

Misspecified modeling of subsequent waves during COVID-19 outbreak: A change-point growth model

Paolo Girardi¹ | Luca Greco²  | Laura Ventura³

¹ Department of Developmental and Social Psychology, University of Padova, Padova, Italy

² University Giustino Fortunato, Benevento, Italy

³ Department of Statistical Sciences, University of Padova, Padova, Italy

Correspondence

Luca Greco, University Giustino Fortunato, Viale R. Delcogliano 12, 82100 Benevento, Italy.
Email: l.greco@unifortunato.eu

Funding information

Università degli Studi di Padova,
Grant/Award Number: BIRD197903



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

In the analysis of cumulative counts of SARS-CoV-2 infections, such as deaths or cases, common parametric models based on log-logistic growth curves adapt well to describe a single wave at a time. Unfortunately, in Italy, as well as all over the globe, from February 2020 to March 2021 more than one wave has been observed. In this paper, we propose a method to fit more than one wave in the same model. In particular, we discuss an approach based on a change-point model in a pseudo-likelihood framework that takes into account some model misspecification issues, such as those concerning the assumption of Poisson marginals and those relating to overdispersion and autocorrelation. An application to data collected in Italy is discussed.

KEYWORDS

HAC, independence log-likelihood, log-logistic curve, overdispersion, Poisson

1 | INTRODUCTION

COVID-19 syndrome soon turned into a global-scale crisis. Its worldwide large diffusion is due to the high rates of virus transmission and the appearance of several variants of the original SARS-CoV-2 coronavirus. After a first epidemic outbreak between March and April 2020, Italy has brought a renewed spread of COVID-19 during the summer 2020 and in the following months, as reported by other countries around the world (Xu & Li, 2020).

The larger mobility combined with a relaxed easing of the previously imposed restrictive measures to allow a return to *normality* and to support the economy, made numbers of new cases of infection, hospitalizations and deaths grow again, also in those countries where the epidemic stood on very low numbers for several consecutive weeks, such as in Italy (Girardi et al., 2020b). In Italy, the growth of the contagion has been massive starting from autumn and assessed on extremely large values during all winter. The number of daily cases peaked in November 13, whereas the daily hospitalizations and deaths reached the maximum value 10 and 20 days later, respectively. In order to mitigate the diffusion of the pandemic, the Italian Government introduced progressive and differentiate strategies, such as targeted lockdowns and other severe actions, as school closure and a prolonged stop to catering and unessential goods selling activities, in addition to the strict restrictions aimed to regulate social behaviors, transportation, sport events, and all those circumstances

characterized by a high risk of gathering. Nowadays, we are still running after the contagion. All containment measures aimed to manage the pandemic turned out to be effective only to a limited extent and often exceedingly costly for their economic and social consequences. The beginning of the vaccine campaign seems to be the only reliable remedy to reduce the spread of the contagion and the burden of severe infections, hospitalizations, and deaths.

The availability of public data about the COVID-19 outbreak has soon represented a crucial modeling challenge for statisticians all over the world, in order to provide meaningful descriptions and predictions. In particular, there is a widespread interest in modeling the past and current wave of infections or deaths and nowcasting possible future waves. A simple way to model a single epidemic wave at a time is to resort to common parametric models, often based on log-logistic growth curves (see among others, Cabras, 2020; Di Loro et al., 2020; Girardi et al., 2020a). Other nowcasting strategies have been discussed in Farcomeni et al. (2021), who proposed an ensemble approach to predict intensive care admissions, Schneble et al. (2021), who developed a model to fit and predict mortality rates related to COVID-19 infections, Kaxiras and Neofotistos (2020) who described a multiple wave forced-SIR mode, Günther et al. (2021), who suggested a hierarchical Bayesian model aimed to account for delays between disease onset and case reporting, Kim et al. (2020) who presented an overdispersed polynomial Poisson regression model (also with covariates) and discussed several approaches to obtain reliable prediction intervals.

Here, we propose a change-point growth model to fit cumulative incidence data, such as infections and deaths, that is able to catch subsequent waves of the pandemic. More precisely, we discuss an approach based on a change-point model in a pseudo-likelihood framework, that allows us to account for model misspecification issues, both with respect to the assumptions regarding marginal distributions and independence. The model is meant to describe the main features of the observed trends in the data and, in particular, to give evidence about the time when different waves were more likely to be originated. The latter estimate could aid the investigation of the main causes leading to different waves.

The rest of the paper is organized as follows. The data are described in Section 2. Section 3 introduces our model, while Section 4 discusses some inferential issues. An application to Italian data is shown in Section 5. Section 6 provides some conclusions and possible extensions.

2 | DATA

Italian COVID-19 epidemic data are available since February 24, 2020 from a GitHub repository daily updated by the *Dipartimento della Protezione Civile*.¹ The data give information about a large number of variables, such as the cumulative and daily counts of laboratory-confirmed infections, the number of swabs, the daily number of hospitalizations and persons hospitalized in intensive care units, the cumulative deaths' toll, and the number of hospitalizations. Data are available both at the national and regional level, so allowing comparisons across the country. When looking at the time series of both incidence (cases and deaths) and prevalence (hospitalization) data, we can observe some common patterns. Let us focus on the cumulative number of reported deaths due to COVID-19. The data are displayed in the left panel of Figure 1 from February 24, 2020 to March 10, 2021. Their behavior can be depicted as a sequence of a couple of S-shaped curves, each of them representing one wave of fatalities due to COVID-19: the second wave starts after that the first wave has reached a sort of plateau. Moreover, in each wave, we first observe a quick exponential increase in the counts, but then their growth decelerates and becomes logistic from some point on, so moving toward an upper bound. The picture of subsequent waves appears more evident when looking at daily counts, as displayed in the right panel of Figure 1. The first exponential growth determines the first raise, then the logistic increase corresponds to the decrease in daily incidence after the peak.

3 | METHODS

Here, we assume a parametric model that allows to fit the subsequent waves simultaneously but also to estimate the time in which is more likely that the transition from the first to the second wave occurred. We believe this is a crucial aspect in the comprehension of the evolution of the pandemic, to evaluate the effectiveness of past policies and to design new interventions.

¹ <https://github.com/pcm-dpc/COVID-19>

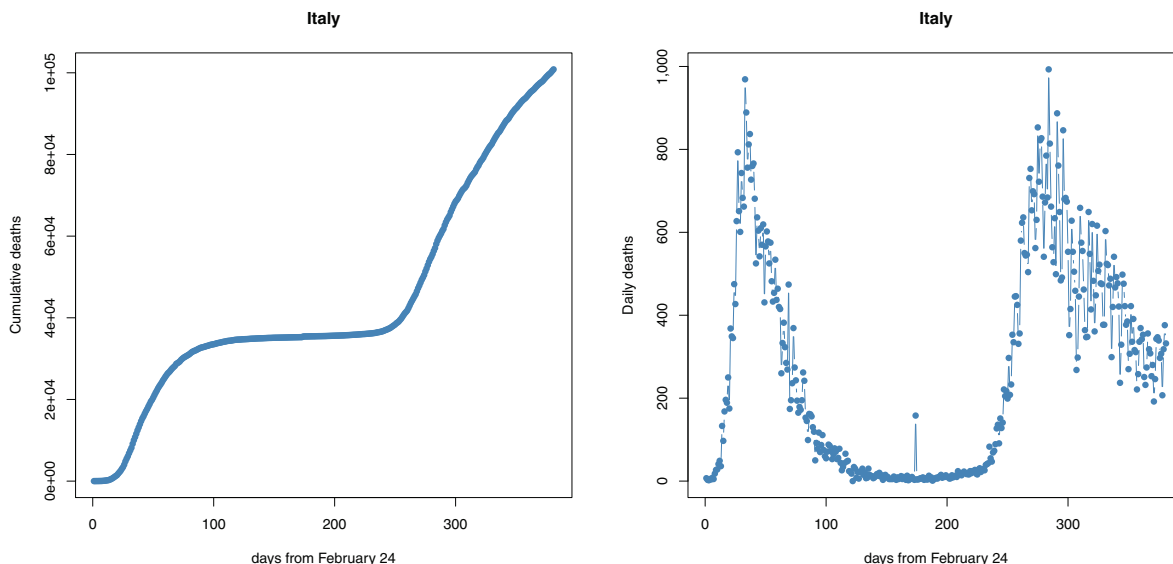


FIGURE 1 Time series of the number of cumulative (left) and daily (right) deaths in Italy from February 24, 2020 to March 10, 2021

