mathematics

MDPI

Article

Expert System to Model and Forecast Time Series of Epidemiological Counts with Applications to COVID-19

Beatriz González-Pérez [1,2,*], Concepción Núñez [3], José L. Sánchez [1], Gabriel Valverde [1] and José Manuel Velasco [4]

[1] Department of Statistics and Operations Research, Complutense University of Madrid (UCM), 28040 Madrid, Spain; joseluis.sanchezmaronas@gmail.com (J.L.S.); gvalverd@ucm.es (G.V.)
[2] Interdisciplinary Mathematics Institute (IMI), Complutense University of Madrid (UCM), 28040 Madrid, Spain
[3] Laboratory of Research in Genetics of Complex Diseases, Hospital Clínico San Carlos, IdISSC, 28040 Madrid, Spain; conchita.npardo@gmail.com
[4] Computer Architecture and Automation Department, Complutense University of Madrid (UCM), 28040 Madrid, Spain; mvelascc@ucm.es
* Correspondence: beatrizg@ucm.es

Abstract: We developed two models for real-time monitoring and forecasting of the evolution of the COVID-19 pandemic: a non-linear regression model and an error correction model. Our strategy allows us to detect pandemic peaks and make short- and long-term forecasts of the number of infected, deaths and people requiring hospitalization and intensive care. The non-linear regression model is implemented in an expert system that automatically allows the user to fit and forecast through a graphical interface. This system is equipped with a control procedure to detect trend changes and define the end of one wave and the beginning of another. Moreover, it depends on only four parameters per series that are easy to interpret and monitor along time for each variable. This feature enables us to study the effect of interventions over time in order to advise how to proceed in future outbreaks. The error correction model developed works with cointegration between series and has a great forecast capacity. Our system is prepared to work in parallel in all the Autonomous Communities of Spain. Moreover, our models are compared with a SIR model extension (SCIR) and several models of artificial intelligence.

Keywords: artificial intelligence; machine learning; non-linear regression; error correction model; SIR

Citation: González-Pérez, B.; Núñez, C.; Sánchez, J.L.; Valverde, G.; Velasco, J.M. Expert System to Model and Forecast Time Series of Epidemiological Counts with Applications to COVID-19. *Mathematics* **2021**, *9*, 1485. https://doi.org/10.3390/math9131485

Academic Editors: Victoria López and Laureano González Vega

Received: 28 May 2021
Accepted: 21 June 2021
Published: 24 June 2021

1. Introduction

The coronavirus disease 2019 (COVID-19), caused by the so-called SARS-CoV-2 virus, has spread throughout the world leading to a terrible pandemic. Starting in China in December 2019, the following two countries were Italy and Spain. The infection's high transmissibility led some regions to suffer a special impact. Such is the case of the Autonomous Region of Madrid in Spain. On March 11, the World Health Organization (WHO) declared COVID-19 a pandemic. On the same date, Madrid was already in an extremely serious situation and all educational centers were closed. Three days later, on March 14, the state of alarm was declared throughout the country. On March 30, freedom of activity outside the home was reduced to essential services. These conditions were relaxed on April 13, with the permission of some non-essential activities. From April 26, children could go outside for 1 h every day, and one week later this measure was extended to the general population. Measures of varying magnitude have been taken in the following three waves that have followed one after the other so far.

Mathematical models to track changes in the behavior and patterns of infection appear to be essential tools for making future decisions.

Some authors (see, e.g., [1,2]) were pioneers in collecting solutions based on artificial intelligence and expert knowledge, among the thousands of articles published on the subject.

Artificial intelligence is useful to monitor the evolution of the pandemic even in real time, either through expert systems or with a predictive approach based on machine learning. Researchers have been able to validate the effectiveness of these models with different illnesses. For example, through a dynamic neural network, it was possible to understand the evolution of Zika (see [3]). The same has happened with Ebola or the common flu. Currently, models are being retrained with new data related to the COVID-19 (see [4]). Abhari et al. [5] used a previously developed agent-based artificial intelligence simulation platform (EnerPol) coupled with big data.

It is worth highlighting the need for artificial intelligence tools to be easy to use by those who want to operate with them. This is why, around the idea of monitoring and forecasting, projects have been generated to visualize the information collected. With this approach, in [6], we can find an ordered list of the most interesting sites with dashboards: UpCode, NextStrain, CSSE (Johns Hopkins), Thebaselab, the BBC, the New York Times, HealthMap and COVID-19 Tracker (Microsoft).

The SIR model ("Susceptible", "Infected", "Recovered") and its extensions are traditional epidemiological models. They are compartmental models of differential equations that relate the variations of different population groups (compartments) through the infection rate and the average infectious period. Most recent studies are based on modifications of the SIR model (see, e.g., [7–9]). The underlying idea is to model the waves of a pandemic as exponential increases and decreases to the left and right of a peak of maximum incidence. In Spain, it is worth highlighting the work carried out by the Interdisciplinary Group of Complex Systems at UC3M [10] and the work carried out by the MOMAT Group at UCM [11]. For comparisons purposes, we implement the SIR model extension developed by Castro et al. [10], SCIR: a SIR model with "confinement".

In this paper, we approach the problem from another point of view: non-linear regression predictive models. Researchers from the Andalusian School of Public Health of Granada have developed a predictive model of the COVID-19 epidemic in Spain with an adjustment to a Gompertz curve [12]. For the adjustment of the Gompertz curve to the observed accumulated data of cases and deaths, they used the Nelter–Mead algorithms [13] implemented by Nash [14]. The software used for the calculations was R via drc package. Our strategy extends this approach by allowing greater flexibility in fitting to Gompertz curves, especially in the distribution tails. Another Gompertz approximation was proposed by Català et al. [15]. Our expert system automatically chooses the best fit from a variety of models, including the Gaussian, double exponential and double Pareto curves. In addition, all programming, the optimization algorithm and the heuristic are original.

Moreover, we develop an Error Correction Model (ECM). This approximation belongs to a category of multiple time series models for data where the underlying variables have a long-run common stochastic trend.

Our research group registered in the "Mathematical action against coronavirus", a cooperative prediction initiative of the Spanish Mathematics Committee (CEMat). (A meta-predictor has been built to provide authorities with information on the short-term behavior of variables of great interest in the spread of the COVID-19 virus. The method uses the predictions from different models/algorithms, provided by the participating researchers, and constructs optimized combinations of them, disaggregated by Autonomous Communities.) Within this initiative, we have participated together with other research groups in the "Cooperative Prediction" action [16], providing daily predictions with our preliminary model since March 2020, during the entire first wave of the pandemic. All models that participate in the construction of the metapredictor developed by the cooperative prediction action promoted by the CEMat have been validated continuously since the beginning of the pandemic.

The results obtained in this paper are reproducible using the code from our public repository. The code for the developed graphical interface that allows the user to interact with our system is also included in: https://github.com/mikiNadal/covid19_article_reproducible (accessed on 22 June 2021).

Section 2 introduces the non-linear regression model. Section 3 introduces the error correction model. Section 4 introduces the SCIR model. Section 5 compares the three models with different metrics. Finally, the conclusions are presented in Section 6.

2. Non-Linear Regression Model

We aim to develop a theoretical framework that allows us to detect peaks and make short- and long-term monitoring and forecasting of the number of people infected, people requiring hospitalization and deaths during an infectious disease. With short-term prediction, we refer to the task that we performed for the CEMat during the first wave, consisting of giving predictions every day with a horizon of 8 days. From the second wave, we were asked for predictions every week with a horizon of 14 days. With long-term forecast, we refer to the prediction of the peak, the total number of infected at the end of a wave under study and giving commitment dates for which only a small percentage of the area under the model remains. These values are monitored day by day and are an indication, for example, of when a wave is exhausted.

This model is implemented with an expert system of artificial intelligence based on non-linear regression and is extremely useful to estimate the effectiveness of the interventions prompted by the governments and to advise on how to proceed in future outbreaks. Furthermore, the machine learning algorithm developed allows parallel running and introduction of new data in real time, and it is scalable.

Our model is based on directly estimating the distribution function of each of the series under study and on the duality between the distribution function and the density function. Since those two functions fully characterize the probability distribution of a continuous variable, our model is able to capture the main characteristics of epidemic outbreaks. To this, we can add its simplicity, since it is formulated only through three parameters. Hereinafter, we refer to our first proposed epidemiological model as the MATGEN model in honor of our group enrolled in the *Mathematical action against coronavirus* [16], an initiative of CEMat (Spanish Mathematics Committee).

2.1. The Model

The notation employed in this work is as follows:

Let the well-known density function of a normal variable of mean be μ and variance σ^2, $N(\mu, \sigma)$ (see Figure 1a), given by

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$$

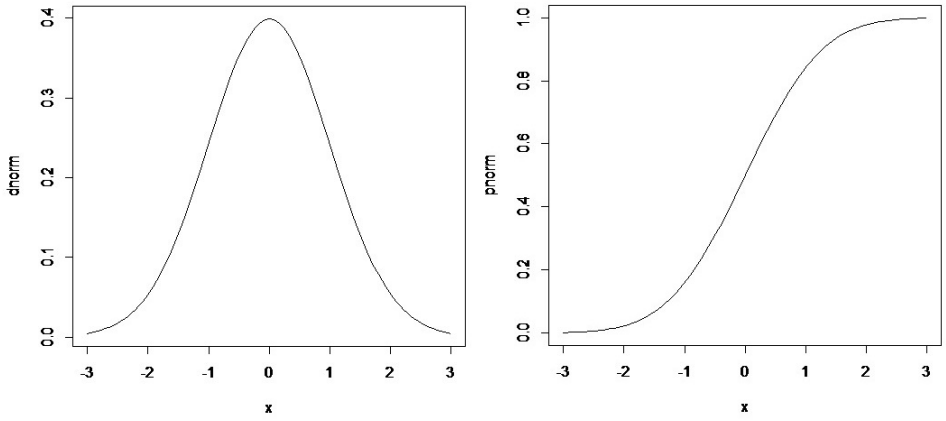
and both its distribution function and the right tail as follows:

$$F(t) = \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

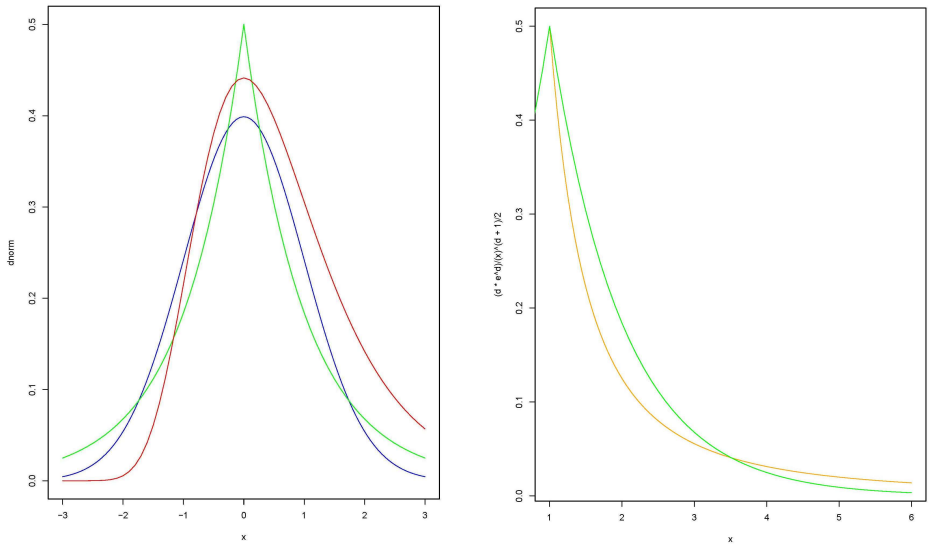
$$1 - F(t) = \int_{t}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Note that the density peak of $N(\mu, \sigma)$ is reached in μ, and it is a point of inflection of its distribution function. Furthermore, it verifies that $F(t) = 1 - F(t) = 0.5$.

For a review of the properties of the distribution function, the density function, the characteristic function of a random variable and the relationships between them, see the work of Quesada and Pardo [17].



(**a**) Normal probability distribution: N(0,1). (Left) density function. (Right) distribution function



(**b**) Blue, normal; green, double exponential; red, Gompertz; orange, double Pareto.

Figure 1. Probability distributions for pandemic wave modeling

Data series of COVID-19 in Spain include day by day the cumulative number of people infected, people requiring hospitalization and deaths. These data can be downloaded from ISCIII [18].

We denote both the relative and cumulative frequencies at time t as follows:

$$N_t \text{ cumulative per day,}$$

$$n_t = N_t - N_{t-1} \text{ new cases per day,}$$

$$f_t = \frac{n_t}{n},$$

$$F_t = \sum_{i=1}^{t} f_i,$$

where n is the total number of cases at the end of the pandemic.

Furthermore, we introduce the average of the cumulative frequencies at time t given by

$$Av_t = \frac{1}{t} \sum_{i=1}^{t} F_i.$$

In this context, we work with the following non-linear regression model:

$$F_i = F(i) + \varepsilon_i, \varepsilon_i \sim N(0,\ \tau),\ i = 1, \cdots, t,$$

where the parameters n, μ and σ are estimated by the least squares method.

For an introduction to frequentist and bayesian regression, see the work of Gómez Villegas [19].

2.2. Other Wave Models

The next subsections detail a basic guide for the correct implementation of the least squares method and the algorithm designed for the detection of pandemic peaks.

2.3. The Algorithm: Peak Detection

Initialize t in t_0, the current moment.

Compute the mean squared error as follows:

$$ECM(t, n, \mu, \sigma | n_1, \ldots, n_t) = \frac{1}{t} \sum_{i=1}^{t} (F(i) - F_i)^2,$$

the total variance

$$SCT(t, n, \mu, \sigma | n_1, \ldots, n_t) = \frac{1}{t} \sum_{i=1}^{t} (F_i - Av_t)^2$$

and the coefficient of determination

$$R^2(t, n, \mu, \sigma | n_1, \ldots, n_t) = 1 - \frac{ECM(t, n, \mu, \sigma | n_1, \ldots, n_t)}{SCT(t, n, \mu, \sigma | n_1, \ldots, n_t)}.$$

In statistics, the coefficient of determination is the proportion of the variance in the dependent variable F_i that is predictable from the independent variable $F(i)$. It is a statistic used in the context of goodness of fit and provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. This coefficient takes values between 0 and 1, and, between two models, the one with the highest determination coefficient is preferred. Furthermore, with this criterion, the best model is the one that maximizes the coefficient of determination within a plausible family of models:

$$\max_{n,\mu,\sigma} R^2(t, n, \mu, \sigma | n_1, \ldots, n_t)\ .$$

On the other hand, we control at the same time the adjustment of the observed frequencies by means of the theoretical density function. To this end, the criterion that we follow is to perform a linear regression

$$\frac{f_i}{f(i)} = a + bi + e_i, e_i \sim N(0, \nu), i = 1, \ldots, t,$$

and to introduce the constraint

$$p - value(F_{obs}) = P(F_{1,t-2} > F_{obs}) > 0.1,$$

where F_{obs} is the observed value of the test statistic for testing $H_0 : b = 0$ vs. $H_1 : b \neq 0$ and $F_{1,t-2}$ is its theoretical distribution under H_0, that is a Snedecor's F distribution with 1 and $t - 2$ degrees of freedom.

We propose to solve the multicriteria optimization problem by obtaining the values of $n(t)$, $\mu(t)$ and $\sigma(t)$, so that

$$\max_{n,\mu,\sigma} R^2(t, n, \mu, \sigma | n_1, \ldots, n_t)\ ,$$

under the constraint

$$p - value(F_{obs}) = P(F_{1,t-2} > F_{obs}) > 0.1.$$

Now, stop if $F(t) = F_t = 0.5$, otherwise incorporate the data t_1, do $t = t_1$ and repeat.

In parallel, fit a model for each of the series, namely the number of new positive cases per day, the number of new deaths per day and the number of new ICU admissions per day, and choose the model that simultaneously maximizes the three values of R^2.

Stop when $F(t) = F_t = 0.5$ is accomplished in all three series.

It is important to note that the algorithm allows the introduction of new data in real time, and it is scalable.

2.4. The Algorithm: Commitment Dates

Let t_p be the first day that $F(t_p) = F_{t_p} = 0.5$, and $n(t_p)$, $\mu(t_p)$ and $\sigma(t_p)$ are the optimal values of the parameters at that time point.

If $f_{t_p+1} \geq f_{t_p}$, do $\mu = t_p + 1$ and compute n and σ so that $F(t_p + 1) = F_{t_p+1} = 0.5$.

Otherwise, determine the value of t_{max} so that $f_{t_{max}} = \max_{t \leq t_p} f_t$, do $\mu = t_{max}$ and compute n y σ_{left} so that $F(t_{max}) = F_{t_{max}} = 0.5$ and σ_{right} fit the series for $t \geq t_{max}$.

