A new regression model for the forecasting of COVID-19 outbreak evolution: an application to Italian data

Davide Sisti
University of Urbino Carlo Bo
Ettore Rocchi
University of Urbino Carlo Bo
Sara Peluso
University of Urbino Carlo Bo
Stefano Amatori (s.amatori1@campus.uniurb.it)
University of Urbino Carlo Bo https://orcid.org/0000-0001-7497-755X
Margherita Carletti
University of Urbino Carlo Bo

Research Article

Keywords: Novel coronavirus (SARS-CoV-2), COVID-19, curve growing, epidemic modelling, forecasting

Posted Date: June 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34576/v1

A new regression model for the forecasting of COVID-19 outbreak evolution: an application to Italian data

Davide Sisti [1], Ettore Rocchi [1], Sara Peluso [1], Stefano Amatori [1], Margherita Carletti [2]

[1] Department of Biomolecular Sciences, Unit of Biostatistics and Biomathematics, University of Urbino Carlo Bo, Urbino, Italy

[2] Department of Pure and Applied Sciences, University of Urbino Carlo Bo, Urbino, Italy

Correspondence:

Stefano Amatori

Email: s.amatori1@campus.uniurb.it

Tel. +39-349-2115150

Abstract

The novel coronavirus SARS-CoV-2 was first identified in China in December 2019. In just over five months, the virus affected over 4 million people and caused about 300,000 deaths. This study aimed to model new COVID-19 cases in Italy using a new curve. A new empirical curve is proposed to model the number of new cases of COVID-19. It resembles a known exponential growth curve which has a straight line as an exponent, but in the growth curve proposed, the exponent is a logistic curve multiplied for a straight line. This curve shows an initial phase, the expected exponential growth; then rises to the maximum value and finally reaches zero. We characterized the epidemic growth patterns for the entire Italian nation and for each of the 20 Italian regions. The estimated growth curve has been used to calculate the expected time of the beginning, the time related to peak, and the end of the epidemics. Our analysis explores the development of the epidemics in Italy and the impact of the containment measures. Data obtained are useful to forecast future scenario and the possible end of the outbreak.

Keywords: Novel coronavirus (SARS-CoV-2), COVID-19, curve growing, epidemic modelling, forecasting

Introduction

The novel coronavirus SARS-CoV-2, responsible of COVID-19 epidemic, was first identified in Wuhan, China, in December 2019 among a cluster of patients that presented with an unidentified form of viral pneumonia (Peeri et al. 2020); it is reported to cause a range of symptoms including fever, cough, and shortness of breath (Huang et al. 2020). At the time of writing (data updated on May 11th, 2020, GMT 07:46), the number of persons infected worldwide is 4.193.381 with 283,992 reported deaths; 212 countries have confirmed cases. In Italy, the first official cases appeared on February 21st in the Lombardia region; eleven municipalities in northern Italy were identified as the centres of the two main Italian clusters and placed under quarantine. The majority of positive cases in other regions traced back to these two clusters. On March 8th, the quarantine was expanded to all of the Lombardia region and 14 other northern provinces, and on the following day to all of Italy, placing more than 60 million people in a de-facto quarantine mode. On March 11th, the Prime Minister prohibited nearly all commercial activity except for supermarkets and pharmacies, and on the same day, the World Health Organization recognized COVID-19 as a pandemic. On March 21st, the Italian government closed all non-essential businesses and industries, with additional restrictions to the movements of people. On May 11th, the total number of cases in Italy were 219.070, with 30.560 deaths; the total of tests performed was over 2 and half million.

Mathematical models are powerful tools for understanding infectious diseases (Gandon et al. 2016). Identifying signature features of the growth kinetics of an outbreak can be useful to design reliable models of disease spread and understand important details of the transmission dynamics of an infectious disease (Chowell et al. 2015). The force of infection in mathematical transmission models is typically estimated using time-series data that describe epidemic growth as a function of time (Viboud et al. 2016). Usually, an outbreak follows an exponential growth at an early stage, peaks, and then the growth rate decays, as containment measures to limit the transmission of the virus are introduced. Most of the

studies in the literature used simple exponential growth models and focused on the initial growing process, but on the other hand, there are also many works arguing that the number of infected people follows a trajectory different from a simple exponential growth (Wu et al. 2020). Moreover, most of the models adapt to the time series of the new infected and not to the daily variation of the same. However, it is undoubtedly useful trying to predict the trend of newly infected people on a day-by-day basis, as this type of time series is much more useful in predicting the end of the epidemic. Through this time series, it is easily identifiable the day with the maximum increase of the number of the infected, while in the functions that consider the total infected number according to the time it is necessary to define the inflexion point.

