{\partial}{\partial t}\,\eta(t)} = \widehat{\beta}_1 + 2\widehat{\beta}_2 t + 3\widehat{\beta}_3 t^2. \tag{3}$$

Similarly, the (relative) second derivative (useful to identify inflection points) is estimated by

$$\frac{\widehat{\frac{\partial}{\partial t^2}\,h(t)}}{h(t)} = \left(2\widehat{\beta}_2 + 6\widehat{\beta}_3 t\right) + \left(\widehat{\beta}_1 + 2\widehat{\beta}_2 t + 3\widehat{\beta}_3 t^2\right)^2. \tag{4}$$

Note that, by invariance, these estimators are the mles of the quantities they estimate. We did not pursue it here, but their large-sample distribution can be obtained by the delta method from the approximate distribution of $\widehat{\boldsymbol{\beta}}$. In particular, they are also approximately normally distributed, and a consistent estimator for the asymptotic variance can be constructed.

While we have described the third-degree polynomial model above, the first-degree and the second-degree models follow by dropping both the quadratic and the cubic term, or just the cubic term. The model-based fitted rates for these models thus follow as well, and their estimated (relative) first and second derivatives are equal to

$$\widehat{\frac{\partial}{\partial t}\,\eta(t)} = \widehat{\beta}_1 + 2\widehat{\beta}_2 t; \quad \frac{\widehat{\frac{\partial}{\partial t^2}\,h(t)}}{h(t)} = 2\widehat{\beta}_2 + \left(\widehat{\beta}_1 + 2\widehat{\beta}_2 t\right)^2 \tag{5}$$

for the second-degree model and

$$\widehat{\frac{\partial}{\partial t}\,\eta(t)} = \widehat{\beta}_1; \quad \frac{\widehat{\frac{\partial}{\partial t^2}\,h(t)}}{h(t)} = \widehat{\beta}_1^2 \tag{6}$$

for the first-degree model. Note that the expression of $\widehat{\lambda}_t$ in Equation (2) should be

adjusted by dropping the appropriate term(s) too. Moreover, one can trivially obtain the expressions for the corresponding first and second derivatives of $\widehat{Y_t}$ by multiplying Equations (3)–(6) by $N\,\widehat{h(t)}$, i.e., $\widehat{\frac{\partial}{\partial t}\,Y_t} = N\widehat{h(t)}\,\widehat{\frac{\partial}{\partial t}\,\eta(t)}$.

These formulas allow us to estimate the rate of change and the curvature in the reported infected rate at each time t; however, our interest is generally concerned with the most recent time point. In what follows, we exploit these formulas to produce a map of the provinces that reflects such trends at the most recent date.

Note that while one can obtain estimates for the (local or global) maxima and minima from the polynomial models, we do not recommend producing such quantities as they might be over-interpreted. This is particularly true when one models the cumulative counts, as that quantity can clearly only increase. For completeness, we describe the formulas for the corresponding point estimators in Appendix A.

Cumulative counts can only increase, and they can be expected to reach a plateau, at least at the end of each wave of the epidemic. As a consequence, we have also considered a logistic curve as a way to describe such data. The functional form of the hazard function is

$$h(t) = K\frac{\exp\left(\beta_0 + \beta_1 t\right)}{1 + \exp\left(\beta_0 + \beta_1 t\right)}, \tag{7}$$

where $K = \exp(\gamma)/\left[1 + \exp(\gamma)\right]$ is a parameter in $(0, 1)$, corresponding to the upper asymptote of the logistic curve. We exclude the degenerate case in which $\beta_1 = 0$. Since the logistic log-hazard is a nonlinear function of the parameters $\boldsymbol{\theta}' = [K, \beta_0, \beta_1]$, we estimate the model parameters by maximizing the Poisson log-likelihood

$$\ln\,\mathcal{L}\left(\boldsymbol{\theta}\,|\,Y_t\,, N\,\right) \propto \sum_t \left[Y_t\,\ln\left(h(\boldsymbol{\theta}; t)\right) - N\,h(\boldsymbol{\theta}; t)\right]\,. \tag{8}$$

In `R`, this can be performed using the `maxLik` package (Henningsen and Toomet 2011).

As for the log-polynomial models, the slope and the curvature of the logistic model at a given time t can be computed from the (relative) first and second derivatives with respect to time:

$$\begin{aligned}
\widehat{\frac{\partial}{\partial t}\,\eta(t)} &= \frac{\widehat{\beta}_1}{1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 t\right)} \\
\frac{\widehat{\frac{\partial}{\partial t^2}\,h(t)}}{h(t)} &= \widehat{\beta}_1^2\,\frac{\left[1 - \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 t\right)\right]}{\left[1 + \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 t\right)\right]^2}\,.
\end{aligned} \tag{9}$$

An important feature of the logistic function, which we will exploit for the model selection, is its inflection point $\breve{t}$. This can be computed by letting the (relative) second derivative be equal to zero:

$$\frac{\widehat{\frac{\partial}{\partial t^2} h(t)}}{h(t)} = 0 \quad \Leftrightarrow \quad \breve{t} = -\frac{\widehat{\beta}_0}{\widehat{\beta}_1}. \tag{10}$$

Note that in Equation (10) we do not consider the (here non-interesting) case of $h(t) = 0$.

We compute the 95% confidence interval for the inflection point by the delta method, i.e., from the gradient of $h(t)$ with respect to its parameters and the variance-covariance matrix of the estimated parameters.