Finally, the percentiles $qnorm_{0.99}$ and $qnorm_{0.999}$ of the normal probability distribution $N(t_{max}, \sigma_{right})$ are the commitment dates to lift the restrictions from least to most conservative.

2.5. The Heuristic

To perform an effective optimization, we opt for an ambitious heuristic that we detail below.

Let $\sigma = \sigma_0$, starting at $\sigma_0 = 15$.

Move σ between $\sigma_0 - 14$ and $\sigma_0 + 14$.

At this point, it is important to note that the incubation period of the disease is between 2 and 14 days (see [20]). In addition, the delay between the time of infection and the report as a positive case is considered.

Let $\mu = \mu_0$, starting at $\mu_0 = t_0$, t_0 being the current moment.

Move μ between the first day of each of the series and $t_0 + 2\sigma_0$. For example, the first day of the series of the number of people infected in the Region of Madrid is Day 26, which corresponds to February 25.

Generate k = 10,000 values of a uniform random variable between 0 and 1.

Compute $n = Np$ for each value p generated in the last step; N = approximately 6,550,000 in the Region of Madrid.

Discard the values of $n < N_{t_0}$.

Find the feasible models with p-value > 0.1 for the noise and select the one with the largest R^2 of the fit in the cumulative frequencies.

In practice, the running of the heuristic generates a .csv file that contains several columns. The columns corresponding to the fitted parameters, μ, σ and n, the coefficient of determination and the p-value are included. Moreover, two columns are added to register every day the moment of the real peak, which corresponds with the day with the highest frequency observed to date, and the day when the cutoff between the models fitted to both sides is observed, i.e., when the distribution becomes positively skewed. The algorithm tries to match the value of the real peak, the cutoff and the parameter μ. It also allows fitting a different σ to the left and right of the cutoff. The last two columns include the commitment dates corresponding to percentiles $qnorm_{0.99}$ and $qnorm_{0.999}$ of the fitted model to the right.

In the next subsections, we present the results that are obtained from the run of the previous algorithm programmed through our expert system. To do this, we consider the data series of COVID-19 in Spain, which are published in [18]. Specifically, we study the case of the Region of Madrid.

2.6. COVID-19 Data Sets

It is necessary to consider the time required to test the presence of the infection and obtain a report to test positive for the virus. This is especially relevant when there are problems with access to care and with bottlenecks in laboratory testing. At some moments, this led the health system in Madrid to test only people with severe symptoms. In addition, the delay of up to several weeks in the notification of positive cases by the laboratories led to changes in the data history depending on the day the data were downloaded (see Tables 1 and 2).

Table 1. The .csv file with the data (continues in Table 2).

		Cases 11/05/2020	**Cases 18/05/2020**	**Cases 21/05/2020**	**Deaths 11/05/2020**	**ICUs 21/05/2020**
25/02/2020	26	1	2	2		
26/02/2020	27	5	6	6		
27/02/2020	28	9	10	10		
28/02/2020	29	19	20	20		
29/02/2020	30	26	27	27		
01/03/2020	31	51	53	53		
02/03/2020	32	93	96	96		
03/03/2020	33	139	142	142		
04/03/2020	34	193	199	199		
05/03/2020	35	305	311	311		
06/03/2020	36	508	515	515		
07/03/2020	37	729	738	738		
08/03/2020	38	992	1003	1003	16	61
09/03/2020	39	1495	1508	1508	21	120
10/03/2020	40	2198	2213	2213	31	184
11/03/2020	41	2922	2943	2943	56	238
12/03/2020	42	3705	3732	3732	81	307
13/03/2020	43	4645	4672	4672	86	370
14/03/2020	44	5544	5576	5576	213	469
15/03/2020	45	6356	6392	6392	213	566
16/03/2020	46	7615	7653	7653	355	702
17/03/2020	47	9561	9601	9601	390	850
18/03/2020	48	11,309	11,356	11,356	498	1011
19/03/2020	49	13,353	13,399	13,399	628	1196
20/03/2020	50	15,676	15,722	15,722	804	1401
21/03/2020	51	17,346	17,397	17,397	1021	1532
22/03/2020	52	18,848	18,900	18,900	1263	1664
23/03/2020	53	21,516	21,569	21,569	1535	1813
24/03/2020	54	24,404	24,473	24,475	1825	1962
25/03/2020	55	27,344	27,420	27,422	2090	2117
26/03/2020	56	30,711	30,794	30,796	2412	2272
27/03/2020	57	33,068	33,160	33,162	2757	2369
28/03/2020	58	34,087	34,186	34,189	3082	2423
29/03/2020	59	34,959	35,058	35,061	3392	2464
30/03/2020	60	37,504	37,604	37,607	3603	2554
31/03/2020	61	39,303	39,404	39,409	3865	2627
01/04/2020	62	41,075	41,191	41,199	4175	2694
02/04/2020	63	42,896	43,027	43,038	4483	2764

Even when the data come from official sources, they may present inconsistencies that must be taken into account. The portal to access the European Union open data [21] publishes data on the evolution of COVID-19 by continent and broken down by country. It can be verified that only positive PCRs are counted in the series of cases on this portal.

On the other hand, Spain (see [18]) and Italy (see [22–24]) offer more detailed information through their national institutional portals. For example, on April 17, Spain introduces two columns, PCR and TestAc, and TestAc is empty until April 18, when the government introduces this type of test into the count. On the recommendation of the Spanish Mathematics Committee, we chose to consider confirmed cases as PCR+TestAc. This situation has been remedied since the second wave of the pandemic.

Table 2. The .csv with the data.

		Cases 11/05/2020	**Cases 18/05/2020**	**Cases 21/05/2020**	**Deaths 11/05/2020**	**ICUs 21/05/2020**
03/04/2020	64	44,613	44,768	44,779	4723	2821
04/04/2020	65	45,496	45,660	45,671	4941	2854
05/04/2020	66	46,016	46,186	46,197	5136	2879
06/04/2020	67	47,568	47,749	47,763	5371	2958
07/04/2020	68	48,945	49,139	49,160	5586	3002
08/04/2020	69	50,357	50,556	50,580	5800	3038
09/04/2020	70	51,296	51,505	51,535	5972	3061
10/04/2020	71	52,146	52,367	52,400	6084	3091
11/04/2020	72	52,680	52,909	52,944	6278	3105
12/04/2020	73	53,017	53,250	53,285	6423	3122
13/04/2020	74	53,988	54,241	54,287	6568	3153
14/04/2020	75	55,062	55,343	55,398	6724	3180
15/04/2020	76	55,959	56,245	56,304	6877	3203
16/04/2020	77	56,792	57,084	57,157	7007	3214
17/04/2020	78	57,598	57,899	57,978	7132	3228
18/04/2020	79	60,558	60,859	60,941	7239	3238
19/04/2020	80	60,952	61,254	61,336	7351	3248
20/04/2020	81	61,568	61,882	61,972	7460	3278
21/04/2020	82	62,440	62,801	62,904	7577	3283
22/04/2020	83	63,558	63,921	64,050	7684	3288
23/04/2020	84	64,120	64,496	64,634	7765	3305
24/04/2020	85	64,785	65,163	65,310	7848	3307
25/04/2020	86	65,015	65,396	65,546	7922	3308
26/04/2020	87	65,477	65,857	66,007	7986	3309
27/04/2020	88	66,241	66,631	66,784	8048	3338
28/04/2020	89	66,884	67,293	67,460	8105	3355
29/04/2020	90	67,332	67,747	67,942	8176	3377
30/04/2020	91	67,714	68,154	68,356	8222	3392
01/05/2020	92	67,830	68,291	68,537	8292	3404
02/05/2020	93	67,947	68,408	68,654	8332	3421
03/05/2020	94	68,056	68,520	68,766	8376	3431
04/05/2020	95	68,447	68,924	69,178	8420	3442
05/05/2020	96	68,745	69,249	69,509	8466	3465
06/05/2020	97	69,110	69,627	69,885	8504	3485
07/05/2020	98	69,323	69,856	70,125	8552	3493
08/05/2020	99	69,566	70,132	70,407	8598	3508
09/05/2020	100	69,697	70,238	70,516	8644	3520
10/05/2020	101	69,730	70,292	70,570	8683	3529

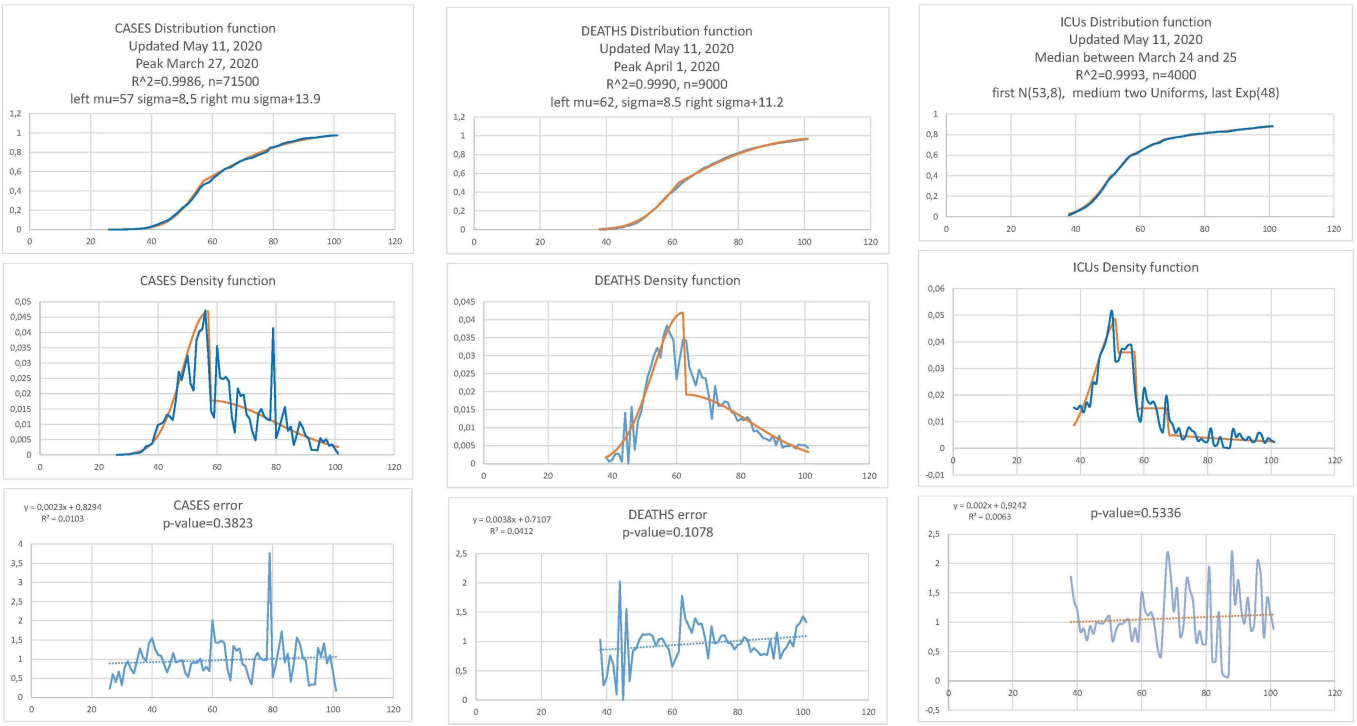
Another controversial point is the notification of deaths due to COVID-19 and the real deaths registered by the undertaking services of the Autonomous Communities of Spain. This suggests that only a percentage of the real deaths due to COVID-19 were recorded. As a matter of fact, many elderly people died in nursing homes before being tested for COVID-19 and in most cases they were not included in the number of deaths.

One more problem that we have had to face is related to the series of ICUs in the Region of Madrid, in which we found an anomaly. Since April 28, the Region of Madrid offers cumulative data on the number of people with coronavirus who have gone through the ICU. Before this date, the data were those of daily occupation. Furthermore, the number of ICU beds varied throughout the course of the epidemic. The maximum number was changing due to the increase in ICU beds in the large hospitals in Madrid and especially the provision of new hospital beds in the IFEMA hospital.

2.7. Expert System

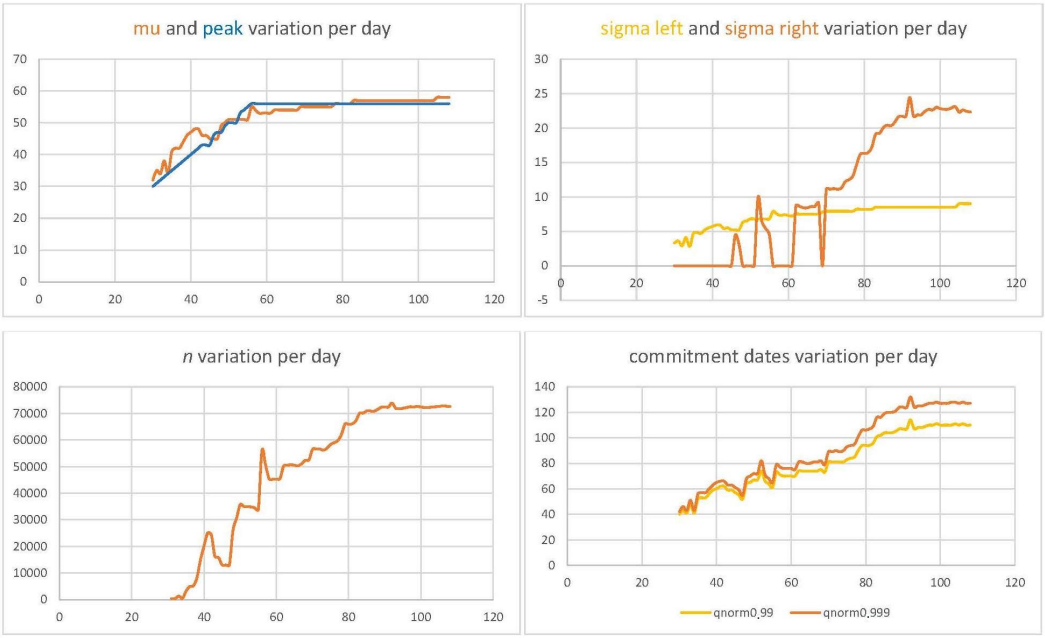
Our expert system was designed to facilitate obtaining the results and it consists of several parts. First, it allows downloading and updating the data file in real time every day. Once the data file has been updated, the expert system allows us to run the algorithm for all the series in parallel or one in particular, as well as for all the provinces of Spain or a specific one. Once the algorithm has been run, our expert system returns three reports:

1. A .pdf file with three graphs for the current time: one of the fitted distribution function, one of the fitted density function and one of the adjustment to white noise. Figure 2a,b shows these reports for cases and deaths in the Region of Madrid.
2. A .csv file (see Tables 3 and 5) with the results of the entire day-to-day history of the process, from which the following can be extracted: (i) the optimal parameters; (ii) the coefficient of determination of the fit to the distribution function; and (iii) the *p*-value of the fit to white noise of the relative errors of the fit to the density function. In addition, this .csv file also contains the day-to-day commitment dates.
3. A .csv file (see Table 6) with an 8-day horizon of the forecast made with the fitted model and a graph with the future model.



(**a**) Report of cases (**b**) Report of deaths (**c**) Report of ICUs

Figure 2. *Cont.*



(**d**) Parameter variation per day until May 11. This report has been automatically generated by our expert system. From top to bottom: peak, sigma left, sigma right and n.

Figure 2. Cases (**a**), Deaths (**b**), ICUs (**c**) and parameter variation per day (**d**). (**a**–**c**) From up to bottom: distribution function adjustment, density function adjustment and noise on May 11.

2.8. New and Cumulative Confirmed Cases per Day Series

Figure 2 and Tables 3–6 summarize the results for the new and cumulative confirmed cases per day in the Region of Madrid.

The real peak of the series is one of the most difficult values to predict. Some days after it has taken place, it is easy to know that the peak was already reached on March 26. Table 3 and Figure 2 indicate that μ matches the value of the real peak between April 17 and 21. However, as the curve of μ variation per day (see Figure 2) begins to flatten from March 20, the model alerts us in advance of the possibility that the real peak appears at any time after that date.

In a virus-free transmission situation, the model would fit to a perfect Gaussian distribution and μ would be equal to the real peak. Therefore, small deviations from the model to the left of the real peak may indicate a change in the evolution of the virus. For example, between March 8 and 14, the curve of the relative frequencies (blue) is above the model fitted for the density function (orange) (see Figure 2a), which may indicate the dangerous situation present in Madrid before March 8. This dangerous situation could be a consequence of individual events increasing close contact between people. This was already evident in the fitted models until this date. The period prior to March 11 can be considered free of disease transmission because interventions have not been applied. Nevertheless, that virus-free model is observed several days after that date. In fact, on March 17, there is a small peak of cases, which could be due to a high number of contagions during different massive events in Madrid on March 8. On March 22, the measures imposed by the government started to be noticed, since the observed data lie below the fitted model, and that situation continues until the global peak of cases is reached around March 26. On March 30, after the peak of cases, freedom of activity outside the home was reduced to essential services. The effect of this intervention is noted on April 9, when the real data lie below the fitted model, and that trend remains until April 12. It seems that interventions take around 11 days to be noted.