In this paper, we employed a new growth model, starting from the known exponential growth. The novelty lies in modelling the exponential exponent with a logistic function so that the exponent can decrease over time, rather than remain constant. Our analysis explores the outbreak development in Italy and the impact of the containment measures, both at the national level and within the 20 Italian regions. We employed simple models to quantitatively document the effects of the Italian control measures against the COVID-19, and provide informative projections on the development of the outbreak, and provide informative implications for the coming pandemic.

Methods

Mathematical assumption

A new empirical curve is proposed to modelling the number of new cases of COVID-19 as a function of time. The simple curve of exponential growth is defined by two parameters: *a* (is the number of times per unit time of growing by a factor *e, i.e.* the base of the natural logarithm) and *b* (related to abscissa values translation). The exponential growth curve is $N(t) = e^{a(t-b)}$, where *N(t)* is the number of new COVID-19 infected per day, *t* is the number of days since the start of registration of the infected (Fig. 1a).

[Insert Figure 1]

Fig. 1 Theoric growth curves: exponential curves varying for *b* parameter (panel A), transformed exponential curves varying for *a* parameter (panel B) and transformed logistic curves, with *c* as a constant and *d* being varied (panel C)

The exponential growth curve can be rewritten as follows:

Considering logarithmic transformation:

$$LnN(t) = lne^{a(t-b)} \qquad (1)$$

then

$$LnN(t) = a(t-b) \qquad (2)$$

and

$$\frac{lnN(t)}{t-b} = a \qquad (3)$$

So, considering Eq. 3, exponential growth, irrespectively to *b* values, can be visualized as straight lines parallel to the x-axis; for example, the exponential growth plotted in Fig. 1a has been reported in Fig. 1b.

In real epidemics, the exponential growth phase is limited to the first period; after this first phase, there is a peak of new infections, only to see a progressive decrease in new infections, down to zero. Therefore, considering the transformation (see above), the graph in Fig. 2 must not have a line parallel to the x-axis, but rather a decreasing trend, where at the two ends there is an almost straight trend. The number of new infected, considering the transformation reported above, can be modelled on a logistic curve (Fig. 1c).

The 3-parameters logistic curve of transformed data has the following equation:

$$\frac{lnN(t)}{t-b} = \frac{a}{1+e^{c(t-d)}} \qquad (4)$$

Solving for *N(t),* we obtained:

$$lnN(t) = \frac{a(t-b)}{1+e^{c(t-d)}} \qquad (5)$$

and finally

$$N(t) = e^{\frac{a(t-b)}{1+e^{c(t-d)}}} \qquad (6)$$

With a factor correction:

$$N(t) = e^{\frac{a(t-b)}{1+e^{c(t-d)}}}\text{-1} \qquad (7)$$

[Insert Figure 2]

Fig. 2 Examples of curve growth proposed; only *a* parameter is varied

The new curve growth proposed (7) has several useful characteristics:

Considering the first growth phases of the infection, the exponent denominator (which is a logistic) approximates to 1, so that the whole exponent approximates at the straight line a(t-b). In other terms, the function approximates at

$N(t) = e^{a(t-b)} - 1$, which is a classic exponential function.

For large values of t, i.e. at the end of the outbreak, the curve tends to zero.

For $t \to +\infty \;\; e^{\frac{a(t-b)}{1+e^{c(t-d)}}} - 1 \; = e^{0} - 1 = 0$

The insertion of the corrective term "-1" in (7) is justified since at the end of the epidemic the number of newly infected people must reach zero value.

The first derivative is:

$$N'(t) = e^{\frac{a(t-b)}{1+e^{c(t-d)}}} \left(\frac{a}{1+e^{c(t-d)}} - \frac{ac(t-b)e^{c(t-d)}}{\left(1+e^{c(t-d)}\right)^2} \right)$$

Unfortunately is not possible to calculate the exact value of t in order to $N'(t)$=0 (time of both peak value), but it is possible to obtain a numerical approximation, with the desired precision, considering:

$$\left(1 + e^{c(t-d)}\right)^2 - \; c(t-b)e^{c(t-d)} = 0$$

This equation can be solved using, for example, the bisection algorithm.