3.1 | Pseudo-likelihood modeling of one wave of cumulative counts

The behavior of cumulative incidence data can be modeled by the well-known five parameters log-logistic growth function (Ritz et al., 2015), given by

$$\mu(t; \theta) = c + \frac{d - c}{\left[1 + \left(\frac{t}{e}\right)^b\right]^f}, \quad \theta = (b, c, d, e, f), \quad b < 0, \quad c, d, e, f > 0, \quad (1)$$

which is expressed as a function of time t . The presence of five parameters in (1) allows a great modeling flexibility. The parameters b , e , and f determine the shape of the growth function, c returns the lower asymptote, in the sense that $c = \lim_{t \rightarrow 0} \mu(t; \theta)$, and d , on the opposite, represents the upper asymptote, that is a direct measure of the final size of the pandemic, with $d = \lim_{t \rightarrow \infty} \mu(t; \theta)$. When $f = 1$, the model is such that $\mu(e; \theta) = (d + c) - \mu(e; \theta) = (d + c)/2$ and the log-logistic model is said to be symmetric. The first derivative $\mu'(t; \theta) = \partial \mu(t; \theta) / \partial t$ describes the behavior of the growth rate over time and allows to model the pattern of daily incidence data. Actually, according to a first-order Taylor expansion, we have that

$$\mu(t; \theta) - \mu(t^0; \theta) = \mu'(t^0; \theta)(t - t^0) + o(|t - t^0|). \quad (2)$$

From (2), after setting $t^0 = t - 1$, it follows that the approximation $\mu(t; \theta) - \mu(t - 1; \theta) \approx \mu'(t - 1; \theta)$ holds.

3.2 | Misspecification I: Independence

Let Y_t^c be the random variable that describes the cumulative number of counts data, with $t = 1, \dots, T$. Let $y^c = (y_1^c, y_2^c, \dots, y_T^c)$ denote the observed series of cumulative counts. The first assumption of our model concerns the specification of the marginal model for Y_t^c . Here, it is assumed that Y_t^c obeys a Poisson distribution with expected value $\mu(t; \theta)$ given by the log-logistic growth curve (1).

Since the nature of the data is such that $y_{t+1}^c \geq y_t^c, \forall t$, the assumption of independence is questionable. In this respect, in the following, we pursue an approach based on a suitable *composite* log-likelihood function, based on marginal probability or density functions (Varin et al., 2011). In particular, we consider the simplest composite marginal likelihood, that is the

pseudo-log-likelihood constructed under the working independence assumption, that is,

$$\ell_I(\theta) = \sum_{t=1}^T \log p(y_t^c; \theta), \quad (3)$$

sometimes referred to in the literature as the *independence* log-likelihood (Chandler & Bate, 2007). The independence likelihood permits inference only on marginal parameters. A related approach has been also adopted in Girardi et al. (2020a), while the reader is pointed to Di Loro et al. (2020) for a likelihood-based nowcasting strategy well suited for count incidence data.

The validity of inference about θ using the independence log-likelihood (3) can be justified invoking the general theory of unbiased M-estimating functions (see Varin et al., 2011). Actually, $\ell_I(\theta)$ shares the properties of a log-likelihood function stemming from a misspecified model. In particular, the maximum composite likelihood estimate (MCLE)

$$\hat{\theta}^I = \operatorname{argmax}_{\theta} \ell_I(\theta) \quad (4)$$

can be also defined as the root of the composite score equation

$$u_I(\theta) = \sum_{t=1}^T u_t(y_t^c; \theta) = \frac{\partial \ell_I(\theta)}{\partial \theta} = 0. \quad (5)$$

The corresponding estimator is asymptotically normally distributed with mean θ and covariance matrix

$$V(\theta) = G(\theta)^{-1} = K(\theta)^{-1} J(\theta) K(\theta)^{-1}, \quad (6)$$

where $G(\theta)$ is the Godambe information matrix, with

$$K(\theta) = E(-\partial u_t(\theta) / \partial \theta^\top), \quad J(\theta) = \operatorname{var}(u_t(\theta)) = E(u_t(\theta) u_t(\theta)^\top). \quad (7)$$

Composite likelihood versions of Wald, score and suitably adjusted likelihood ratio statistics can be obtained that all share the classical asymptotic chi-squared distribution (see Pace et al., 2011, and references therein). Standard error evaluation requires consistent estimation of the matrices $J(\theta)$ and $K(\theta)$: for large T , they may be estimated by

$$\hat{J} = \frac{1}{T} \sum_{t=1}^T u_t(y_t^c; \hat{\theta}^I) u_t(y_t^c; \hat{\theta}^I)^\top, \quad \hat{K} = \frac{1}{T} \sum_{t=1}^T \left. \frac{\partial u_t(y_t^c; \theta)}{\partial \theta^\top} \right|_{\theta = \hat{\theta}^I}. \quad (8)$$

Moreover, when it is possible to simulate from the model, the matrices $J(\theta)$ and $K(\theta)$ can be estimated through Monte Carlo samples (Cattelan and Sartori, 2016; Varin et al., 2011).

3.3 | Misspecification II: Overdispersion and autocorrelation

The classical sandwich estimator of $V(\theta)$ may not be able to catch some important aspects in the data and lead to underestimate uncertainty. First, the assumption of Poisson marginals may lead to neglect possible overdispersion in the data. Furthermore, the nonstationarity of the series of cumulative counts data suggests some adjustments to take into account autocorrelation, but also some type of heteroskedasticity that is not completely caught by the Poisson assumption.

In order to deal with these two issues, we propose a couple of possible adjustments. The first correction term comes from the general theory of regression models for counts data (Agresti, 2015). By paralleling the approach based on quasi-likelihood inference, one could take into account overdispersion by inflating the sandwich variance-covariance matrix $V(\theta)$ by a dispersion parameter estimate, obtained as

$$\hat{\phi} = \frac{1}{T-p} \sum_{t=1}^T \frac{[y_t^c - \mu(t; \hat{\theta}^I)]^2}{\mu(t; \hat{\theta}^I)}, \quad (9)$$

where $p = \dim(\theta)$. Then, the estimated variance–covariance matrix is

$$\widehat{\text{var}}_a(\hat{\theta}^I) = \hat{\phi} V(\hat{\theta}^I). \quad (10)$$

The second robust covariance estimation consists in the evaluation of the Newey–West heteroskedasticity and autocorrelation consistent (HAC) sandwich estimate of the variance–covariance matrix (Newey & West, 1987; Zeileis, 2004, 2006). This procedure allows us to take into account the unknown autocorrelation and heteroskedasticity patterns of the cumulative counts at hand (Dorward et al., 2021; Hardin, 1998). In this case, the fitted variance–covariance matrix is denoted as

$$\widehat{\text{var}}_b(\hat{\theta}^I) = V_{HAC}(\hat{\theta}^I). \quad (11)$$

3.4 | Change-point growth model

Let us consider the situation with two waves. Assuming that each wave can be modeled according to (1), the expected value of Y_t^c is given by

$$\mu(t; \tau) = \begin{cases} \mu(t; \theta_1), & t \leq t_0, \\ \mu(t - t_0; \theta_2), & t > t_0, \end{cases} \quad (12)$$

where $\tau = (\xi, t_0)$, with $\xi = (\theta_1, \theta_2)$. The function (12) is characterized by one change point at an unknown time t_0 where the mean switches from $\mu(t; \theta_1)$ to $\mu(t; \theta_2)$. Moreover, in the second branch the lower asymptote is fixed as $c_2 = \mu(t_0; \theta_1)$ so that $\mu(t; \theta_2) \geq \mu(t; \theta_1) \forall t$, and equality holds at $t = t_0$. Therefore, a fourparameter log-logistic model is fitted in the second wave, with $\theta_2 = (b_2, d_2, e_2, f_2)$, while $\theta_1 = (b_1, c_1, d_1, e_1, f_1)$.

The independence log-likelihood function for τ is

$$\ell_I(\tau) = \sum_{t=1}^T [z_t \log p(y_t^c; \mu_1(t; \theta_1)) + (1 - z_t) \log p(y_t^c; \mu_2(t - t_0; \theta_2))], \quad (13)$$

where $z_t = 1$ for $t \leq t_0$ and zero otherwise. The change point can be estimated by a composite log-likelihood profile approach as

$$\hat{t}_0^I = \operatorname{argmax}_{t_0} \ell_{Ip}(t_0), \quad (14)$$

where $\ell_{Ip}(t_0) = \ell_I(\hat{\xi}_{t_0}^I, t_0)$ and $\hat{\xi}_{t_0}^I$ is the constrained MCLE of the branches parameters for fixed change point. Then, the unconstrained MCLE of ξ is obtained as $\hat{\xi}_{\hat{t}_0}^I$.