2.3 Automatic model selection for the parametric models

We start by fitting the three nested models (first-, second-, and third-degree log-polynomials) to the data and recover the convergence status for all three models. This results in eight possible combinations of convergence (i.e., no models converged, only Model 1 converged, only Model 2 converged, and so on), which we analyze separately as follows:

- No models converged: we output the message "No models converged."
- One model converged: we select that model.
- Two models converged: we consider the significance of the higher-order log-polynomial model through either the appropriate Wald test statistic (if the two models differ by just one term) or the likelihood ratio test statistic (if the two nested models that converge are the first-degree and the third-degree models).
- All models converged: we perform a backward elimination process starting from the third-degree polynomial model. If the cubic term is statistically significant, then we retain it; if not, we drop the third-degree model and compare the second-degree model to the first-degree model by looking at the Wald statistic of the quadratic term.

For all model selection tests, we use a p-value cutoff equal to 0.05.

When dealing with cumulative counts, we also consider the logistic model. In general, we expect the polynomial model to provide an accurate fit at the beginning and in the middle of the growth of the epidemic, but we expect the logistic curve to provide a better fit in the latest stages of a wave of the pandemic.

Thus, in the analysis of cumulative data, we compare the selected polynomial model

from the procedure above with the logistic model in two ways. First, we compare the models in terms of the Bayesian Information Criterion (BIC, Schwarz 1978). The BIC is a measure often employed to evaluate model differences in terms of trade-off between model parsimony and accuracy: lower values of the BIC are associated with better models. We thus choose the model with a lower BIC. However, if the selected model is polynomial, and it is characterized by a downward trend at the current date, we consider this an indication that a logistic pattern has been achieved, and thus we select the latter model. However, to avoid selecting the logistic curve too early in time, we retain such a model only if the 95% confidence interval of the estimated inflection point falls entirely to the left of seven days prior to the last observed data point. If this criterion is not met, we simply report that no adequate model could be found.

2.4 Nonparametric modeling

A potential limitation of modeling the cumulative number of cases is that, in later stages of an epidemic, moderately large to increases in new cases would not appear as large proportionate increases in the cumulative hazard, as this latter value is already large. To overcome this limitation, we provide users of `Epilocal` the choice of modeling the daily number of new COVID-19 cases and deaths. This metric has the advantage of being less dependent on previous observations, and as such, it is a more appropriate measure to monitor eventual resurgences of the epidemic after the early stages. For these measures, we implement the nonparametric P-splines approach of Eilers and Marx (1996), which in our applications often provides a better fit to such noncumulative data than parametric approaches. To further improve the goodness-of-fit for the most recent observations (those that we are most interested in), we complement the nonparametric approach with the minimization of the out-of-sample forecasting error. The P-splines methodology proposed by Eilers and Marx (1996) has been vastly used in demographic studies (see, e.g., Currie, Durban, and Eilers 2004; Camarda 2008; Ouellette and Bourbeau 2011; Ebeling 2018; Basellini and Camarda 2019).

Specifically, we model the linear predictor $\eta(t)$ in Equation (1) using a linear combination of B-spline bases:

$$\eta(t) = \sum_{k=1}^{K} B_k^q(t)\, \alpha_k, \tag{11}$$

where $B_k^q(t)$ are K equally spaced cubic B-splines (i.e., degree $q = 3$) with associated coefficients α_k. Following Eilers and Marx (1996), we select a large number of B-splines to capture relevant trends in the mortality pattern. Smoothness is achieved by a roughness

penalty acting on the associated coefficients α_k. For a fuller treatment of P-splines for mortality data, we refer the interested reader to Section 2 of Camarda (2019).

P-splines in the Poisson setting are readily implemented in the `MortalitySmooth` package in `R` (Camarda 2012). Since our interest is primarily concerned with the most recent observations, we select the roughness parameter using an out-of-sample forecast accuracy procedure (rather than minimizing the BIC criterion, as in the package default option). We consider an out-of-sample window containing the last n observed data points. For each out-of-sample value y_t, we compute the corresponding forecasts $\hat{y}_t$ by: (i) fitting the P-splines to a subset of the observations (from the first data point until ℓ observations before y_t) and (ii) extrapolating the fitted Equation (11) ℓ steps ahead. We then compute the root mean square error (RMSE) for a given roughness parameter λ:

$$\text{RMSE}\,(\lambda) = \sqrt{\frac{\sum_{t=1}^{n} \left(\hat{y}_t(\lambda) - y_t\right)^2}{n}} \tag{12}$$

where y_t is the observed out-of-sample value, $\hat{y}_t(\lambda)$ is the corresponding ℓ step-ahead forecast based on the value λ for the roughness parameter, and n is the length of the out-of-sample window. We repeat this procedure for a series of roughness parameters ranging from small to large values (from 1 to 10^5), and we select the one that minimizes the RMSE. As default values, we choose a window of $n = 10$ last observed values and $\ell = 3$ steps ahead for the forecasts. However, we leave these inputs as customizable options in the `R` code and the shinyapp.

Finally, given the estimated coefficients $\hat{\alpha_k}$, the rate of change of the linear predictor can be directly derived as follows:

$$\widehat{\frac{\partial}{\partial t}\eta(t)} = \frac{1}{h}\sum_{k=1}^{K} B_k^{q-1}(t)\,[\hat{\alpha}_k - \hat{\alpha}_{k+1}], \tag{13}$$

where $B_k^{q-1}(t)$ are B-splines of a lower degree, quadratic in our case, and h is the knot distance. For additional details, see Eilers and Marx (1996) and Camarda (2019).

2.5 Output and graphical displays

For both parametric and nonparametric analysis, we produce a separate plot for each province and region. In each plot, we show the reported cases over time together with a line showing the fitted values from the selected model. For parametric models, the

plot also contains the indication of the degree of the polynomial that was selected (or the logistic model for cumulative responses), the estimated rate of increase and curvature at the last time point, and the possible messages signaling that no modeling was possible due to the low number of cases in that province or lack of convergence of all models.

