A COVINDEX based on a GAM beta regression model with an application to the COVID-19 pandemic in Italy

Luca Scrucca
Dipartimento di Economia
Università degli Studi di Perugia
Via A. Pascoli 20, 06123 Perugia, Italy
luca.scrucca@unipg.it
https://orcid.org/0000-0003-3826-0484

Abstract Detecting changes in COVID-19 disease transmission over time is a key indicator of epidemic growth. Near real-time monitoring of the pandemic growth is crucial for policy makers and public health officials who need to make informed decisions about whether to enforce lockdowns or allow certain activities. The effective reproduction number R_t is the standard index used in many countries for this goal. However, it is known that due to the delays between infection and case registration, its use for decision making is somewhat limited. In this paper a near real-time COVINDEX is proposed for monitoring the evolution of the pandemic. The index is computed from predictions obtained from a GAM beta regression for modelling the test positive rate as a function of time. The proposal is illustrated using data on COVID-19 pandemic in Italy and compared with R_t. A simple chart is also proposed for monitoring local and national outbreaks by policy makers and public health officials.

Keywords pandemic surveillance; GAM beta regression; COVINDEX; public-health decision-making

1. Introduction

The World Health Organization (WHO) declared coronavirus disease (COVID-19) a pandemic on 11 March 2020. Since then, most countries around the world have addressed this threat by implementing various strategies to fight the pandemic. From simple preventive measures, such as case identification and contact tracing, quarantine and isolation, to more severe strategies based on general lockdowns of all non-essential economical and social activities. Since public health decision-making requires the balancing of numerous, and often conflicting, factors, a timely and data-informed decision making process appears crucial.

Several studies have been recently devoted to the analysis of COVID-19 data. Referring to the Italian situation, Sebastiani et al. (2020) evaluated the impact of government measures on the evolution of pandemic. Girardi et al. (2020) used robust dose-response curves to predict the contagion dynamics of COVID-19, while Alaimo Di Loro et al. (2021) proposed an extended Generalized Linear Model based on the Richards' curve to model and predict incidence indicators. A Poisson autoregressive model was discussed by Agosto et al. (2021) to monitor the time evolution of the COVID-19 contagion curve, while Bartolucci and Farcomeni (2021) introduced a spatio-temporal model based on discrete latent variables for the analysis of weekly positive rates. Finally, Farcomeni et al. (2021) investigated an ensemble approach for short-term prediction of occupancy of intensive care units due to COVID-19 outbreak.

The basic reproduction number, R_0, is an indicator of the epidemic's virulence. It is defined as the average number of infections caused by an infected person when the whole population is susceptible, and for SARS-CoV-2 is between 2 and 3 (Li et al., 2020; Hilton and Keeling, 2020). As the pandemic evolves, the effective reproduction number R_t is a more useful measure. This is the average number of infections that an infected person will cause. An R_t above 1.0 indicates that the outbreak is growing, and below 1.0 means that it is shrinking. As a simple understood measure, R_t is regularly published and discussed by the media, and it has been used in many countries, including Italy, to decide whether to tighten or loosen control measures. However, R_t suffers from several drawbacks when used to monitor the transmission of the disease over time, the main one being the delay with which it signals the evolution of the pandemic (Gostic et al., 2020; Adam, 2020). Therefore, with a delay on the estimate of R_t between ten days to two weeks, the use of R_t as a near real-time decision-making tool appears rather pointless. For further discussion on the risks caused by the misuse of the

reproduction number in the COVID-19 surveillance see Maruotti et al. (2021).

This paper introduces a COVID-19 index, called COVINDEX, which tries to assess whether the epidemic is growing, shrinking, or holding steady. The proposed index is estimated by modelling the test positive rate (TPR) with a GAM beta regression model. TPR is an easily computed statistic, defined as the fraction of all COVID-19 tests performed on a given day that are actually positive. This metric can be used to understand the spread of the virus, but it also offers a measure of how adequately a country is testing. TPR can be high if the number of positive tests is too high, but also if the number of total tests is too low. Most developed countries faced limited testing capacity during the initial phase of the pandemic, which resulted in high TPR values due to testing conducted primarily on symptomatic individuals. In the following months the ability to administer tests using PCR (polymerase chain reaction) or molecular swabs largely increased, leading to a situation that allows both symptomatic and asymptomatic individuals to be tested. Although TPR can't be used for estimating incidence of the virus in the general population, a fundamental epidemic parameter that would require a carefully designed sampling plan, it can be used for monitoring the evolution of infection and transmission in the community. Higher positive rates suggest the need for further restrictions, such as wearing masks and physical distancing, to slow down the spread of the disease. As a rule of thumb, World Health Organization recommended 5% as the threshold for the percent positive rate to declare the COVID-19 transmission under control.

The main advantages of the proposed COVINDEX is the use of data routinely collected and its timely estimation which provides a near real-time tool to assess the effectiveness of interventions and to inform policy. Furthermore, since it is based on a statistical model, the associated uncertainty can be estimated.

The paper is organized as follows. Section 2 introduces the GAM beta regression model, its estimation and uncertainty assessment. Section 3 describes the proposed COVINDEX and its usage for monitoring the pandemic evolution. Section 4 includes a detailed analysis of COVID-19 pandemic in Italy, including the estimation of COVINDEX, from early March 2020 to the end of March 2021. Section 5 contains a comparison between the proposed COVINDEX and the effective reproduction number, showing the advantages of COVINDEX as near real-time monitoring tool. The final section provides some concluding remarks.

2. Statistical Model for the Test Positive Rate

2.1. GAM beta regression

Let y_t be the test positive rate (TPR), defined as the ratio of the number of new positive cases P_t to the number of tests T_t at time t, with t assuming integer values between 1 and n, respectively, for the first and last day of the analyzed period. As a proportion TPR is naturally limited in the range $[0,1]$. Several approaches and models can be used for response variables that are expressed as proportions (Douma and Weedon, 2019), and perhaps the most popular statistical model is the beta regression model (Ferrari and Cribari-Neto, 2004; Zeileis and Cribari-Neto, 2010).

Assume that TPR can be modelled by a beta distribution written as

$$y_t \sim \text{Beta}(\mu_t, \phi),$$

with mean and variance of the beta distribution given, respectively, by

$$\mathbb{E}[y_t] = \mu_t,$$

and

$$\mathbb{V}[y_t] = \frac{\mu_t(1-\mu_t)}{1+\phi}.$$

Strictly speaking, the beta distribution can only model data in the open set $(0,1)$. If extreme values 0 and 1 can actually be observed, the inflated zero- and/or one beta distribution of Ospina

and Ferrari (2010) could be used. Since in practice TPR rarely assumes a value of 0 and almost never 1, if needed, the simple approach proposed by Smithson and Verkuilen (2006) can be adopted by applying the data transformation $(y_t(n-1)+0.5)/n$. The latter is the approach followed in this paper.

The mean μ_t can be expressed as a function of the linear predictor $\eta_t = \boldsymbol{\beta}^\top \boldsymbol{x}_t$, where $\boldsymbol{\beta}$ is a $(p+1)$-dimensional vector of unknown regression coefficients (including the intercept), and $\boldsymbol{x}_t$ is the vector of observed values on p predictors plus the intercept. Usually, the logistic function is used in beta regression, so we can write

$$\mu_t = \text{logistic}(\eta_t) = \frac{\exp(\eta_t)}{1+\exp(\eta_t)} = \frac{1}{1+\exp(-\eta_t)}.$$

The inverse of the logistic function is the logit function, the so-called *link* function in GLM terminology (McCullagh and Nelder, 1989), given by:

$$\text{logit}(\mu_t) = \log\left(\frac{\mu_t}{1-\mu_t}\right) = \eta_t.$$

Generalized Additive Models (GAMs; Hastie and Tibshirani, 1990) allows to model the dependence of the response variable in a flexible way using smooth functions of the predictors by defining the linear predictor as

$$\eta_t = \beta_0 + \sum_{j=1}^{p} f_j(x_{tj}),$$

where $f_j(x_{tj}) = \sum_{k=1}^{K_j} \beta_{jk} B_{jk}(x_{tj})$ is the smoothing term for the jth predictor with $\{B_{jk}()\}_{k=1}^{K_j}$ a set of known basis functions associated to unknown parameters β_{jk}.