4 | INFERENCE ISSUES

Standard errors for the components of $\hat{\theta}_j^I$ and for the fitted values $\mu_j(t, \hat{\theta}_j^I)$, $j = 1, 2$, can be evaluated conditionally on \hat{t}_0^I and using the asymptotic distribution of $\hat{\theta}_j^I$, that is a normal distribution centered at the MCLE $\hat{\theta}_j$ with variance–covariance matrix $\widehat{\text{var}}_h(\hat{\theta}_j^I)$, where $\widehat{\text{var}}_a(\hat{\theta}_j^I) = \hat{\phi}_j V(\hat{\theta}_j^I)$ for $h = a$ and $\widehat{\text{var}}_b(\hat{\theta}_j^I) = V_{HAC}(\hat{\theta}_j^I)$ for $h = b$, $j = 1, 2$.

Wald-type asymptotic confidence intervals (CIs) for the mean function $\mu_j(t; \theta_j)$ and its first derivative $\mu_j'(t; \theta_j)$, $j = 1, 2$, can be obtained using the delta method. For instance,

$$\text{var}_h \left[\mu(t; \hat{\theta}_j) \right] = \left(\frac{\partial \mu_j(t; \theta_j)}{\partial \theta^\top} \right)^\top \text{var}_h(\hat{\theta}_j^I) \left(\frac{\partial \mu_j(t; \theta_j)}{\partial \theta^\top} \right), \quad h = a, b. \quad (15)$$

CIs around \hat{t}_0^I can be evaluated according to the inverse function

$$\mu^{-1}(\mu; \theta) = e \left[\left(\frac{d-c}{\mu-c} \right)^{1/f} - 1 \right]^{1/b} \quad (16)$$

at $\mu = \mu(\hat{t}_0^I; \hat{\theta}_1^I)$ and the delta method, as well.

In addition to CIs for the mean function, prediction intervals are fairly derived through a parametric double bootstrap procedure (Di Loro et al., 2020; Efron, 2004, 2012). This strategy allows us to take into account two sources of uncertainty: one stemming from parameter estimation and the other from the distribution of the data. In the double bootstrap procedure, first parameters values are drawn from the asymptotic distribution of $\hat{\xi}^I = (\hat{\theta}_1^I, \hat{\theta}_2^I)$, and then cumulative data are simulated according to a Poisson–Gamma mixture. This strategy obeys the assumptions about marginals and allows us to take into account overdispersion. It is worth noting that bootstrap is performed conditionally on the estimated change point \hat{t}_0^I . This means that prediction intervals are obtained separately for each branch of the model and uncertainty around \hat{t}_0^I is not taken into account. This strategy is coherent with the estimation technique described in Subsection 3.4. In details, the procedure can be summarized as follows:

1. Generate B realizations $\hat{\xi}_1^I, \dots, \hat{\xi}_B^I$ from the asymptotic distribution

$$N_8 \left(\hat{\xi}^I, \widehat{\text{var}}_h(\hat{\xi}^I) \right), \quad (17)$$

where $\widehat{\text{var}}_h(\hat{\xi}^I) = (\widehat{\text{var}}_h(\hat{\theta}_1^I), \widehat{\text{var}}_h(\hat{\theta}_2^I))$ is block diagonal, for $h = a, b$. The matrix is block diagonal since parameters in each branch, accordingly to the independence log-likelihood (13) for τ , have been estimated independently from each other, conditionally on the fitted change point;

2. Simulate B series of cumulative counts $y^c = (y_1^c, y_2^c, \dots, y_T^c)$ from a Poisson–Gamma mixture;
3. Prediction intervals are obtained computing pointwise percentiles.

In Step 2, cumulative counts are simulated from a Poisson–Gamma mixture with linear variance function, consistently with the adjustment described in Subsection 2.2, in order to account for overdispersion in the data. The procedure is structured as follows:

- (a) Simulate $m_t \sim \text{Gamma}(\hat{\mu}_t, \nu_t)$, with expected value

$$\hat{\mu}_t = \begin{cases} \mu(t; \hat{\theta}_1^I), & t \leq \hat{t}_0^I \\ \mu(t - t_0; \hat{\theta}_2^I), & t > \hat{t}_0^I \end{cases} \quad (18)$$

and shape parameter

$$\nu_t = \begin{cases} \lambda_1 \mu(t; \hat{\theta}_1^I), & t \leq \hat{t}_0^I \\ \lambda_2 \mu(t - t_0; \hat{\theta}_2^I), & t > \hat{t}_0^I. \end{cases} \quad (19)$$

- (b) Simulate $Y_t^c | m_t \sim \text{Pois}(m_t)$.

The values for (λ_1, λ_2) can be obtained according to

$$\begin{aligned} \text{var}(Y_t^c) &= \text{var}_{m_t} [E_{Y_t^c}(Y_t^c | m_t)] + E_{m_t} [\text{var}_{Y_t^c}(Y_t^c | m_t)] \\ &= \frac{\hat{\mu}_j}{\lambda_j} + \hat{\mu}_t = \left(\frac{1 + \lambda_j}{\lambda_j} \right) \hat{\mu}_t = \hat{\phi}_j \hat{\mu}_t \end{aligned} \quad (20)$$

with $j = 1$ for $t = 1, 2, \dots, \hat{t}_0$ and $j = 2$ for $t = \hat{t}_0 + 1, \dots, T$. We get $\lambda_j = (\hat{\phi}_j - 1)^{-1}$.

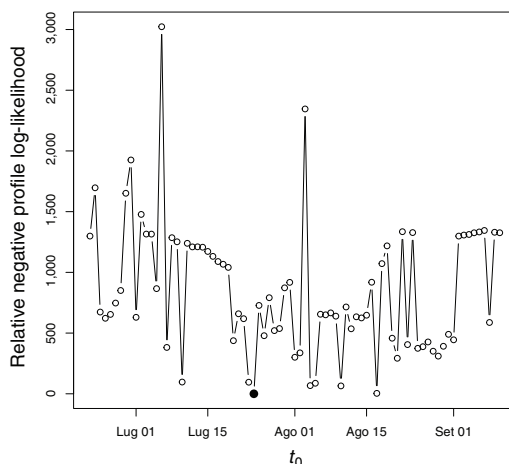


FIGURE 2 Selection of t_0 . The plain black circle gives the fitted change point

A similar approach can be used to evaluate prediction intervals for daily incidence counts y_t^d . To this end, we further assume the daily counts also stem from a (overdispersed) Poisson distribution with expected value $\hat{\mu}_t^d = \hat{\mu}_t - \hat{\mu}_{t-1}$ and a variance inflated by the parameter ϕ_j , with $j = 1$ for $t = 1, \dots, \hat{t}_0^I$ and $j = 2$ for $t = \hat{t}_0^I + 1, \dots, T$. Then, overdispersed daily incidence counts can be generated from a Poisson–Gamma mixture as follows:

- (a) simulate $m_t^d \sim \text{Gamma}(\hat{\mu}_t^d, \nu_t^d)$
- (b) simulate $Y_t^d | m_t^d \sim \text{Pois}(m_t^d)$,

with $\nu_t^d = \hat{\mu}_t^d / (\hat{\phi}_j - 1)$ and ϕ_j is estimated from the cumulative data. Actually, here we assume that $\text{var}(Y_t^d) = \phi_j \mu_t^d$ and $\text{var}(Y_t) = \text{var}(\sum_{k=1}^t Y_k^d) = \phi_j \mu_t$, and thus the dispersion parameter is the same. In view of this, we suggest to use the estimate of ϕ_j evaluated over the cumulative data, since we fit the model over cumulative and not daily counts. Furthermore, daily data exhibit a severe autocorrelation that does not allow us to obtain a reliable estimate of the dispersion parameter.

4.1 | Computational details

The analyses have been carried out using the R software environment (R Core Team, 2020). The independence log-likelihood function in (3) has been optimized according to the function `drm` available from package `drc`. The sandwich estimator has been evaluated with the function `sandwich`, whereas the sandwich estimator HAC can be obtained with the command `vcovHAC`, both available from the package `sandwich`. Wald-type CIs around the mean function, its first derivative, and its inverse function based on the delta method can be evaluated on the basis of the function `n1Confint` from package `n1WaldTest`.

