The Covid-19 pandemic in Italy

Franco Peracchi*
Georgetown University, EIEF, and University of Rome Tor Vergata

This version: March 30, 2020

1 Introduction

In this note I use the data from COVID-19 Italia, the website of Dipartimento della Protezione Civile (DPC), which at the moment represent the main source of quantitative information on the Covid-19 pandemic in Italy and are described in some more detail in Section 2. These data are updated daily, so I will also update my note daily and post it to EIEF's Covid-19 Forum shortly after the DPC data have been released.

What I try to do here is just to describe and summarize some basic aspects of the pandemic in Italy with no attempt at structural modeling. Neverheless, I hope that the stylized facts presented in Section 3 will help in sorting out possible interpretations of the process that is still unfolding. However, in Section 4, I will venture into a simple forecasting exercise based on the extrapolation of the observed trends.

2 Data

The DPC website contains several daily time-series at the national, regional and provincial level. I focus on the following five series: the number of currently positive cases ("totale attualmente positivi"), the number of new positive cases ("nuovi attualmente positivi"), the number of discharged ("dimessi guariti"), the number of deaths ("deceduti"), and total number of confirmed cases ("totale casi"). Notice that the number of currently positive cases is not equal to the number of currently infected members of the population, but only to the number of those who have been tested positive. The number of currently infected people is likely to be an order of magnitude larger (see Li et al., 2020). Further notice that the ratio of currently positive to the currently infected should not be taken as approximately constant, as the criteria and the intensity of the testing process is likely to vary both over time and across regions. The new positive cases is the daily variation in the number of

* E-mail address: fp211@georgetown.edu. I thank Luca Anderlini, Giuseppe De Luca, Raffaela Giordano, Luigi Guiso, Caterina Peracchi, Marta Peracchi, and Daniele Terlizzese for comments and helpful discussions.

currently positive cases, while the total number of confirmed cases is the sum of the currently positive cases, the deaths and the discharged. These data are available, starting with February 24, 2020, both at the national and the regional level, and are updated daily around 6:00 pm CET. Daily data on the number of total confirmed cases by province are also available from the DPC website but are not used in this note.

The DPC data are very noisy and currently offer no breakdown of the cases by age and gender (for a discussion of the value of this information in the present context see Beam Dowd et al., 2020). The noise reflects a number of factors that have been stressed by various commentators, such as delays in recording and transmitting the information, uncertainty in the classification of cases, clerical errors, etc. This is unsurprising, as people these days are battling for saving human lives, often in dramatic conditions (see Nacoti et al., 2020), not for perfecting statistical information. However, because the noise is large and pervasive, especially at the regional and provincial level, smoothing the data becomes essential in order to detect underlying trends that would otherwise be obscured. Of course, different smoothing methods may sometimes lead to different conclusions.

While I rely a lot on smoothing, though simple moving averages and polynomial trends, I do not pre-process the original data except for merging the autonomous provinces of Trento and Bolzano into a single region, "Trentino-Alto Adige", resulting in a geographical disaggregation of the country into 20 regions.

I produce the descriptive statistics in Section 3 and the forecasts in Section 4 using Stata/MP version 16.1. The Stata code is available upon request.

3 Descriptive statistics

Figures 1 and 2 summarize the Covid-19 pandemic at the Italian level, and various versions of them have appeared in the media.

Figure 1 shows the daily changes in the number of positive cases (blue profile), deaths (red profile), discharged (green profile) and in the total number of confirmed cases (purple profile). In addition to the original series (represented by the thinner profile), I also present a 5-day centered moving average (represented by the thicker profile) as a visual aid in highlighting the trends. The time profile of the changes in the number of positive cases is approximately hump-shaped, suggesting that the pandemic may be slowing down.

Figure 2 shows the percentage daily changes in the number of positive cases, dead, discharged, and total confirmed cases. It reveals an almost steady downward trend after the very high growth rates of the number of deaths and positive cases in late February and early March.