We then plot the estimated slopes on a map of Italy as obtained from Stamen Design (data by OpenStreetMap) using the `ggmap` package (Kahle and Wickham 2013). For each province or region, we plot an empty circle centered at the representative coordinates, with the area of the circle proportional to the estimated counts at the most recent time, divided by the population, i.e., $\widehat{Y}_t/N$. When no appropriate model is found, we average the last two observed reported cases over the resident population.

Furthermore, we use color-coded circles to indicate the value of the growth rates, estimated as described in Subsection 2.2. We employ the perceptually uniform color palette from the `viridis` package to map the growth rates to each circle (Garnier 2018). When no models are found, the circles are drawn in gray.

In addition, we provide the option of visualizing only a single metric of interest, i.e., either the magnitude of cases or their rate of change. This is achieved by coloring the area of each province or region using the geospatial vector maps provided by the Eurostat (2020).

Lastly, we export a text output containing all the information for each geographical unit, i.e., details about the selected model (parametric form, estimated model coefficients with standard errors, and p-values), the estimated number of cases, and the rate of growth and curvature at the last data point.

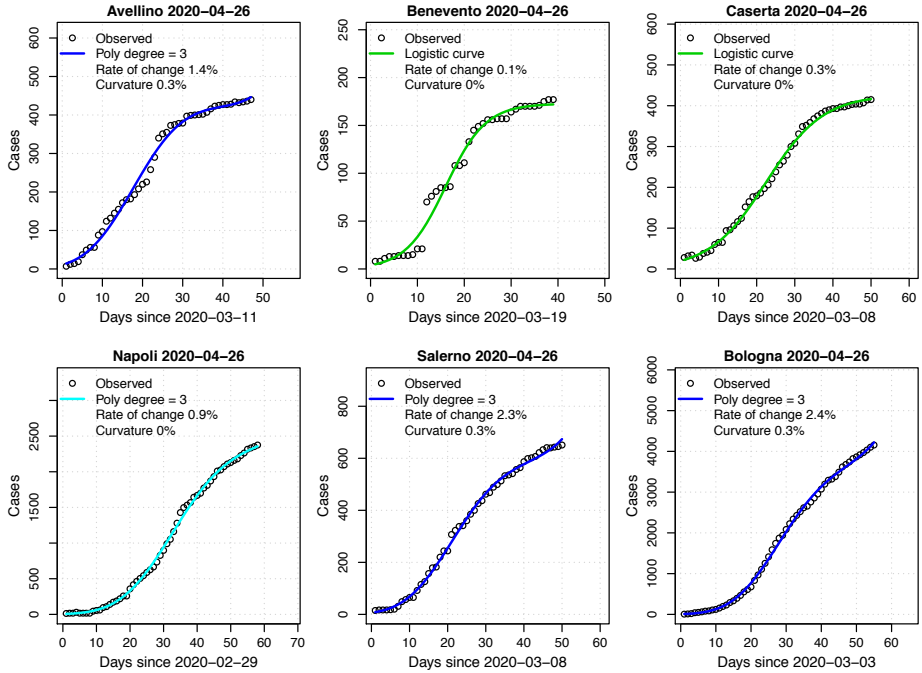
3. Results

We ran our software at two time points, which are indicated at the top of the plots produced by `Epilocal`: April 26 and October 27. For provinces, the complete output contains one plot for each province, one figure for the whole country, and a detailed text output. For regions, the same output is produced for the selected variable of interest (reported cases, deaths, or hospital admissions). These outputs are obtainable by running the provided `R` code.

We start by showing the results of the parametric analysis, which we run until April 26, 2020, to illustrate the automatic adaptation of the model selection procedure to the different shapes typical of the early stages of epidemics. Figure 1 shows a representative example of some province-specific plots. The plots show the starting date of analysis (in the x-label), the last date of the analysis (in the title), and the selected model (with its degree for log-polynomial ones). Here we model cumulative counts, and the observed fit is rather accurate, even from a purely visual inspection. Estimated slopes with delta-

method-based 95% confidence intervals for all provinces at the last day of the analysis are reported in Appendix B.

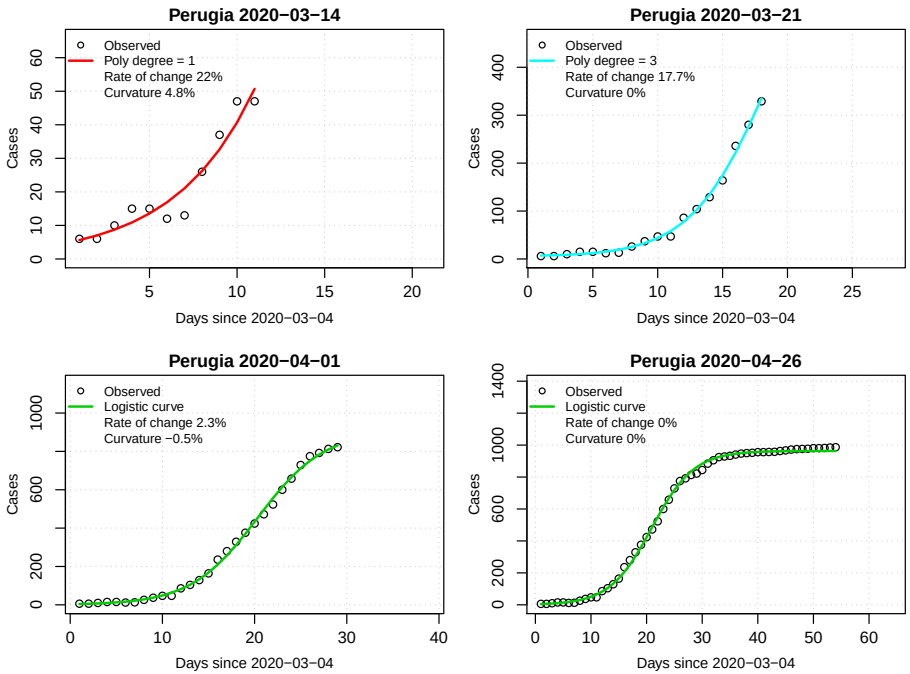
Figure 1: Sample of plots for six Italian provinces, showing the cumulative number of reported infected SARS-CoV-2 cases over time since the day when the reported cases reached the number five



Note: The curves show the fitted values from the model selected for each province. Dark (or light) blue lines correspond to log-polynomial models with a positive (negative) curvature, and green lines correspond to the logistic model.