5 | APPLICATION: ITALIAN DEATH COUNTS

In this section, we consider the cumulative death counts collected from February 24, 2020, until March 10, 2021. As previously stated, in this period we have observed two epidemic waves, characterizing the trajectories of infections, hospitalizations, and deaths.

The analysis aims to estimate the change point and to model the shape of the two waves for Italy, taking into account model misspecifications. Here, in order to avoid unpleasant optimization convergence issues, we set $c_1 = 0$ in $\mu(t; \theta_1)$. This choice is reasonable in our study, since we assume there were not any COVID-19 death before the date of the first reported cases.

The composite log-likelihood profile criterion in (14) is displayed in Figure 2. The change-point growth model for cumulative death counts locates the structural break \hat{t}_0^I on July 24, 2020 (see last line in Table 1). There is evidence that the second wave of deaths has started well before the end of summer 2020. This result confirms that the number of infections growth

TABLE 1 Parameter estimates with 99% confidence intervals based on the overdispersed-inflated sandwich and the HAC sandwich covariance matrix estimate

		MCLE	Overdispersed	HAC
θ_1	b_1	-3.18	-3.37 to -3.00	-3.53 to -2.84
	d_1	35,892.07	35,628.85–36,155.30	35,351.91–36,432.24
	e_1	40.20	37.72–42.68	36.57–43.82
	f_1	1.33	1.16–1.50	1.05–1.60
θ_2	b_2	-8.61	-9.97 to -7.26	-9.63 to -7.60
	d_2	106,591.97	100,717.85–112,466.09	105,019.25–108,164.69
	e_2	257.46	246.05–268.86	230.59–284.33
	f_2	3.39	2.45–4.33	1.10–5.68
t_0		07/24	07/14–08/03	07/03–08/14

Note. The entries in the last row are dates in the form month/day.

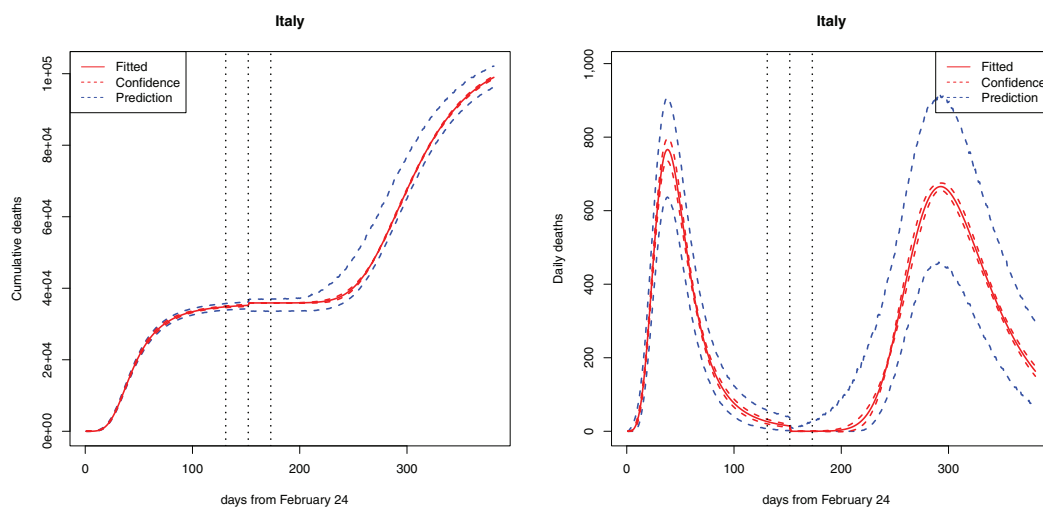


FIGURE 3 Left: cumulative deaths. Right: daily deaths. Fitted model with 0.99-level confidence and prediction intervals based on the HAC sandwich covariance matrix estimate. The dotted vertical lines give the fitted change point with the corresponding 0.99-level confidence interval

again during the first half of summer 2020. The entries in Table 1 give parameters estimates $(\hat{\theta}_1^I, \hat{\theta}_2^I)$, with 99% CIs based on both $\widehat{\text{var}}_a(\hat{\theta}_j^I)$ and $\widehat{\text{var}}_b(\hat{\theta}_j^I)$, $j = 1, 2$. On the basis of the results, it can be noted that the fitted model estimates a total amount of deaths in the second wave about three times the amount in the first wave. The observed cumulative number of deaths on March 10 was 100,842. Then, according to the fitted model there are about 7000 COVID-19 deaths that are still to happen, assuming that a third wave would have not occurred. Unfortunately, a third wave affected Italy and, at the time of writing, the deaths toll is about 120,000.

The fitted curve (12) is given in the left panel of Figure 3, together with pointwise 0.99-level Wald-type asymptotic CIs based on the HAC sandwich covariance matrix estimate and double bootstrap prediction intervals. Prediction intervals are based on 10,000 simulated samples. The right panel of Figure 3 shows the daily counts along with the first derivative of each branch of (12) and the corresponding confidence and prediction intervals. The figures stemming from the employ of the overdispersed-inflated sandwich estimate $\widehat{\text{var}}_a(\hat{\theta}^I)$ were very similar and are not displayed here. We report a satisfactory coverage accuracy equal to 99% for cumulative counts (99% also using the overdispersed-inflated sandwich estimate of the covariance matrix) and about 90% for daily counts (87% with the alternative estimate of the covariance matrix).

The benefits of the proposed misspecified modeling approach are evident when comparing the results about empirical coverages stated above with those stemming from methods that do not account for: (i) the independence adjustment made by the composite likelihood-based sandwich estimate of the variance–covariance matrix; (ii) overdispersion or autocorrelation. The entries in Table 2 give empirical coverages of 99% double bootstrap prediction intervals based on the

TABLE 2 Empirical coverages of nominal 99% double bootstrap prediction intervals based on the overdispersed-inflated sandwich, the HAC sandwich, the classical MLE, and the noninflated sandwich, over cumulative counts

	Cumulative	Daily
MLE	0.661	0.446
Sandwich	0.696	0.459
Overdispersed	0.990	0.874
HAC	0.990	0.900

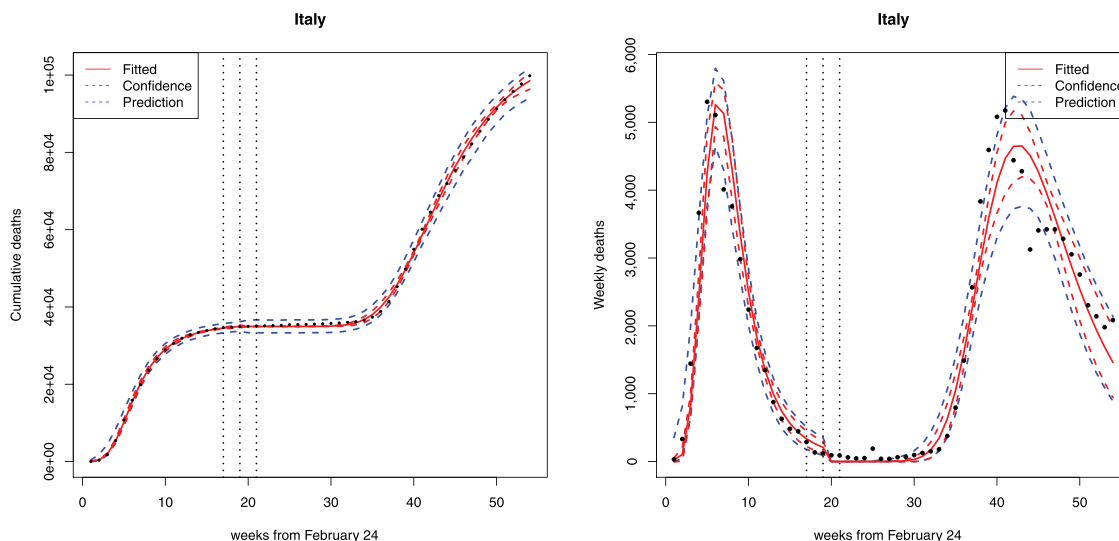


FIGURE 4 Left: cumulative weekly deaths. Right: weekly deaths. Fitted model with 0.99-level confidence and prediction intervals based on the HAC sandwich covariance matrix estimate. The dotted vertical lines give the fitted change point with the corresponding 0.99-level confidence interval

overdispersed-inflated sandwich, the HAC sandwich, the classical maximum likelihood estimate (MLE), and the noninflated sandwich, over cumulative counts. The classical MLE stems from the genuine likelihood function and violates both misspecification adjustments, while the noninflated sandwich does not account for overdispersion or autocorrelation. The gain of the proposed techniques is overwhelming.

