Predicting the evolution of the Covid-19 epidemic using a SEIR model

David Aleja *, Regino Criado and Miguel Romance

Laboratory of Mathematical Computing in Complex Networks and their Applications
Department of Applied Mathematics, Materials Science and Engineering and Electronic Technology, Universidad Rey Juan Carlos, C/ Tulipn s/n, Móstoles 28933 (Madrid), Spain
Data, Complex Networks and Cybersecurity Research Institute, Pl. Manuel Becerra 14, 28028 Madrid (Spain)

April 11th, 2020

---

1 Introduction and description of the methodology

The aim of this report is to describe a methodology to provide a daily prediction of the number of deaths in Spain in the Covid-19 coronavirus pandemic based on a SEIR model adjusted with the official data provided daily by the Instituto de Salud Carlos III [10]. In any case, the methodology is equally applicable to the temporal prediction of other parameters of the pandemic, such as the number of people infected, hospitalized, admitted to ICU or recovered, both in the whole of the national territory and broken down by Autonomous Communities, with only the adjustment of the different parameters of the model being adapted.

To establish the notation, we start with a time series $f_1, \cdots, f_T \in [0, +\infty)$, so that f_i is the number of deaths up to day i (included) of the pandemic, according to data provided daily by the Instituto de Salud Carlos III [10]. Based on this data, we want to predict the $\tilde{f}_{T+1}, \cdots, \tilde{f}_{T+k} \in [0, +\infty)$ corresponding to an estimate of the number of deaths in days $T+1$, $\cdots$, $T+k$ respectively, with fixed $k \in \mathbb{N}$. The following steps will be followed:

1. A mathematical model of epidemic spread based on differential equations of the SEIR type (a variant of the classical SIR model) is constructed to fit the $f_1, \cdots, f_T \in [0, +\infty)$ of the official number of deaths on each day of the pandemic.

2. The different parameters of the SEIR model (such as the basic reproduction rate and others) are optimized, so that the relative error between the real data $f_1, \cdots, f_T$ and those predicted by the model is as low as possible.

3. Using the optimized propagation model obtained in the previous step, the prediction accuracy of the model is adjusted by performing a cross validation with the differences of the real data and the forecasts given by the model. This adjustment is done through a functional optimization process.

Let us review each of these steps in detail in the particular case where $k = 1$, i.e. in the case where, from the official data of the number of deaths $f_1, \cdots, f_T$ we want to estimate the number of deaths up to the day $(T+1)$ which we will denote $\tilde{f}_{T+1}$.

The starting point is the SEIR model of differential equations [2, 5, 6], which is a variant of the classical SIR model proposed in [7] by W. O. Kermarck and A. G. McKendrick, in which, given a population of fixed size N in which an epidemic has been triggered that spreads by contagion in a time t (measured in days), individuals can be in four different states:

*E-mail addreses: david.aleja@urjc.es.

- Susceptible $S(t)$: number of individuals who can contract the disease.
- Exposed $E(t)$: number of individuals who have been infected but cannot infect.
- Infected $I(t)$: number of infected (each of which can infect β individuals).
- Recovered $R(t)$: number of individuals who have overcome the disease or have died

Considering t_I the time that an individual is in the infected phase, that is, the recovery rate is $\gamma := 1/t_I$, t_0 the time of beginning of the study and $t_0 + T$ the final time, we have the SEIR model:

$$\begin{cases} S'(t) = -\beta I(t)S(t)/N, & t_0 < t \leq t_0 + T, \\ E'(t) = \beta I(t)S(t)/N - \sigma E(t), \\ I'(t) = \sigma E(t) - \gamma I(t), \\ R'(t) = \gamma I(t), \\ S(t_0) = S_0, \quad E(t_0) = E_0, \quad I(t_0) = I_0, \quad R(t_0) = R_0, \end{cases} \tag{1.1}$$

where the parameter $\sigma := 1/t_E$ is the incubation rate of the disease and S_0, E_0, I_0 and R_0 are the initial data at time t_0. The model considered is similar to the one used in [4] also to model the Covid-19 epidemic, but with a different adjustment to model the influence of containment measures. A relevant parameter in the study of epidemic propagation is the*basic reproduction rate* [2, 6], denoted by R_0, which represents the number of new infections produced by a single infected person throughout his stage of infection, i.e,