Using the same equation is also possible to model the number of new dead of COVID-19 as a function of time (days). The estimated growth curve can be directly used to calculate both the expected time of both the beginning and the end of epidemic curves.

Data sources

The temporal resolution of the datasets was day-to-day, starting from February 24th, 2020. New COVID-19 cases and new deaths were extracted from the database of the Italian government, at the following link: https://github.com/pcm-dpc/COVID-19/blob/master/dati-regioni/dpc-covid19-ita-regioni.csv. We characterized the epidemic growth patterns for the entire Italian nation and for each of the 20 Italian regions (the autonomous provinces of Trento and Bolzano were considered instead of the whole Trentino Alto Adige region).

Statistical data analyses

Parameters (*a, b, c, d*) can be jointly estimated through nonlinear least-square curve fitting to the case incidence curve modelled by the equation above reported. Nonlinear regressions were performed determining unweighted least squares estimates of parameters using the Levenberg-Marquardt method. The goodness of fit was quantified using adjusted R square; normality of residuals was tested using the D'Agostino-Pearson test; finally, a Run test was performed to test autocorrelation among residuals. Initial values were chosen using a simulated curve on a Microsoft Excel spreadsheet. The trend of new cases and deaths is reported at the graphic level only in relation to the entire Italian nation; the data for each region are available as Supplementary materials. All statistical analyses were performed using Excel 365 or SPSS 20.0; the significance threshold was fixed at 0.05.

Results

In Figure 3, modelling of a new growth model for the forecasting of COVID-19 outbreak evolution is shown. The data were related to the entire Italian nation. Growth curve proposed showed significant goodness of fitting; r square adjusted ranged from 0.3453 (Basilicata) to 0.8937 (Lazio), with a mean r square = 0.7061. On the abscissa axis are the days starting from the beginning of the epidemiological data records, that is from 24/02/2020. The graph has the secondary axis in y with maximum value = ¼ of the value of the primary axis, in order to be able to compare from a visual point of view the relationship between the number of new deaths and the number of new cases per day. The curve relative to the new cases/day has parameters: a=291.1; b=0.02125; c=-312.4; d=-12.44; the curve relative to the new deaths/day has parameters: a=110634; b=0.02470; c=505.4; d=-2.549. Both the new cases/day and new deaths/day curves showed good R^2 values (0.8806 and 0.9041, respectively). Residuals satisfied normality and absence of autocorrelation assumptions for the new cases (D'Agostino-Pearson test; p=0.2016; Run test; p<0.0001), whilst the new deaths/day showed a deviation from homoscedasticity (D'Agostino-Pearson test; p=0.0042) but met the assumption of autocorrelation absence (Run test; p=0.0009). The smallest regions showed a much lower total number of infected, so the new cases time-series is subjected to significant sampling noise; R^2 value is lower than for bigger regions. Using a regression curve is possible to hypothesize that the numbers of the new infected will be less than 1 per day on day 230 (October 11th, 2020) from February 24th.

[Insert Figure 3]

Fig. 3 Regression model results related to Italian data. White dots represent new cases per day, while grey diamonds represent new deaths per day. The black line is the regression relative to the new cases, grey line the regression relative to new deaths (data updated on May 10th, 2020)

In Figure 4a, the number of days from the (estimated) start to the peak (x-axis), and from the peak to the (estimated) end of the outbreak (y-axis), for each of the 19 Italian regions, and the 2 autonomous provinces have been plotted. This plot shows the heterogeneity in the spreading speed of the virus between the different regions. In particular, it has been possible to calculate the average number of days between the beginning of the epidemic and the peak (31 ± 9 days); the minimum occurred for the Basilicata (15 days), while the maximum for the Lombardia region (45 days). Moreover, considering an estimated time-interval between the day of the maximum number of new cases and the end of the epidemic, an average number of days of 101 ± 46 has been calculated; the shortest time frame (26 days) is has been found for the Molise region, while the longest (197 days) for the Piemonte region. Figure 4b shows the estimated date of the onset of the epidemic (in x-axis) and the total number of infected (data updated on May 10th). Is noteworthy that the regions where the outbreak started first, had the highest number of total infected (Lombardia, Piemonte, Emilia Romagna and Veneto, in

chronological order); in the remaining regions, where contagion began between February 21st and March 13th, the total number of cases in sensibly lower.