In order to validate the findings of our analysis, we also considered aggregated data at weekly level. This strategy allows to mitigate the effect of several measurement issues in the data collection process that has given place to wide daily oscillations and an evident weekly seasonality (Bartolucci & Farcomeni, 2021). Starting from February 24, we now consider 54 consecutive weeks. The results are in strong agreement with those discussed above and are displayed in Figure 4. A 99%CI around the week t_0^t goes from 08–14/06 to 20–26/07, giving again evidence supporting a second epidemic wave already during the first half of July. Moreover, the weekly estimate (with corresponding CIs) of the parameter d_2 is close to that given in Table 1.

5.1 | Model validation

Since the choice of a log-logistic curve may result in a robust but rigid model, in particular for future prediction, we considered two types of validation analysis. In order to test the ability of the model to predict the evolution of the epidemic, the first analysis was built to verify the error size of a short to medium window forecast. For this purpose, we calculated the out-of-sample root mean squared prediction error (RMSPE) for:

- a model fitted on different time windows $t = 1, \dots, t^*$, where t^* goes from April 13, 2020 up to March 10, 2021;
- an increasing forecast horizon $K \in 1, 3, 7, 14$ days.

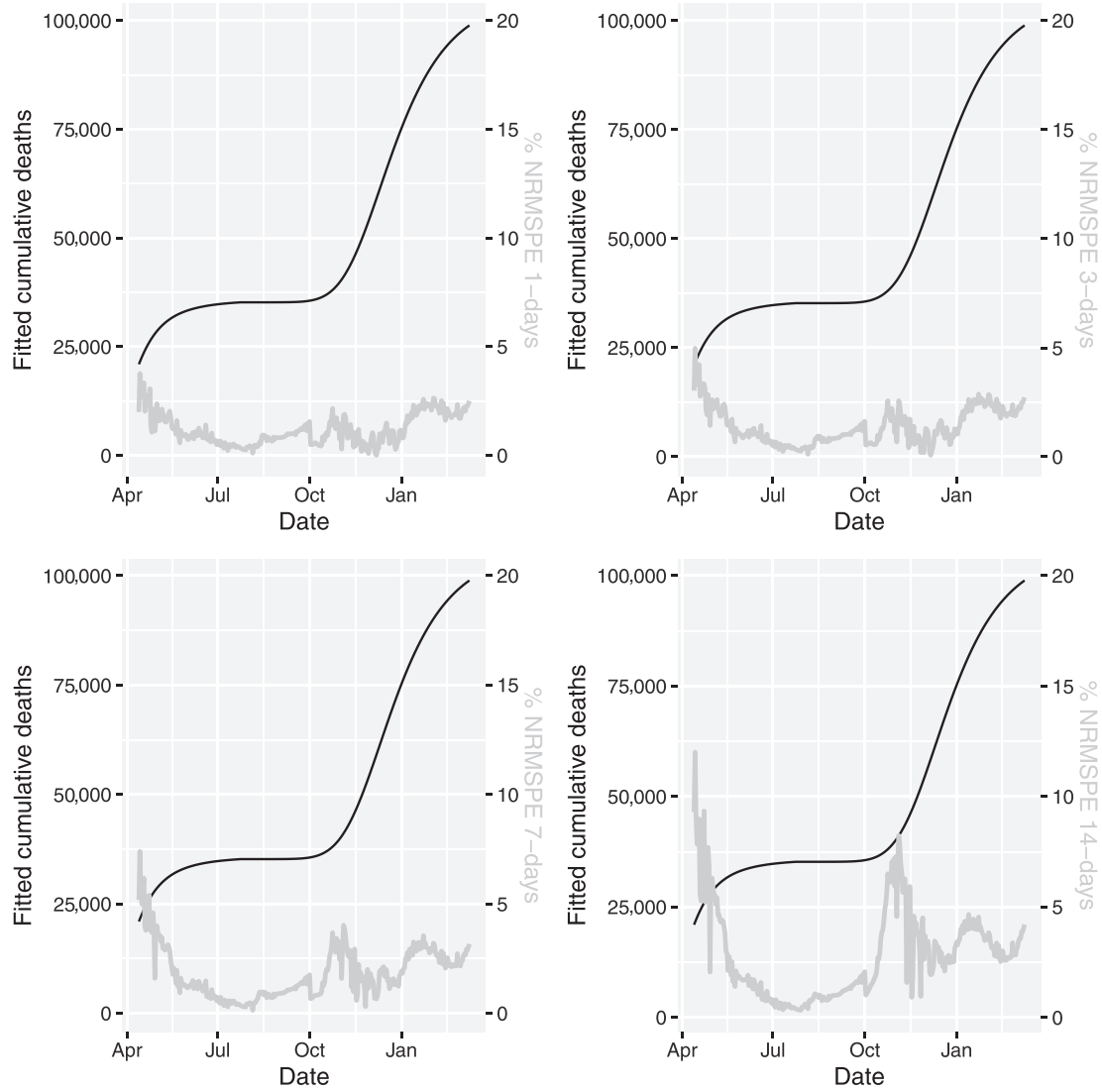


FIGURE 5 Black: fitted cumulative deaths. Gray: % NRMSPE. Results presented for 1, 3, 7, and 14 days step-ahead forecast

For each subset $t = 1, \dots, t^*$, we estimated a log-logistic model $\hat{\mu}(t; \hat{\tau}_{t^*})$ and, for each value of K , we calculated the RMSPE as

$$RMSPE(t^*, K) = \sqrt{\frac{1}{K} \sum_{i=1}^K (\hat{\mu}(t^* + i; \hat{\tau}_{t^*}) - y_{t^*+i}^c)^2}. \quad (21)$$

Moreover, in order to relate this quantity with the process average, we calculated its normalized version (Normalized RMSPE) as

$$NRMSPE(t^*, K) = \frac{RMSPE(t^*, K)}{\mu(t^*; \hat{\tau}_{t^*})}, \quad (22)$$

that can be expressed as pure number or its relative percentage. The % NRMSPEs for each step-ahead are presented in Figure 5. As we can observe, the % NRMSPE was relatively high at the beginning of the considered period, while relatively low values, below the 5% in both 1-day step ahead and 3-days steps ahead predictions, were reported. In the 7-day and even more in the 14-day steps ahead case, a generalized increase of the % NRMSPE was observed for the whole period, but the values were still below the 5% with the exception of the 14 days steps ahead, in which the NRMSPE percentage

FIGURE 6 Fitted (line) and observed (points) cumulative deaths and 95% and 99% confidence interval

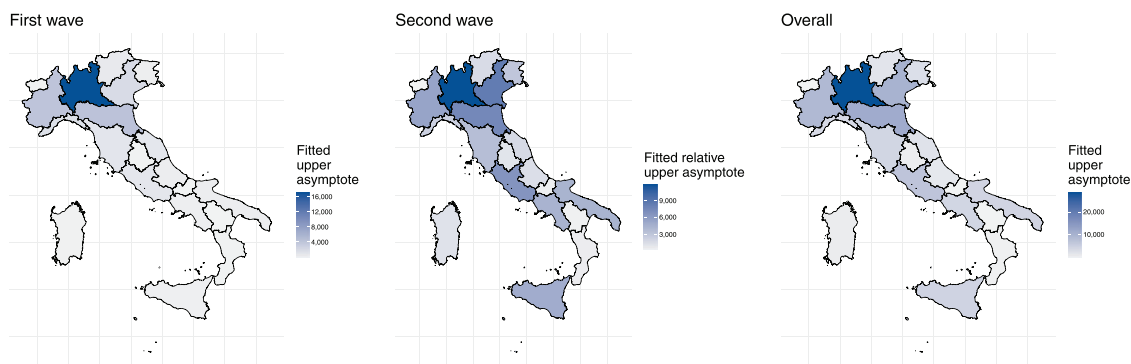
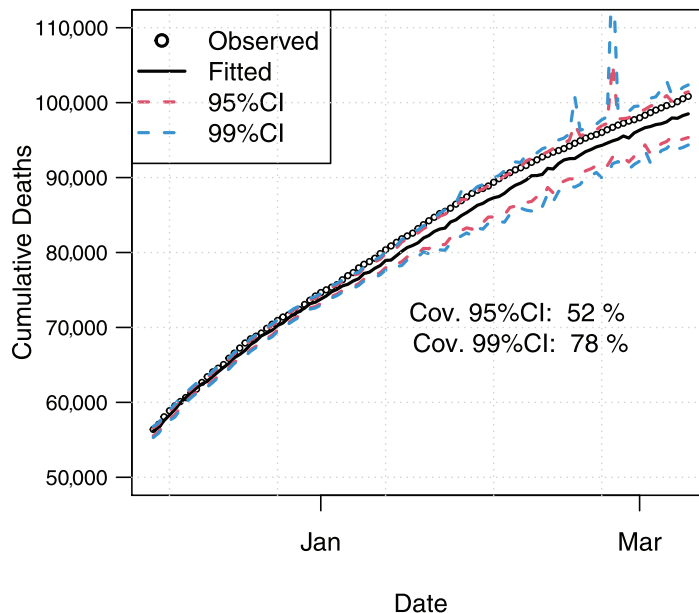