$$R_0 = \frac{\beta}{\gamma} = \beta t_I.$$

It is necessary to modify this classic model to take into account the effects of the containment decreed by the Spanish Government on 15 March 2020, which means that the containment measures (protection and isolation) taken mean that the parameter β can change over time. To this end, a new parameter $\alpha \in [0, 1]$ is introduced to represent the influence of containment on the basic reproduction rate (see [1]), so that

$$R_0(\alpha) := \beta_\alpha t_I \quad \text{where} \quad \beta_\alpha := \alpha\beta.$$

Once the type of mathematical model to be considered has been established, [1]shows how, given a time series $f_1, \cdots, f_T \in [0, +\infty)$ corresponding to the (actual) number of cumulative deaths each day of the pandemic, the different parameters can be optimized to obtain a SEIR model like the previous one so that the relative error between the actual data $f_1, \cdots, f_T$ and those predicted by the model for the same days is as low as possible. In this way, from the time series $\sigma = (f_1, \cdots, f_T) \subseteq [0, +\infty)$, a function $\phi_\sigma : \mathbb{N} \longrightarrow \mathbb{R}^4$ has been constructed in such a way that

$$\phi_\sigma(t) = (S(t), E(t), I(t), R(t)),$$

so that if the number of deaths on day t is $F(t) := \tau \cdot R(t)$, where τ is the proportion of deaths among those recovered, then the fourth component of the function $\phi(\cdot)$ minimizes the error between $f_1, \cdots, f_T$ and $F(1), \cdots, F(T)$.

To estimate the reliability of the function $\phi(\cdot)$ as a predictor of the value of the number of deaths at the instant $T + 1$ (i.e. $F(T + 1)$), a cross validation procedure is performed as follows: We start from the time series $f_1, \cdots, f_T \in [0, +\infty)$ which corresponds to the (actual) number of cumulative deaths each day of the pandemic

- For every $1 \leq i \leq T - 1$ you get a SEIR model optimized ϕ_i for the number of deaths given by the time sub-series $\sigma_i = (f_1, \cdots, f_i)$, i.e. the fourth component of the function $\phi(\cdot)$ minimizes the error between $f_1, \cdots, f_i$ and $F_i(1), \cdots, F_i(i)$. This optimization procedure is described in [1].
- For each $t_* \leq i \leq T - 1$, with $t_* \geq 1$, we compare $\hat{x}_i = F_i(i + 1) - F_i(i)$ and the official (actual) number of deaths on day $i + 1$, i.e., $x_i = f_{i+1} - f_i$.