Figures 3 and 4 are similar to Figures 1 and 2 but look separately at nine Italian regions selected either for their size or for the particular intensity of the pandemic: Campania, Emilia-Romagna, Lazio,

Lombardia, Marche, Piemonte, Puglia, Sicilia, and Veneto (the figures for the other eleven regions are available upon request). These two figures are meant to give a sense of the large differences between Italian regions.

Figure 3 presents the daily changes in the number of positive cases, deaths, discharged and total confirmed cases by region. Notice the very different vertical scales across regions. The regional differences in the number of new positive cases are huge, with the Northern regions of Emilia-Romagna, Lombardia, Piemonte and Veneto being the hardest hit. However, the qualitative differences in the time profiles are much smaller.

Figure 4 presents the percentage daily changes by region over the last 20 days. Although the regions with a lower incidence of the pandemic display much larger fluctuations, the similarity of the process across regions is quite striking, with the observed peaks in the rate of increase of mortality following by about 10–15 days the peaks in the rate of increase of new positive cases.

4 Forecasts

Figures 5 and 6 present the observed number of new positive cases at the national level (thin blue profile) together with in-sample predictions and out-of-sample forecasts of the number of new positive cases (thick orange profile) up to 50 days from the current date (March 30, 2020) using a common model but different fitting criteria. The observed number of new positive cases is exactly the same shown in Figure 1. To predict or forecast the number of new positive cases, I employ a polynomial time trend model for the natural logarithm of the number of new positive cases at time (day) t, denoted in what follows by Y_t. Bonetti (2020) uses a similar strategy for analyzing the data on the fraction of currently positive at the provincial level in Italy. This model provides a simple way of allowing for the fact that Y_t is a nonnegative number but can assume very large values.

Specifically, my statistical model for Y_t is

$$\ln Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k + U_t, \quad t = 0, 1, \ldots, \tag{1}$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are unknown parameters, U_t is an unobservable random error with zero mean and finite variance, and $t = 0$ corresponds to February 24, 2020. The statistical model (1) is just a "black box" that I use for forecasting the future number of new positive cases under the strong assumption of time invariance of the process underlying the Covis-19 pandemic in Italy. Given estimates $\widehat{\beta}_{0|T}, \widehat{\beta}_{1|T}, \widehat{\beta}_{2|T}, \ldots, \widehat{\beta}_{k|T}$ of the model parameters based on all the data available up to the current date T, my predictions of Y_t at any date up to the current date T and my forecasts of Y_t at any future date $t = T + 1, T + 2, \ldots$, are of the form

$$\widehat{Y}_{t|T} = \exp(\widehat{\beta}_{0|T} + \widehat{\beta}_{1|T} t + \widehat{\beta}_{2|T} t^2 + \cdots + \widehat{\beta}_{k|T} t^k). \tag{2}$$

The inverse exponential transformation in the forecast rule (2) automatically enforces the non-negativity constraint on the predictions. After experimenting with various specifications, I found that either a quadratic or a cubic trend model ($k = 2$ or 3) provide a reasonable compromise between goodness of fit and parsimony. My choice is also supported by the results of a more formal model selection method based on minimization of the Schwarz Information Criterion, also know as BIC. My two fitting criteria are ordinary least squares (OLS), in which case $\widehat{Y}_{t|T}$ is interpreted as an estimate of the mean of the probability distribution of new positive, and least absolute deviations (LAD), in which case $\widehat{Y}_{t|T}$ is interpreted as an estimate of the median of the probability distribution of new positives. Although the two estimates should be about the same if the probability distribution of new positives is approximately symmetric and has moderate tails, I employ both because the LAD estimates are much less sensitive than the OLS estimates to outliers, which represent a big problem in these data.