It is interesting to observe how the automatic model selection chooses different models over time for the same province. Figure 2 shows the observed and fitted number of cumulative reported infected SARS-CoV-2 cases in the province of Perugia. As for several other provinces, the log-polynomial model of degree 1 (red curve) is chosen at the first stages of the epidemic. Then, a higher order polynomial is selected, and finally the model selection transitions into the logistic curve.

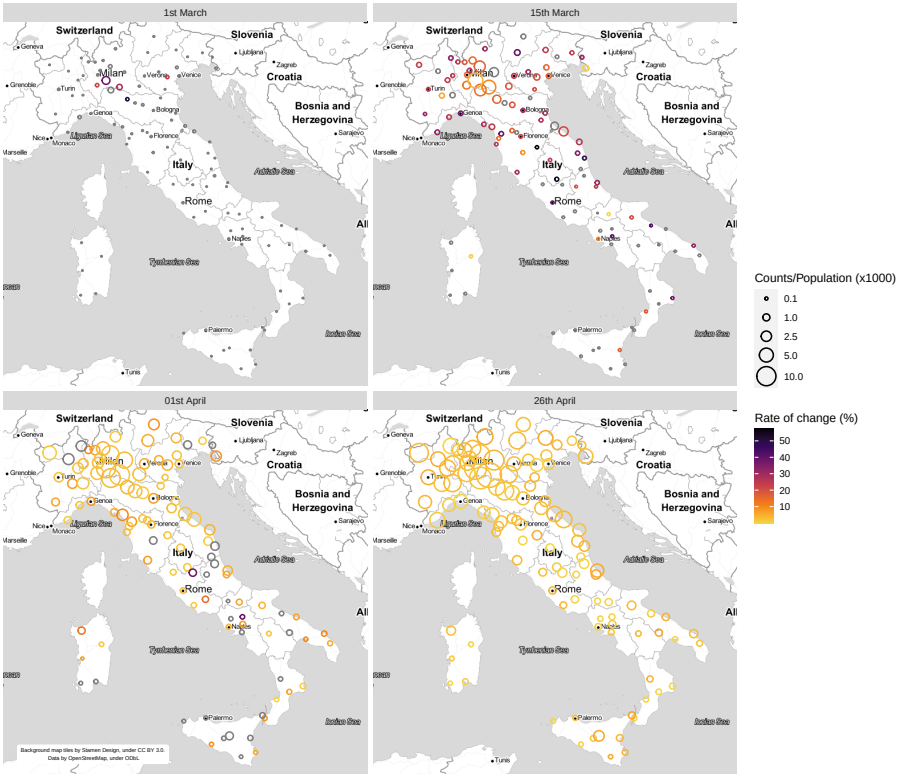
Figure 2: Observed and fitted cumulative number of reported SARS-CoV-2 infections at four different dates for the Italian province of Perugia



To better assess the evolution of the pandemic over time in the entire country, we run `Epilocal` on all provinces at different previous dates. Figure 3 shows the overall results plotted on a map of the whole country at these dates. The area and color of each circle correspond to the estimated magnitude and the rate of growth of the reported cases, respectively, as described in Subsection 2.5.

This figure provides an immediate visualization of the different stages of the COVID-19 pandemic at the Italian subnational level over time. In the first weeks of March, few provinces in the north of Italy were affected by the virus, with a fast rate of diffusion. As time went by, northern provinces were characterized by higher numbers of counts (corresponding to larger circles), especially in the Padana Valley, with relatively lower rates of change. Conversely, provinces in the center and south of Italy showed lower counts, but with higher rates of increase in later weeks.

Figure 3: **Maps of Italy at four different time points with circles indicating, for each province, the magnitude of reported SARS-CoV-2 infections (circle area) and its rate of change (circle color)**



Note: Gray circles correspond to observed data (no appropriate model was found).

One should expect variations in fit across different regions to be likely due to the different magnitude of the counts. Diverse shapes can be related to the different stages of the epidemic that each region is experiencing (including the starting date), but also to the different containment measures, region sizes, population densities, and the effect of local and national transportation infrastructure.