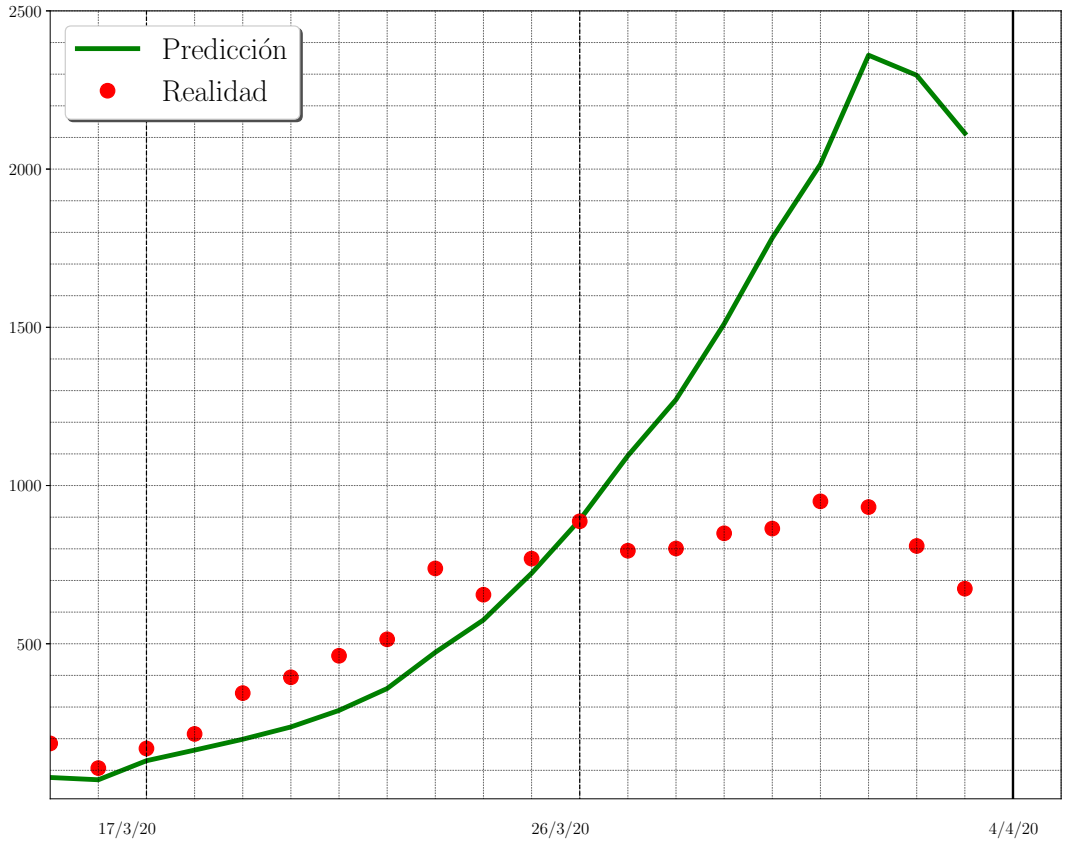

Figure 1: Comparison between actual data on deaths and estimates given by the raw SEIR model.

Figure 1 shows the expected increase for the next day (green) and what actually happened (red). Specifically, in this case, we have chosen $t_* \equiv$ March 15th and $T \equiv$ April 4th, 2020,, and the curves indicate the following:

- Red $\rightarrow i$ front $x_i = f_{i+k} - f_i$.
- Green $\rightarrow i$ front $\hat{x}_i = F_i(i+1) - F_i(i)$.

In view of Figure 1 it appears that the raw $\phi(\cdot)$ function was valid as a predictor of the value of the number of deaths on the following day until March 26th, 2020, although from that moment on the raw model developed from the SEIR model does not conform to the official data, probably because the model does not correctly take into account the effects of the confinement decreed by the Spanish government on March 15th, 2020, and the tightening of the measures taken on March 27th, 2020. At this point, in addition to abandoning the SEIR modeling work, there are several alternatives for improving the results, including the following:

- Modify the original SEIR model to more efficiently take into account the containment measures of March 15th, 2020, and the tightening of the measures taken on March 27th, 2020. This would mean having a more precise model, although it would increase the number of parameters to be optimized and therefore the calculation times.

- Adjust the data given by the optimized SEIR model, so that the values given by the predictor are adjusted to the real data.

If we follow the second proposal, in Figure 1 we see that the predictor in the considered time window has two different behaviors:

(i) From 03/15/2020 to (approximately) 03/26/2020: The predictor is essentially consistent with the actual data. In any case, to adjust the predicted values a little more to the real data, we can transform them using a function of the form

$$g_1(x) = x + c, \tag{1.2}$$

with c an independent constant (such as $c = 100$), so if the optimized SEIR model gives us an estimate $\hat{x}_i$, then we can rescale it giving as an estimate $\tilde{x}_i = g_1(\hat{x}_i)$.

(ii) From 03/26/2020: The predictor is considerably separated from the actual data and the separation between them grows as we move away from the day 03/26/2020. If we denote this cut-off point as $t_o = 26/03/2020$, the optimized SEIR model gives us an estimate $\hat{x}_i$ para $t_o < i \leq T - 1$ such that we can adjust, for example, as

$$\tilde{x}_i = g_2(\hat{x}_i) = \frac{\hat{x}_i}{\sqrt{i - t_o + 1}}. \tag{1.3}$$

Figure 2 shows the actual data (in red), the raw prediction given by the optimized SEIR model (in green) and the prediction given by the rescaled model (in blue, i vs $\tilde{x}_i$) using the $g_1(\cdot)$ functions until 03/26/2020 and $g_2(\cdot)$ from 03/27/2020, showing a considerable improvement in the prediction accuracy. Therefore, we should choose

$$\tilde{f}_{T+1} = f_T + \tilde{x}_T = f_T + \frac{\hat{x}_T}{\sqrt{T - t_o + 1}}$$

as the prediction for $T + 1$ day.