FIGURE 7 Fitted upper asymptote in the first wave (left), in the second (middle), and overall (right) at the regional level. Note that data from the P.A. of Trento and P.A. of Bolzano have been merged

reached values equal to 12% and 8.2% at the beginning of the considered window and of the second wave, respectively. To assess the coverage of our confidence band, we performed a second analysis in which we calculated the coverage rate in 1-day-ahead prediction. In order to stress the model, we considered the last 100 days (from November 11, 2020 to March 10, 2021) calculating for each estimated model $\mu(t; \hat{\tau}_{t^*})$, with $t = 1, \dots, (T - t^*)$ and $t^* = 1, \dots, 100$. We performed a 1-day step ahead prediction estimating the relative 95% and 99%CI via double bootstrap and checking if the observed value falls in the interval. Figure 6 reported the two confidence bands, the fitted and observed deaths. Results were discrete, that is, the coverage in 1-step ahead calculated in the last 100 days prediction reported a 52% in the 95%CI and 78% in the 99%CI. In addition, looking to the estimated trajectory, the fitted curves seems to underestimate the number of cumulative death especially in the second part of the considered temporal window.

5.2 | A study across Italian regions

The spread of the contagion across Italy showed some heterogeneity at the regional level. Actually, the COVID-19 infection emerged in February 2020 in the two regions of Lombardia and Veneto and then severely hit also all the other Northern regions mainly. In contrast, the second wave was widely spread all over the country. In order to investigate more in-depth the dynamics of the epidemic, the same analysis carried out for the deaths series at the national level has been conducted for each region. Fitted models at the regional level are given in Figures A.1 and A.2. As a summary, Figure 7 shows the fitted upper asymptote for deaths in each region at the end of the first wave (d_1^I), the relative upper asymptote only concerning

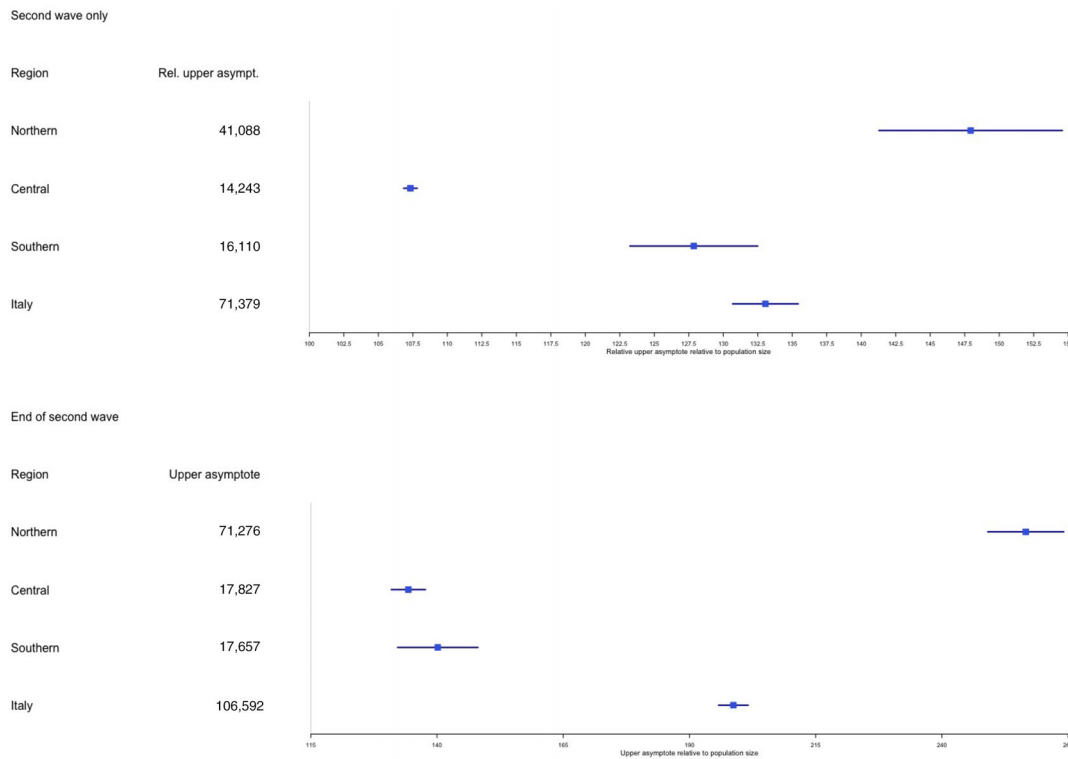
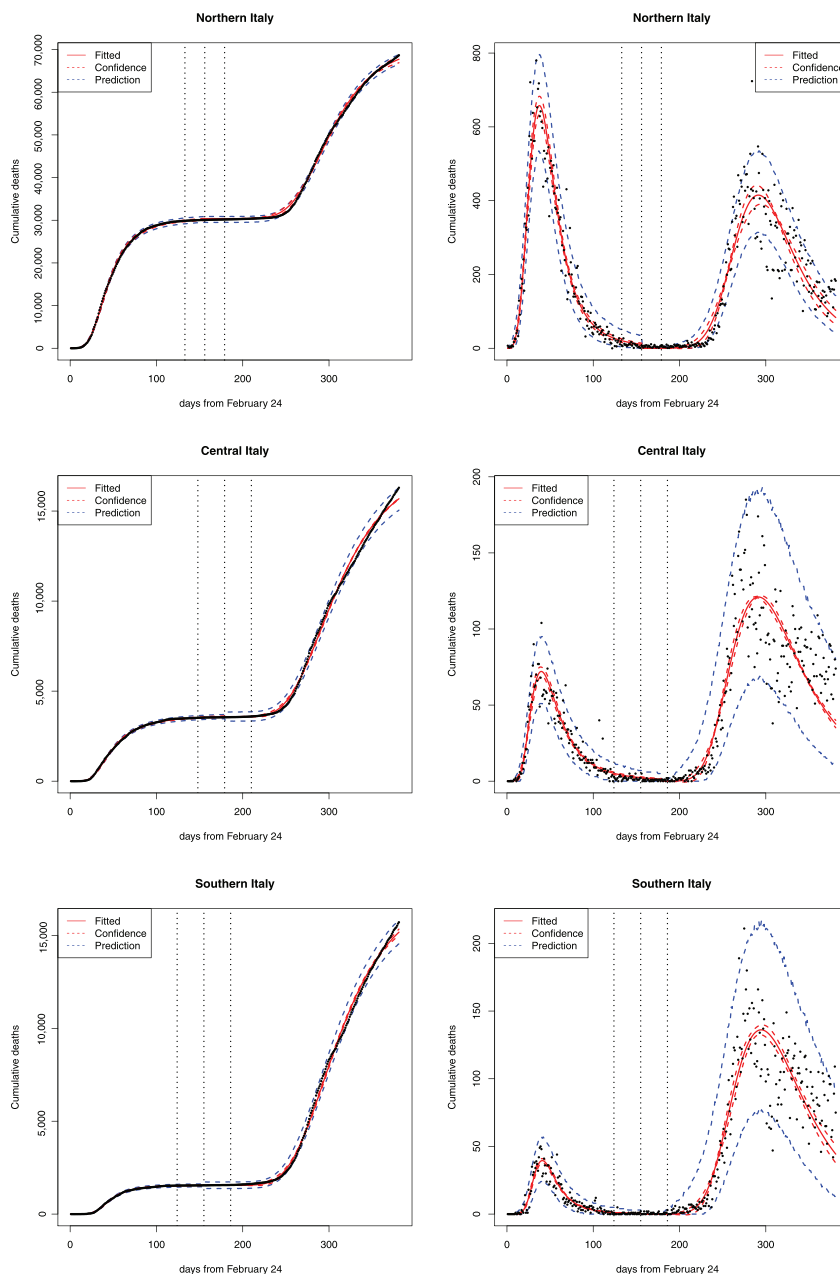