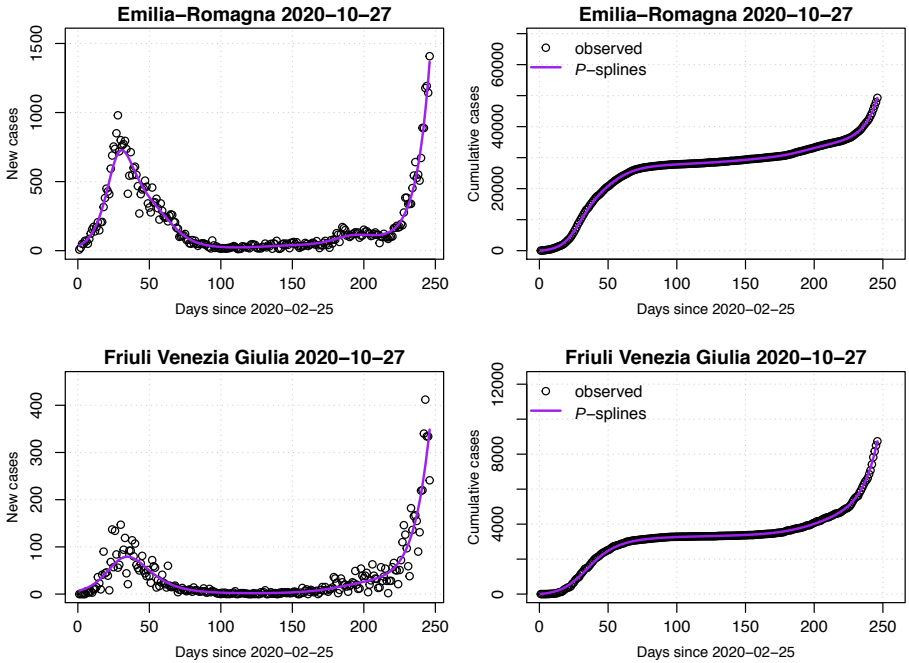
One may also note a few jumps in the absolute frequencies in some of the plots. This is likely due to gaps and bunching in the processing of, e.g., swabs, as well as in

the reporting of these administrative data from different parts of the country. Indeed, weekend effects have been routinely observed in Italy.

Similar results as those shown in Figures 1–3 can be obtained using data at the regional level. In addition to the reported cases, the region-level analysis further allows us to monitor the evolution of cumulative deaths and current hospitalizations (separately for intensive care units or not). Sample plots at the regional level are provided in Appendix B.

We now show the results of the nonparametric analysis, which we run until October 27, 2020. Figure 4 shows the output for two specific regions: Emilia-Romagna and Friuli Venezia Giulia. The figure reports the new as well as the cumulative number of positive infected individuals in the two regions, starting from the first available date in the dataset. The goodness-of-fit of the flexible P-spline approach clearly emerges from the figure.

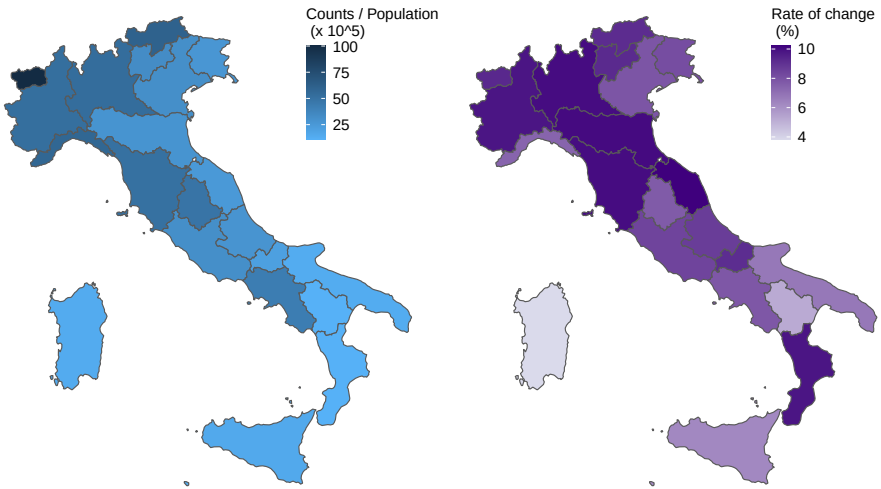
Figure 4: **Sample of plots for two Italian regions, showing the new (left panels) and cumulative (right panels) number of reported infected SARS-CoV-2 cases over time since the first date available (February 25, 2020)**



Note: The curves show the fitted values from the nonparametric P-spline approach.

Similar to the parametric analysis, we can extract and visualize the full results of `Epilocal` on the map of Italy. Figure 5 shows the two metrics of interest – magnitude of new cases and their rate of change – in two separate maps. This visualization allows for a direct assessment of the metrics free of the overlaps of the circles in Figure 3. Important regional variations clearly emerge from these maps.

Figure 5: **Maps of Italy showing, for each region, the magnitude of newly reported SARS-CoV-2 infections (left) and their rates of change (right) on October 27, 2020**



4. Conclusions

We have introduced `Epilocal`, a software tool that allows for a direct visualization and monitoring of the spread of the COVID-19 pandemic at the subnational level in Italy. As in any modeling framework, our analyses are also influenced by the available data, and here we discuss some potential limitations of our study.

First, it should be noted that the data provided by the Dipartimento della Protezione Civile (2020) refer to the number of reported cases, which clearly underestimates the true number of positives in the population. Therefore, the use of the terms *prevalence* and *incidence* would not be appropriate when analyzing these data. One might consider the ratio of the reported cases to the number of performed tests instead. However, our analyses are necessarily based on public data whose very definition and sampling processes are not well documented, and that are subject to changes over time. For example, the targets of swabs change over time, as one moves from the initial emergency situation, in which only symptomatic subjects are tested, to the one in which contact tracing is per-

formed, to the one in which local population-wide screening is performed in some areas. As a consequence, it is not clear how one could interpret such ratios (positives on tested) reliably. Note also that such ratios can clearly be increasing or decreasing, so that models for noncumulative binary data would seem more natural for such analyses.

Furthermore, the data do not allow us to ascertain the date of onset of the infection, so any inference at the reported level will suffer from a delay relative to the trend of the infections in the population (see, e.g., Wu and McGoogan 2020). Lastly, the precise attribution of deaths as COVID-related and the lack of details (home vs. hospital) on the flows in and out of the categories, such as number hospitalized and number recovered/dismissed, all suggest some caution in the interpretation of these data.