In this case we see that, if we add up all the deaths from 16/03/2020 to 04/04/2020, both those that appear in the official data and those predicted by the optimized SEIR model and by the rescaled optimized SEIR model, then, as shown in Table 2, with the raw SEIR model an accumulated error of almost 54% is committed, while if the rescaled SEIR model is used the accumulated error is reduced to 2.02%.

Table 1: Comparison of the cumulative number of deaths in both the official data and the models proposed from 03/16/2020 to 04/04/2020.

	N° of Deceased	**Diff. from actual data**	**% error**
Actual data	12112	0	0%
Model SEIR raw	18631	6519	53.82%
Model SEIR rescaled	11867	-245	2.02%

Obviously in this introduction we have illustrated in detail the methodology proposed for the prediction of the number of deaths on day $T + 1$ from the official data on all the previous days, but this procedure can be optimized and extended both to predictions for a longer horizon and to other parameters (number of infected, hospitalized, number of people hospitalized in Intensive Care Units or recovered) or other environments (such as different Autonomous Communities), for which the following observations should be taken into account:

- Both the $g_1(\cdot)$ and $g_2(\cdot)$ functions and the t_o point can (and should) be optimized, proposing as families of functions to be optimized

$$f_1(x, t) = x + c, \qquad f_2(x, t) = \frac{x}{(t - t_o + a)^b},$$

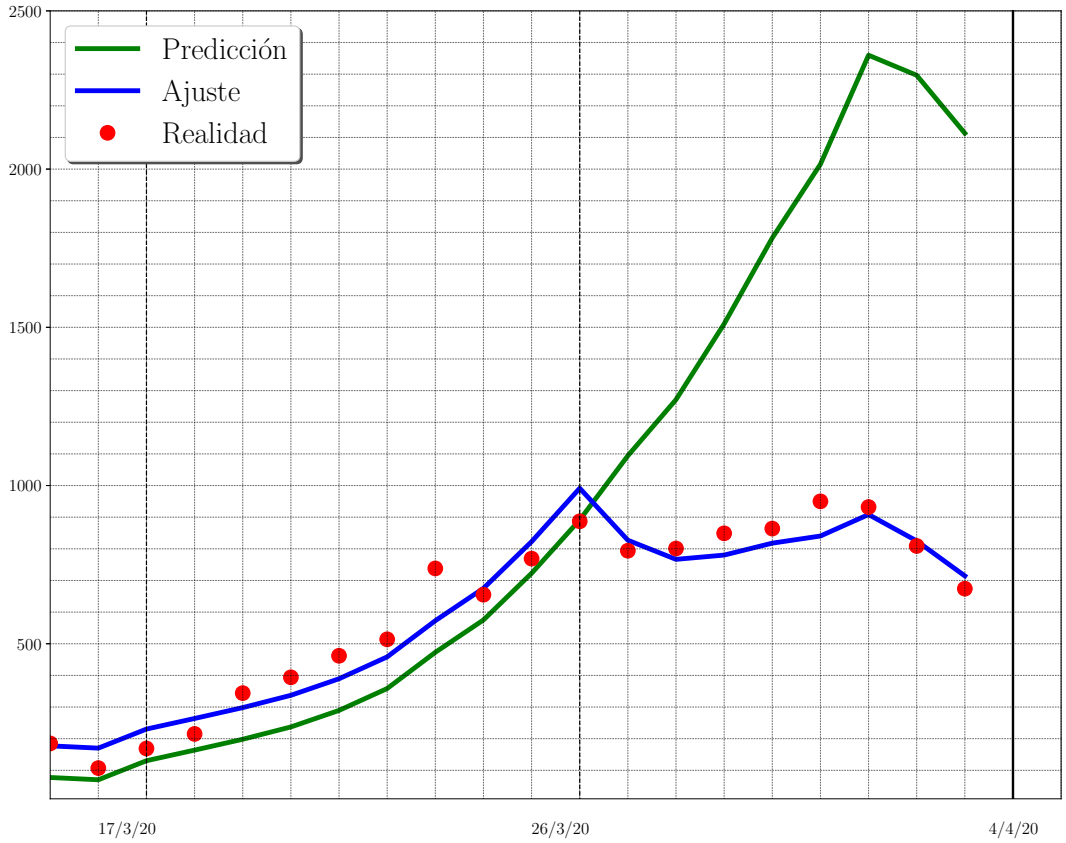

Figure 2: Comparison between actual death data, estimates given by the raw SEIR model and predictions given by the SEIR model rescaled using $g_1(\cdot)$ and $g_2(\cdot)$.

being the optimization parameters $a, b, c \in \mathbb{R}$ and $t_* \leq t_o \leq T-1$.