[Insert Figure 4]

Fig. 4 Panel A: number of days from the (estimated) start to the peak and from the peak to the (estimated) end, for each of the 19 Italian regions and 2 autonomous provinces (P.A.); dots dimension is proportionate to the maximum of daily cases. Panel B: number of total cases for each of the 19 Italian regions and 2 autonomous provinces (P.A.), in the function of (estimated) starting time (data updated on May 10th, 2020); dots dimension is proportionate to the total cases.

Discussion

To improve the ability of epidemiological models to capture the trajectory and impact of pandemics and epidemics, we have introduced a generalized-growth model to characterize the epidemic ascending phase, maximum and descending phase until zero. To our knowledge, this new growth curve is proposed for the first time; it adapts to modelling new cases (and new deaths) per day. It shows some useful features: it is simple, uses only 4 parameters, so the risk of overfitting is minimized. The Levenberg-Marquardt algorithm reaches convergence with few iterations (5-10), while the initial values to be used, once found for a specific curve, are also suitable for all the others, both for new cases and for new deaths. This curve can be added to other many already present in the literature. The type of curve that is proposed is empirical, as it does not derive from a model of differential equations but from the *ad hoc* construction of an equation that shows the desired shape characteristics. This can be useful because, being freed from the assumption of the models, it can better adapt to the changes recorded in the time series of new cases of COVID-19. It shows similarities with the first derivative of Richard's function (Wu et al. 2020; Guescini et al. 2008). Richard's function, even if it takes into account the asymmetry between the ascending and the subsequent descending phases, cannot exceed a limit value; on the contrary, the function proposed in this paper overcomes this limitation, adapting to any type of asymmetry present.

This type of curve is important since it allows to foresee the amount of time it takes for the epidemic to reach its end. This is especially useful in times of emergency, in order to rationalize resources, to monitor the heterogeneity of the outbreak evolution in the different regions and to plan the necessary public health interventions in the period of maximum crisis. It should be noted that the numbers of new cases used for these epidemiological growth curves are obtained from nasopharyngeal swabs. In Italy, this screening procedure has been carried out mainly on subjects with evident severe symptoms. On the other hand, in the various regions, there has been a heterogeneous approach to screening procedures, for which the results may have been affected by bias and random errors due to the different criteria used for performing nasopharyngeal swabs. Nevertheless, the estimated values of the onset of the epidemic are fairly consistent with each other and quantify a scenario already reported by the official sources of information.

This curve could also be useful to assess shifts in epidemic growth patterns resulting from mitigation during an outbreak, as a consequence of population behaviour changes or public health control interventions that affect transmission.

Conclusions

In this paper, we have calibrated the new growth curve to model the reported number of infected and death cases in the COVID-19 epidemics, from February 24th to May 11th, for the whole of Italy and its 20 regions. Our analysis explores the development of the epidemics in Italy and the impact of the containment measures both at the aggregate level and within each region. We documented the early stages of infection, which follow the exponential pattern, estimated the peak of new infections and made future scenario projections of the possible end of the outbreak. We quantified the initial reactions and ramping up of control measures on the dynamics of the epidemics and unearthed an inverse relationship between the number of days from peak to the quasi- end and the duration from start to the peak of the epidemic among the 20 analysed Italian regions. We identified heterogeneity of the development of the epidemic and responses across the regions. The goodness of forecasting will establish the goodness of this new equation and its possible use in the future.

References

Chowell, G., Viboud, C., Hyman, J. M., & Simonsen, L. (2015). The Western Africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. *PLoS Curr, 7*, doi:10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261.

Gandon, S., Day, T., Metcalf, C. J. E., & Grenfell, B. T. (2016). Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases. *Trends Ecol Evol, 31*(10), 776-788, doi:10.1016/j.tree.2016.07.010.

Guescini, M., Sisti, D., Rocchi, M. B., Stocchi, L., & Stocchi, V. (2008). A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC Bioinformatics, 9*, 326, doi:10.1186/1471-2105-9-326.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet, 395*(10223), 497-506, doi:10.1016/S0140-6736(20)30183-5.