Notice that my out-of-sample forecasts are based on all the data available up to the current date T. As soon as new DPC data are released, I produce a new set of estimated parameters $\widehat{\beta}_{0|T+1}, \widehat{\beta}_{1|T+1}, \widehat{\beta}_{2|T+1}, \ldots, \widehat{\beta}_{k|T+1}$ and a new set of forecasts $\widehat{Y}_{T+2|T+1}, \widehat{Y}_{T+3|T+1}$, etc. Over time, with the availability of more and more data, and possibly also with the revision of previously released data, the uncertainty associated with the forecasts is likely to be reduced, especially if the pandemic (or the data collection effort by the DPC) continues without major structural breaks.

Given the forecast rule (2), I can predict the day when the pandemic will end. This is just the day t_0 such that $\widehat{Y}_{t_0|T} \approx 0$. Since $\widehat{Y}_{t|T}$ can only be zero in the limit, as $t \to \infty$, I predict the end date of the pandemic as the first day when the forecasted number of new positive cases is rounded to zero. Notice that the predicted date t_0 inherits the large statistical uncertainty associated with the entire sequence of forecasts $\widehat{Y}_{T+1|T}, \widehat{Y}_{T+2|T}, \ldots$

Figures 5 and Figure 6 present in-sample predictions and out-of-sample forecasts based on a quadratic time trend model (selected on the basis of the BIC) and, respectively, the OLS and the LAD criterion. The vertical black line marks the current date (March 30, 2020), while the vertical orange line marks the predicted end date of the pandemic in Italy. Symmetric pointwise standard-error bands of varying width are included to provide an indication of the large statistical uncertainty associated with these forecasts. Since the standard errors increase rapidly with t, the shaded regions tend to "curl-up" fast. For simplicity, I employ classical standard errors based on the homoskedasticity assumption (see, for example, Stock and Watson, 2019). The LAD standard errors are larger than the OLS standard errors, so the forecasts based on the LAD estimates, though more robust to outliers, are subject to greater uncertainty.

Figure 7 presents in-sample predictions and out-of-sample forecasts of the new positive cases up to 40 days from the current date by region, based on a cubic trend model, i.e. $k = 3$ in (1). My forecast

rule for the number Y_{rt} of new positives in region r, where r is any of the 20 Italian regions, has the same form as (2), namely $\widehat{Y}_{rt|T} = \exp(\widehat{\beta}_{0r|T} + \widehat{\beta}_{1r|T}t + \widehat{\beta}_{2r|T}t^2 + \widehat{\beta}_{3r|T}t^3)$, where $\widehat{\beta}_{0r|T}, \widehat{\beta}_{1r|T}, \widehat{\beta}_{2r|T}$ and $\widehat{\beta}_{3r|T}$ are parameters specific to region r estimated from the data available up to the current date T. For a few regions the out-of-sample forecasts are not presented because my model performs poorly. Figure 8 compares the aggregate OLS and LAD in-sample predictions and out-of-sample forecasts with those obtained by aggregation of the regional LAD predictions and forecasts.

Finally, Table 1 presents the dates when the Covid-19 pandemic is predicted to end, both for Italy as a whole and for each region for which the model can be estimated. The aggregate model (1) gives different dates depending on whether it is estimated by OLS (more sensitive to outliers) or LAD (less sensitive to outliers). The predicted end date for the country based on the regional disaggregation is the latest end date predicted at the regional level.

References

Beam Dowd, J., V. Rotondi, L. Andriano, D. M. Brazel, et al. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. Technical report, Leverhulme Centre for Demographic Science, University of Oxford & Nuffield College. Available at https://www.medrxiv.org/content/10.1101/2020.03.15.20036293v1.

Bonetti, M. (2020). Epilocal: a real-time tool for local epidemic monitoring. Technical report, Carlo F. Dondena Research Center, Bocconi University. Available at https://arxiv.org/abs/2003.07928.

Li, R., S. Pei, B. Chen, Y. Song, et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* (March 16, 2020). Available at https://science.sciencemag.org/content/early/2020/03/24/science.abb3221.abstract.

Nacoti, M., A. Ciocca, A. Giupponi, P. Brambillasca, et al. (2020). At the epicenter of the Covid-19 pandemic and humanitarian crises in italy: Changing perspectives on preparation and mitigation. *NEJM Catalyst: Innovations in Care Delivery* (March 21, 2020). Available at https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0080.