Nonetheless, we believe that our analyses have the potential for capturing useful trends for the monitoring and management of the disease burden on the health care system. In addition to modeling the number of reported cases, our software also produces analyses for the other counts available at the regional level, such as the cumulative number of COVID-positive deceased subjects and the current number of COVID-positive patients in ICU and non-ICU departments at any given time. For the hospitalization counts, one clearly does not expect a monotonic trend. So the model selection for these variables does not prevent the algorithm from selecting log-polynomials with negative rate of changes as the best models, as we have done for the cumulative counts.

There are many possible alternative approaches to modeling these reported infections data. One option, which allows for forecasting from dynamic models, is described in Chiogna and Gaetan (2002) and was recently implemented at github.com/cgaetan/COVID-19. Furthermore, Agosto et al. (2020) proposed a Poisson auto-regression for the daily number of new reported cases.

Note that here we have assumed independence in the observed counts across different provinces or regions, although the spatial spread of diseases is widely documented. In the absence of detailed data on the inflow and outflow of population migration, one could try to improve the model fit by including the baseline infrastructure of railways, highways, and airports across provinces. Not accounting for such factors might falsely attribute transportation-related factors to time. One possibility could be to entertain models with spatial correlation across provinces, but the spread of the epidemic likely follows transportation routes rather than pure distance among provinces, so it would not be trivial to implement such an extended, global model over a country. This would be further complicated by the large number of geographical units present on the Italian territory. In addition, the starting times of the local epidemics in different regions clearly seem to differ. Indeed, as shown in our results (Figure 1), the choice of fitting separate regression models on the log scale allows one to capture local epidemics that may (and do in our case) start at different times in different locations. The approach that we have followed here uses potentially different starting times for the different areas while providing raw estimates of growth for the same (final) day separately for each province or region.

Nonetheless, extending our software to explicitly consider spatial autocorrelation seems to be a promising avenue for future work, which would require additional lower-level geographic information as well as methodological extensions.

As a general comment, non-pharmaceutical measures (such as school closures, travel bans, and lockdowns) are expected to induce changes in the observed curves. A growing body of literature is indeed investigating the effects of such interventions on the spread of COVID-19 (see, e.g., Flaxman et al. 2020; Dehning et al. 2020). Such analyses are typically based on complex (and delicate) compartmental models and/or agent-based models, which focus on individual-level dynamics. Specifically, extensions of well-established Susceptible-Infectious-Recovered (SIR) and Susceptible-Exposed-Infectious-Recovered (SEIR) models have been proposed to model the spread of COVID-19 (see, e.g., Flaxman et al. 2020; Giordano et al. 2020; Lin et al. 2020). This is a rather different approach as compared to our monitoring tool, which takes a more demographic perspective based on a Poisson regression at the aggregate level. Moreover, we do not include covariates in our regression, other than the time component of the epidemic (captured by parametric or nonparametric functional forms of the hazard function) and the exposed population (included as an offset term in the regression). We decided not to include additional covariates in our framework as our goal is to describe, rather than explain, the time series of COVID-19 cases and deaths. The flexibility of the parametric and nonparametric functions is powerful enough to capture the observed data well (see, e.g., Figures 1, 2, and 4).

Notably, our results show that the selected models fit the data very well. As the dataset grows, however, the parametric approach may become too rigid to describe the observed data, and the misfit may bias the rate of growth estimates reported in the maps. In such instances, especially when analyzing longer time series, we recommend switching to the nonparametric analysis. It is further worth mentioning that the estimation of the log-polynomial models is achieved practically every time (except in the presence of very few observed data points). However, any model used in such ongoing monitoring will need to be assessed closely as the days go by, so that they may continue to provide a satisfactory fit for all provinces. In particular, the models for the cumulative counts should be closely monitored for their switching over time to higher-degree polynomials and to the logistic curve formulation beyond what is suggested from the automatic model selection. Such counts may also be modeled with other shapes of rescaled cumulative distribution functions, should the data suggest that. This would be particularly true if the data were to not satisfy the expected symmetry implicit in the logistic curve or in other cumulative distribution functions. Alternatives could be the complementary log-log curve or the generalized extreme value distribution.

As with any statistical model, there exists a risk for misuse and/or over interpretation of the results. Users of `Epilocal` should keep this point in mind, so that the interpretation of the results does not go beyond the description of the spread of COVID-19. Specif-

ically, assessment of causal effect of specific interventions should not be pursued here. Moreover, users should resist the temptation to produce forecasts beyond a few days, as such complex phenomena can easily reveal very quick changes in pace. To help users avoid this risk, we have limited the length of the step-ahead forecast in the out-of-sample forecasting error minimization to a maximum of seven days. The software also provides warning messages whenever the model-fitting procedure fails to converge.

An important advantage of our tool is that the software runs in a few seconds. Repeating the analyses daily allows for the production of a new map for each day, and the comparison of such maps allows for the appreciation of the spatial dynamics of the epidemic across Italy over time (see, e.g., Figure 3). For this purpose, and to provide a comprehensive visualization of the software's results, our shinyapp is available at `https://ubasellini.shinyapps.io/EPILOCAL/`. In the app, users are prompted to select which type of output they wish to analyze (map of the whole country at the region or province level, or individual plots for any geographical unit), as well as to customize their choices in terms of dates for the analysis, variable of interest, and type of analysis (parametric or nonparametric).

Finally, users can download the `R` codes to reproduce the results presented in this article and monitor the evolution of the pandemic (see supplementary materials for this article). Moreover, the `R` codes could be adapted without much work to monitor the COVID-19 epidemic in other countries, or for future outbreaks of other infectious diseases.

5. Acknowledgments

We thank Marcello Pagano, Elena Savoia, Rino Bellocco, Alessia Melegaro, Marília Nepomuceno, and Giancarlo Camarda, as well as three anonymous reviewers and the associate editor, for useful discussions and comments on previous versions of this work.