Table 3. Cases updated on May 18: the .csv file with the results of the entire day-to-day history of the process (continues in Table 4).

Date	Day	F.obs	f.obs	Peak	μ	σ_{left}	σ_{right}	n	R^2	p-Value	$qnorm_{0.99}$	$qnorm_{0.999}$
29/02/2020	30	27	7	30	32	3.3	0	100	0.982587668	0.892176923	40	42
01/03/2020	31	53	26	31	35	3.6	0	400	0.978671691	0.622704742	43	46
02/03/2020	32	96	43	32	34	2.9	0	400	0.981876806	0.175458205	41	43
03/03/2020	33	142	46	33	38	4.1	0	1300	0.995770675	0.392234347	48	51
04/03/2020	34	199	57	34	34	2.8	0	400	0.997745553	0.137856774	41	43
05/03/2020	35	311	112	35	41	4.7	0	3100	0.996817645	0.240883244	52	56
06/03/2020	36	515	204	36	42	4.8	0	4900	0.987584896	0.284897495	53	57
07/03/2020	37	738	223	37	42	4.7	0	5200	0.994676239	0.124339758	53	57
08/03/2020	38	1003	265	38	44	5.2	0	8100	0.998540043	0.239198136	56	60
09/03/2020	39	1508	505	39	46	5.5	0	14,900	0.996779356	0.131424278	59	63
10/03/2020	40	2213	705	40	47	5.7	0	20,200	0.992783337	0.300331128	60	65
11/03/2020	41	2943	730	41	48	5.9	0	25,000	0.997812351	0.315691322	62	66
12/03/2020	42	3732	789	42	48	5.9	0	24,400	0.998798116	0.270683941	62	66
13/03/2020	43	4672	940	43	46	5.4	0	16,200	0.999136218	0.111431486	59	63
14/03/2020	44	5576	904	43	46	5.5	0	15,800	0.998952129	0.571789971	59	63
15/03/2020	45	6392	816	43	45	5.2	0	13,000	0.998944934	0.235342367	57	61
16/03/2020	46	7653	1261	46	45	5.2	4.4	13,000	0.999269903	0.363550861	55	59
17/03/2020	47	9601	1948	47	45	5.2	3.1	13,000	0.998123674	0.579021251	52	55
18/03/2020	48	11,356	1755	47	49	6.3	0	26,000	0.996732574	0.272478617	64	68
19/03/2020	49	13,399	2043	49	50	6.5	0	30,600	0.996779095	0.490069959	65	70
20/03/2020	50	15,722	2323	50	51	6.8	0	35,700	0.997098073	0.121977009	67	72
21/03/2020	51	17,397	1675	50	51	6.8	0	35,000	0.998182881	0.180665458	67	72
22/03/2020	52	18,900	1503	50	51	6.8	9.9	35,000	0.998601989	0.120079419	74	82
23/03/2020	53	21,569	2669	53	51	6.8	6.5	34,800	0.998561382	0.138391236	66	71
24/03/2020	54	24,473	2904	54	51	6.8	5.4	34,500	0.998229605	0.121359895	64	68
25/03/2020	55	27,420	2947	55	51	6.8	4.5	33,800	0.997254468	0.105764493	61	65
26/03/2020	56	30,794	3374	56	55	7.9	0	56,000	0.996014729	0.11940529	73	79
27/03/2020	57	33,160	2366	56	54	7.5	0	50,600	0.9969215	0.529316564	71	77
28/03/2020	58	34,186	1026	56	53	7.3	0	45,400	0.997913273	0.18053051	70	76
29/03/2020	59	35,058	872	56	53	7.4	0	45,200	0.998173357	0.105651154	70	76

Table 4. Cases updated on May 18: the .csv file with the results of the entire day-to-day history of the process (continues in Table 5).

Date	Day	F.obs	f.obs	Peak	μ	σ_{left}	σ_{right}	n	R^2	p-Value	$qnorm_{0.99}$	$qnorm_{0.999}$
30/03/2020	60	37,604	2546	56	53	7.3	0	45,300	0.998303486	0.172127509	70	76
31/03/2020	61	39,404	1800	56	53	7.2	0	45,500	0.998245308	0.470549706	70	75
01/04/2020	62	41,191	1787	56	54	7.5	8.7	50,200	0.998433035	0.754592417	74	81
02/04/2020	63	43,027	1836	56	54	7.5	8.6	50,500	0.998428145	0.50656996	74	81
03/04/2020	64	44,768	1741	56	54	7.5	8.4	50,700	0.998184198	0.284094123	74	80
04/04/2020	65	45,660	892	56	54	7.5	8.4	50,500	0.998541295	0.349922514	74	80
05/04/2020	66	46,186	526	56	54	7.5	8.6	50,300	0.998878631	0.598595831	74	81
06/04/2020	67	47,749	1563	56	54	7.5	8.6	51,100	0.998464545	0.183556346	74	81
07/04/2020	68	49,139	1390	56	54	7.5	9.1	52,400	0.997432118	0.058691187	75	82
08/04/2020	69	50,556	1417	56	55	7.8	0	52,500	0.996017911	0.111161442	73	79
09/04/2020	70	51,505	949	56	55	7.9	11.1	56,500	0.998957163	0.054509831	81	89
10/04/2020	71	52,367	862	56	55	7.9	11.1	56,600	0.998954379	0.037195949	81	89
11/04/2020	72	52,909	542	56	55	7.9	11.2	56,600	0.999093668	0.052789849	81	90
12/04/2020	73	53,250	341	56	55	7.9	11.1	56,200	0.999270096	0.110086342	81	89
13/04/2020	74	54,241	991	56	55	7.9	11.3	56,900	0.999050173	0.030039666	81	90
14/04/2020	75	55,343	1102	56	55	7.9	12.2	58,300	0.998466827	0.009881759	83	93
15/04/2020	76	56,245	902	56	55	7.9	12.5	59,000	0.997872056	0.002736576	84	94
16/04/2020	77	57,084	839	56	55	7.9	13	59,800	0.997263961	0.000860309	85	95
17/04/2020	78	57,899	815	56	56	8.2	14.7	62,100	0.999148019	0.051330147	90	101
18/04/2020	79	60,859	2960	56	56	8.2	16.2	66,000	0.994984648	0.016081334	94	106
19/04/2020	80	61,254	395	56	56	8.2	16.3	65,900	0.995736332	0.024215511	94	106
20/04/2020	81	61,882	628	56	56	8.2	16.4	66,100	0.995644281	0.020793633	94	107

Table 4. *Cont.*

Date	Day	F.obs	f.obs	Peak	μ	σ_{left}	σ_{right}	n	R^2	p-Value	$qnorm_{0.99}$	$qnorm_{0.999}$
21/04/2020	82	62,801	919	56	56	8.2	17.2	67,200	0.994675811	0.009949682	96	109
22/04/2020	83	63,921	1120	56	57	8.5	19.1	70,000	0.996894671	0.031130121	101	116
23/04/2020	84	64,496	575	56	57	8.5	19.2	70,100	0.997000217	0.031958511	102	116
24/04/2020	85	65,163	667	56	57	8.5	20	70,900	0.996864121	0.026171537	104	119
25/04/2020	86	65,396	233	56	57	8.5	20.4	70,900	0.997486764	0.055788057	104	120
26/04/2020	87	65,857	461	56	57	8.5	20.3	70,800	0.997614211	0.060271108	104	120
27/04/2020	88	66,631	774	56	57	8.5	20.8	71,500	0.997206052	0.034021395	105	121
28/04/2020	89	67,293	662	56	57	8.5	21.6	72,300	0.996908785	0.02468125	107	124
29/04/2020	90	67,747	454	56	57	8.5	21.7	72,400	0.996957114	0.025102979	107	124

Table 5. Cases updated on May 18: the .csv file with the results of the entire day-to-day history of the process.

Date	Day	F.obs	f.obs	Peak	μ	σ_{left}	σ_{right}	n	R^2	p-Value	$qnorm_{0.99}$	$qnorm_{0.999}$
30/04/2020	91	68,154	407	56	57	8.5	21.7	72,400	0.997054178	0.026233746	107	124
01/05/2020	92	68,291	137	56	57	8.5	24.4	73,900	0.997145056	0.101063259	114	132
02/05/2020	93	68,408	117	56	57	8.5	21.7	71,900	0.997874745	0.12149971	107	124
03/05/2020	94	68,520	112	56	57	8.5	21.9	71,800	0.99812767	0.225740331	108	125
04/05/2020	95	68,924	404	56	57	8.5	21.9	71,900	0.99809795	0.174545762	108	125
05/05/2020	96	69,249	325	56	57	8.5	22.4	72,200	0.998125597	0.188507265	109	126
06/05/2020	97	69,627	378	56	57	8.5	22.7	72,500	0.998056675	0.160258625	110	127
07/05/2020	98	69,856	229	56	57	8.5	22.6	72,400	0.998141899	0.166678645	110	127
08/05/2020	99	70,132	276	56	57	8.5	23	72,600	0.998176115	0.183701463	111	128
09/05/2020	100	70,238	106	56	57	8.5	22.8	72,400	0.998286944	0.246999497	110	127
10/05/2020	101	70,292	54	56	57	8.5	22.7	72,200	0.998403616	0.382879386	110	127
11/05/2020	102	70,482	190	56	57	8.5	22.7	72,200	0.998435983	0.369879957	110	127
12/05/2020	103	70,775	293	56	57	8.5	22.9	72,400	0.998409686	0.270053998	110	128
13/05/2020	104	70,964	189	56	57	8.5	23.1	72,500	0.998425921	0.276756083	111	128
14/05/2020	105	71,280	316	56	58	9	22.3	72,600	0.999084054	0.158835446	110	127
15/05/2020	106	71,572	292	56	58	9	22.6	72,800	0.999097634	0.10264384	111	128
16/05/2020	107	71,590	18	56	58	9	22.4	72,700	0.999115033	0.164105561	110	127
17/05/2020	108	71,595	5	56	58	9	22.3	72,600	0.99913069	0.284532441	110	127

Table 6. Cases: The .csv file with the results of the forecast with the fitted model updated on 11 May 2020 with data until May 10 and the real data updated on 21 May 2020.

Date	Day	Cases Forecast	Deaths Forecast	ICUs Forecast	Cases 21/05	Deaths 21/05	ICUs 21/05
11/05	102	69,907.52664	8760.086075	3522.268127	70,764	8720	3543
12/05	103	70,069.38546	8785.675876	3532.117916	71,064	8760	3555
13/05	104	70,217.06763	8808.931051	3541.764624	71,273	8779	3564
14/05	105	70,351.54658	8830.015338	3551.212438	71,616	8809	3574
15/05	106	70,473.75895	8849.086841	3560.465458	71,932	8826	3577
16/05	107	70,584.60256	8866.297502	3569.527702	71,956	8847	3584
17/05	108	70,684.93486	8881.792689	3578.403102	71,995	8863	3594
18/05	109	70,775.57183	8895.710879	3587.09551	72,121	8894	3600

It is important to note that the parameter σ of the model changes to the right of the real peak of cases. This indicates that the containment measures are in fact effective. This results in an increase in the variance of the model to the right of the turning point μ, which indicates a slowdown in infections (see Figure 2). On April 18, with the addition of the column TestAc to the dataset, an explosion in the graph of the density of cases is observed (see Figure 2a). Except for this incident, the model remains quite stable to the right of the peak of cases and commitment dates for the progressive lifting of mobility restrictions can be proposed, as Figure 2 and Table 4 show. For example, May 11 shows commitment dates

between May 19 and June 5 on the basis of 0.01 and 0.001 for the right tail area of the model (with a forecast of 72,200 for the total of confirmed cases at the end of the pandemic).

This report concludes with a forecast for cumulative cases over the 8-day horizon. For example, Table 6 shows the forecast from May 11 to 18 generated on May 11.

2.9. New and Cumulative Deaths per Day Series

Figure 2b summarizes the results for the new and cumulative deaths per day series in the Region of Madrid.

A similar analysis to the one explained for cases can be done for deaths. Some days after the real peak of deaths has taken place: it is easy to know that is was already reached on March 28. The peak of the model, μ, was reached around April 1 (see Figure 2b).

The update on May 11 shows two early peaks on March 14 and 16. This situation was followed by an increase with respect to the fitted model between March 22 and 28 and then between April 2 and 9. Between these two periods of time, the situation of ICUs is dramatic, as explained in the following subsection.

It is important to highlight that the model changes to the right side due to the effectiveness of the containment measures. This translates into an increase in the variance of the model on the right side, which indicates a slowdown in deaths. On May 11, the fitted model forecasts a total of around 9000 deaths at the end of the pandemic.

2.10. New and Cumulative ICUs per Day Series

Figure 2c summarizes the results of new and cumulative ICUs in the Region of Madrid.

Although the official data are confusing, the fitted model for new and cumulative ICUs (to fit a suitable model to the series of ICUs, it was necessary to modify the algorithm considering two uniform and one exponential models to adequately describe the consecutive situations of plateau to the right of the normal model) reveals that the median of the model occurred between March 24 and 25 (see Figure 2c). It is important to note that the cumulative frequencies of new ICUs evolve very slowly, as the first graph in Figure 2c shows. It can be seen that 10% of the probability distribution of the model remains after May 11.

Taking into account that the duration of ICU stay depends on each patient and it usually ranges between 8 and 28 days, one can understand the saturation of the ICUs in the Region of Madrid.

The real peak of the series is one of the most difficult features to forecast. It may have occurred after the peak of deaths on March 28. The date when more new cases were incorporated into ICUs was March 20 with 205. This situation is detected with the value of the parameter μ of the normal model fitted to the left, which corresponds to March 21 (see the second graph of Figure 2c). However, that date does not correspond to the real peak because ICUs became saturated. On April 2, the reported number of ICUs reached the highest value: 1528. The two dates cited are around the worst moments in terms of numbers of deaths.

On May 11, the fitted model forecasts that a total of 4000 people will have gone through the ICU at the end of the pandemic.

2.11. Second Wave

After the first wave, the format in which the data were provided changed and their quality increased, although the series continues to change from one day to the next.

Figure 3 shows the fitted model for Madrid on December 12, for the second wave of all series, cases, deaths, hospitalizations and ICU admissions. Figure 4 shows the monitoring of the parameters: μ, σ_{left}, σ_{right} and n. Note how the effect of the interventions is manifested in a preview of the peak, in the jump that σ_{left} experiences with respect to σ_{right} and in the stabilization of n after reaching the peak. In addition, the peak is predicted in the future from the end of August.