Stock, J. H. and M. W. Watson (2019). *Introduction to Econometrics*. New York: Pearson.

Table 1: Dates when the Covid-19 pandemic is predicted to end.

Region	Predicted date
Abruzzo	April 11
Basilicata	April 6
Calabria	April 9
Campania	.
Emilia-Romagna	April 24
Friuli-Venezia Giulia	April 10
Lazio	April 13
Liguria	April 12
Lombardia	April 22
Marche	.
Molise	.
Piemonte	April 15
Puglia	April 12
Sardegna	.
Sicilia	April 14
Toscana	May 1
Trentino-Alto Adige	April 10
Umbria	April 7
Valle d'Aosta	April 8
Veneto	April 15
Italy, LAD aggr	May 6
Italy, OLS aggr	May 10
Italy, LAD disaggr	May 1

Note: Predicted dates are set to missing (.) when the model produces diverging forecasts.

Figure 1: Daily changes in the number of cases: original series (thinner profile) and 5-day centered moving average (thicker profile).

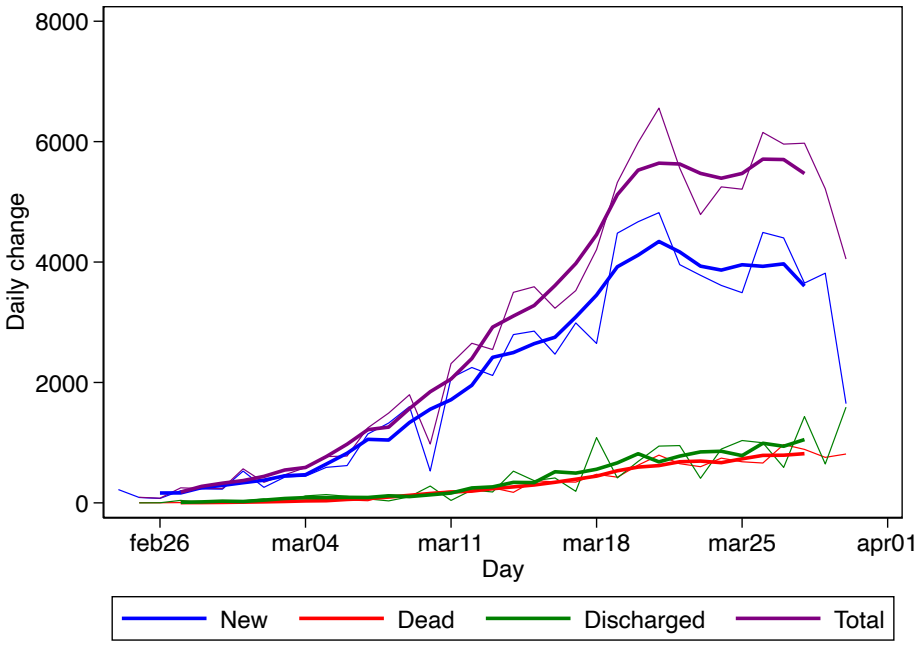

Figure 2: Percentage daily changes in the number of cases: original series (thinner profile) and 5-day centered moving average (thicker profile).