Peeri, N. C., Shrestha, N., Rahman, M. S., Zaki, R., Tan, Z., Bibi, S., et al. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *Int J Epidemiol*, doi:10.1093/ije/dyaa033.

Viboud, C., Simonsen, L., & Chowell, G. (2016). A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics, 15*, 27-37, doi:10.1016/j.epidem.2016.01.002.

Wu, K., Darcet, D., Wang, Q., & Sornette, D. (2020). Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. *medRxiv*, 2020.2003.2011.20034363, doi:10.1101/2020.03.11.20034363.

Declarations:

Competing interests: The authors declare no competing interests.

Figures

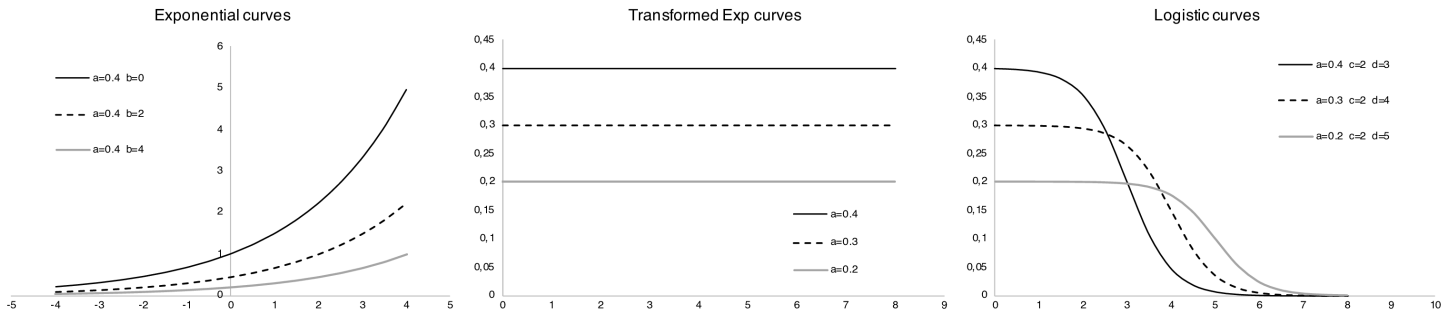

Figure 1

Theoric growth curves: exponential curves varying for b parameter (panel A), transformed exponential curves varying for a parameter (panel B) and transformed logistic curves, with c as a constant and d being varied (panel C)