References

Agosto, A., Campmas, A., Giudici, P., and Renda, A. (2020). Monitoring Covid-19 contagion growth in Europe. Brussels: Centre for European Policy Studies (CEPS working paper).

Basellini, U. and Camarda, C.G. (2019). Modelling and forecasting adult age-at-death distributions. *Population Studies* 73(1): 119–138. doi:10.1080/00324728.2018.1545918.

Brillinger, D.R. (1986). A biometrics invited paper with discussion: The natural variability of vital rates and associated statistics. *Biometrics* 42(4): 693–734. doi:10.2307/2530689.

Camarda, C.G. (2012). MortalitySmooth: An R package for smoothing poisson counts with P-splines. *Journal of Statistical Software* 50: 1–24. doi:10.18637/jss.v050.i01.

Camarda, C.G. (2008). Smoothing methods for the analysis of mortality development [PhD thesis]. Madrid: Universidad Carlos III de Madrid, Department of Statistics.

Camarda, C.G. (2019). Smooth constrained mortality forecasting. *Demographic Research* 41(38): 1091–1130. doi:10.4054/demres.2019.41.38.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R*. R package version 1.4.0.2.

Chiogna, M. and Gaetan, C. (2002). Dynamic generalized linear models with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51(4): 453–468. doi:10.1111/1467-9876.00280.

Currie, I.D., Durban, M., and Eilers, P.H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* 4(4): 279–298. doi:10.1191/1471082x04st080oa.

Dehning, J., Zierenberg, J., Spitzner, F.P., Wibral, M., Neto, J.P., Wilczek, M., and Priesemann, V. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* 369(6500): 1–9. doi:10.1126/science.abb9789.

Dipartimento della Protezione Civile (2020). Dataset of COVID-19 infected cases in Italy by province [electronic resource]. github.com/pcm-dpc/COVID-19.

Ebeling, M. (2018). How has the lower boundary of human mortality evolved, and has it already stopped decreasing? *Demography* 55: 1887–1903. doi:10.1007/s13524-018-0698-z.

Eilers, P.H. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science* 11(2): 89–102. doi:10.1214/ss/1038425655.

Eurostat (2020). GISCO: Geographical information and maps [electronic resource]. Brussels: European Commission. ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts.

Flaxman, S., , Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., Monod, M., Ghani, A.C., Donnelly, C.A., Riley, S., Vollmer, M.A.C., Ferguson, N.M., Okell, L.C., and Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* 584(7820): 257–261. doi:10.1038/s41586-020-2405-7.

Garnier, S. (2018). *viridis: Default color maps from 'matplotlib'*. R package version 0.5.1.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Filippo, A.D., Matteo, A.D., and Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine* 26(6): 855–860. doi:10.1038/s41591-020-0883-7.

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics* 26(3): 443–458. doi:10.1007/s00180-010-0217-1.

Istat (2020). Resident population for year 2019 [electronic resource]. demo.istat.it/pop2019/index1.html.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal* 5(1): 144–161. doi:doi:10.32614/rj-2013-014.

Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S.S., Wang, M.H., Cai, Y., Wang, W., Yang, L., and He, D. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases* 93: 211–216. doi:10.1016/j.ijid.2020.02.058.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models (2nd ed.)*. Boca Raton: Chapman and Hall/CRC.

Ouellette, N. and Bourbeau, R. (2011). Changes in the age-at-death distribution in four low mortality countries: A nonparametric approach. *Demographic Research* 25(19): 595–628. doi:10.4054/DemRes.2011.25.19.

R Core Team (2020). R: A language and environment for statistical computing [electronic resource]. www.R-project.org/.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2): 461–464. doi:10.1214/aos/1176344136.

WHO (2020). Coronavirus disease (COVID-2019) situation reports. Situation report – 55. March 15, 2020 [electronic resource]. www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.

Wu, Z. and McGoogan, J.M. (2020). Characteristics of and important lessons from the Coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* 323(13): 1239–1242. doi:10.1001/jama.2020.2648.

Appendix A: Additional formulas

In this section, we provide some additional formulas related to Section 2.2 of the manuscript.

The formulas for the first derivatives of the log-polynomials models of degrees two and three allow for the immediate calculation of the time(s) $\tilde{t}$ at which the local epidemic will reach a maximum (or minimum). These local maxima (or minima) are, for the two models,

$$\tilde{t} = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2} \text{ and } \tilde{t}_{1,2} = \frac{-2\widehat{\beta}_2 \pm \left[4\widehat{\beta}_2^2 - 12\widehat{\beta}_1\widehat{\beta}_3\right]^{1/2}}{6\widehat{\beta}_3}, \tag{A-1}$$