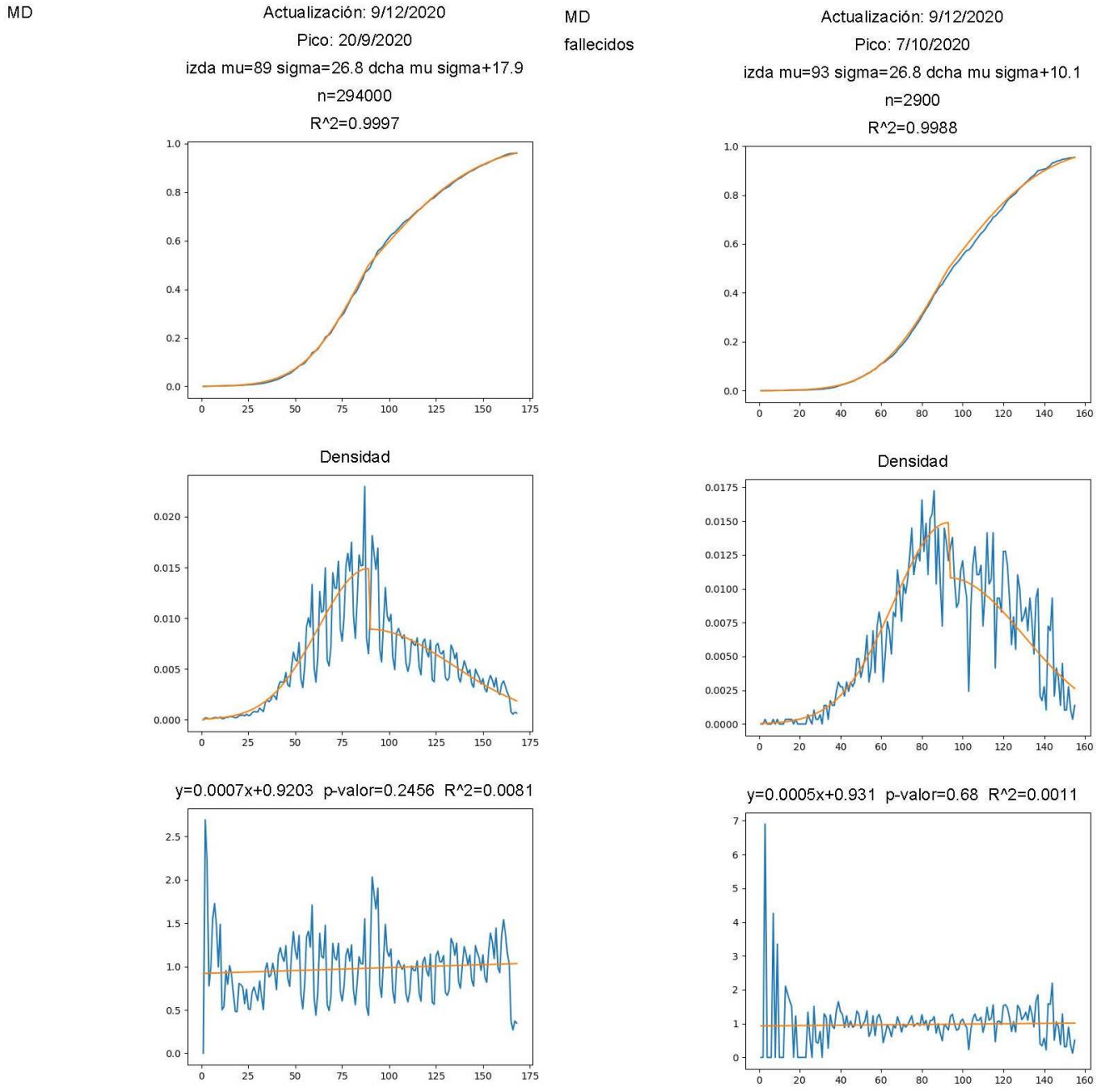


Figure 3. *Cont.*

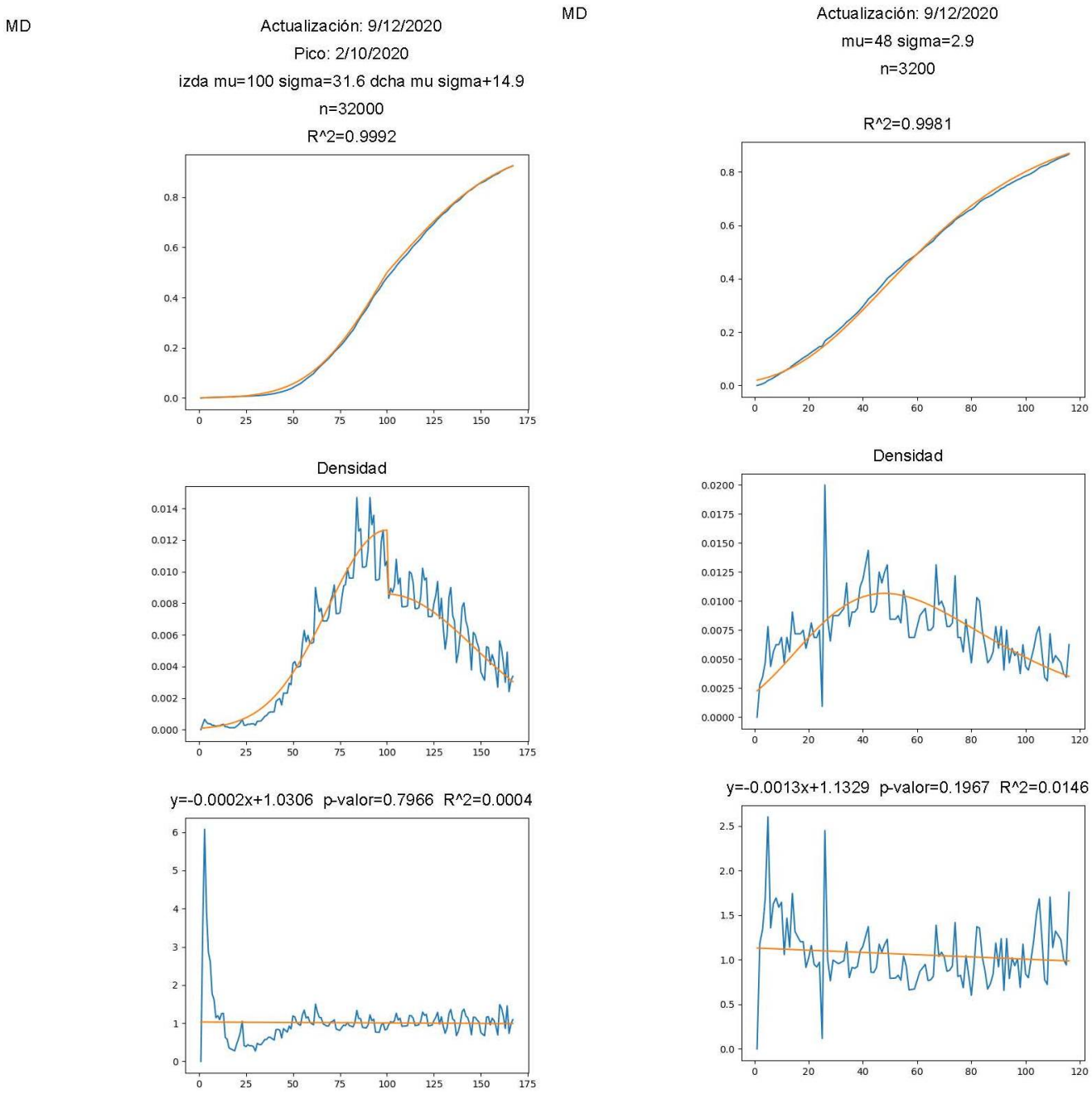


Figure 3. Second wave: cases (normal) (**top-left**); deaths (normal) (**top-right**); hospitalizations (normal) (**bottom-left**); and ICUs (Gompertz) (**bottom-right**).

Figures 5 and 6 shows the fitted models of the second wave for Asturias. These figures make evident the usefulness of testing different regression functions. Unlike in Madrid, the expert system selects as best fits those made with a double exponential or double Pareto model.

During the third and fourth waves, we incorporated the error correction model into our expert system that increased our predictive capacity. This is the model with which we currently give our 14-day predictions to the Spanish Mathematics Committee. We found in the comparison tool implemented on the website of this initiative that this new model remains among the top three over time with respect to all error metrics. Another of its advantages is that it adjusts the four data series at the same time: cases, deaths, hospitalized and ICUs.

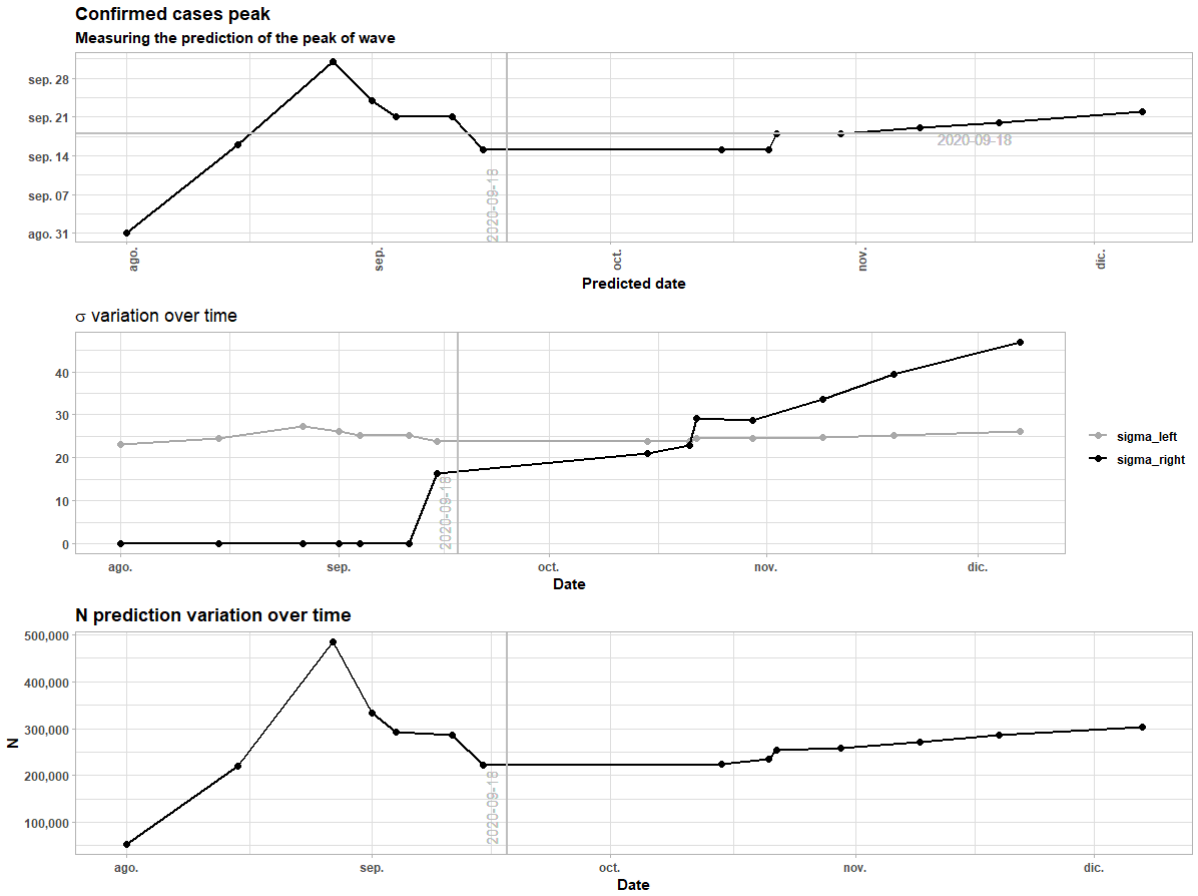


Figure 4. Second wave: cases parameter monitoring.

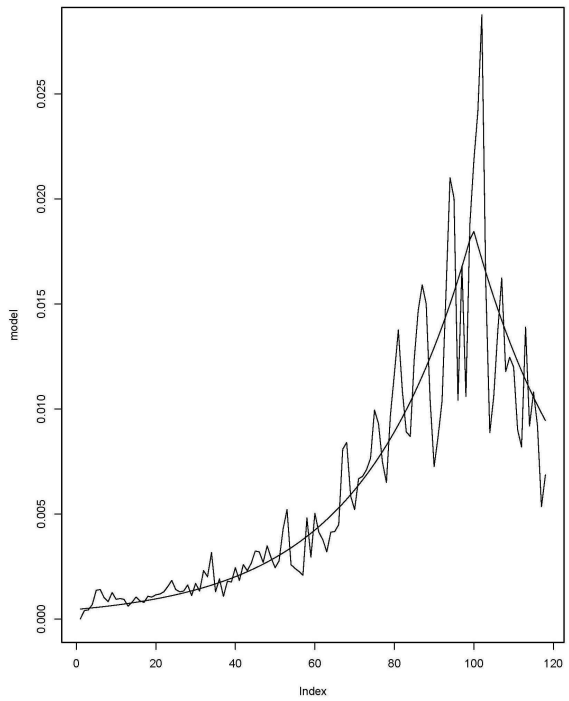


Figure 5. Second wave cases Asturias: double exponential.

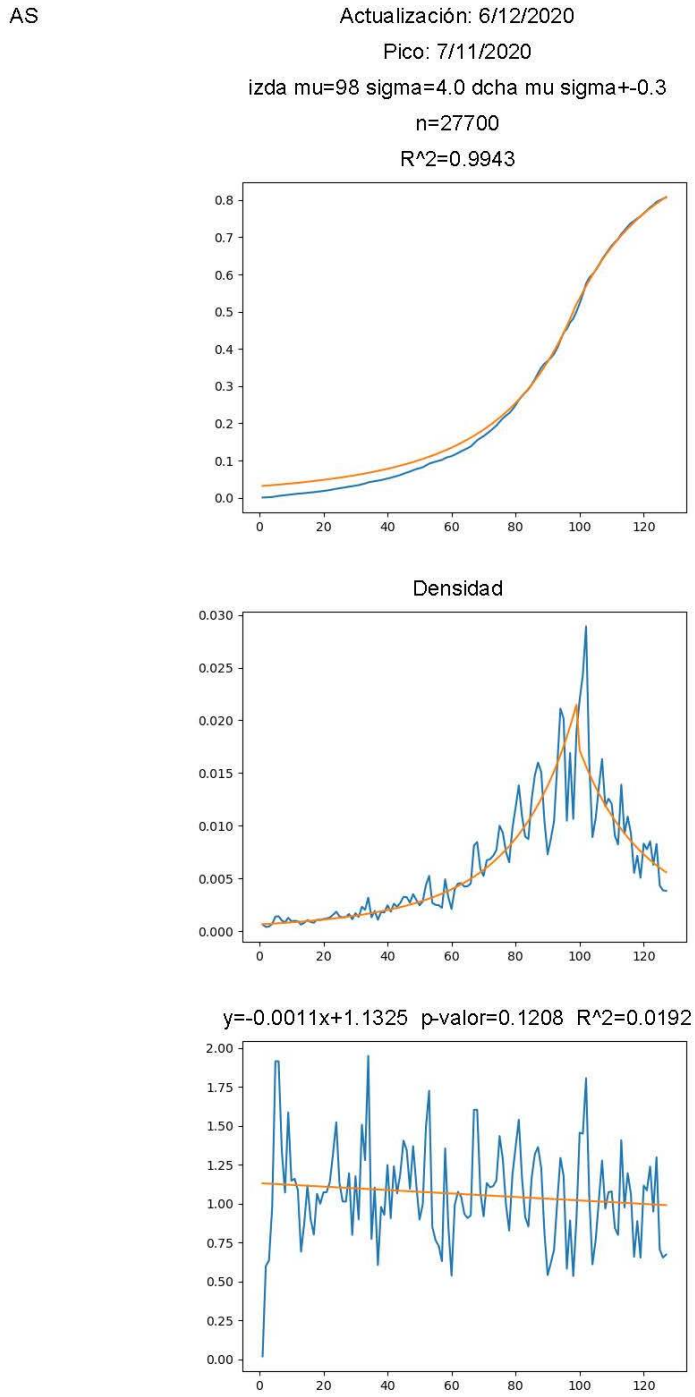


Figure 6. Second wave cases Asturias: double Pareto.

3. Error Correction Model

3.1. Model Definition

It is intuitive to think that a peak in the confirmed series will increase the other three data series with some delay. This type of relationship is usually modeled including the displacement of the confirmed series p times to the future and using this feature to predict the present of other series. This method for one series is called auto-regressive model and is generalized to more than one series in the vector auto-regressive model. This type of model needs the time series to be stationary, which means that the process has the first two moments constant along time. This restriction is a problem in this case, where we want to predict future changes in trends.

When one has a set of stationary time series, $y_t = (y_{1t}, \ldots, y_{Kt})$, one can define the stable auto-regressive vector model of order p as:

$$y_t = \nu + A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t$$

where A_i are matrices of parameters, $\nu = (\nu_1, \ldots, \nu_K)$ represents the mean of each time series and u_t represents the random error term of the model, with $u_t \sim \mathcal{N}(0, \Sigma_u)$.

The usual procedure when working with this model with non-stationary series is to differentiate them until they become stationary, but this procedure clouds the inference about the model.

There will be a long-term relationship. The stochastic trend of the series will be shared between the series as they will go down and up in the same way, keeping a certain distance between them.

We can say that the set of time series y_t have a long equilibrium if there exists a vector β such that $\beta' y_t = \beta_1 y_{1t} + \ldots + \beta_K y_{Kt} = 0$ and define the process $z_t = \beta' y_t$ as the deviations from this relation.

Thus, the series in the set y_t are said to be cointegrated if y_{it} is a series of order 1 for all $i = 1, \ldots, K$ and there exists a vector β such that $z_t = \beta' y_t$ is a stationary process.

With this relationship in mind, it is possible to define a model based on the vector auto-regresive for this type of data, which is the error correction model and is given by

$$\begin{aligned} \Delta y_t &= \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{p-1} \Delta y_{t-p+1} + \phi d_t + u_t & (1) \\ &= \alpha \beta' y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{p-1} \Delta y_{t-p+1} + \phi d_t + u_t & (2) \end{aligned}$$

where Γ_i are matrices of parameters, d_t is the deterministic part of the model, $u_t \sim \mathcal{N}(0, \Sigma_u)$ and $\Pi = \alpha \beta'$ with α the loading matrix and β the cointegration matrix.