FIGURE 8 Fitted upper asymptote relative to population size during the second wave only (top) and overall at the end of the second wave (bottom), for Northern, Central, Southern Italy, and overall. The estimated size \hat{d}^I is given in the second column: $\hat{d}_2^I - \mu(\hat{t}_0^I; \hat{\theta}_1^I)$ in the top panel, \hat{d}_2^I in the bottom panel

the second wave ($\hat{d}_2^I - \mu(\hat{t}_0^I; \hat{\theta}_1^I)$), and the overall fitted size of the number of deaths at the end of the second wave (\hat{d}_2^I). The tribute in human lives paid by Northern regions, and in particular by Lombardia, was impressive in the first wave of the contagion. During this period, the Central and Southern regions suffered a relatively low death incidence. The fitted upper asymptote in the first wave for Lombardia only is 17,103, whereas it amounts to 3675 and 1566 for all Central and Southern regions, respectively. The large differences between Northern Italy and the rest of the country were mitigated during the second wave due to the spread of the contagion all over the country. In particular, Central and Southern regions experimented with records of infections and deaths never met during the first wave. Then, Italian regions have been grouped in the three macroareas corresponding to Northern (Valle d'Aosta, Piemonte, Liguria, Lombardia, P.A. Trento, P.A. Bolzano, Veneto, Friuli Venezia Giulia, and Emilia Romagna), Central (Toscana, Umbria, Marche, Abruzzo, and Lazio), and Southern (Molise, Campania, Basilicata, Puglia, Calabria, Sicilia, and Sardegna) Italy. Figure 8 gives the final size for the number of deaths fitted over the three areas during the second wave only and overall at the end of the second wave, relative to population size. The top panel confirms the spread of the epidemic all over the country with 146, 107, and 126 deaths for 100,000 inhabitants, in Northern, Central, and Southern regions, respectively. The bottom panel, however, remarks the different overall incidence between Northern regions and the rest of the country, with 257, 134, and 140 deaths for 100,000 inhabitants, respectively. Figures have been produced based on the facilities of the R package `forestplot`. The behavior of cumulative and daily deaths in the three macroareas is displayed in Figure 9, along with fitted models, 99% pointwise confidence and prediction intervals. The panels on the left give cumulative incidence, whereas those in the right column give daily incidence. The first wave was mainly driven by Northern Italy where people suffered an unsurpassed deaths toll; then, the second wave did not exceed the first daily peak. In contrast, the second wave of deaths in Central and Southern Italy was superior to the first. However, Northern Italy still had to deal with a remarkably larger number of deaths (and infections) with respect to Central and Southern Italy. We also notice some common features: the two waves both peaked in the same period all over Italy and inferences about the change point are consistent among the three macroareas. It is worth to remark the large uncertainty that characterizes prediction interval over daily death counts for Central and Southern Italy. Actually, they account for the large variability in the observed data, that is likely due to some heterogeneity in the collection process of records among regions.

FIGURE 9 Left: cumulative deaths for Northern, Central, and Southern Italy. Right: daily deaths for Northern, Central, and Southern Italy. Fitted model with 0.99-level confidence and prediction intervals based on the HAC sandwich covariance matrix estimate. The dotted vertical lines give the fitted change point with the corresponding 0.99-level confidence interval



6 | DISCUSSION

We proposed a methodology to model cumulative counts in an epidemic characterized by two distinct subsequent waves. The method is based on the combination of two growth curves employing a change-point model. Our proposal is consistent with counting data and the presence of a potential overdispersion/serial dependence. Actually, the proposed technique is robust in the sense that we relaxed assumptions about independence, but two adjustments in the evaluation of the standard errors were suggested, in order to account for overdispersion and heteroskedasticity and serial dependence, respectively.

We considered an application to Italian death counts. The results confirm the renewed spread of the contagion during the beginning of summer. Evidence for a second wave of deaths already during July, means that the growth of infections can be placed 2–3 weeks before, as soon as many restrictions have been removed. However, the two waves were very different in nature, since the first was mainly driven by Northern regions whereas the second had spread all over the country. The main differences, but also the common features, across Italy have been studied in a regional-level analysis.

The method performed satisfactory both in terms of goodness of fit and prediction ability, even if there is still room for improvement.

As a future line of research, the present methodology can be developed further to include more than two epidemic waves and allow some comparisons between models characterized by a different number of waves. Moreover, one could consider the possible inclusion of covariates and joint analysis of infections and deaths. We remark that we considered the series of deaths, since they exhibit a more regular behavior, due to a likely more homogeneous collection process of the records.

ACKNOWLEDGMENTS

The authors wish to thank the Associate Editor and two anonymous referees whose comments and suggestion helped to improve the paper. This research work was partially supported by University of Padova (BIRD197903).


CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in a GitHub repository daily updated by the Dipartimento della Protezione Civile at <https://github.com/pcm-dpc/COVID-19>.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Luca Greco  <https://orcid.org/0000-0002-1511-1657>

REFERENCES

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. Wiley.
- Bartolucci, F., & Farcomeni, A. (2021). A spatio-temporal model based on discrete latent variables for the analysis of COVID-19 incidence. *Spatial Statistics*. <https://doi.org/10.1016/j.spasta.2021.100504>
- Cabras, S. (2020). A Bayesian-deep learning model for estimating COVID-19 evolution in Spain. *arXiv preprint arXiv:2005.10335*.
- Cattelan, M., & Sartori, N. (2016). Empirical and simulated adjustments of composite likelihood ratio statistics. *Journal of Statistical Computation and Simulation*, 86, 1056–1067.
- Chandler, R. E., & Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94, 167–183.
- Di Loro, P. A., Divino, F., Farcomeni, A., Lasinio, G. J., Lovison, G., Maruotti, A., & Mingione, M. (2020). Nowcasting COVID-19 incidence indicators during the Italian first outbreak. *Statistics in Medicine*, 40, 3843–3864.
- Dorward, J., Khubone, T., Gate, K., Ngobese, H., Sookrajh, Y., Mkhize, S., Jeewa, A., Bottomley, C., Lewis, L., Baisley, K. (2021). The impact of the COVID-19 lockdown on HIV care in 65 South African primary care clinics: An interrupted time series analysis. *Lancet HIV*, 8, e158–e165.
- Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99, 619–632.
- Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Annals of Applied Statistics*, 6, 1971–1997.
- Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., & Lovison, G. (2021). An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biometrical Journal*, 63(3), 503–513.
- Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., & Ventura, L. (2020a). Robust inference for non-linear regression models from the Tsallis score: Application to coronavirus disease 2019 contagion in Italy. *Stat*, 9, e309.
- Girardi, P., Greco, L., Musio, M., Racugno, W., & Ventura, L. (2020b). The evolution of the endemic stage of the COVID-19 outbreak in Italy during summer 2020. *Significance Magazine*, <https://www.significancemagazine.com/690>.
- Günther, F., Bender, A., Katz, K., Küchenhoff, H., & Höhle, M. (2021). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*, 63(3), 490–502.
- Hardin, J. W. (1998). Newey-West standard errors for probit, logit, and Poisson models. *Stata Technical Bulletin*, 7, 32–35.
- Kaxiras, E., & Neofotistos, G. (2020). Multiple epidemic wave model of the COVID-19 pandemic: Modeling study. *Journal of Medical Internet Research*, 22, e20912.
- Kim, T., Lieberman, B., Luta, G., & Pena, E. (2020). Prediction regions for Poisson and over-dispersed Poisson regression models with applications to forecasting number of deaths during the COVID-19 pandemic. *arXiv e-prints (arXiv:2007.02105)*.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703–708.
- Pace, L., Salvan, A., & Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21, 129–148.

- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. *PLoS One*, *10*, e0146021.
- Schneble, M., De Nicola, G., Kauermann, G., & Berger, U. (2021). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*, *63*(3), 471–489.
- Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5–42.
- Xu, S., & Li, Y. (2020). Beware of the second wave of COVID-19. *Lancet*, *395*, 1321–1322.
- Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, *11*, 1–17.
- Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*, 1–16.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Girardi, P., Greco, L., Ventura, L. Misspecified modeling of subsequent waves during COVID-19 outbreak: A change-point growth model. *Biometrical Journal*. 2021;1–17.
<https://doi.org/10.1002/bimj.202100129>

APPENDIX

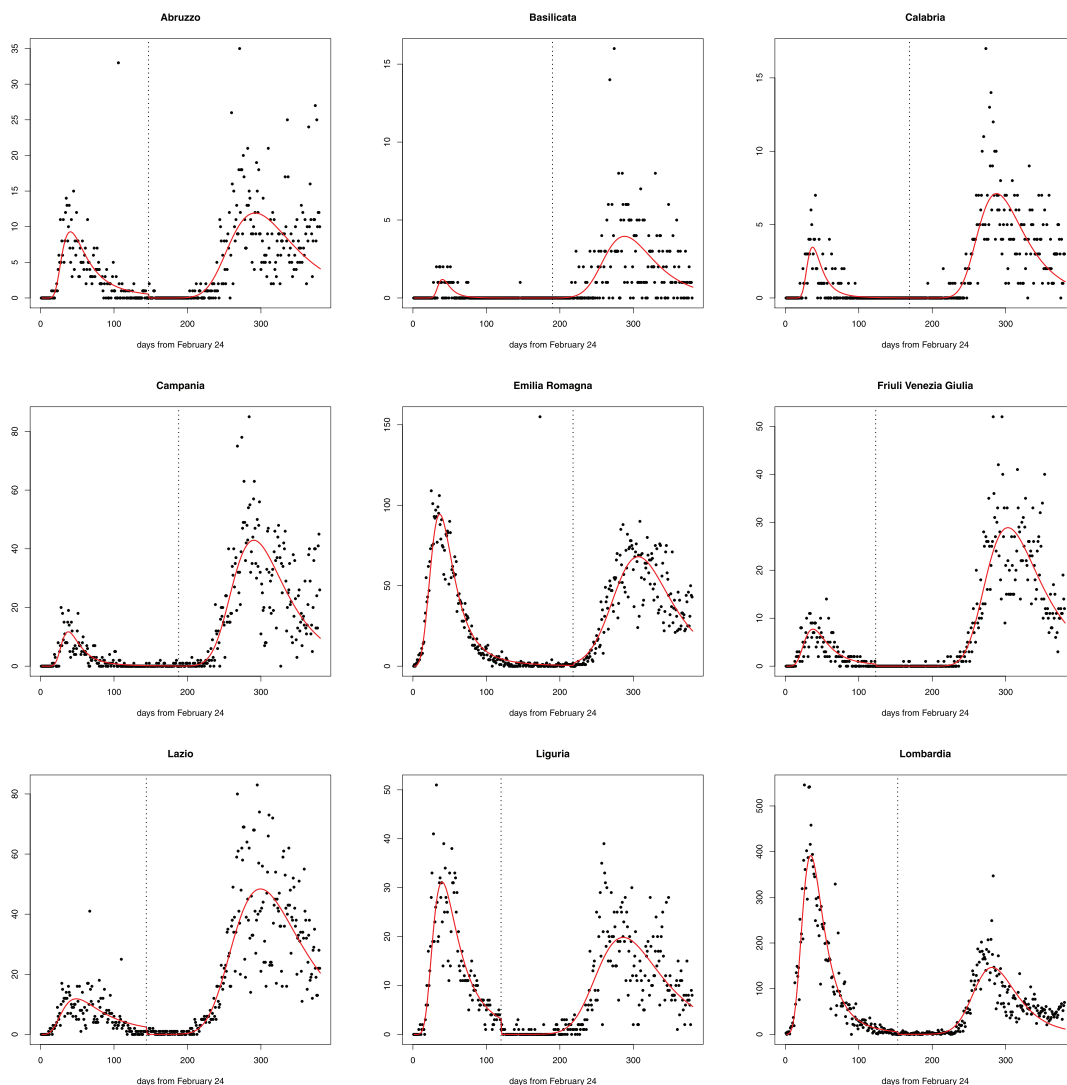


FIGURE A.1 Daily deaths and first derivative of the fitted change-point growth model at the regional level

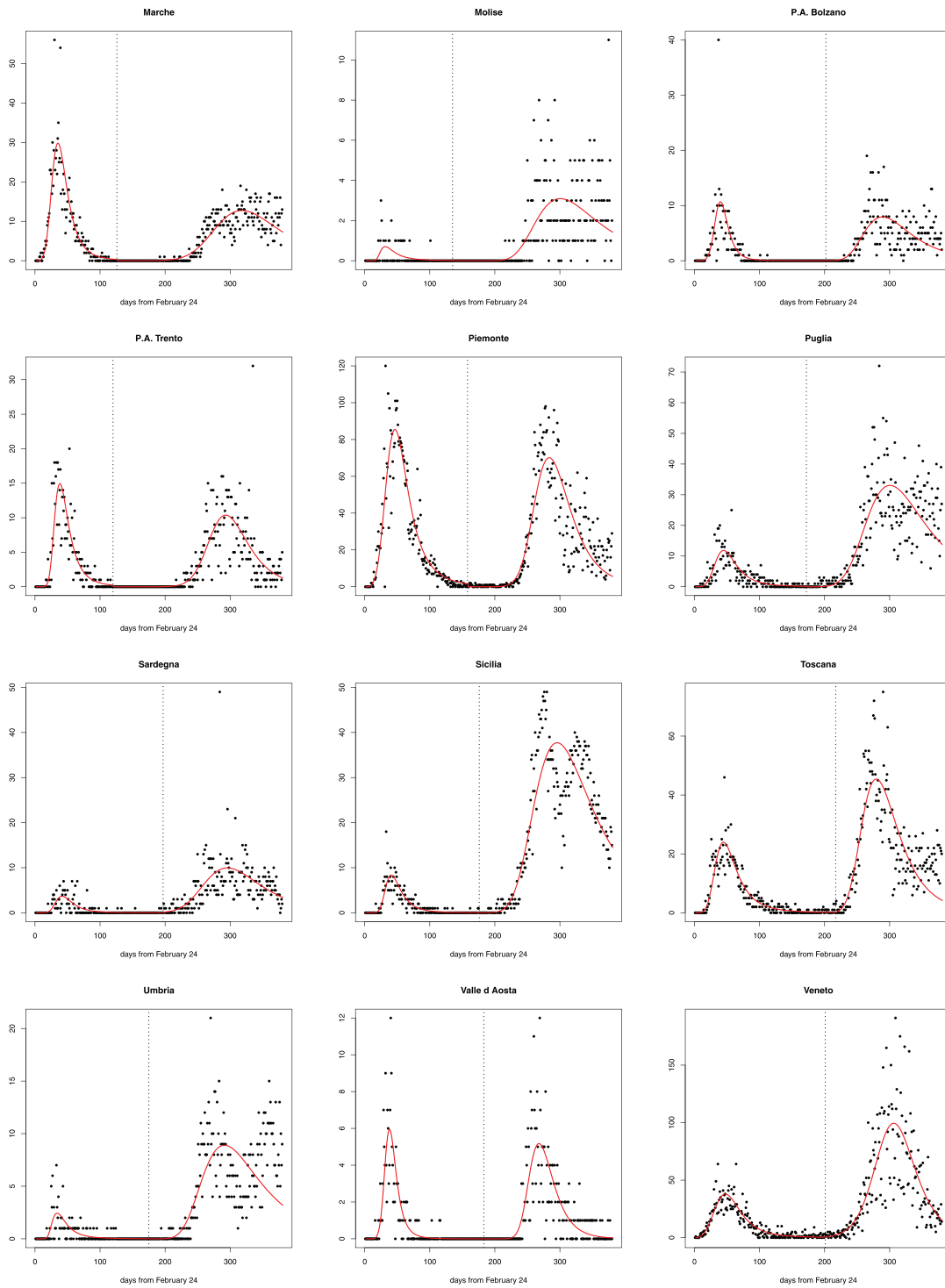


FIGURE A.2 Daily deaths and first derivative of the fitted change-point growth model at the regional level