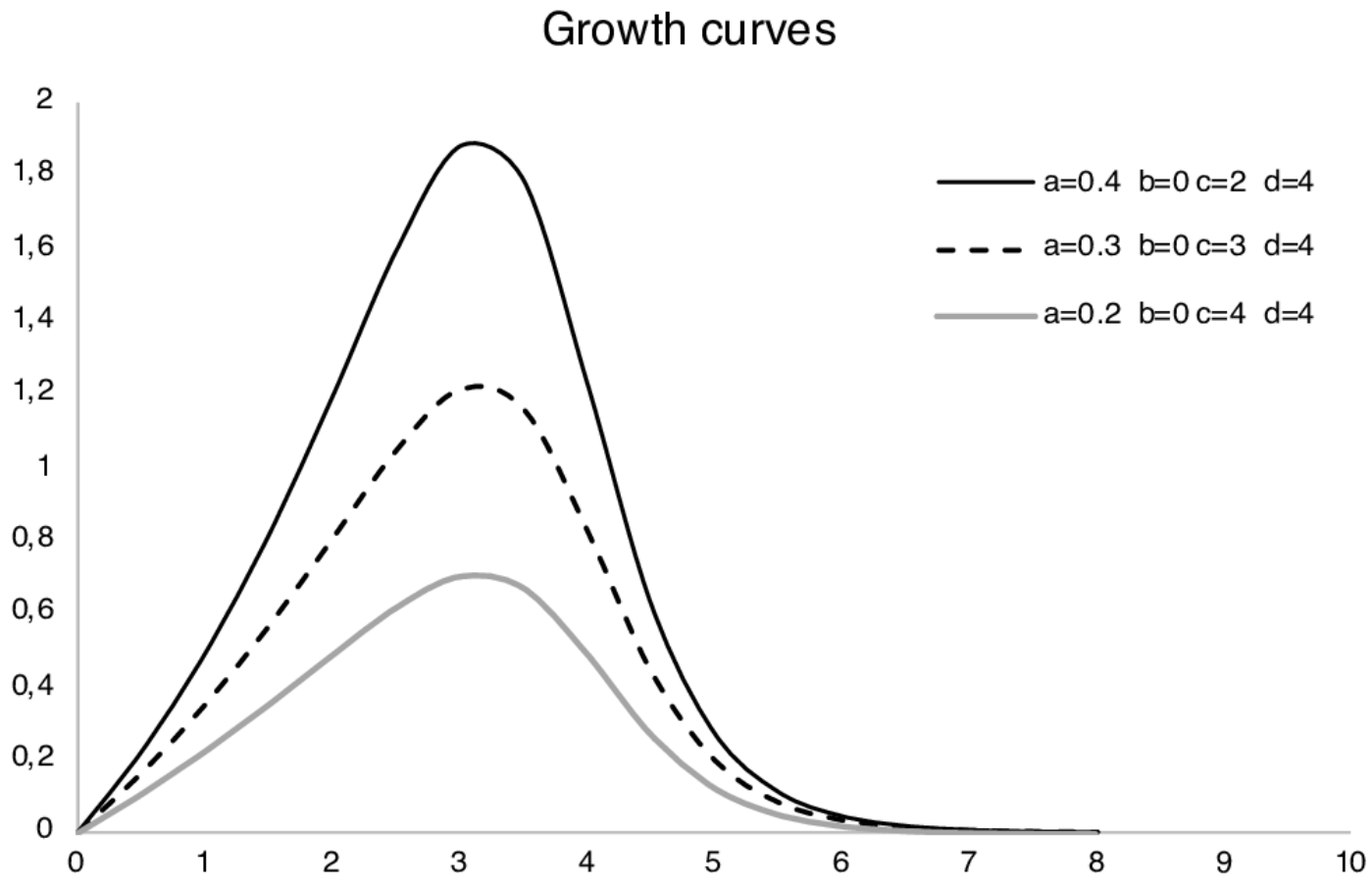

Figure 2

Examples of curve growth proposed; only a parameter is varied

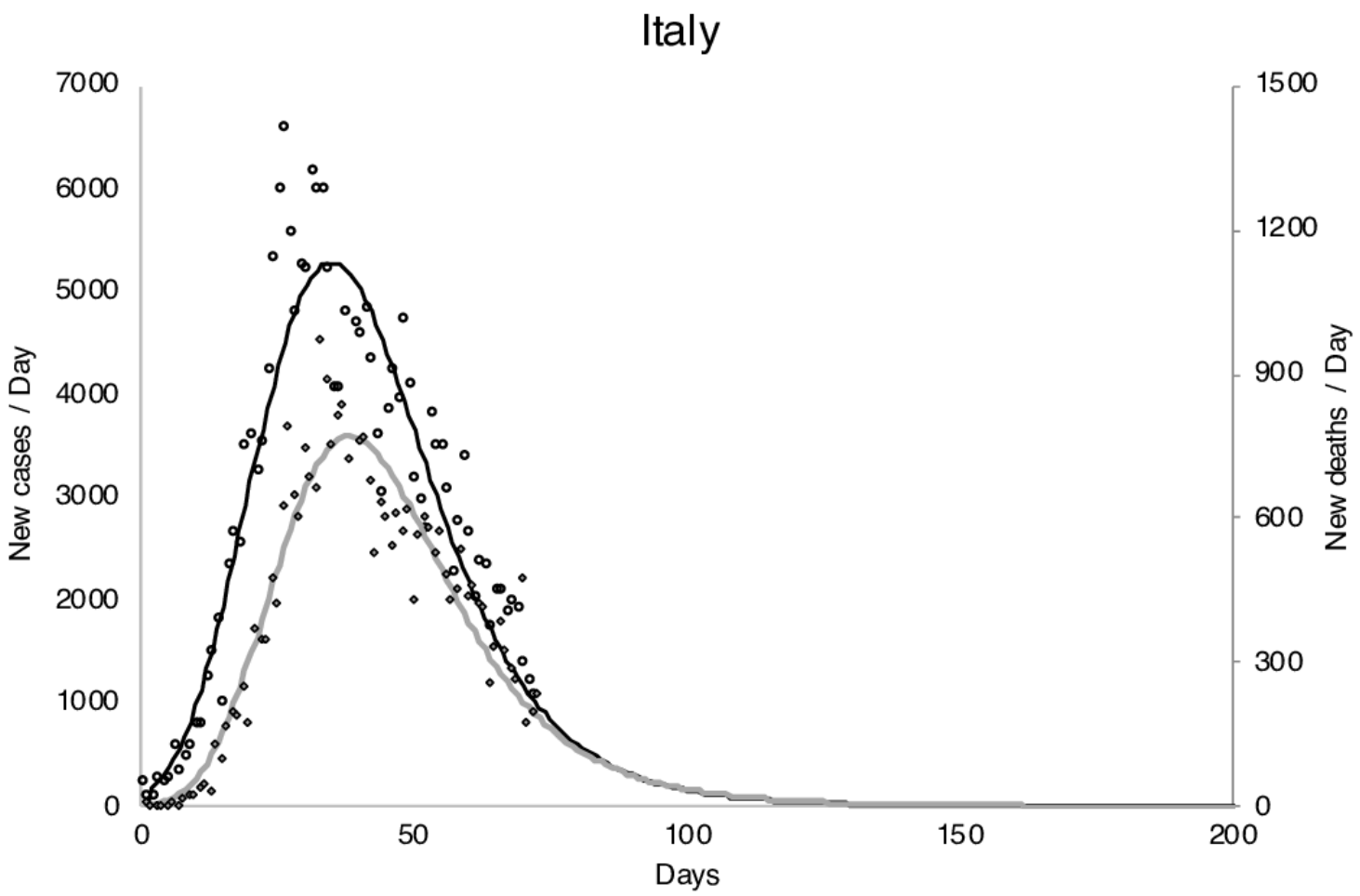

Figure 3

Regression model results related to Italian data. White dots represent new cases per day, while grey diamonds represent new deaths per day. The black line is the regression relative to the new cases, grey line the regression relative to new deaths (data updated on May 10th, 2020)

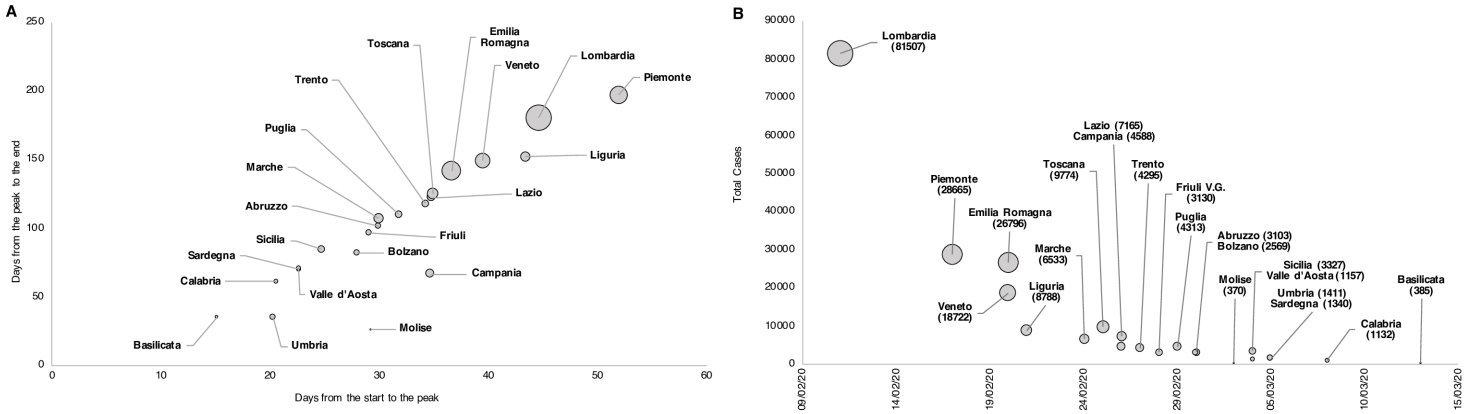

Figure 4

Panel A: number of days from the (estimated) start to the peak and from the peak to the (estimated) end, for each of the 19 Italian regions and 2 autonomous provinces (P.A.); dots dimension is proportionate to

the maximum of daily cases. Panel B: number of total cases for each of the 19 Italian regions and 2 autonomous provinces (P.A.), in the function of (estimated) starting time (data updated on May 10th, 2020); dots dimension is proportionate to the total cases.