Several smoothers can be defined by adopting different basis functions, such as penalized regression splines, cubic regression splines, etc. For an overview of the several smoothing functions available using splines bases see Wood (2017, Chapter 5). Among the various possibilities, thin plate regression splines (TPRS; Wood, 2003) represents a convenient form because TPRS (i) do not require to specify the "knots", (ii) use a low rank approximation of the full basis expansion, and (iii) are isotropic smoothers, so they are unaffected by any rotation or reflection of the covariates.

In our application the only feature included as smoothing term in the linear predictor is time, so x_1 is an integer counting the days since the first day of the analysis. To some extent, the coding of such feature has no practical consequence, and other equivalent forms could have been used as well. In addition, to account for the reduced tracing activity during weekends (Saturday and Sunday) and holidays, a dummy variable x_2 is included taking value 1 for data referring to weekends or holidays, and 0 otherwise. The rationale behind the inclusion of such term is that the number of swabs processed is noticeably limited during weekends and holidays, so a significant increase in the test positivity rate is often observed due to the limited testing capacity and the higher probability of testing only symptomatic cases.

Thus, in our case the linear predictor of GAM simplifies to

$$\eta_t = \beta_0 + \sum_{k=1}^{K} \beta_{1k} B_{1k}(x_{t1}) + \beta_2 x_{t2},$$

where $\{B_{1k}\}_{k=1}^{K}$ represents the basis of thin plate regression splines. Note, however, that other smooth functions would have given nearly equivalent results.

2.2. Estimation

Estimation of the GAM model introduced in previous section can be pursued by REstricted Maximum Likelihood (REML), which amounts to maximize the penalized log-likelihood

$$\ell_P(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\lambda\boldsymbol{\beta}^\top \boldsymbol{S}\boldsymbol{\beta}, \tag{1}$$

where $\ell(\boldsymbol{\beta}) = \sum_{t=1}^{n} \ell(y_t|\boldsymbol{\beta})$ is the log-likelihood for the observed values y_t of the response variable. The last term in the right-hand side represents the smoothing penalty, with λ a smoothing parameter and $\boldsymbol{S}$ a known penalty matrix.

As reported in Wood (2011) and Wood et al. (2016), REML is equivalent to marginal likelihood when the model contains Gaussian random effects, and it leads to more stable estimates of λ and much reduced risk of under-smoothing compared to GCV. Furthermore, as discussed in the next section, the REML estimates of regression coefficients have an asymptotically MAP Bayesian interpretation that is very useful for obtaining simulated credible intervals for predictions. For a recent review on inference and computation in GAMs see Wood (2020).

The selection of the smoothing parameter can be obtained, among many other proposals, by minimizing the conditional Akaike's information criterion (AIC). This version of AIC for GAMs uses the log-likelihood evaluated at the penalized MLE, and with the effective degrees of freedom computed as discussed in Wood et al. (2016).

However, because the number of administered swabs is not constant over time, we must take into account this fact when modelling the test positive rate. There are several reasons for this empirical evidence. First of all, during the weekends (particularly on Sundays) and holidays the number of swabs drops drastically. Furthermore, during periods of strong expansion of the pandemic, the monitoring system is unable to carry out effective surveillance and only symptomatic patients are likely to be tested. Accounting for the different number of swabs in the model for the positive rate can be achieved by adopting a weighted penalized log-likelihood criterion. This amounts to replace the log-likelihood $\ell(\boldsymbol{\beta})$ in (1) with the weighted version

$$\ell_W(\boldsymbol{\beta}) = \sum_{t=1}^{n} w_t \ell(y_t|\boldsymbol{\beta}),$$

where w_t are prior weights specifying the contribution of each data point to the log-likelihood. In particular, indicating with $\bar{T}$ the average number of administered swabs over the period, weights can be defined as $w_t = T_t/\bar{T}$ so that positive rates y_t computed from number of swabs larger than the average have proportionally larger weights, and vice versa for those rates based on number of swabs smaller than the average. Furthermore, with the adopted definition for the weights the contribution of each datum is specified without changing the overall magnitude of the log-likelihood.

Once the model is fitted, the predicted TPR can be computed as

$$\widehat{\mu}_t = \text{logistic}\left(\widehat{\beta}_0 + \sum_{k=1}^{K} \widehat{\beta}_{1k} B_{1k}(x_{t1}) + \widehat{\beta}_2 x_{t2}\right). \tag{2}$$

On certain occasions, for instance when computing the COVINDEX discussed in Section 3, we may want to compute predictions for the TPR with the weekends/holidays effect ruled out. This is easily accomplished by setting $x_{t2} = 0$ for all t.

2.3. Uncertainty and inference

The penalized likelihood approach described above has also a Bayesian interpretation by assuming an improper multivariate normal prior on $\boldsymbol{\beta}$. In this case, the REML estimates of $\boldsymbol{\beta}$ coefficients are asymptotically the maximum a posteriori (MAP) of the Bayesian posterior distribution, with the latter given by

$$\boldsymbol{\beta}|(\boldsymbol{y},\lambda) \sim N(\widehat{\boldsymbol{\beta}},(\widehat{\boldsymbol{I}}+\lambda\boldsymbol{S})^{-1}), \tag{3}$$

where $\widehat{\boldsymbol{I}}$ is the observed information matrix (Hessian of the negative log-likelihood) at $\widehat{\boldsymbol{\beta}}$ (Wood, 2017, Section 6.10). This result is useful for computing approximate credible intervals for any function of $\boldsymbol{\beta}$ by simulating from the posterior (Gelman and Hill, 2006, Section 7.2). Wood (2017, p. 294) reported good frequentist coverage properties for such Bayesian credible intervals, with empirical coverage close to the nominal level when averaged across the domain of the function.

In practice, coefficients $\boldsymbol{\beta}^*$ are simulated from (3), and then plugged in equation (2) to get the simulated means μ_t^*. The process is replicated a large number of times, say 10 000 or more, and the percentiles of the simulated distributions at different values of x_t can be used to compute the limits of approximate credible intervals for the mean. To compute approximate credible intervals for the single prediction we simulate response values as $y_t^* \sim \text{Beta}(\mu_t^*, \widehat{\phi})$, where μ_t^* is the simulated mean as described above, $\widehat{\phi}$ is the model estimate of the precision parameter, and then we compute the percentiles of the simulated distribution of predicted values for the response. The empirical coverage of the prediction intervals will be assessed in Section 4.2.

3. COVINDEX as a Monitoring and Decision-Making Tool

The COVINDEX proposed in this paper is an attempt to compute a synthetic index summarizing the evolution of the COVID-19 pandemic, which can be useful to policy makers and public health officials for monitoring local and national outbreaks. In our proposal this is simply computed as

$$\text{COVINDEX}_t = \frac{\widehat{\mu}_t}{\widehat{\mu}_{t-7}}, \tag{4}$$

the ratio of the predicted positive rate at time t to the prediction 7 days earlier. The value of 7 is chosen because it is approximately the expected incubation time for COVID-19 (Nazar and Elfadil, 2021), and because it corresponds to the observed weekly fluctuation in testing. A COVINDEX value larger than 1.0 means that the pandemic is growing, while a value smaller than 1.0 indicates that new infections are slowing down.

The COVINDEX estimate is clearly affected by uncertainty and to account for it the approach outlined in Section 2.3 for TPR can be used here as well. In particular, for each simulated series of values μ_t^*, simulated COVINDEX series can be obtained by applying equation (4) to get the simulated values $\text{COVINDEX}_t^* = \mu_t^*/\mu_{t-7}^*$. Approximate credible intervals can then be computed from the percentiles of the simulated distribution.