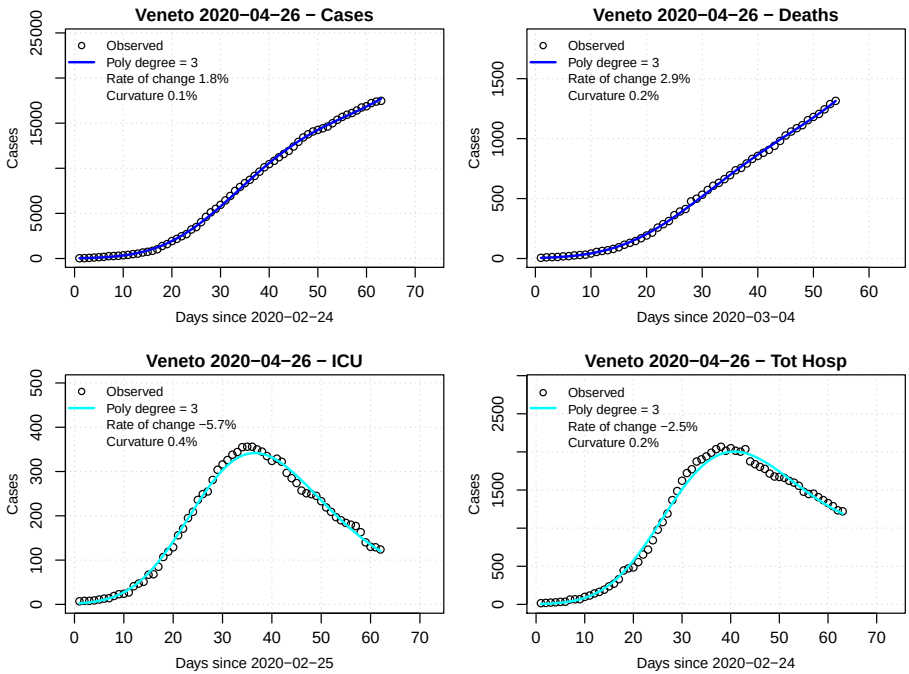
when $4\widehat{\beta}_2^2 - 12\widehat{\beta}_1\widehat{\beta}_3 > 0$. These predicted local or global maxima and minima are likely to be quite unstable until enough data are collected, and they should be interpreted with caution. As time goes by, one expects such estimated times to have increasing precision. More generally, due to the polynomial shape that we have chosen for the regression component, we expect these models to only be adequate to describe the epidemic during the initial growth phase for the first-degree model, and only up until the maximum prevalence of the epidemic for the second-degree and third-degree models. In particular, any predicted decreases in the cumulative number of cases are clearly meaningless since recovered cases are not accounted for in the data. In general, any forecasting beyond the last observed time point should be performed with caution.

Appendix B: Additional results

Here we provide some additional results of the analysis presented in the manuscript or obtainable with `Epilocal`.

Figure B-1 shows the observed and fitted SARS-CoV-2 numbers for four variables available at the region-level analysis (here Veneto), namely reported cumulative cases, cumulative deaths, current number of ICU admissions, and current number of total hospitalizations.

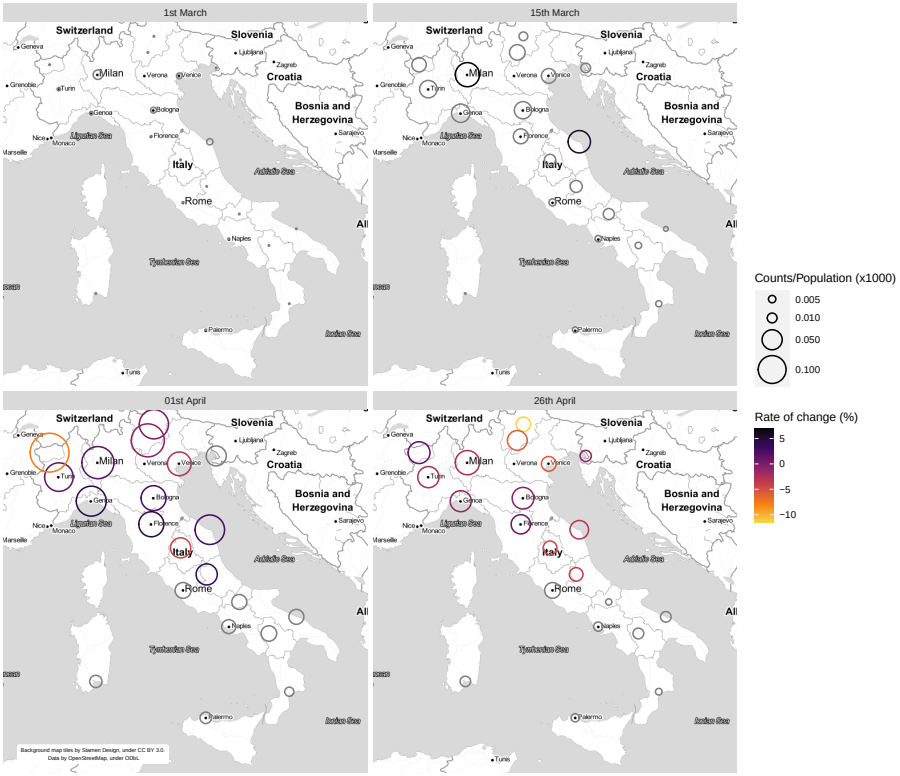
Figure B-1: Observed and fitted numbers of SARS-CoV-2 cumulative reported infected cases (top left), cumulative deaths (top right), current admissions to intensive care units (ICU, bottom left), and current hospitalisations (bottom right) for the Italian region of Veneto



Similarly to what is shown in Figure 3 of the manuscript, we can plot results for all regions on a map of the entire country. Figure B-2 shows the regional evolution of the current admissions to ICU departments at four different dates. Since here we are mod-

eling a noncumulative response variable, negative growth rates are observable (see also bottom left panel of Figure B-1). The figure shows that ICU admissions first increased at fast rates in the first weeks of March and then decreased in magnitude (with negative growth rates) in April.

Figure B-2: Maps of Italy at four different time points with circles indicating, for each region, the magnitude of SARS-CoV-2 ICU admissions (circle area) and its rate of change (circle color)



Note: Gray circles correspond to observed data (no appropriate model found).

Finally, we provide the results for the estimated first derivatives from the parametric approach. For these estimates, we compute 95% confidence intervals using the delta method on Equations (3) and (9) in the manuscript. Figure B-3 shows the results of this procedure for all the provinces on the shorter time series analyzed in our paper (i.e., until

April 26, 2020). The estimated first derivatives at the end of the time series are rather precise, and while in some instances the intervals overlap, they generally allow us to distinguish province-specific results in several cases.

Figure B-3: Estimated first derivatives with delta method based on 95% confidence intervals at the end of the time series (April 26, 2020) for 107 Italian provinces