- In the case that we want to make estimates in higher order horizons (i.e. predictions for $T+k$, with any $k \in \mathbb{N}$), we have to take into account that the optimization of the $g_1(\cdot)$ and $g_2(\cdot)$ must be done for each of the obtained k, , giving different values, as we will see in Section 2.

2 Optimal Predictor Setting

Given $k \in \mathbb{N}$, , our goal is to obtain a prediction of deaths in time $T+k$ by the method described in Section1. First, for $1 \leq i \leq T$, we consider $\hat{f}_i$ and $\hat{f}_{i+k}$ the estimates given by the SEIR model optimized with f_1, f_2, ..., f_i and we calculate their difference

$$\hat{x}_i = \hat{f}_{i+k} - \hat{f}_i. \tag{2.1}$$

From an instant $t_* \in [1, T]$ (fixed), we define

$$\tilde{x}_i = g_1(\hat{x}_i) = \hat{x}_i + c \quad \text{para} \;\; t_* \leq i \leq t_o$$

and

$$\tilde{x}_i = g_2(\hat{x}_i) = \frac{\hat{x}_i}{(i - t_o + a)^b} \quad \text{para} \;\; t_o < i \leq T$$

for certain parameters $a, b, c \in \mathbb{R}$ and $t_* \leq t_o \leq T-k$, we would like to find these parameters that minimize

$$Error = \sum_{i=t_*}^{T-k} \left| \frac{x_i - \tilde{x}_i}{x_i} \right|, \quad \text{con } x_i = f_{i+k} - f_i,$$

in order to determine $\tilde{x}_T$ and, therefore, to obtain the prediction of deaths in an instant $T+k$ as follows:

$$\tilde{f}_{T+k} = f_T + \tilde{x}_T.$$

Note that these parameters can change according to the k value and the moment T .

Next we illustrate an example considering

$$k \in \{1, 2, ..., 7\}, \qquad i = 1 \rightarrow \text{March 8th} \qquad t_* \rightarrow \text{March 15th} \qquad \text{y} \qquad T \rightarrow \text{April 9th},$$

providing for each value of k the parameters $a, b, c \in \mathbb{R}$ and $t_* \leq t_o \leq T-k$, as well as giving a prediction for the 7 days following April 9th. Assuming that the parameters vary as

- $a \in [0, 1]$ with step 0.01,
- $b \in [0, 1]$ with step 0.01,
- $c \in [0, 2000]$ with step 10,
- t_o varying between March 23 and March (both included).

The results obtained are given in Table 2.