We argue that decisions made by policy makers should be based both on the COVINDEX, which provides an outlook on the likely behaviour of the pandemic in the near future, and on the level of the estimated TPR, which represents its current status. Following this idea, a TPR-COVINDEX risk quadrant chart can be drawn (see Figure 1). This chart illustrates four potential scenarios which represent a useful tool for a decision maker. The quadrants are defined by the dashed lines drawn at selected threshold values. For COVINDEX the natural reference value is 1.0, with values below it indicating a shrinking outbreak, and values higher than 1.0 indicating epidemic situations that are increasingly worrying and out of control. Note that, since the index is a ratio, the y-axis is expressed in logarithmic scale. For the positive rate, the threshold value can be set according to the World Health Organization, which published a set of criteria to inform whether the epidemic is under control. In particular, one criterion states that "[...] less than 5% of samples positive for COVID-19, at least for the last 2 weeks, assuming that surveillance for suspected cases is comprehensive" (World Health Organization, 2019).

According to the above mentioned threshold values, the upper-right quadrant represents the worst-case scenario, with high values of both TPR and COVINDEX. On the contrary, the best-case scenario is the lower-left quadrant which has both low TPR and COVINDEX less than 1.0 indicating a decreasing circulation of the virus. The remaining quadrants are intermediate cases. Typical situations will move in a clockwise direction, moving from the worst-case, represented by the red quadrant on top-right, to the orange quadrant at bottom-right, and eventually reaching the yellow quadrant indicating an outbreak under control. However, in some cases the pandemic could regain strength by getting COVINDEX values greater than 1.0, thus moving towards the top-left orange quadrant or directly towards the worst-case situation described by the red quadrant. A description of the Italian situation since March 2020 is discussed in Section 4.

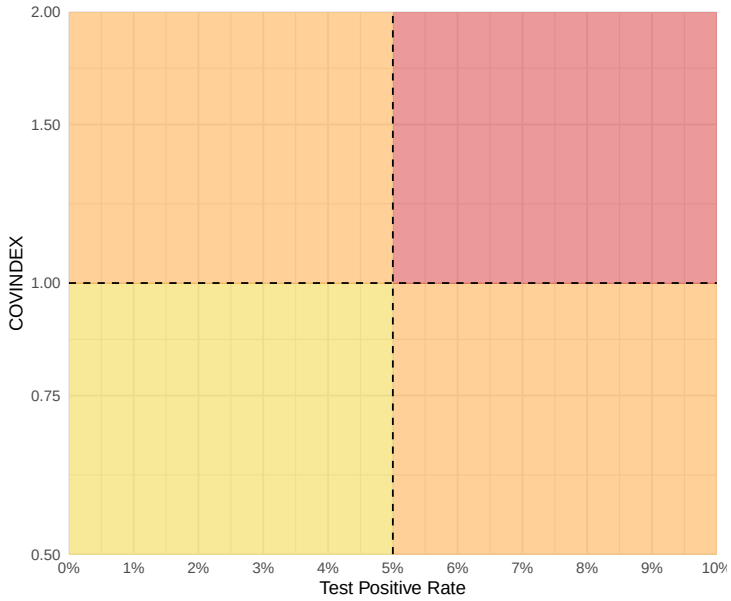

Figure 1. TPR-COVINDEX risk quadrant chart.

4. Application to Italian COVID-19 Pandemic

4.1. Data

The Italian Department of Protezione Civile provides daily information on the COVID-19 pandemic, both at the national and the regional level, in a public GitHub repository (Presidenza del Consiglio dei Ministri – Dipartimento della Protezione Civile, 2020). Among the data contained in this repository, the cumulative number of naso-pharyngeal or molecular swabs and the corresponding positive tests are provided. Starting with January 15th, 2021, antigen tests are also officially recorded, while previously only some regions included them in the recorded statistics since autumn 2020. The reliability of such information is at best questionable and not available uniformly for the year 2020. For these reasons, in our analyses we considered the information from daily molecular swabs (not persons tested) to compute the test positive rate (TPR), a commonly used screening and diagnostic tool for COVID-19 (World Health Organization, 2020).

we consider the total number of daily swabs, not the persons tested

The plot on Figure 2 shows the observed TPR over time with points proportional to the administered swabs.

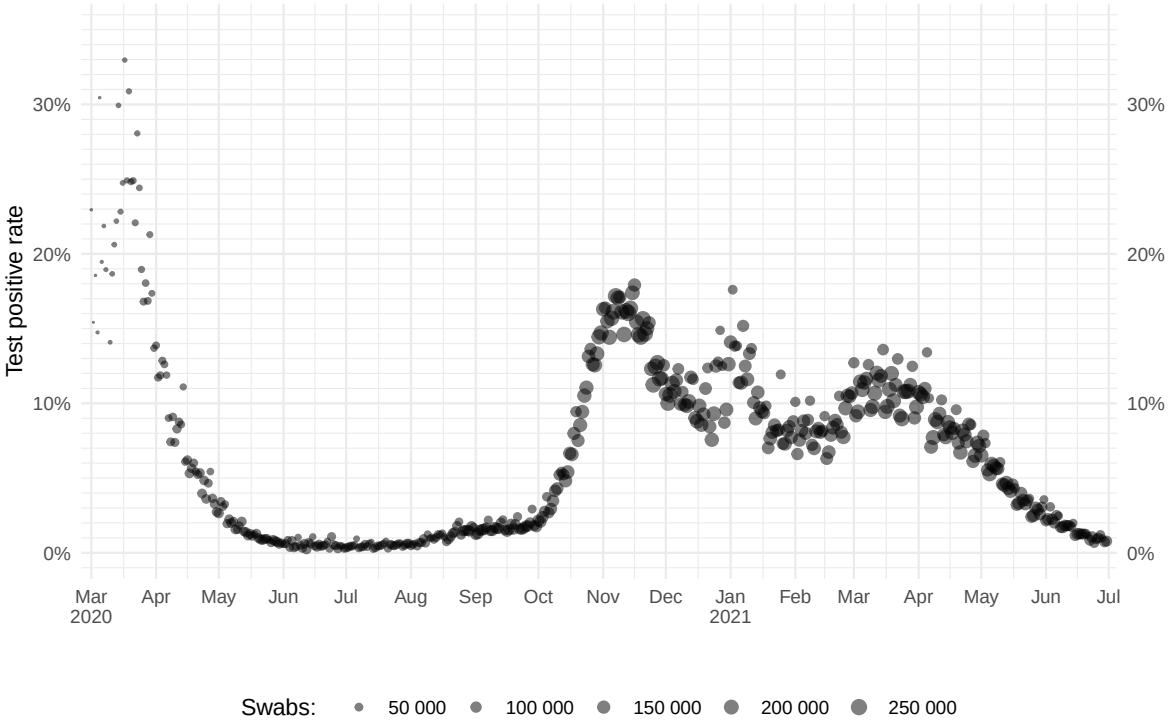

Figure 2. Plot of test positive rate from beginning of COVID-19 pandemic in Italy to the end of observational period with size of points proportional to the number of molecular swabs administered.

4.2. GAM beta regression model estimate

Table 1 reports the summary output of the estimated GAM beta regression model for the test positive rate in Italy from March 1st 2020 to June 30th 2021. The parametric terms include the intercept and a dummy variable for the days following the weekends (Saturday and Sunday) and holidays. The smooth term captures the evolution of underlying trend in the observed test positive rate. The amount of smoothing applied to the time predictor is selected by minimizing the AIC. The graphs of the autocorrelation and partial autocorrelation functions for the deviance residuals in Figure 3 show no significant remaining correlation at different lags.

Table 1. GAM beta regression model summary.