Integrated and cointegrated systems must be interpreted cautiously. A tool to make inferences about the model is the impulse response function. This tool describes the evolution of the model variables in reaction to a shock in one or more of them. The impulse response function is developed for VAR models, but we can express the error correction model as a VAR model, as mentioned in [25].

For the application of the error correction model to the COVID-19 data, the following modifications are made:

- It is first necessary to apply the logarithm to each of them; this is because seasonality is multiplicative while the model is additive, so it is necessary to transform this relationship and capture it with the proposed model.
- The data series show a strong seasonality due to the data collection and publication policy of each region. To capture this seasonality in the model, the deterministic component, d_t, a dummy variable relative to each day of the week except one, is introduced (to avoid collinearity).
- A last change in the model is carried out using a dummy variable for a change of scenario: that of the test policy, since tests were not available at the beginning of the pandemic. While at the beginning of the first wave only some suspected cases could be tested for SARS-CoV-2, later the scenario changes and diagnostic testing can be extended to all suspected cases, close contacts and even several mass screenings are performed.

With these changes, the final model formula is given by Equations (3) and (4).

$$\begin{aligned} \Delta y_t &= \Pi_1 y_{t-1} + \Gamma_{1,1} \Delta y_{t-1} + \ldots + \Gamma_{1,p-1} \Delta y_{t-p+1} + \phi_1 d_t + u_t \\ &= \alpha_1 \beta_1' y_{t-1} + \Gamma_{1,1} \Delta y_{t-1} + \ldots + \Gamma_{1,p-1} \Delta y_{t-p+1} + \phi_1 d_t + u_t \text{ for } t \leq T_1 \end{aligned} \quad (3)$$

and

$$\begin{aligned}\Delta y_t &= \Pi_2 y_{t-1} + \Gamma_{2,1}\Delta y_{t-1} + \ldots + \Gamma_{2,p-1}\Delta y_{t-p+1} + \phi_2 d_t + u_t \\ &= \alpha_2\beta_2' y_{t-1} + \Gamma_{2,1}\Delta y_{t-1} + \ldots + \Gamma_{2,p-1}\Delta y_{t-p+1} + \phi_2 d_t + u_t \text{ for } t > T_1\end{aligned} \tag{4}$$

Once the model is defined, there is only one parameter to decide, the regression order p, that is, the number of lags of each of the series in the equation. For the decision of this parameter and given the ease of calculating this model, a cross-validation procedure is carried out with different p, taking the order p with the smallest mean absolute percentage error.

3.2. Application to the Case Study

To assemble the model, we proceed as described in the previous section.

A logarithmic transformation is applied to the data and the two-to-two cointegration of the series under study is checked with the Engle and Granger [26] procedure and the Phillips–Ouliaris test [26]. Both tests work with the null hypothesis of non-existence of cointegration and the p-values of series two by two are shown in Table 7.

Table 7. Cointegration tests for Madrid series.

Test	Combination	p-Value
Phillips–Ouliaris	confirmed–hosp	<0.01
Engle and Granger	confirmed–hosp	<0.01
Phillips–Ouliaris	confirmed–icu	<0.01
Engle and Granger	confirmed–icu	<0.01
Phillips–Ouliaris	confirmed–deaths	<0.01
Engle and Granger	confirmed–deaths	<0.01
Phillips–Ouliaris	hosp–icu	<0.01
Engle and Granger	hosp–icu	<0.01
Phillips–Ouliaris	hosp–deaths	<0.01
Engle and Granger	hosp–deaths	<0.01
Phillips–Ouliaris	icu–deaths	<0.01
Engle and Granger	icu–deaths	<0.01

The cointegration relationship shows us that there is a long-term equilibrium between the series. Another present relationship between these is the regressive part. In order to confirm it, a cross-correlation study was carried out among the series of ICU, hospitalized and deaths against the series of confirmed patients. To avoid spurious correlations, a differentiation process is carried out, to transform these into stationary series individually, applying the logarithm and differentiating once.

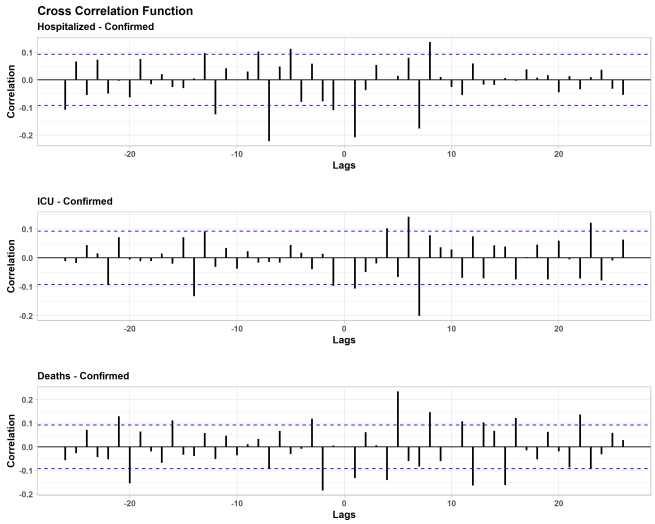
It is observed in Figure 7a that there are important shifts of the series in the confirmed series (as well as between all of them, outside the scope of this article) indicated by entering the rejection region of the test for correlation 0.

To obtain the optimal order for all of them simultaneously, the auto-regressive order p is optimized, obtaining in this case that the optimal p is 9.

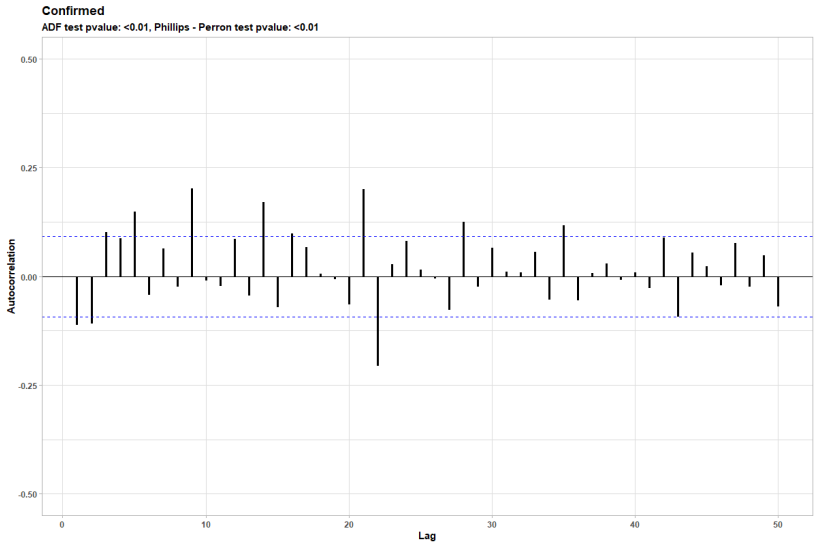
To ensure the adequacy of the model, the independence of the errors is checked. The p-values for the augmented Dickey–Fuller test [26] and the Phillips–Perron test [26] are included in the auto-correlation graph of the residuals in Figure 7b.

No pattern is observed in the auto-correlation graph of the residuals in Figure 7b, and the p-values of the tests allow us to reject the hypothesis of the presence of a unit root, which is why it is concluded that the errors are stationary.

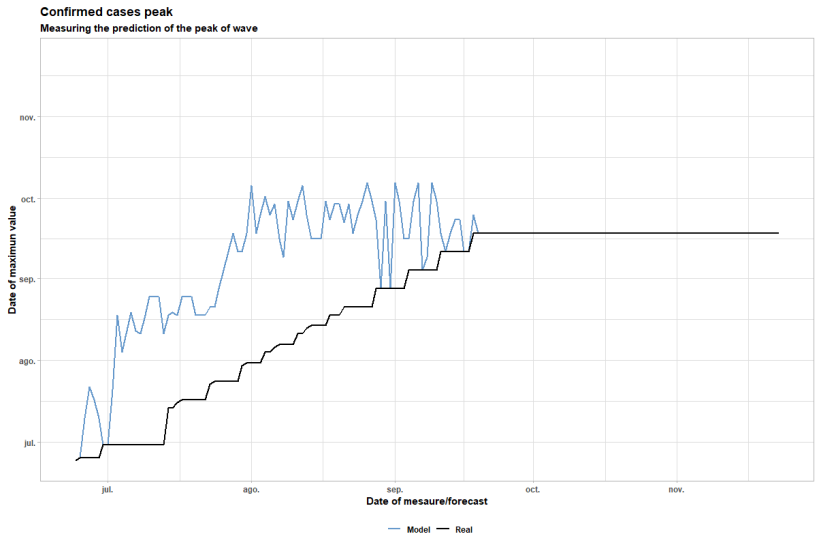
Another good check in the case study is to see how the peak of the wave is predicted. To this end, Figure 7c presents the current date on the x-axis and the peak date on the y-axis. The black line represents the maximum value of the series up to the real date, while the blue line represents the date of the peak predicted by the model.



(**a**) CCF of ICU, hospitalized and deaths versus confirmed series



(**b**) ACF of confirmed model residuals



(**c**) Peak prediction of the second wave in Madrid

Figure 7. Second Wave in Madrid.

It can be observed in the graph that, while in the first half of July the peak was predicted in mid-August, in the second half of July the prediction of the peak rises sharply until mid-September, remaining stable in this prediction until the actual date of the peak (18 September 2020).

In addition, it is appreciated that the model is capable of detecting when the peak has passed. This is observed in Figure 7c by noticing how from the real date of the peak (18 September 2020) the model predicts the peak to pass.

To check the adequacy of the model for prediction, error metrics are calculated [27] using cross-validation techniques [28] for fitted and forecast value and presented for frequency (Table 8) and cumulative data (Table 9).

Finally, the results of applying the impulse response function with two different histories are presented. The first represents the data from the beginning of the pandemic to before the second wave. The second represents the data from the beginning of the pandemic to before the third wave. This graph can be observed in Figure 8a.

In this figure, the maximum influence and the moment at which its effects decrease can be observed, the curves being less pronounced.

Table 8. Metrics of error correction model for punctual data.

Metric	Train	Test
MAPE		28.5
MPE		8.76
R^2	0.949	0.798

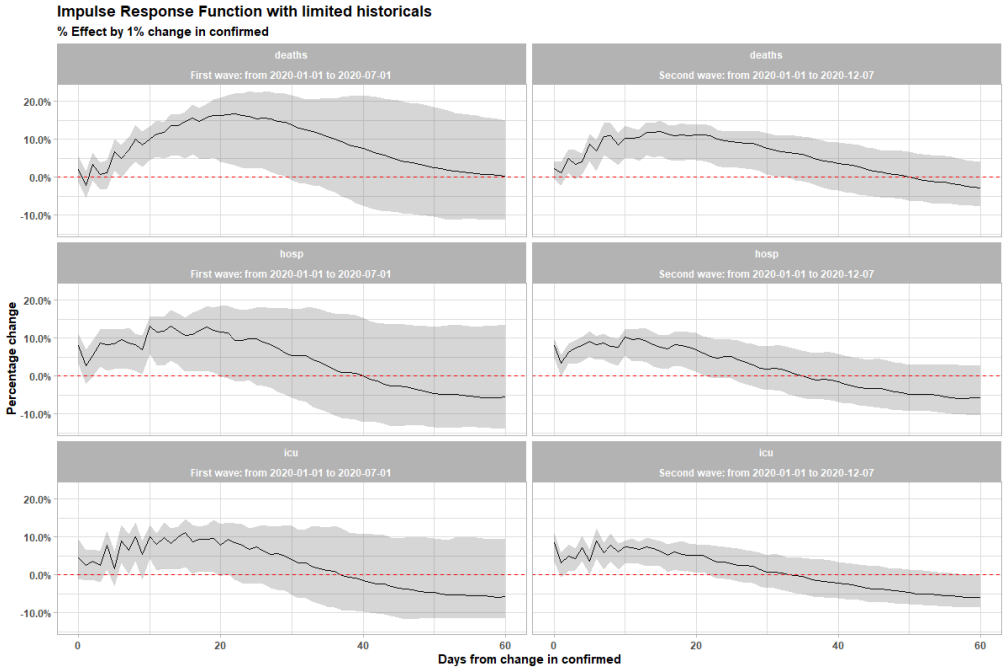
Table 9. Metrics of the error correction model for cumulative data.

Metric	Train	Test
MAPE	14.5856	3.3848
MPE	11.1601	1.6564
R^2	0.9998	0.9913

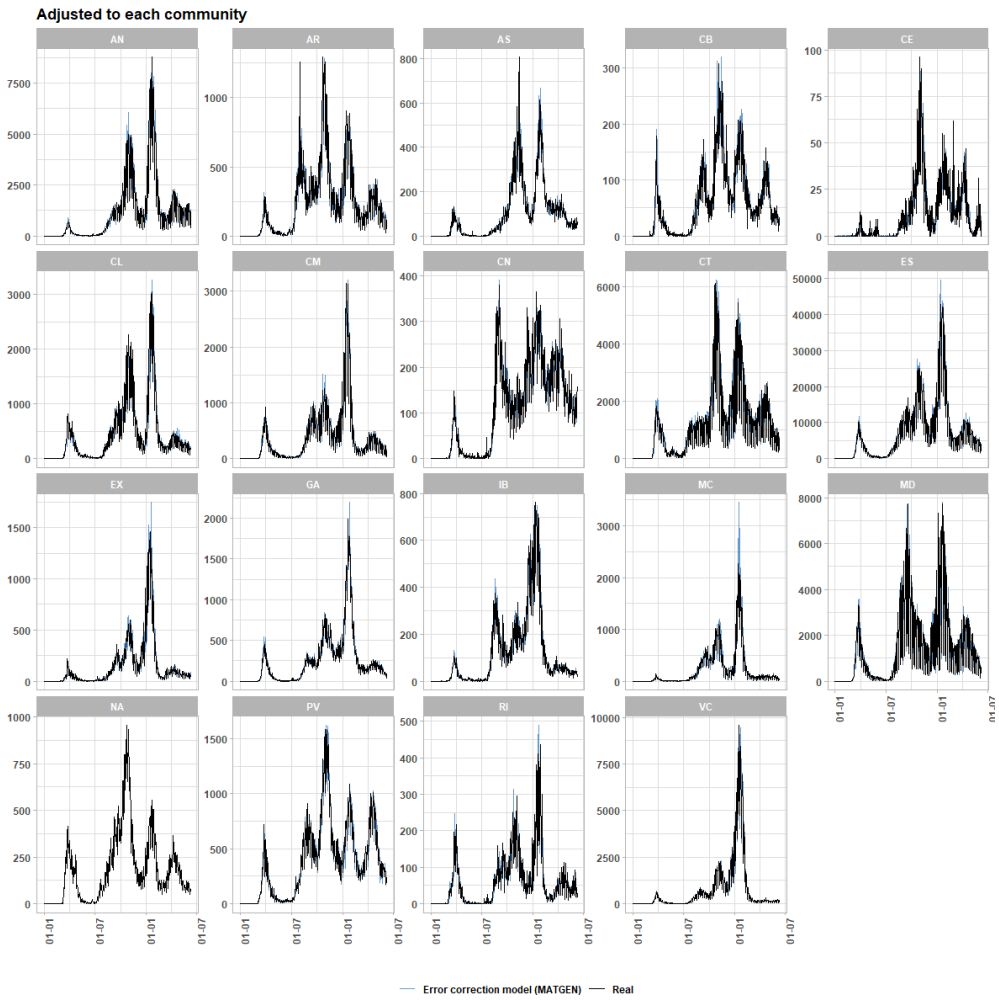
It should be noted that, initially, this model was not developed for the prediction of confirmed cases, but rather for the prediction of the other three series that depend closely on it. Despite this, the model works really well with this series, but it works better in the inpatient series.

The implementation of the model allows quick execution of the fit and forecast for all the Autonomous Communities (Figure 8b).

For the comparisons between models, we considered the following extension of the SIR model proposed by Castro et al. [10] and an automatic implementation of ARIMA forecast model by Hyndman and Khandakar [29]. We chose the SCIR model because it incorporates the compartment of the deaths into its definition and is formulated through only five parameters. Different extensions of SIR models can be found in [30,31]. The automatic ARIMA forecast model was chosen due to its widespread use in time series forecasting. Some example can be found in [32,33].