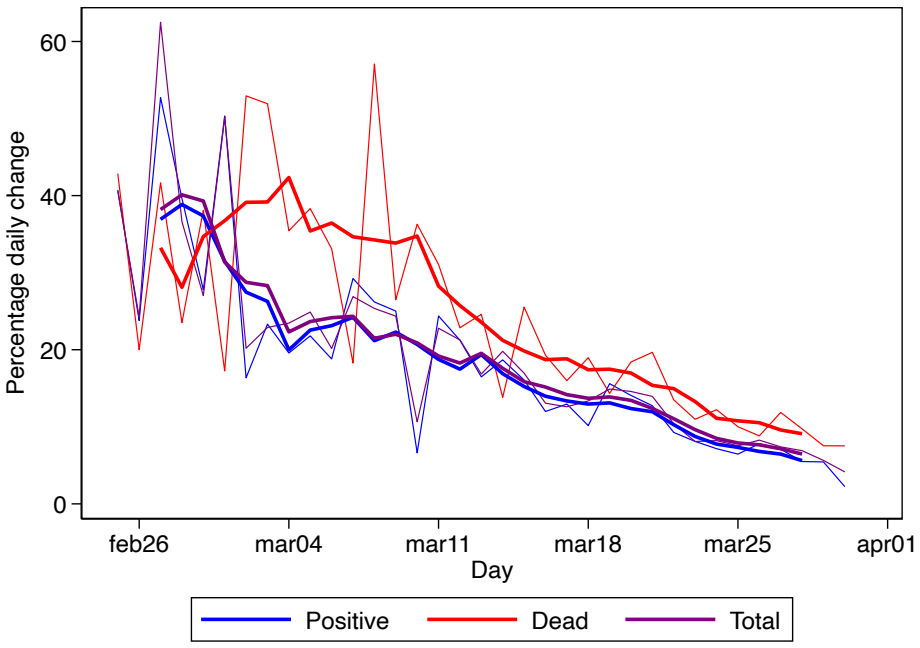

Figure 3: Daily changes in the number of cases by region: original series (thinner profile) and 5-day centered moving average (thicker profile).

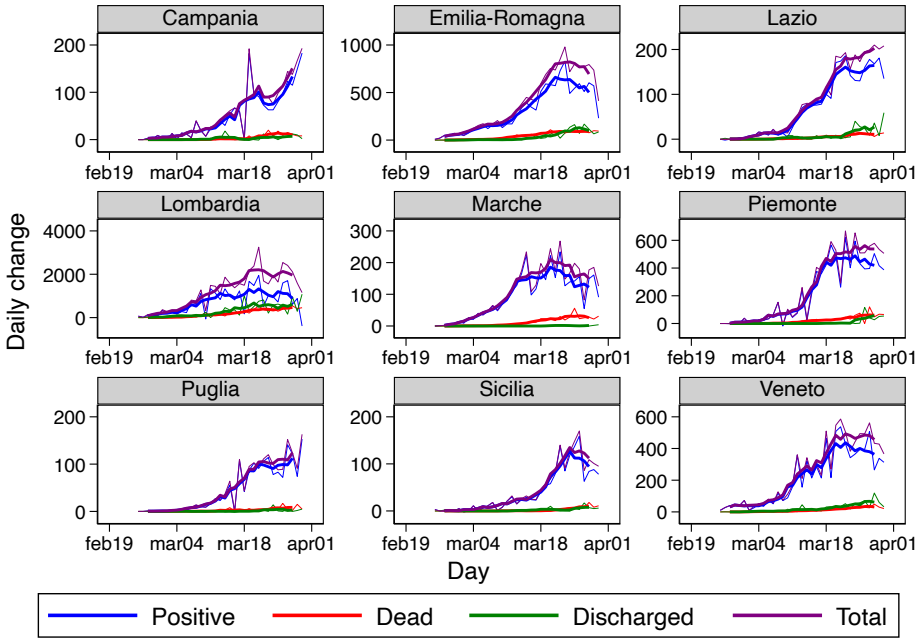

Figure 4: Percentage daily changes in the number of cases by region (last 20 days): original series (thinner profile) and 5-day centered moving average (thicker profile).