Table 2: The prediction and adjustment parameters for the days between 04/10/2020 and 04/162020.

k	a	b	c	t_o	$\tilde{f}_{T+k}$
1	0.01	0.45	50	25/3/20	16290
2	0.38	0.46	150	25/3/20	16733
3	0.47	0.53	300	25/3/20	16981
4	0.27	0.56	550	24/3/20	17248
5	0.32	0.62	730	24/3/20	17377
6	0.18	0.64	1080	23/3/20	17586
7	0.19	0.7	1390	23/3/20	17616

Finally, in Figure 3 3 we show only the first adjustment made for $k = 1$ so that it can be seen much better and the rest in Figure 4. Keep in mind that, in each of them, the three curves indicate the following:

- Red $\rightarrow i$ vs $x_i = f_{i+k} - f_i$.
- Green $\rightarrow i$ vs $\hat{x}_i = \hat{f}_{i+k} - \hat{f}_i$.
- Blue $\rightarrow i$ vs $\tilde{x}_i$.

References

[1] Aleja, D., Criado, R., & Romance, M., "SEIR Model for Covid-19 Coronavirus" (2020).

[2] Diekmann, O., Heesterbeek, H., & Britton, T. ,"Mathematical tools for understanding infectious disease dynamics" (Vol. 7). Princeton University Press (2012).

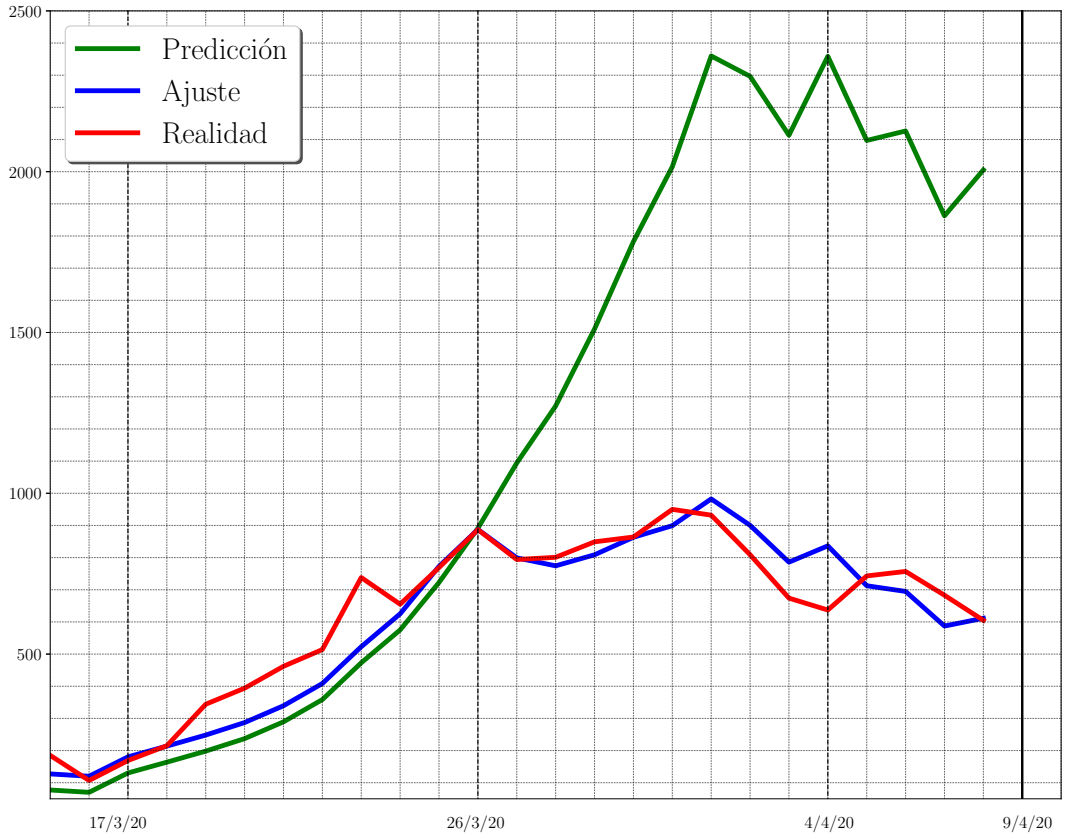

Figure 3: Optimal Predictor Adjustment for $k = 1$.

[3] Gandhi, K.R.R., & Casella, F. , "Non-Pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality". Available at SSRN 3560688 (2020).