Num. of obs. = 487			Dispersion par. = 1467.3	
Log-likelihood = 1828.7			Deviance expl. = 0.9891	
REML = 1726.3			AIC = -3583.1	
Parametric coefficients:				
	Estimate	Std. error	z-value	p-value
(Intercept)	-3.1904	0.01177	-271.07	< 0.001
Weekend	0.1681	0.01023	16.44	< 0.001
Smooth terms:				
	edf	Ref. df	ChiSq-value	p-value
$s(t)$	35.11	39.67	16820	< 0.001

Figure 4 reports the estimated curve for the test positive rate with 95% credible intervals for the mean and the single value obtained by simulating from the posterior distribution as described in Section 2.3. The highest positive rates are achieved in March 2020 during the first wave of pandemic, and on November 2020, corresponding to the second wave. A resurgence of spread during the end of 2020 is followed by a quick decrease in earlier 2021. Subsequently, the situation remained stable

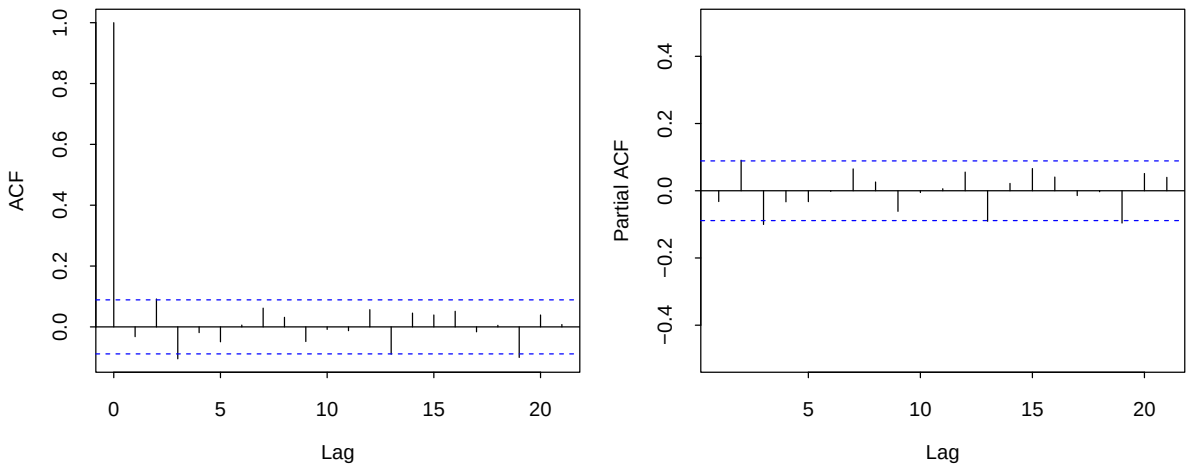

Figure 3. Autocorrelation and partial autocorrelation functions for the deviance residuals of the estimated GAM beta regression in Table 1

for about a month, but in the second part of February another sharpe increase occurred due to the appearance of COVID-19 variants in the Italian territory (in particular the Alpha or english variant). Starting from the beginning of April, a marked decline in the TPR can be observed, likely favored by the increased full vaccination coverage of the Italian population.

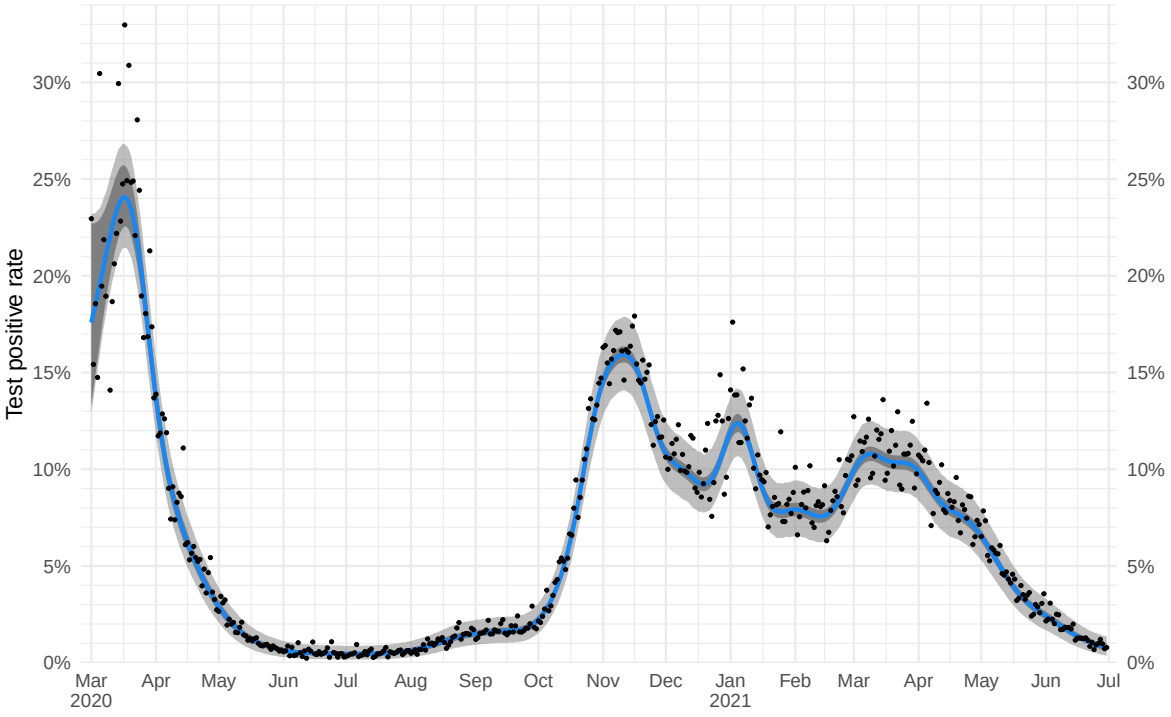

Figure 4. GAM estimate of test positive rate (blue line) in Italy during 2020 and first half of 2021, with 95% simulated credible intervals for the mean (dark grey area) and for the single value (light grey area).

4.3. Empirical coverage of credible intervals for predictions

To investigate the accuracy of the simulated prediction intervals, we considered all the dates from February 1st 2021 to May 31st 2021 and the time horizon for the predictions from 1 day to 14 days. For each date we fitted the GAM beta regression model using all the data available up to that day, and then we used the estimated model to compute the simulated credible intervals for the following 14 days. This process was replicated for the all the days in the specified range and the empirical coverage calculated. Figure 5 reports on the left panel a graph showing the inclusion or exclusion of the observed values of TPR in the simulated prediction intervals, while on the right a plot of the empirical coverage for the time horizon from 1 up to 14 days. Overall the coverage is close to the nominal level, with all the values above 90% for the forecasts of the first week, and between 85% and 90% for the second week. It is interest to note that most of the coverage errors occur in periods of abrupt changes, for instance at the sharp rise of TPR in the last week of February or at the beginning of TPR decline in mid-March. As expected, the empirical coverage decreases as the prediction horizon increases.

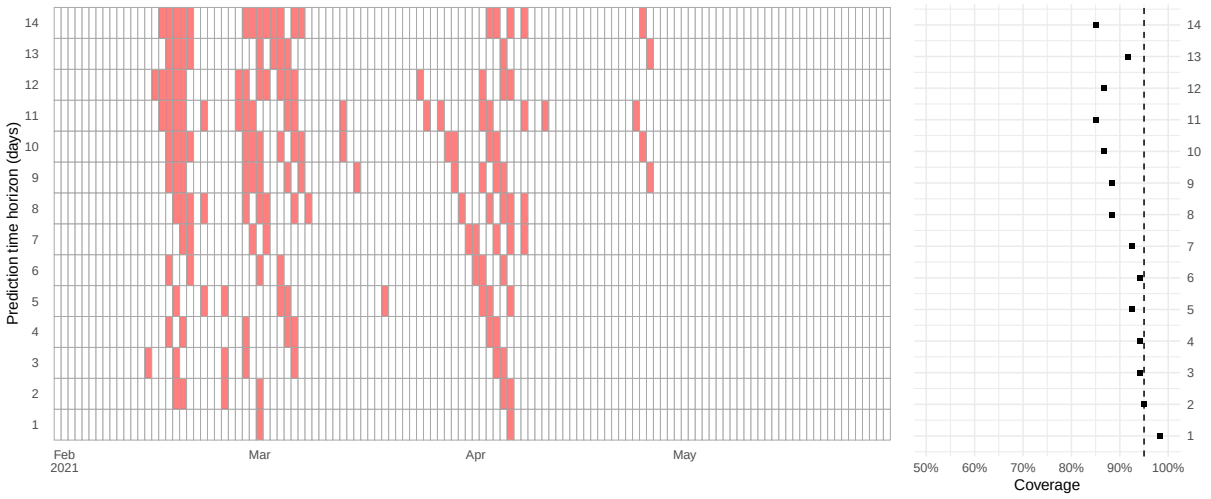

Figure 5. Prediction intervals coverage for the GAM beta regression model of TPR at 95% nominal level. The graph on the left shows the inclusion (blank cells) or exclusion (red cells) of the observed value of TPR for the prediction intervals in subsequent dates corresponding to the time horizon from 1 up to 14 days. The empirical coverage percentages are shown on the right graph, with the vertical dashed line representing the nominal level.