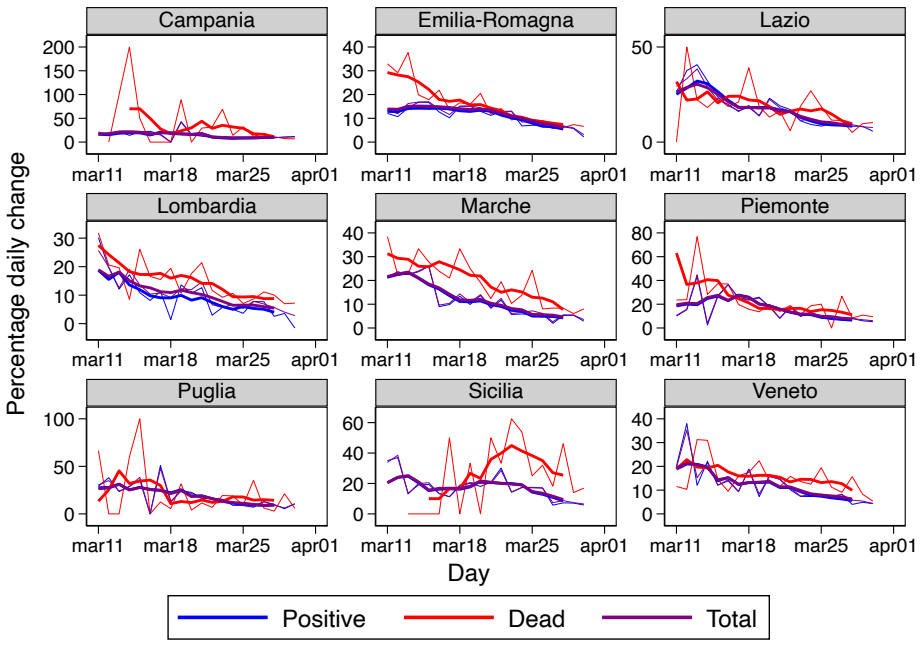

Figure 5: Observed and predicted or forecasted number of new positive cases. OLS predictions and forecasts . The vertical orange line marks the predicted end date of the pandemic in Italy.

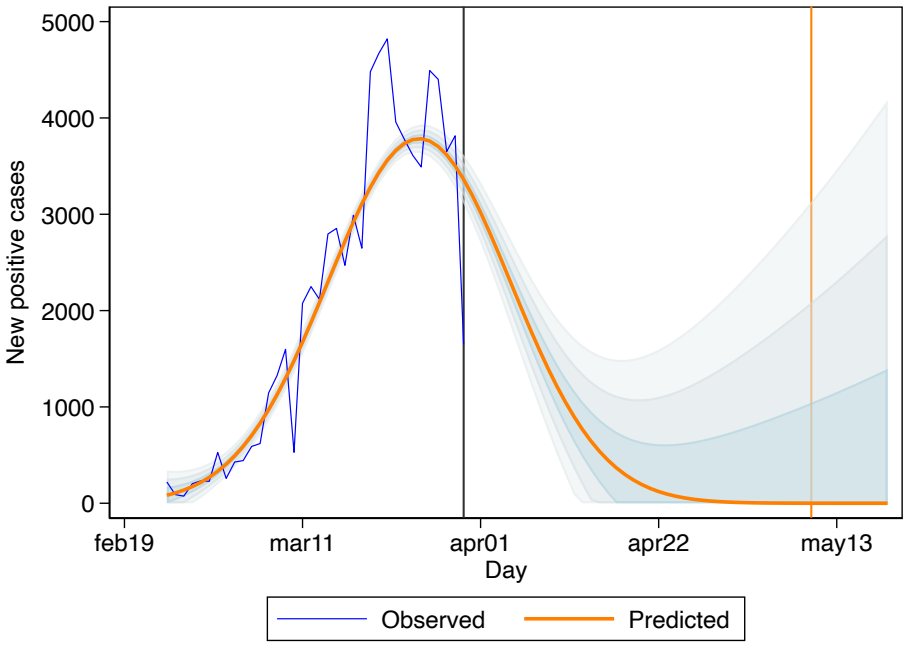

Figure 6: Observed and predicted or forecasted number of new positive cases. LAD predictions and forecasts. The vertical orange line marks the predicted end date of the pandemic in Italy.