[4] Gutiérrez, J.M. & Varona, J.L., "Análisis de la posible evolución de la epidemia de coronavirus COVID-19 por medio de un modelo SEIR", disponible en `https://belenus.unirioja.es/jvarona/coronavirus/SEIR-coronavirus.pdf` (2020).

[5] Hethcote, H. W., "The mathematics of infectious diseases", SIAM Review, 42(4), 599-653 (2000).

[6] Keeling, M. J., & Rohani, P. , "Modeling infectious diseases in humans and animals". Princeton University Press (2011).

[7] Kermack, W. O., & McKendrick, A. G., "A contribution to the mathematical theory of epidemics". Proceedings of the Royal Society of London. Series A, 115(772), 700-721 (1927).

[8] Kucharski, A. J., Russell, et. al.,"Early dynamics of transmission and control of COVID-19: a mathematical modelling study". The Lancet Infectious Diseases (2020).

[9] Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklov, J., "The reproductive number of COVID-19 is higher compared to SARS coronavirus". Journal of travel medicine (2020).

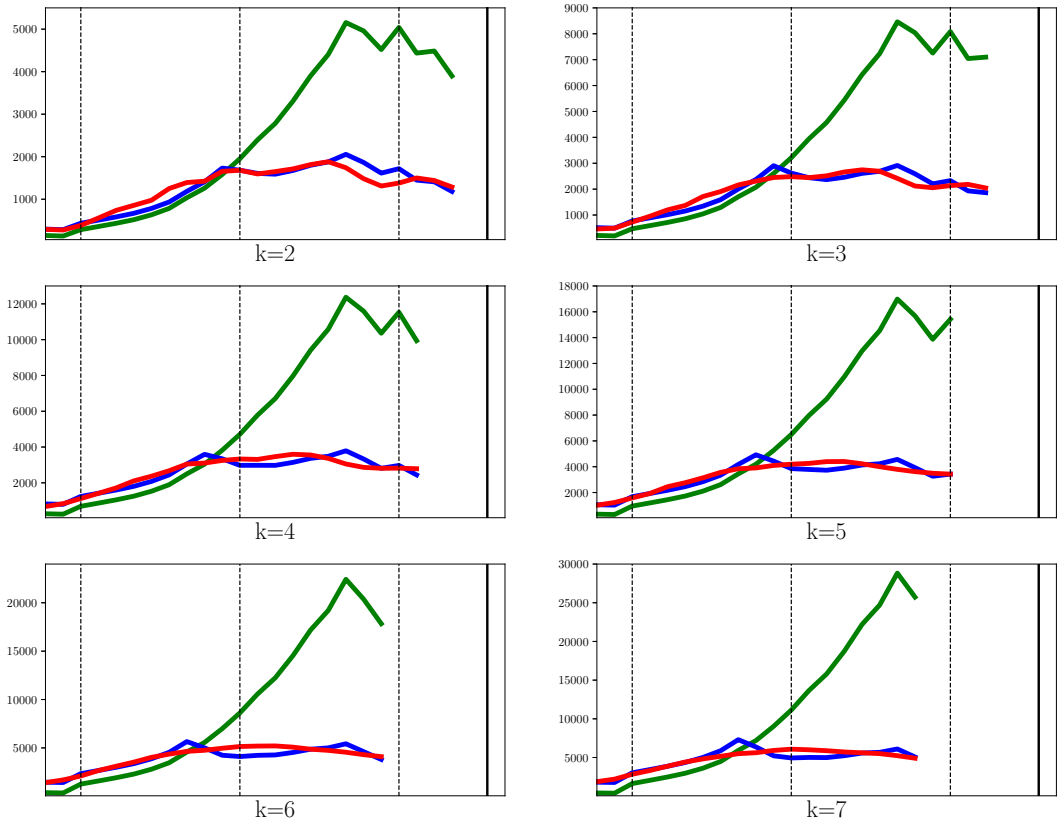

Figure 4: Optimal Predictor Adjustment for $k = 2, 3, ..., 7$.

[10] Official data (in spanish) from the Spanish Government on the spread of Covid-19 provided by the Instituto de Salud Carlos III (ISCIII), updated daily at `https://covid19.isciii.es/resources/serie_historica_acumulados.csv`.