(**a**) IRF



(**b**) Adjusted graph for each Region

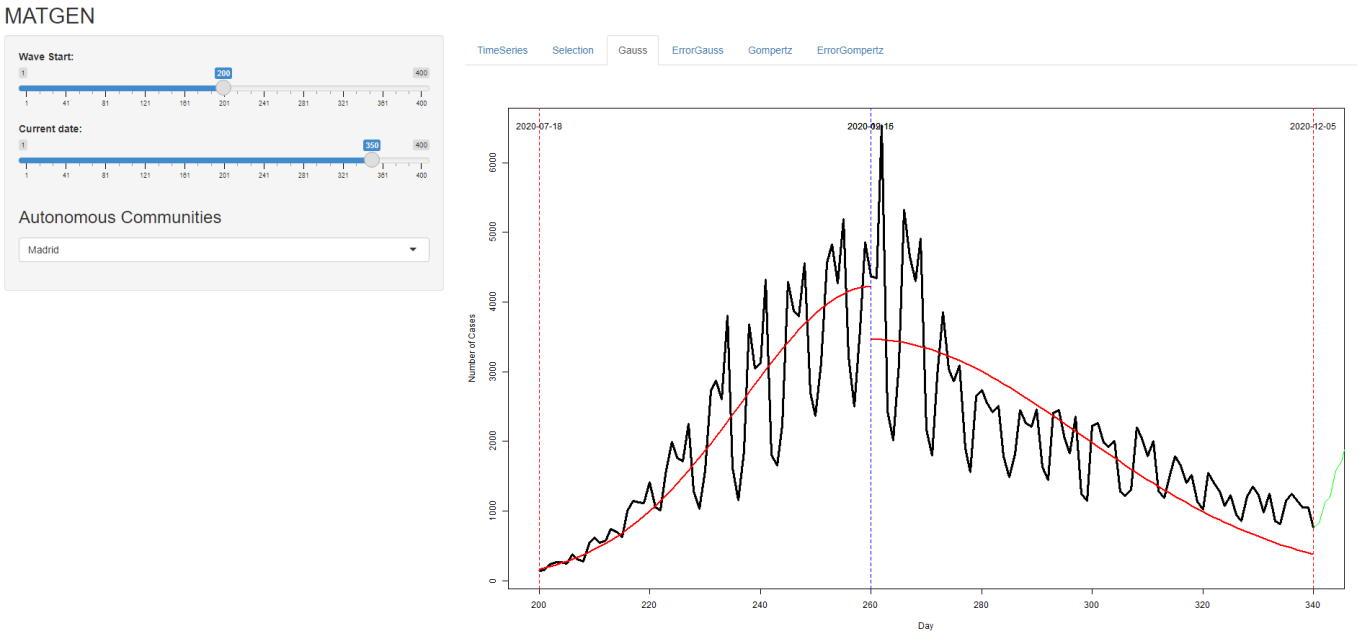
Figure 8. IRF and adjusted graph for each Region.

4. SCIR Model: A SIR Model with Confinement

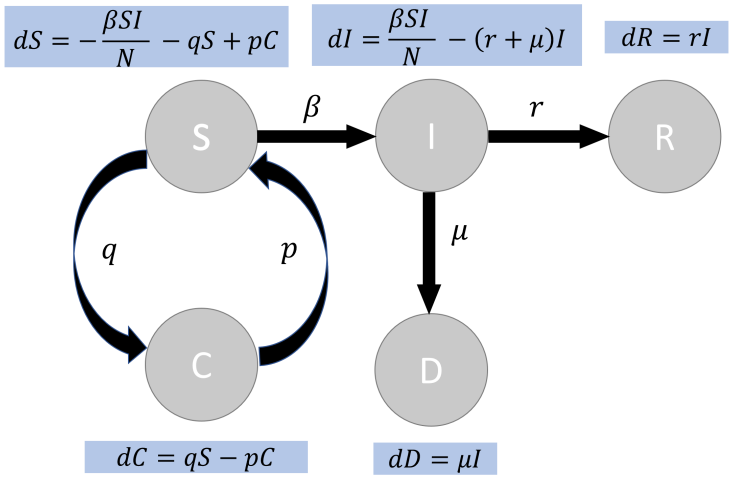
The SCIR model includes the usual states of an SIR model plus a class C for individuals sent to confinement who are susceptible, but not infected. Susceptible individuals (S) can enter and exit confinement (C) or become infected (I). Infected individuals can recover (R) or die (D). Figure 9b shows the diagram of the equations of the SCIR model.

For the optimization of the parameters, we included in the objective function an accuracy measure that combines the determination coefficients of the fits of both the cases and the deaths.

In the next section, the comparison between the three models is illustrated with the second wave for the Region of Madrid.



(**a**)



(**b**)

Figure 9. Expert system and SCIR diagram. (**a**) Expert system graphical interface. (**b**) Diagram of the SCIR model [10].

5. Automatic SARIMA Model

The procedure followed to apply the ARIMA automatic adjustment method given in [29] is:

- Automatic model adjustment, with the selection of the model's hyper-parameters $(p, d, q)(P, D, Q)$ and a Box–Cox transformation [34] with parameter λ.
  The result of this step in the data is the model with: $p = 1$, $d = 1$, $q = 2$, $P = 0$, $D = 1$, $Q = 1$ and $\lambda = 0.2086024$.
  The following model equation was chosen:

$$(1 - \phi_1 \mathrm{B})\ (1 - \mathrm{B})\ (1 - \mathrm{B}^7) y_t = (1 + \theta_1 \mathrm{B} + \theta_2 \mathrm{B}^2)\ (1 + \Theta_1 \mathrm{B}^7)\ (1 + \mathrm{B}^7) \varepsilon_t$$

  where B is the backshift operator.
- In the validation phase, the regression parameters are recalculated with these same hyper-parameters.

6. Other Models

To verify the contribution of the two models developed, other algorithms such as neural networks [35] and machine learning algorithms qwew tested. To present some of the results, the predictions are shown in Figure 10. The same dates for the implementation of random forest and xgboost with a direct forecasting strategy are presented [36].

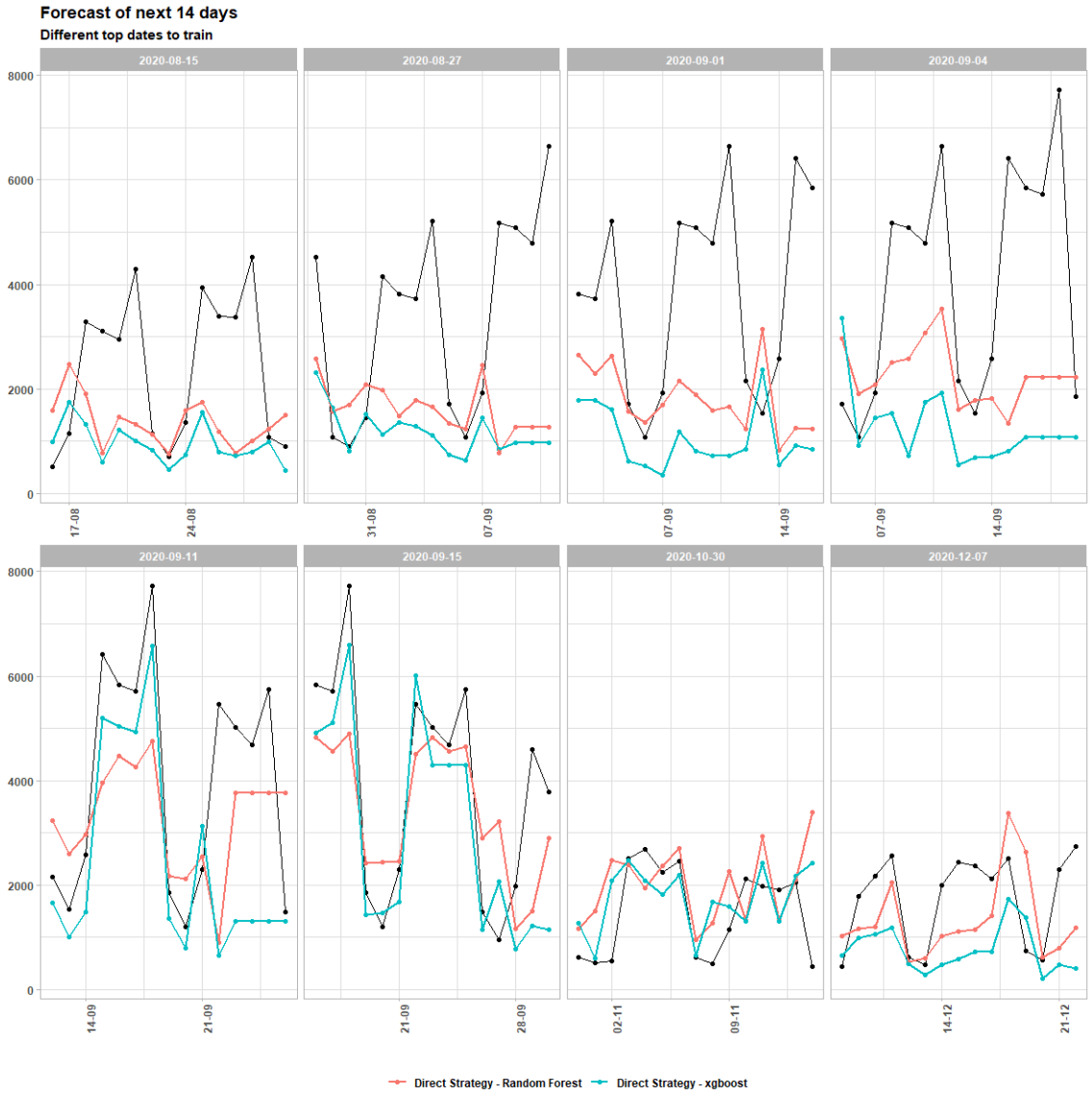


Figure 10. ML models example.

In [37], a deep long short-term memory network is used very successfully for financial time series forecasting. We tried to apply this neural network to our problem, but the result is not satisfactory. This is undoubtedly due to the fact that the number of samples of each of the waves is very insufficient for the training of a deep learning network.

7. Comparisons

Figures 11 and 12 show the fit and 14-day forecasts with the three models for the time interval corresponding to the second wave for eight different endings of the historical time series (all of them starting on June, 24th), respectively, for the cumulative and non-cumulative absolute frequencies of the cases.

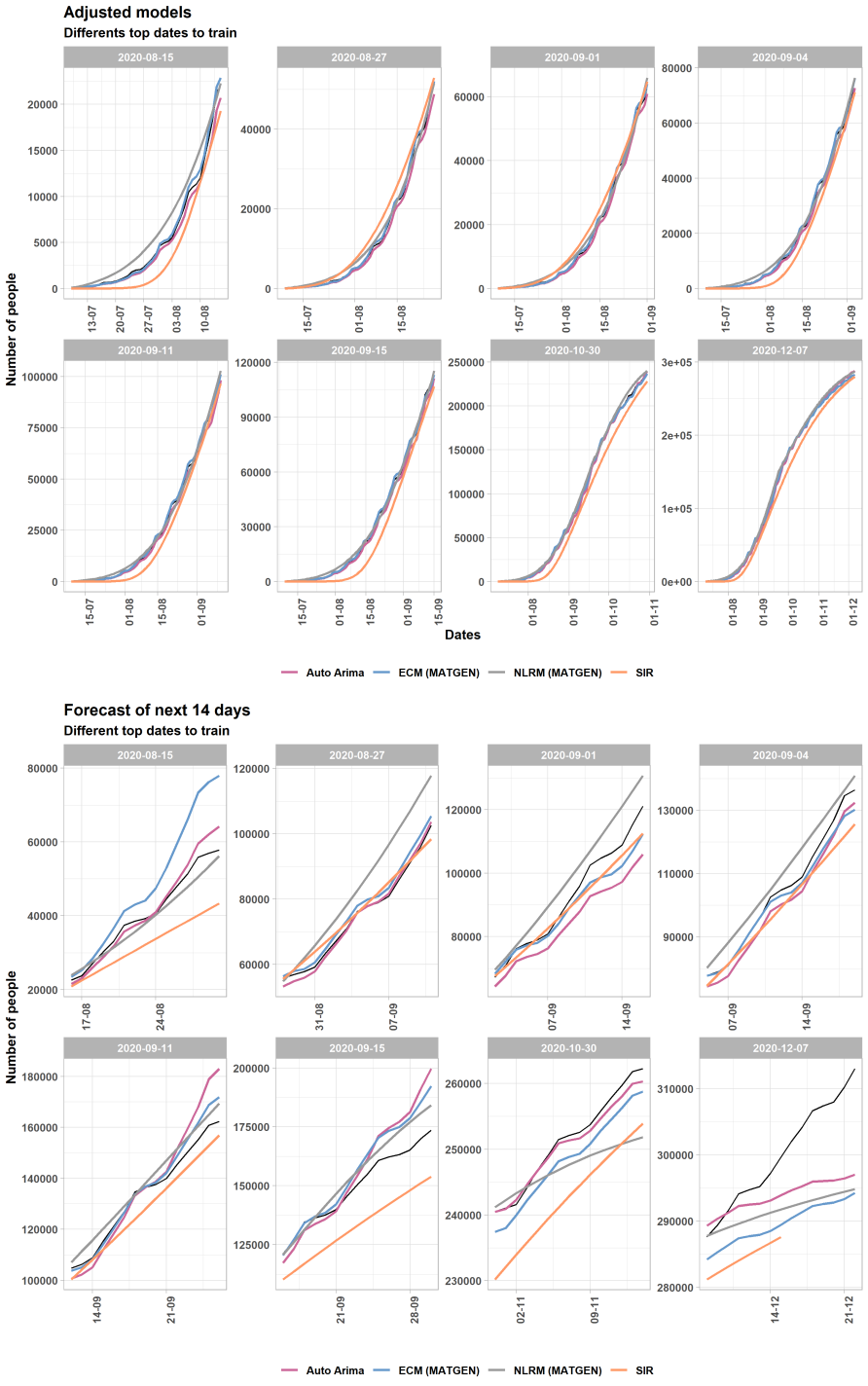


Figure 11. (**Top**) Cumulative cases adjustment. (**Bottom**) Cumulative cases forecast at 14 days.

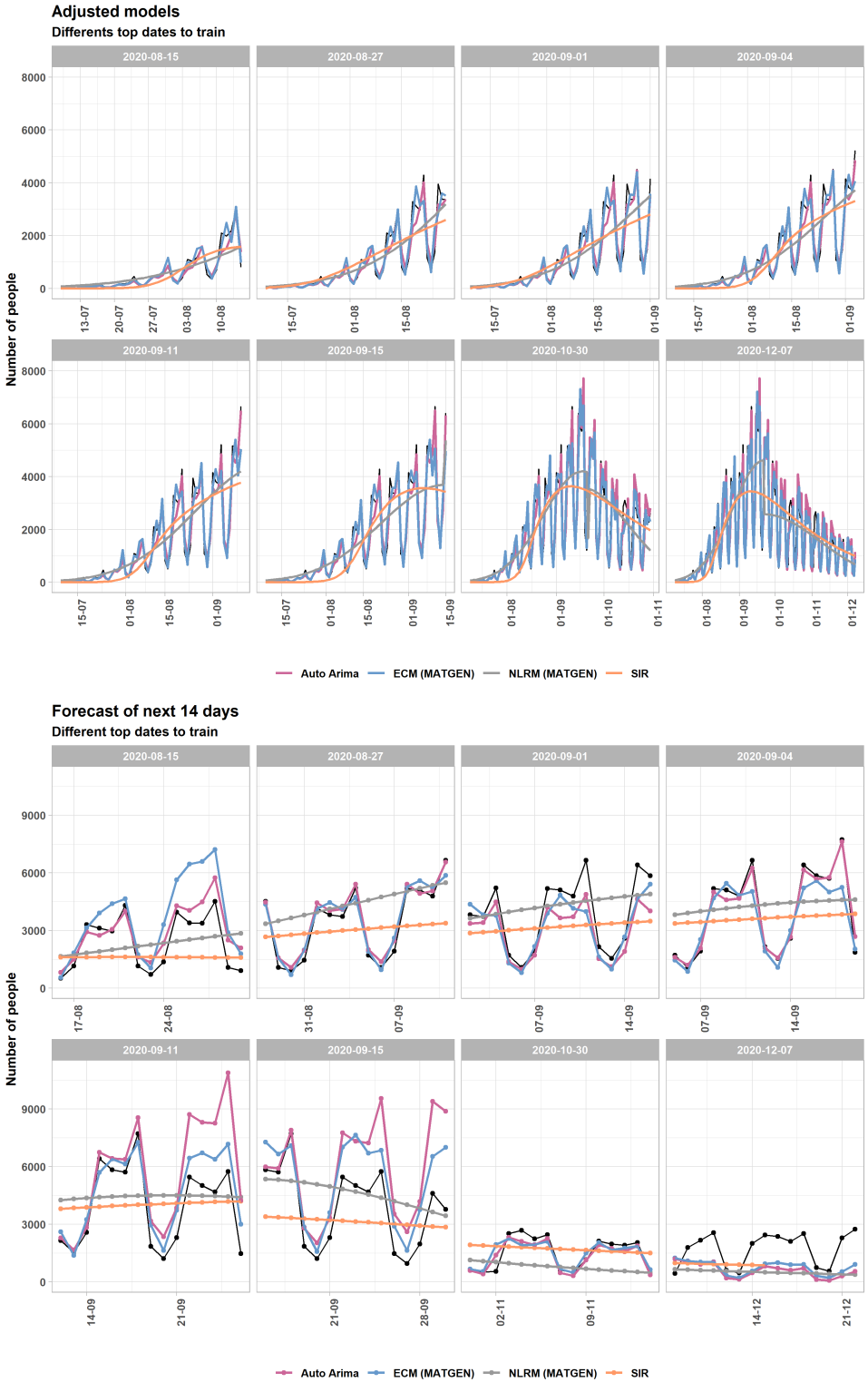


Figure 12. (**Top**) Non-cumulative cases adjustment. (**Bottom**) Non-cumulative cases forecast at 14 days.