4.4. COVINDEX estimate

Based on the estimated model and uncertainty for the test positive rate, the COVINDEX is computed following equation (4). Figure 6 shows the estimated COVINDEX with 95% credible intervals. Notice that the y-axis is expressed on logarithmic scale, the natural scale to visualize ratios (Wilke, 2019, , Sec. 3.2). The index fluctuates widely throughout the year 2020, following the periods of expansion and contraction of the spread of the pandemic. After the first wave in spring 2020 we observe a quick decreasing trend, followed by a slowly increase during the summer, corresponding to a relaxation of the containment measures, with values significantly larger than 1.0 during August. This represents the first signal of a resurgence of the pandemic. Sharpe and large increases are also observed during October in conjunction with the second wave that strongly affected Italy. Two additional peaks are detected at the end of 2020 and on February 2021, corresponding to gradual relaxation of containment measures before Christmas and mid January, with the latter that occurred during the period of political instability associated with the change of government.

From the adopted definition of equation (4), COVINDEX is computed by taking the ratio of the estimated TPR with respect to a 7-days-before estimate. In Section 3 we provide the rationale for this choice. However, it may be interesting to investigate how the index changes assuming different

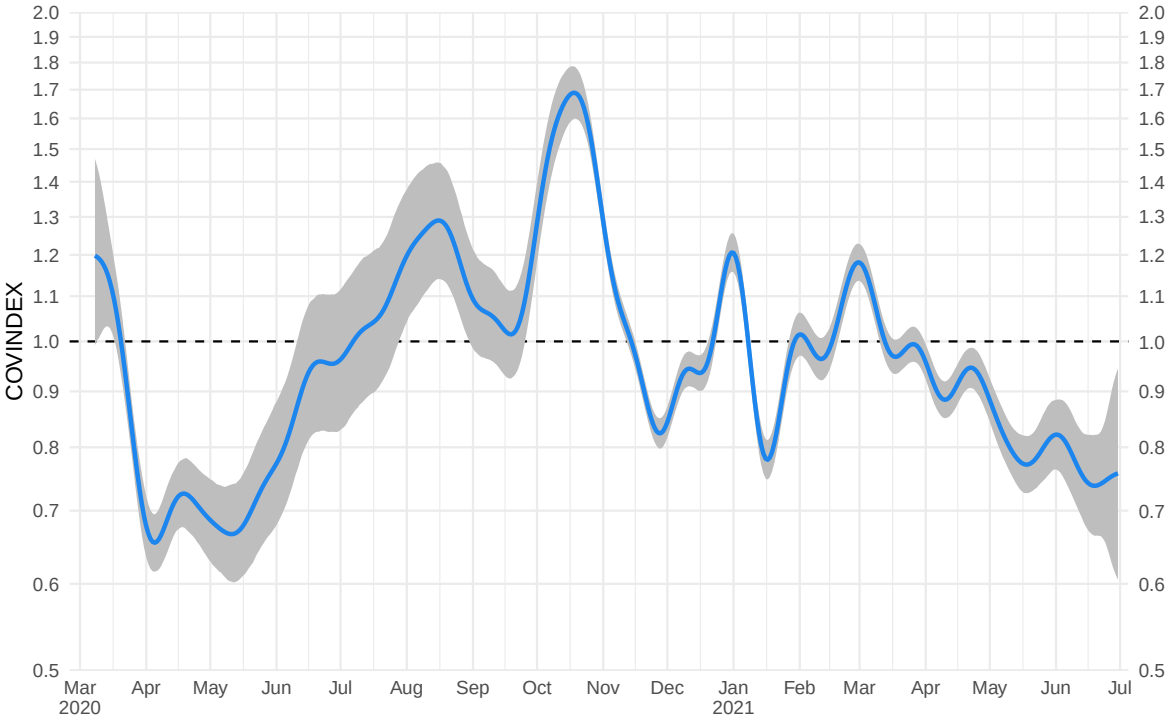

Figure 6. Evolution of COVINDEX (on logarithmic scale) for Italy during 2020 and first half of 2021, with approximate 95% simulated credible intervals.

lags. Figure 7 shows the COVINDEX estimates obtained when different lag values are used. The general behaviour of the curves is similar across different lags, but the amplitude of the oscillation increases as the lag increases. This appears reasonable since, essentially, COVINDEX compares the estimated TPR at time t with the value at time $t-\text{lag}$. Thus, for smaller values of the lag the index fluctuates less and is more stable than at higher lag values. However, lag values that are too small cannot highlight the dynamics of TPR because too close values tend to be quite similar. In this sense, the selected 7-day lag appears to be a sensible choice.

4.5. TPR-COVINDEX risk quadrant chart analysis

Figure 8 shows the TPR-COVINDEX risk quadrant chart for Italy, with points connected following the temporal path, a graph also known as connected scatterplot (Haroz et al., 2015). The curve concisely represents the evolution of both indices during the pandemic. Starting with the critical situation in March 2020, the situation improved in the following months, moving from the red quadrant to the orange bottom-right quadrant and then the yellow quadrant during summer 2020. By the end of summer 2020 we observe a worsening of the situation that lead to the red quadrant in November. In the following months there has been a constant oscillation between the red and right-orange quadrants, indicating a serious pandemic situation.

A scatterplot of TPR vs COVINDEX is also useful for surveillance of the pandemic in different Italian regions. Figure 9 summarizes the status of the pandemic for the Italian regions at selected time points. A high-risk situation is observed at the beginning of November 2020, where all regions belong to the red quadrant. The following month saw an improvement with most regions moving towards the bottom-right orange quadrant. A more complex and varied situation is observed between February and March 2021, with some regions moving from the red to the orange quadrant, and vice versa for other regions.

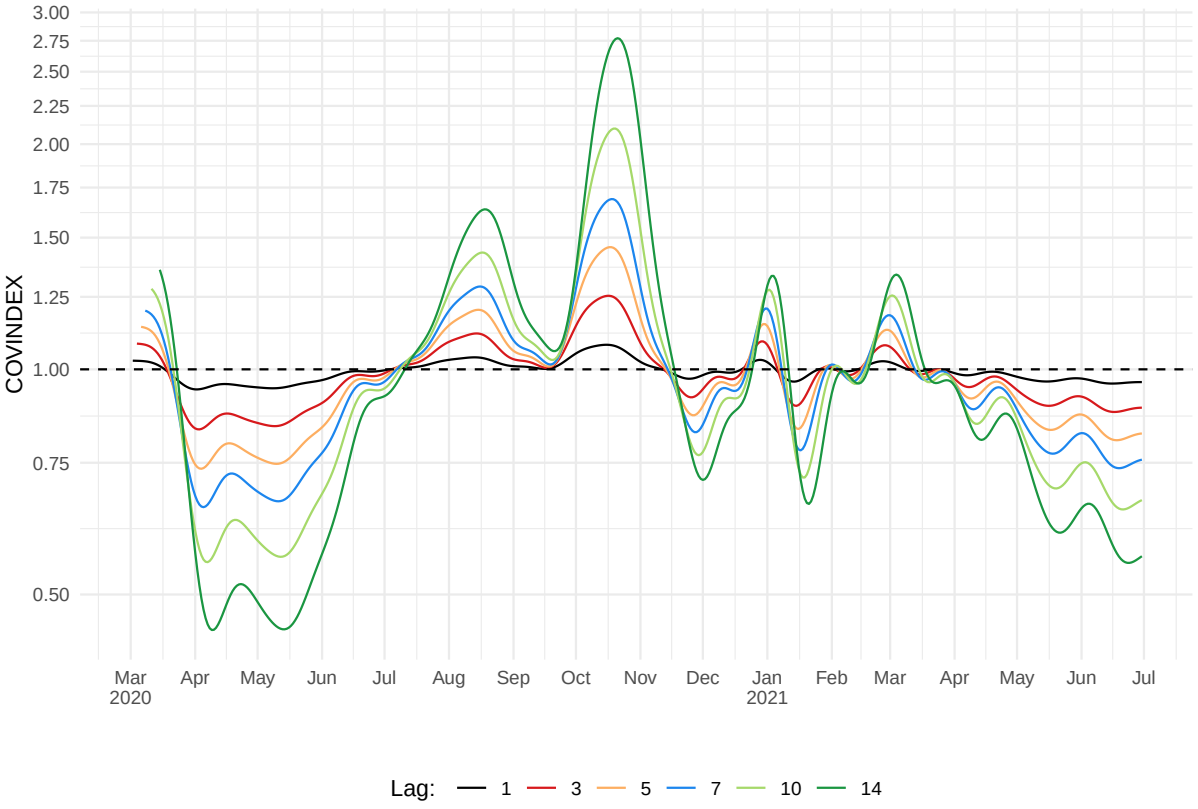

Figure 7. A comparison of COVINDEX (on logarithmic scale) computed at different lags for Italy on 2020 and first half of 2021.