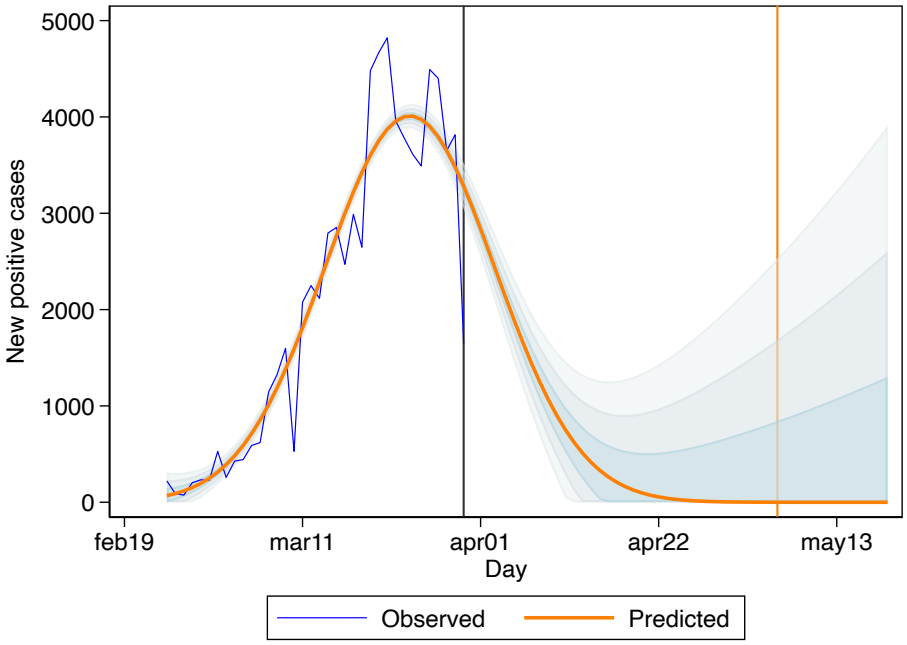

Figure 7: Observed and predicted or forecasted number of new positive cases by region. LAD predictions and forecasts.

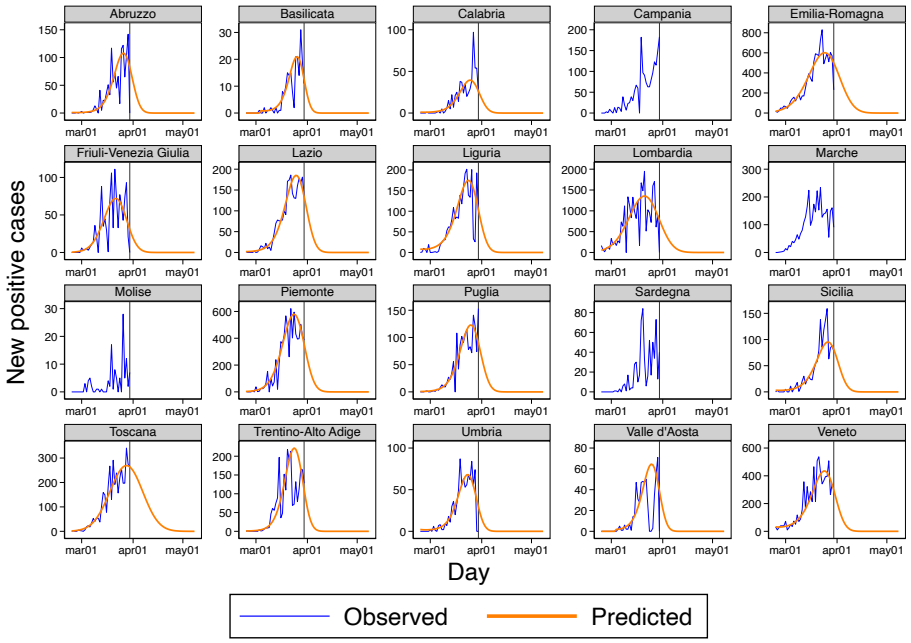

Figure 8: Observed and predicted or forecasted number of new positive cases. Aggregate OLS and LAD (OLS and LAD aggr) and aggregation of regional LAD prediction or forecasts (LAD disaggr).