Figure 13 shows the corresponding box-plots for the metrics values obtained with the three models. The means of the metrics are shown in Tables 10 and 11. The individual values for the eight different data histories are shown in Tables 12–15. In general, it is noted that our two proposed MATGEN models improve the metric values obtained with the SCIR model in both fitting and forecast.

Table 10. Mean of the metrics values for cumulative data. Error Correction Model (ECM) and Non-Linear Regression Model (NLRM).

	Model	Type	ME	RMSE	MAE	MPE	MAPE	R^2
1	Auto ARIMA	adjusted	−1248.54	1728.10	1269.68	−7.85	8.59	1.00
2	Auto ARIMA	prediction	−1112.81	6833.94	4632.59	−1.40	3.81	0.99
3	ECM (MATGEN)	adjusted	81.20	1345.88	894.42	−0.81	5.09	1.00
4	ECM (MATGEN)	prediction	−19.95	6932.53	4867.94	1.44	4.06	0.99
5	NLRM (MATGEN)	adjusted	1348.81	2443.29	1790.66	16.39	17.37	1.00
6	NLRM (MATGEN)	prediction	2073.89	7198.78	5863.88	2.44	4.85	0.99
7	SIR	adjusted	−6007.57	8953.81	6686.95	−17,728.71	17,734.67	0.99
8	SIR	prediction	−6714.01	8943.46	7271.55	−6.84	7.61	0.99

Table 11. Mean of the metrics values for non-cumulative data. Error Correction Model (ECM) and Non-Linear Regression Model (NLRM).

	Model	Type	ME	RMSE	MAE	MPE	MAPE	R^2
1	Auto ARIMA	adjusted	−28.51	340.40	219.37	−4.26	21.07	0.96
2	Auto ARIMA	prediction	283.29	1376.83	891.92	−28.59	56.27	0.71
3	ECM (MATGEN)	adjusted	−14.78	402.27	250.27	−0.67	17.67	0.94
4	ECM (MATGEN)	prediction	186.95	1133.20	853.33	−11.54	39.33	0.74
5	NLRM (MATGEN)	adjusted	6.32	1052.47	766.99	9.02	48.03	0.58
6	NLRM (MATGEN)	prediction	110.35	1785.72	1506.44	−36.81	80.40	0.28
7	SIR	adjusted	−62.88	1073.00	779.05	−13,852.97	13,888.64	0.57
8	SIR	prediction	−402.71	1789.38	1595.02	−18.22	61.42	0.24

Table 12. Metrics' values for the adjustment of the cumulative data. Error Correction Model (ECM) and Non-Linear Regression Model (NLRM).

Set	Model	ME	RMSE	MAE	MPE	MAPE	R^2
1	Auto ARIMA	−485.88	700.55	486.71	−12.34	13.81	0.99
1	ECM (MATGEN)	168.90	400.31	289.59	−2.18	8.24	1.00
1	NLRM (MATGEN)	1260.45	1551.80	1291.45	33.55	33.70	0.94
1	SIR	−1522.24	1823.16	1522.24	−46,288.96	46,288.96	0.92
2	Auto ARIMA	−858.09	1201.04	858.72	−10.84	11.95	0.99
2	ECM (MATGEN)	253.20	477.60	349.93	−1.96	6.66	1.00
2	NLRM (MATGEN)	764.84	1255.90	1035.12	22.35	23.21	0.99
2	SIR	2212.52	2870.74	2220.49	−12,599.75	12,637.41	0.96
3	Auto ARIMA	−978.93	1323.82	979.50	−10.20	11.21	1.00
3	ECM (MATGEN)	302.09	519.49	392.29	−1.71	6.24	1.00
3	NLRM (MATGEN)	589.18	1511.35	1251.04	20.61	22.35	0.99
3	SIR	1628.68	2289.74	1736.60	−11,456.21	11,487.04	0.99
4	Auto ARIMA	−1088.73	1470.23	1089.27	−9.91	10.87	1.00
4	ECM (MATGEN)	389.35	627.22	473.02	−1.26	6.09	1.00
4	NLRM (MATGEN)	642.38	1498.33	1246.00	19.55	21.10	1.00
4	SIR	−3325.18	3955.14	3325.18	−31,652.14	31,652.14	0.97
5	Auto ARIMA	−1375.22	1860.10	1375.71	−9.32	10.18	1.00
5	ECM (MATGEN)	522.39	850.41	642.24	−0.55	5.82	1.00
5	NLRM (MATGEN)	508.56	1823.34	1392.07	16.28	18.56	1.00
5	SIR	−3793.62	4639.00	3817.71	−28,582.67	28,582.70	0.98
6	Auto ARIMA	−1530.03	2051.45	1530.49	−9.01	9.81	1.00
6	ECM (MATGEN)	436.65	910.16	696.56	−0.52	5.60	1.00
6	NLRM (MATGEN)	616.64	1643.56	1324.56	17.57	18.93	1.00
6	SIR	−6159.50	7507.13	6159.50	−22,241.48	22,241.48	0.95
7	Auto ARIMA	−1630.21	2043.77	1630.49	−5.87	6.36	1.00
7	ECM (MATGEN)	39.24	1406.28	1048.43	−0.30	3.92	1.00
7	NLRM (MATGEN)	2051.99	3257.69	2473.12	11.33	12.01	1.00
7	SIR	−11,425.65	13,086.26	11,425.65	−13,595.76	13,595.76	0.98

Table 12. *Cont.*

Set	Model	ME	RMSE	MAE	MPE	MAPE	R^2
8	Auto ARIMA	−1257.97	1796.90	1340.12	−4.42	4.82	1.00
8	ECM (MATGEN)	−513.38	2151.91	1649.35	−0.21	3.33	1.00
8	NLRM (MATGEN)	2271.70	3192.56	2437.28	10.69	10.92	1.00
8	SIR	−10,433.77	11,931.45	10,433.77	−5642.22	5642.22	0.99

Table 13. Metrics values for the forecast of cumulative data. Error Correction Model (ECM) and Non-Linear Regression Model (NLRM).

Set	Model	ME	RMSE	MAE	MPE	MAPE	R^2
1	Auto ARIMA	742.27	2607.38	1972.18	0.22	4.40	0.95
1	ECM (MATGEN)	8078.15	10,443.86	8078.15	13.90	13.90	0.22
1	NLRM (MATGEN)	−1722.69	2601.55	2217.13	−3.68	5.64	0.95
1	SIR	−8332.40	9505.03	8332.40	−24.04	24.04	0.35
2	Auto ARIMA	−361.36	1250.37	1059.50	−0.85	1.61	0.99
2	ECM (MATGEN)	1986.20	2158.57	1986.20	2.48	2.48	0.98
2	NLRM (MATGEN)	9290.25	10,806.19	9418.25	10.11	10.34	0.47
2	SIR	1505.19	2685.56	2262.73	2.22	3.08.62	8.62
3	Auto ARIMA	−7682.18	8583.72	7682.18	−8.62	8.62	0.74
3	ECM (MATGEN)	−3316.37	4683.55	3621.03	−3.27	3.70	0.92
3	NLRM (MATGEN)	6570.61	7325.77	6570.61	6.34	6.34	0.81
3	SIR	−2875.42	4134.21	3236.74	−2.89	3.34	0.94
4	Auto ARIMA	−4195.20	4234.32	4195.20	−4.23	4.23	0.95
4	ECM (MATGEN)	−2057.25	2967.20	2111.82	−1.75	1.81	0.98
4	NLRM (MATGEN)	5299.33	5672.03	5299.33	4.95	4.95	0.92
4	SIR	−4942.52	6137.30	5013.43	−4.62	4.71	0.90
5	Auto ARIMA	3244.98	8616.59	6127.89	1.44	4.07	0.80
5	ECM (MATGEN)	1976.52	4041.05	2769.67	1.12	1.82	0.96
5	NLRM (MATGEN)	4409.42	4942.49	4543.91	3.19	3.29	0.93
5	SIR	−5467.07	5976.66	5467.07	−4.22	4.22	0.90
6	Auto ARIMA	6351.74	11,367.51	8513.56	3.24	4.95	0.51
6	ECM (MATGEN)	6700.05	9047.66	6849.74	3.84	3.96	0.69
6	NLRM (MATGEN)	5235.72	6889.86	5807.06	3.09	3.53	0.82
6	SIR	−16,797.09	17,040.06	16,797.09	−12.63	12.63	−0.09
7	Auto ARIMA	−708.88	1054.81	868.51	−0.28	0.34	0.98
7	ECM (MATGEN)	−3022.85	3068.54	3022.85	−1.21	1.21	0.83
7	NLRM (MATGEN)	−4102.31	5676.48	4633.04	−1.64	1.86	0.41
7	SIR	−8786.20	8826.30	8786.20	−3.63	3.63	−0.42
8	Auto ARIMA	−6293.84	8310.95	6641.69	−2.13	2.25	−0.09
8	ECM (MATGEN)	−10,504.03	11,506.51	10,504.03	−3.61	3.61	−1.09
8	NLRM (MATGEN)	−8389.22	10,053.77	8421.72	−2.86	2.87	−0.60
8	SIR	−9156.26	9312.66	9156.26	−3.21	3.21	−4.44

Table 14. Metrics values for the adjustment of the cumulative data. Error Correction Model (ECM) and Non-Linear Regression Model (NLRM).

Set	Model	ME	RMSE	MAE	MPE	MAPE	R^2
1	Auto ARIMA	−36.55	200.52	117.06	−7.82	25.31	0.92
1	ECM (MATGEN)	20.56	171.19	95.72	0.99	16.63	0.94
1	NLRM (MATGEN)	4.78	435.54	298.20	20.66	47.93	0.64
1	SIR	−74.26	434.98	300.27	−34,245.49	34,267.18	0.64
2	Auto ARIMA	−49.77	288.63	179.84	−5.37	23.94	0.94
2	ECM (MATGEN)	14.12	326.32	181.57	0.53	17.40	0.93
2	NLRM (MATGEN)	4.27	702.71	483.61	13.99	46.54	0.66
2	SIR	32.40	721.61	505.11	−12,598.90	12,658.23	0.64
3	Auto ARIMA	−44.57	295.50	188.00	−4.45	22.92	0.95
3	ECM (MATGEN)	10.46	323.84	194.94	0.15	18.09	0.94
3	NLRM (MATGEN)	46.30	868.08	605.99	13.91	48.45	0.56
3	SIR	24.38	843.97	604.11	−11,454.88	11,513.44	0.58

Table 14. *Cont.*

Set	Model	ME	RMSE	MAE	MPE	MAPE	R^2
4	Auto ARIMA	−58.22	297.94	194.23	−4.64	22.15	0.96
4	ECM (MATGEN)	1.37	351.12	212.71	0.07	17.86	0.94
4	NLRM (MATGEN)	5.11	863.13	606.81	12.21	46.72	0.65
4	SIR	−85.49	874.68	640.00	−23,183.48	23,209.36	0.64
5	Auto ARIMA	−65.79	295.95	196.90	−4.30	20.50	0.97
5	ECM (MATGEN)	−23.47	401.64	250.76	−0.22	17.71	0.94
5	NLRM (MATGEN)	3.02	1021.29	731.38	10.00	46.99	0.64
5	SIR	−93.38	1044.13	769.12	−20,880.91	20,908.29	0.63
6	Auto ARIMA	−58.64	288.40	191.54	−3.86	19.56	0.97
6	ECM (MATGEN)	−30.64	427.64	268.51	−0.25	17.56	0.94
6	NLRM (MATGEN)	2.15	1071.60	798.41	11.13	47.70	0.63
6	SIR	−122.06	1169.80	856.83	−16,662.54	16,688.03	0.56
7	Auto ARIMA	−0.20	405.01	268.60	−3.44	19.76	0.95
7	ECM (MATGEN)	−27.63	476.34	319.69	−1.37	17.78	0.93
7	NLRM (MATGEN)	1.68	1276.71	1001.69	4.26	48.29	0.50
7	SIR	−101.72	1314.40	1011.77	−10,167.64	10,199.50	0.47
8	Auto ARIMA	6.03	396.20	264.21	−3.59	19.91	0.94
8	ECM (MATGEN)	−27.85	436.26	285.29	−1.90	17.74	0.93
8	NLRM (MATGEN)	0.17	1192.96	921.71	3.67	49.28	0.47
8	SIR	−45.29	1183.58	899.62	−4958.01	4995.86	0.48

Table 15. Metrics values for the forecast of the non-cumulative data. Error Correction Model (ECM) and Non-Linear Regression Model (NLRM).

Set	Model	ME	RMSE	MAE	MPE	MAPE	R^2
1	Auto ARIMA	516.72	770.60	656.07	22.21	26.73	0.71
1	ECM (MATGEN)	1292.29	1665.30	1313.82	33.28	33.97	−0.35
1	NLRM (MATGEN)	−123.29	1393.85	1317.77	−6.65	58.84	0.05
1	SIR	−778.54	1584.90	1359.83	−48.49	84.86	−0.23
2	Auto ARIMA	236.05	312.25	278.05	10.58	11.38	0.97
2	ECM (MATGEN)	137.44	442.86	379.73	4.20	13.92	0.94
2	NLRM (MATGEN)	1003.11	1887.68	1462.84	23.83	34.71	−0.00
2	SIR	−385.05	1766.60	1666.21	−11.23	54.82	0.12
3	Auto ARIMA	−848.62	1028.91	848.62	−27.42	27.42	0.70
3	ECM (MATGEN)	−629.06	1028.97	750.27	−22.44	26.12	0.70
3	NLRM (MATGEN)	476.07	1845.24	1564.23	11.08	35.72	0.05
3	SIR	−661.61	1912.08	1765.00	−20.45	55.07	−0.02
4	Auto ARIMA	−51.16	298.94	213.27	0.66	6.42	0.98
4	ECM (MATGEN)	−427.13	902.31	636.70	−11.87	18.96	0.84
4	NLRM (MATGEN)	282.85	2087.31	1943.95	7.62	45.32	0.13
4	SIR	−388.76	2141.24	2020.72	−9.83	55.55	0.09
5	Auto ARIMA	1664.09	2240.33	1664.09	28.02	28.02	−0.07
5	ECM (MATGEN)	738.73	1046.20	912.73	17.54	21.55	0.77
5	NLRM (MATGEN)	459.34	2123.00	1908.62	10.57	43.08	0.04
5	SIR	49.68	2084.13	1929.12	1.33	48.04	0.07
6	Auto ARIMA	2005.82	2518.26	2005.82	35.48	35.48	−0.43
6	ECM (MATGEN)	1372.00	1633.92	1459.07	28.89	30.13	0.40
6	NLRM (MATGEN)	695.40	2054.18	1639.29	15.35	35.80	0.05
6	SIR	−753.06	2119.06	1943.78	−23.21	61.44	−0.01
7	Auto ARIMA	−126.86	317.46	239.17	−12.89	20.96	0.87
7	ECM (MATGEN)	−17.64	459.68	295.66	1.56	17.84	0.72
7	NLRM (MATGEN)	−702.71	1138.53	967.96	−104.96	132.27	−0.71
7	SIR	223.08	897.94	822.94	12.09	47.43	−0.06
8	Auto ARIMA	−1129.68	1365.44	1230.28	−285.34	293.76	−1.42
8	ECM (MATGEN)	−971.04	1206.39	1078.69	−143.49	152.12	−0.89
8	NLRM (MATGEN)	−1207.97	1489.81	1246.88	−251.29	257.50	−1.88
8	SIR	−636.58	1066.63	952.83	−70.27	104.03	−0.39