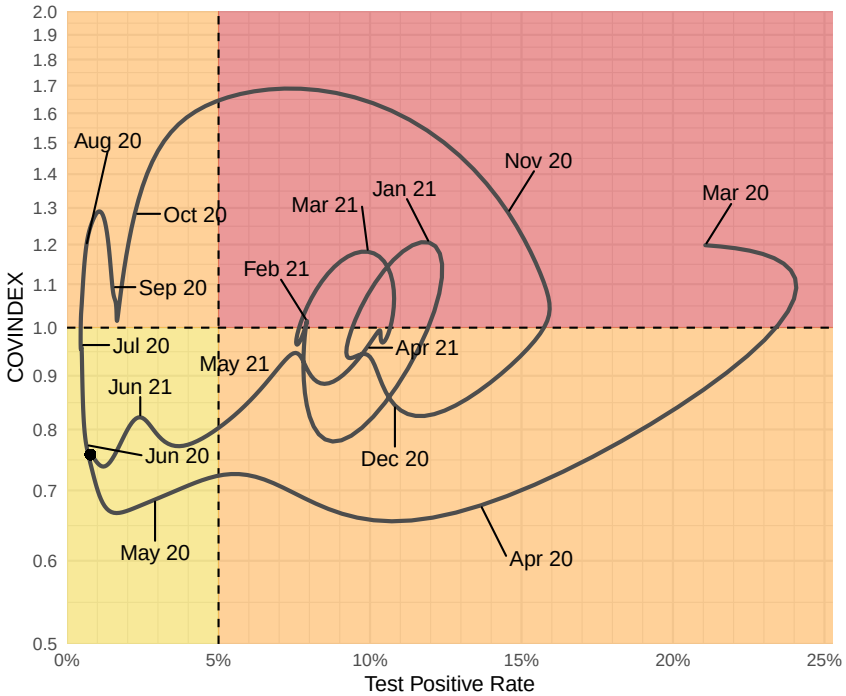

Figure 8. TPR-COVINDEX risk chart as connected scatterplot for Italy. The first day of each month is highlighted to provide a temporal reference.

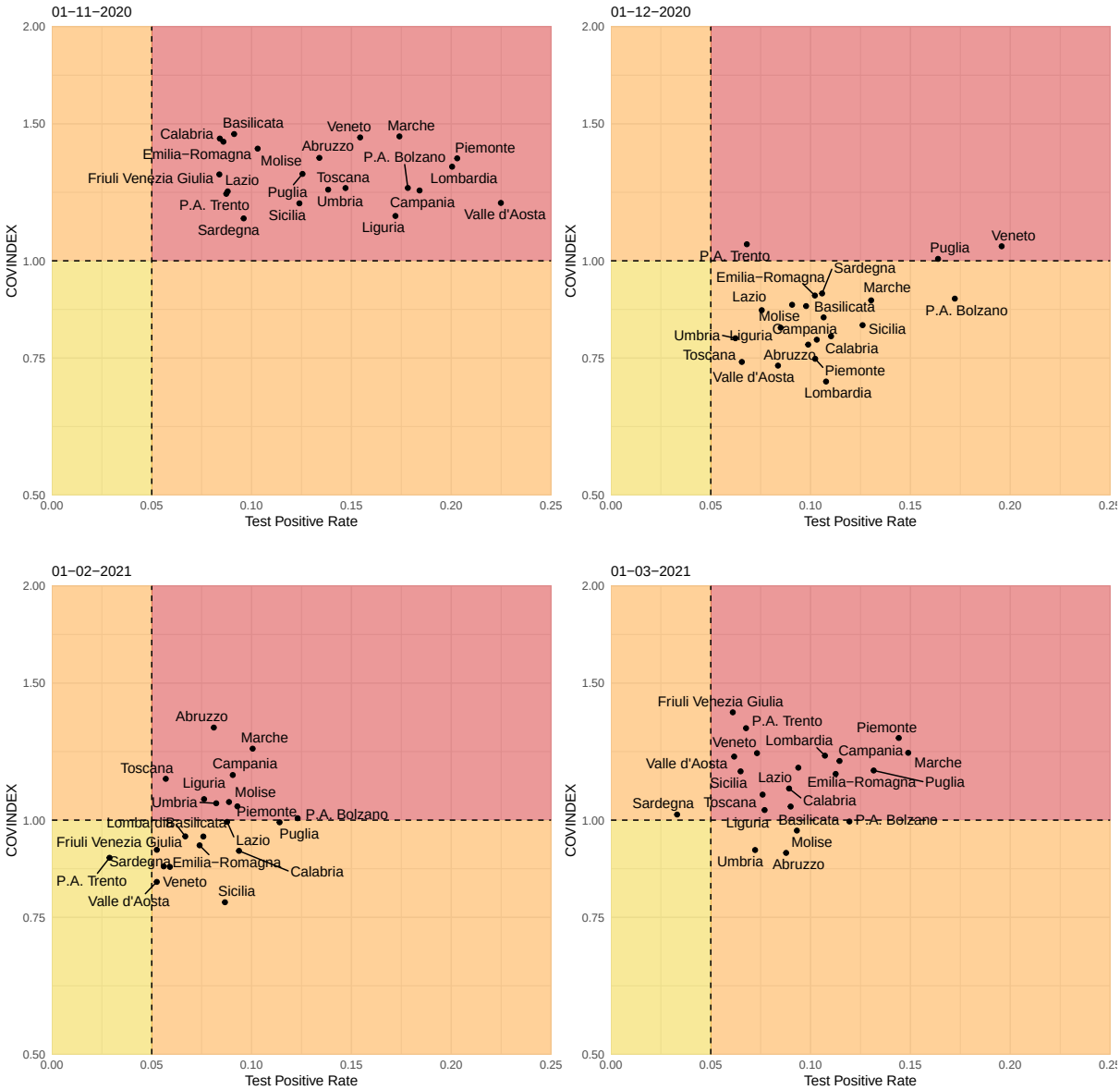

Figure 9. TPR-COVINDEX plot for Italian regions at different time points.

5. A comparison of COVINDEX with the effective reproduction number

The main index used in Italy for pandemic surveillance is R_t, the effective reproduction number. The procedure employed for estimating R_t is described by Guzzetta and Merler (2020) and it is based on the Bayesian methodology of Cori et al. (2013). Details can be found at `https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati`. An archive containing both the data and the R script used by the Italian National Institute of Health (ISS) for computing R_t is available at `https://www.epicentro.iss.it/coronavirus/open-data/calcolo_rt_italia.zip`.

In this Section we provide a comparison of the proposed COVINDEX with the values of R_t estimated following the procedure outlined above for Italy from March 2020 to June 2021. Furthermore, since the effective reproduction number does not provide a timely snapshot of the evolution of the pandemic, we also provide two examples showing the failure of R_t to highlight the likely evolution of the pandemic and we compare its behaviour with the proposed COVINDEX.