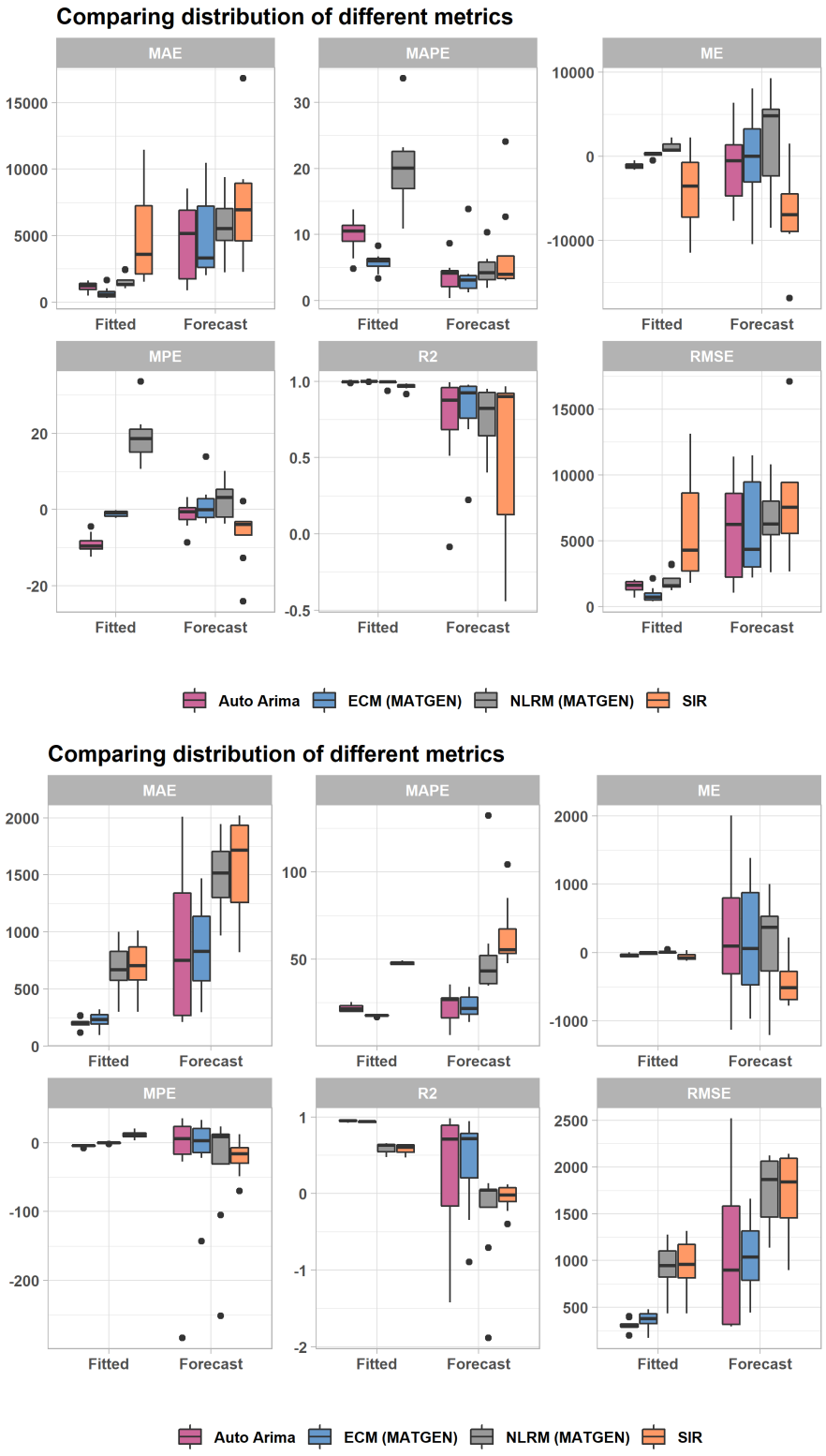


Figure 13. (**Left**) Cumulative cases metric comparisons. (**Right**) Non-cumulative cases metric comparisons.

Figure 14 shows the monitoring of the peak detection achieved with the three models. The model that best detects the peak is the non-linear regression model.

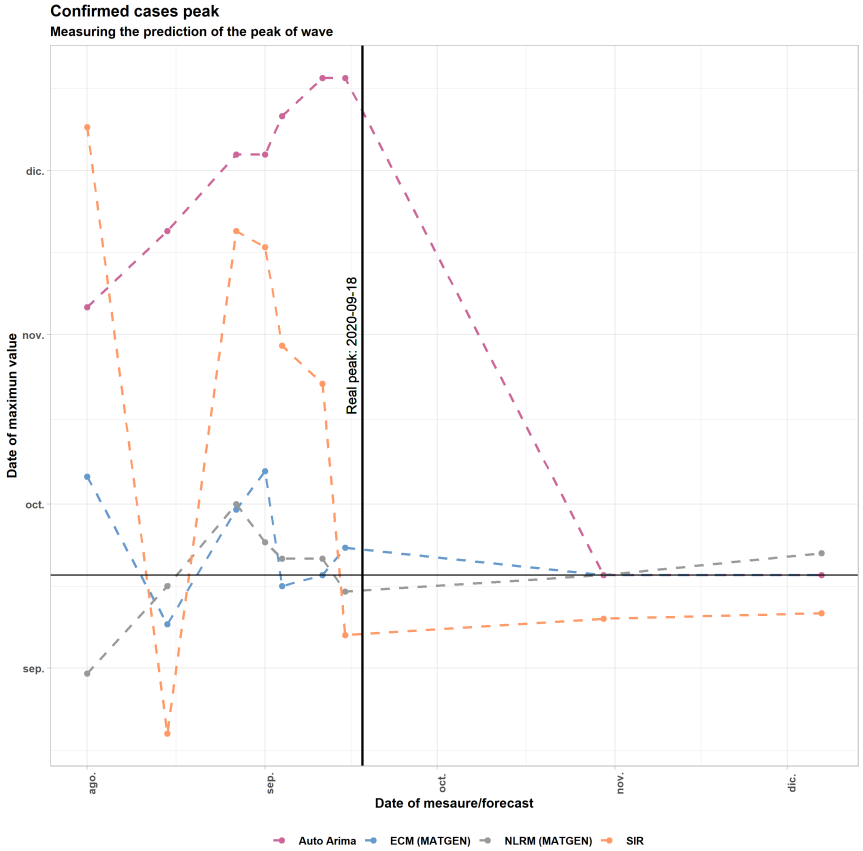


Figure 14. Peak prediction comparisons.

8. Conclusions

A non-linear regression model for count data in time series is developed and implemented by means of an expert system of artificial intelligence. It is based on directly estimating the distribution function of each of the series under study and on the duality between the distribution function and the density function. Since those two functions fully characterize the probability distribution of a variable, our model is able to capture the main characteristics of epidemic outbreaks. The simplicity of MATGEN must also be noted, since it is formulated only through four parameters: μ, σ_{left}, σ_{right} and n. The monitoring of all model parameters makes it possible to easily quantify and detect the effect of interventions over time. Furthermore, the machine learning algorithm developed is scalable, allows parallel running of different data series and is capable of introducing new data in real time.

We apply this model to the COVID-19 series of the Region of Madrid (Spain) during the first and second waves to give an eight-day forecast for the Spanish Mathematical Initiative [16]. This theoretical framework allows us to detect pandemic peaks and make short- and long-term monitoring and forecasting of the number of people infected, people requiring hospitalization and deaths. This expert system proves very useful to estimate the effectiveness of the interventions prompted by the government, which seems to have an impact after 11 days of its implementation during the first wave. Moreover, it is useful to propose commitment dates to lift the mobility restrictions and to advise on how to proceed in future outbreaks. On May 25, the Region of Madrid entered Phase 1 of the de-escalation. The MATGEN update on May 11 showed commitment dates between May 19 and June 5 (with a forecast of 72,200, 9000 and 4000 for the total numbers of confirmed cases, deaths and ICUs at the end of the pandemic, respectively).

The flexibility of our theoretical framework allows us to fit different regression functions (Gaussian, Gompertz, double Pareto, double exponential and uniform) between different dates. During the first wave, the series of new cases per day and deaths in the Region of Madrid only needed one cutoff and normal models in each part. The ICU series adjustment was more difficult: three cutoffs and different models in a wide family of probability distributions were needed. The computational cost of the last situation is higher than that of the first. In this case, the algorithm needed to detect the appropriate number of cutoffs and tried all the possible combinations of the models belonging to the family considered.

At present, in order to provide 14-day forecasts for the Spanish Mathematical Initiative [16], we implement another MATGEN model based on error correction models, useful for estimating both short- and long-term effects of one time series on another. This model is based on the cointegration of the four series in the study, namely cases, deaths, hospitalizations and ICUs, and it was incorporated into the expert system too.

Finally, the comparison from different points of view of our two models with the SCIR model [10] yields the following conclusions:

- Among the advantages of using the SCIR model, we can highlight its simplicity, since it is formulated using five parameters and the interpretability of them from an epidemiological point of view.
- The MATGEN non-linear regression model (formulated with four parameters per series that are easy to monitor) is the most explanatory for studying the peak detection and the effect of interventions. In addition, it is equipped with a control procedure that allows detecting trend changes in the tails that indicate the start of a new wave. Unlike the two other models, the fit of the four series is in parallel.
- The MATGEN error correction model depends on a greater number of parameters but allows us to approximate the four series simultaneously, and it is the best in all the metrics, both in fit and in forecast. In addition, this model incorporates the impulse response function, a method that allows making inference about the impact of one series on the remaining ones.

Therefore, MATGEN combines our two proposed models in one expert system as a new epidemiological tool that can be proved extremely useful in new COVID-19 outbreaks and future epidemics of infectious diseases.

Author Contributions: Conceptualization, B.G.-P., J.L.S. and C.N.; methodology, B.G.-P., G.V. and J.L.S.; software, B.G.-P., J.M.V., G.V. and J.L.S.; validation, B.G.-P., G.V. and J.L.S.; formal analysis, all authors; investigation, all authors; writing—original draft preparation, B.G.-P., J.M.V. and J.L.S.; writing—review and editing, B.G.-P., J.M.V. and C.N.; visualization, B.G.-P. and J.L.S.; supervision, B.G.-P.; and project administration, B.G.-P. and J.M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Complutense University of Madrid, Spain, research group 910395 MÉTODOS BAYESIANOS.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://cnecovid.isciii.es/covid19/.

Acknowledgments: We thank the reviewers for their suggestions that have helped to improve the initial draft of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Naudé, W. Artificial Intelligence Against Covid-19: An Early Review. *IZA Discuss. Pap.* **2020**, *13110*. Available online: https://ssrn.com/abstract=3568314 (accessed on 22 June 2021).
2. Verelst, F.; Willem, L.; Beutels, F. Behavioural change models for infectious disease transmission: A systematic review (2010–2015). *J. R. Soc. Interface* **2016**, *13*, 20160820. [CrossRef] [PubMed]
3. Akhtar, M.; Kraemer, M.U.G.; Gardner, L.M. A dynamic neural network model for predicting risk of Zika in real time. *BMC Med.* **2019**. [CrossRef] [PubMed]
4. Hao, K. This is How the CDC Is Trying to Forecast Coronavirus's Spread. 2020. Available online: https://www.technologyreview.com/2020/03/13/905313/cdc-cmu-forecasts-coronavirus-spread/ (accessed on 22 June 2021).
5. Abhari, R.S.; Marini, M.; Chokani, N. COVID-19 Epidemic in Switzerland: Growth Prediction and Containment Strategy Using Artificial Intelligence and Big Data. *medRxiv* **2020**. [CrossRef]
6. MITTechnologyReview. The Best, and the Worst, of the Coronavirus Dashboards. 2020. Available online: https://www.technologyreview.com/2020/03/06/905436/best-worst-coronavirus-dashboards/ (accessed on 22 June 2021).
7. Ivorra, B.; Ferrández, M.R.; Vela-Pérez, M.; Ramos, A.M. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Commun. Nonlinear Sci. Numer. Simul.* **2020**, *88*, 105303. [CrossRef] [PubMed]
8. Wang, L.; Zhou, Y.; He, J.; Zhu, B.; Wang, F.; Tang, L.; Eisenberg, M.; Song, P.X.K. An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. *medRxiv* **2020**. [CrossRef]
9. Maier, B.F.; Brockmann, D. Effective containment explains sub-exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China. *medRxiv* **2020**. [CrossRef]
10. Castro, M.; Ares, S.; Cuesta, J.A.; Manrubia, S. The turning point and end of an expanding epidemic cannot be precisely forecast. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26190–26196. [CrossRef]
11. Ramos, A.; Ferrández, M.; Vela-Pérez, M.; Kubik, A.; Ivorra, B. A simple but complex enough θ-SIR type model to be used with COVID-19 real data. Application to the case of Italy. *Phys. D Nonlinear Phenom.* **2021**, *421*, 132839. [CrossRef] [PubMed]
12. Sánchez-Villegas, P.; Daponte Codina, A. Predictive models of the COVID-19 epidemic in Spain with Gompertz curves. *Gac. Sanit.* **2020**. [CrossRef] [PubMed]
13. Nelder, J.A.; Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **1965**, *7*, 308–313. [CrossRef]
14. Nash, J.C. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*; Hilger: Bristol, UK; New York, NY, USA, 1990.
15. Català, M.; Alonso, S.; Alvarez-Lacalle, E.; López, D.; Cardona, P.J.; Prats, C. Empiric model for short-time prediction of COVID-19 spreading. *medRxiv* **2020**. [CrossRef]
16. CEMAT. Cooperative Prediction. 2021. Available online: https://covid19.citic.udc.es/ (accessed on 22 June 2021).
17. Quesada, V.; Pardo, L. *Curso Superior de Probabilidades*; Promociones y Publicaciones Universitarias (PPU): Madrid, Spain, 1988.
18. ISCIII. Instituto de Salud Carlos III. 2020. Available online: https://cnecovid.isciii.es/covid19 (accessed on 22 June 2021) .
19. Gómez Villegas, M.A. *Inferencia Estadística*; Díaz de Santos: Madrid, Spain, 2011.
20. Lauer, S.A.; Grantz, K.H.; Bi, Q.; Jones, F.K.; Zheng, Q.; Meredith, H.R.; Azman, A.S.; Reich, N.G.; Lessler, J. The Incubation Period of Coronavirus Disease 2019 (COVID-19) from Publicly Reported Confirmed Cases: Estimation and Application. *Ann. Intern. Med.* **2020**, *172*, 577–582. [CrossRef] [PubMed]
21. OpenDataUE. Portal for access to COVID-19 Open Data of the European Union. 2020. Available online: https://data.europa.eu/euodp/es/data/dataset/covid-19-coronavirus-data (accessed on 22 June 2021).
22. githubItaly. Dati Andmento Nazionale. 2021. Available online: https://github.com/pcm-dpc/COVID-19/tree/master/dati-andamento-nazionale (accessed on 22 June 2021).
23. githubItalyProvince. Dati Province. 2021. Available online: https://github.com/pcm-dpc/COVID-19/tree/master/dati-province (accessed on 22 June 2021).
24. githubItalyRegioni. Dati Regioni. 2021. Available online: https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni (accessed on 22 June 2021).
25. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: Berlin, Germany, 2005.
26. Trapletti, A.; Hornik, K. *Tseries: Time Series Analysis and Computational Finance*; R package version 0.10-47; 2019.
27. Hyndman, R.; Athanasopoulos, G.; Bergmeir, C.; Caceres, G.; Chhay, L.; O'Hara-Wild, M.; Petropoulos, F.; Razbash, S.; Wang, E.; Yasmeen, F. *Forecast: Forecasting Functions for Time Series and Linear Models*; R package version 8.13; University of Bath: Bath, UK, 2020.
28. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **2018**, *120*, 70–83. [CrossRef]
29. Hyndman, R.J.; Khandakar, Y. Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **2008**, *26*, 1–22.
30. Uddin, M.S.; Nasseef, M.T.; Mahmud, M.; AlArjani, A. Mathematical Modelling in Prediction of Novel CoronaVirus (COVID-19) Transmission Dynamics. *Preprints* **2020**. [CrossRef]
31. Pazos, F.; Felicioni, F.E. A control approach to the Covid-19 disease using a SEIHRD dynamical model. *medRxiv* **2020**. [CrossRef]

32. Alghamdi, T.; Elgazzar, K.; Bayoumi, M.; Sharaf, T.; Shah, S. Forecasting Traffic Congestion Using ARIMA Modeling. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 1227–1232. [CrossRef]
33. Yermal, L.; Balasubramanian, B. Application of Auto ARIMA Model for Forecasting Returns on Minute Wise Amalgamated Data in NSE. In Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, 14–16 December 2017; pp. 1–5. [CrossRef]
34. Box, G.E.P.; Cox, D.R. An Analysis of Transformations. *J. R. Stat. Soc. Ser. B Methodol.* **1964**, *26*, 211–243. [CrossRef]
35. Hyndman, R.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018.
36. Taieb, S.B.; Bontempi, G.; Atiya, A.; Sorjamaa, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *arXiv* **2011**, arXiv:1108.3259.
37. Vochozka, M.; Vrbka, J.; Suler, P. Bankruptcy or Success? The Effective Prediction of a Company's Financial Development Using LSTM. *Sustainability* **2020**, *12*, 7529. [CrossRef]