The top graph reported in Figure 10 shows the estimated curves for COVINDEX and R_t. Overall, a similar trend can be observed for the two curves, particularly since October 2020. R_t appears to be more wiggly than COVINDEX, especially during the summer 2020. Likely, this is related to the large uncertainty in that period due to the relative small number of positive cases (around few hundreds) observed in that period. One of main drawbacks of using R_t for real-time monitoring is shown in the final part of the graph. In fact, if at the end of the June the COVINDEX curve seems to suggest a resumption of the pandemic, the R_t index continues to show a decreasing trend This behaviour can also be seen in other time periods, as discussed below.

As mentioned in Section 4.4, the second wave of COVID-19 epidemic hit Italy between the second half of October and the beginning of November 2020, followed by a rapid decrease during the remainder of the month. However, from the beginning of December 2020 it was evident that this decline had stopped and that the situation was starting to get worse. This is clearly indicated by the upward slope of the COVINDEX computed on December 5th and shown in the bottom-left graph of Figure 10. On the contrary, the R_t index calculated on the same day, with estimates considered valid up to 14 days before, produces a curve which erroneously suggests a decline in the spread of the pandemic. However, if the R_t curve is estimated a week later, we begin to see an increase in the spread of the pandemic (see the dotted red curve in bottom-left graph of Figure 10). The main problem is that such alert is reported too late.

A similar situation is also faced at the beginning of March 2021. After a period of almost constant positive rate during February 2021, with both COVINDEX and R_t oscillating around 1.0, by the end of the month there was a clear increase of the test positive rate. This was immediately signalled by COVINDEX computed on February 28th 2021 (see bottom-right graph in Figure 10), but R_t computed on the same day was still signalling a steady state and only after a week an increasing value of R_t would have signalled the resurgence of the pandemic.

The comparison between COVINDEX and R_t can also be conducted at the regional level. Here we present a comparison for two Italian regions, Lombardia and Umbria. These are two very different regions, both in size and geographical position, but also in terms of pandemic history. If Lombardia was the most affected region of Italy during the 1st wave of the COVID-19 pandemic, Umbria suffered only marginal effects in this phase. On the contrary, the so-called 3rd wave that occurred in winter/spring 2021 hit Umbria earlier than in the rest of Italian regions, including Lombardia.

Likewise the national level, there is a substantial similarity between the trend of COVINDEX and R_t for the two regions, with the former which appears to have a smoother behaviour (see Figure 11). Both indices correctly identified the peak of the pandemic in October 2020 and at the end of 2020. But if for Umbria the beginning of 2021 is marked by the arrival of the third wave caused by the circulation of SARS-CoV-2 variants, namely Alpha (or English) and Gamma (or Brazilian), in Lombardia the presence of these variants only occurred from mid-February. Subsequently, starting from spring 2021, a decline in the epidemic can be observed in both regions.

However, there are also differences that are worth pointing out. For Lombardia there are two values of R_t, in mid-February and early June, which appear suspicious as they are placed outside the

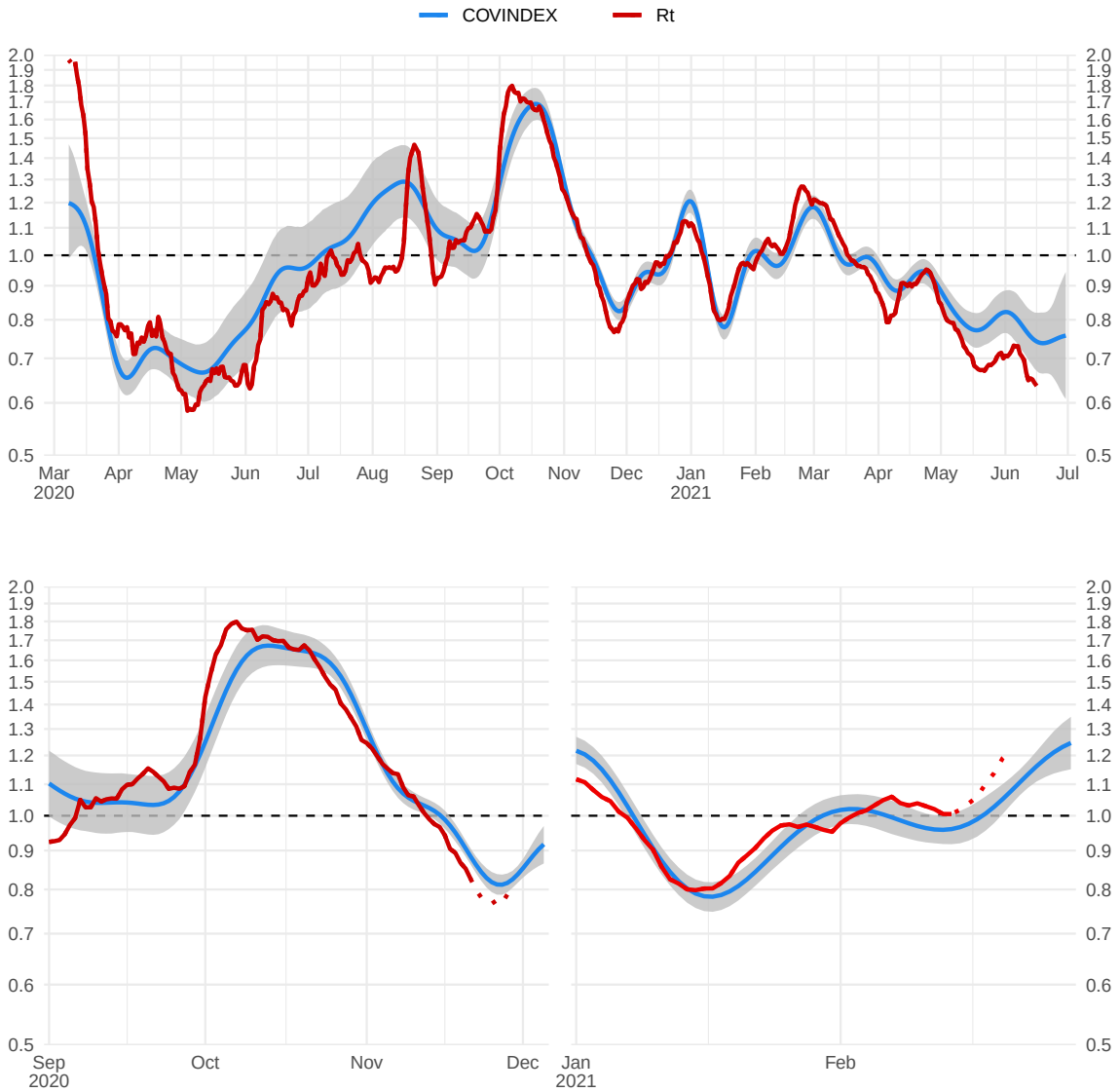

Figure 10. Comparison of COVINDEX and R_t for Italy. Bottom panels show the comparison at early December 2020 (left) and at the end of February 2021 (right).

apparent trend, underestimating in the first case and overestimating the trend in the second case. For Umbria, the R_t seems to increase starting from the last week of May, while the COVINDEX still suggests a decreasing trend. This behaviour of R_t is also suspect as the TPR of the region remains substantially stable or slightly decreasing in this period, with all the TPR values less than 1% in the last 10 days of June.

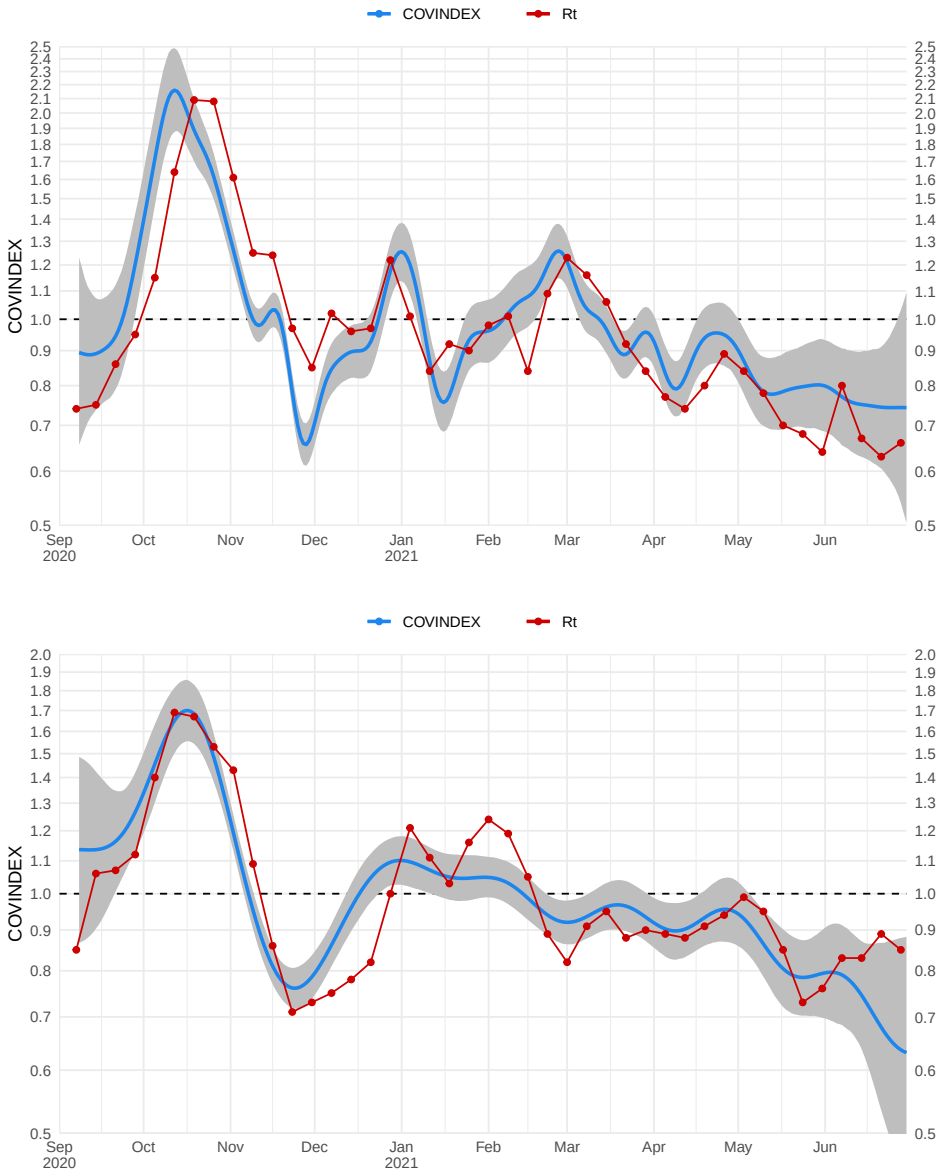

Figure 11. Comparison of COVINDEX and R_t for the Italian regions Lombardia (top) and Umbria (bottom), from September 2020 to June 2021.

6. Final comments

In this paper we have proposed an index, named COVINDEX, that can be used for near real-time monitoring of COVID-19 pandemic. The index is computed as the ratio of the estimated test positive rate on a given day with respect to the value estimated for a week before. Estimation of test positive rates is obtained by statistical modelling the daily empirical positive rates calculated from the observed data. To this end, a GAM beta regression model with weights proportional to the administered tests is fitted. By exploiting the relationship of penalized likelihood for GAMs with MAP Bayesian estimation, credible intervals for COVINDEX can be obtained via simulation to express the associated uncertainty.

We applied the proposed methodology to the Italian COVID-19 outbreak and we compared the trend of COVINDEX to the effective reproduction number R_t. The analyses carried out confirm that R_t is a delayed index of epidemic trend, and for this reason may provide a biased picture of the current pandemic status. On the contrary, COVINDEX seems to provide a more up-to-date information which

can be used as a decision-making tool. This aspect is of crucial importance for all policy makers and public health officials. We defer to future research the evaluation of the implications deriving from the adoption of the proposed index.

Although the main focus of the analysis in this paper was the national level, similar considerations can be made for territorial administrative divisions, such as regions and provinces. In these cases, however, it should be noted that further assumptions are necessary, in particular the independence of the epidemic trend between neighbouring territories.

All the analyses have been performed in `R` version 4.1.0 (R Core Team, 2021), using the package `mgcv` (Wood, 2021) and functions written by the author. Code to reproduce the analyses is available in a GitHub repository at `https://github.com/luca-scr/COVINDEX`.

References

Adam, D. (2020). A guide to R – the pandemic's misunderstood metric. *Nature*, 583(7816):346–348. `https://www.nature.com/articles/d41586-020-02009-w`.

Agosto, A., Campmas, A., Giudici, P., and Renda, A. (2021). Monitoring COVID-19 contagion growth. *Statistics in Medicine*.

Alaimo Di Loro, P., Divino, F., Farcomeni, A., Jona Lasinio, G., Lovison, G., Maruotti, A., and Mingione, M. (2021). Nowcasting covid-19 incidence indicators during the italian first outbreak. *Statistics in Medicine*, 40(16):3843–3864.

Bartolucci, F. and Farcomeni, A. (2021). A spatio-temporal model based on discrete latent variables for the analysis of COVID-19 incidence. *Spatial Statistics*.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512.

Douma, J. C. and Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution*, 10(9):1412–1430.

Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., and 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.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., and Ventura, L. (2020). Robust inference for non-linear regression models from the tsallis score: Application to coronavirus disease 2019 contagion in italy. *Stat*, 9(1):e309.

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., et al. (2020). Practical considerations for measuring the effective reproductive number, R_t. *PLoS Computational Biology*, 16(12):e1008409.

Guzzetta, G. and Merler, S. (2020). Stime della trasmissibilità di SARS-CoV-2 in Italia. EpiCentro - Istituto Superiore di Sanità: `https://www.epicentro.iss.it/coronavirus/open-data/rt.pdf`.

Haroz, S., Kosara, R., and Franconeri, S. L. (2015). The connected scatterplot for presenting paired time series. *IEEE Transactions on Visualization and Computer Graphics*, 22(9):2174–2186.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43. Chapman & Hall/CRC.

Hilton, J. and Keeling, M. J. (2020). Estimation of country-level basic reproductive ratios for novel coronavirus (SARS-CoV-2/COVID-19) using synthetic contact matrices. *PLoS Computational Biology*, 16(7):e1008031.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382:1199–1207.

Maruotti, A., Ciccozzi, M., and Divino, F. (2021). On the misuse of the reproduction number in the COVID-19 surveillance system in Italy. *Journal of Medical Virology*, 93(5):2569–2570.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, CRC, London, 2nd edition.

Nazar, Z. and Elfadil, A. (2021). The estimations of the COVID-19 incubation period: A scoping reviews of the literature. *Journal of Infection and Public Health*, 14(5):638–646.

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, 51(1):111–126.

Presidenza del Consiglio dei Ministri – Dipartimento della Protezione Civile (2020). Dati COVID-19 Italia. GitHub repository: `https://github.com/pcm-dpc/COVID-19`.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sebastiani, G., Massa, M., and Riboli, E. (2020). Covid-19 epidemic in italy: evolution, projections and impact of government measures. *European journal of epidemiology*, 35(4):341–345.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71.

Wilke, C. O. (2019). *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media.

Wood, S. (2021). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.8-36.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized Additive Models: an introduction with R*. Chapman & Hall/CRC, 2nd edition.

Wood, S. N. (2020). Inference and computation with generalized additive models and their extensions. *Test*, 29(2):307–339.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

World Health Organization (2019). Public health criteria to adjust public health and social measures in the context of COVID-19. Annex to Considerations in adjusting public health and social measures in the context of COVID-19, 12 May 2020: `https://apps.who.int/iris/bitstream/handle/10665/332073/WHO-2019-nCoV-Adjusting_PH_measures-Criteria-2020.1-eng.pdf`.

World Health Organization (2020). Considerations for implementing and adjusting public health and social measures in the context of COVID-19. Iterim guidance, 4 November 2020: `https://www.who.int/publications/i/item/considerations-in-adjusting-public-health-and-social-measures-in-the-context-of-covid-19-interim-guidance`.

Zeileis, A. and Cribari-Neto, F. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2):1–24.