A phenomenological model for COVID-19 data taking into account neighboring-provinces effect and random noise

Julia Calatayud [a], Marc Jornet [b], Jorge Mateu [a]

[a] Department of Mathematics,
Universitat Jaume I,
12071 Castellón, Spain.
email: calatayj@uji.es; mateu@uji.es

[b] Department of Mathematics,
Universitat de València,
46100 Burjassot, Spain.
email: marc.jornet@uv.es

Abstract. We model the incidence of the COVID-19 disease during the first wave of the epidemic in Castilla-Leon (Spain). Within-province dynamics may be governed by a generalized logistic map, but this lacks of spatial structure. To couple the provinces, we relate the daily new infections through a density-independent parameter that entails positive spatial correlation. Pointwise values of the input parameters are fitted by an optimization procedure. To accommodate the significant variability in the daily data, with abruptly increasing and decreasing magnitudes, a random noise is incorporated into the model, whose parameters are calibrated by maximum likelihood estimation. The calculated paths of the stochastic response and the probabilistic regions are in good agreement with the data.

Keywords: COVID-19 infections; Generalized logistic differential equation; Parameter calibration; Spatial correlation; Stochastic modeling

AMS Classification 2010: 92-10; 34A55; 34F05; 62M30

1. Introduction

Phenomenological (or statistical) models are often useful to reproduce and forecast the course of an epidemic, when the insight is limited, treatments and interventions rapidly change, and data are scarce, uncertain and vary abruptly [5, 15]. In these circumstances, some mechanistic models, based on specific laws of transmission, may not work well.

The main example of phenomenological model is the logistic growth curve. Devised by P.F. Verhulst in 1838 as an extension of the Malthusian exponential model, it has become an essential tool in biology, ecology and epidemiology for the fit of growth phenomena. Examples of logistic epidemic modeling include Ebola [6] and COVID-19 [21]. Generalizations of the logistic equation, to capture other growth profiles, have been suggested and applied to tumor growth [4, 13, 17, 20] and to diseases such as SARS [7, 9], dengue fever [10], influenza H1N1 [8], Zika [5], Ebola [15], and COVID-19 [1, 12, 14, 22].

Though phenomenological, these models may be extended to incorporate some spatial effects. Extending logistic models to heterogeneous space may be done by including logistic growth as the reaction term in a reaction-diffusion partial differential equation model, or by modeling space as a collection of discrete patches, among which populations can disperse [23]. The second case yields a coupled system of ordinary differential equations, which is simpler than a mechanistic compartmental system.

Phenomenological models may also incorporate stochastic effects, to deal with the uncertainty associated to data collection and the phenomenon itself [19]. For the COVID-19 disease, some examples include a frequentist approach for the derivative of the logistic map with Gaussian

error [18], a Bayesian approach for the Gompertz curve [3], and a phenomenological model based on the spatio-temporal evolution of a Gaussian probability density function [2].

In this paper, the aim is to model COVID-19 data phenomenologically, taking into account spatial and stochastic effects. We base on daily new infections through the derivative of a generalized logistic map, by generalizing somehow the simple logistic map from [18]. By coupling, we include a spatial structure in the phenomenological model of differential equations, with a positive correlation of cases between nearby regions. To our knowledge, such a model has not been utilized in mathematical epidemiology. Finally, we also incorporate a random noise into the deterministic solution, in order to capture the highly irregular dynamics of the data. Our case study is the Spanish autonomous community of Castilla-Leon, divided into 9 regions called provinces. It is the largest community in Spain by area, it is located in the northwest of Spain, and it has a population of around 2.5 million. In Figure 1, we show the location of Castilla-Leon among the autonomous communities of Spain (left map), as well as the nine provinces of Castilla-Leon (right map). In Table 1, codes for the nine provinces are shown, as well as their populations (year 2019, approximated to the nearest thousands). Our aim is to model the first wave of the COVID-19 epidemic, from 1st March 2020 to 22nd June 2020 (114 days), with recorded data on daily new infections for the provinces. The cases have been retrieved from the open data portal of Castilla-Leon: `https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html`.

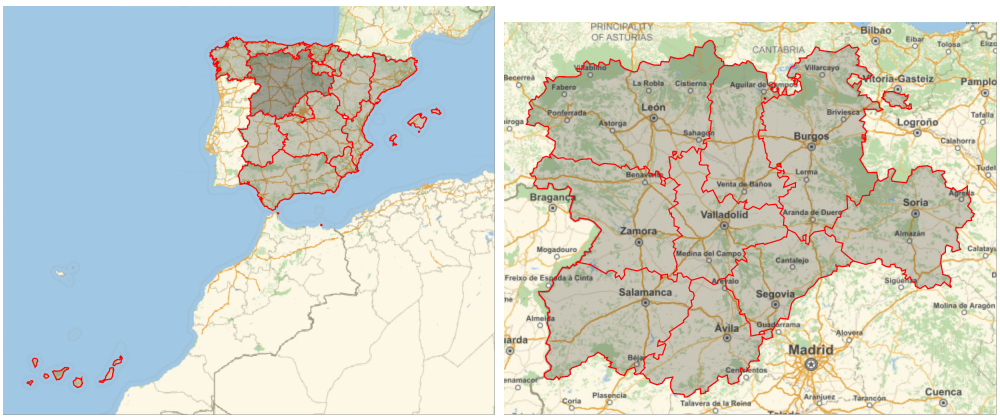

FIGURE 1. Location of Castilla-Leon among the autonomous communities of Spain (left map), and the nine provinces of Castilla-Leon (right map). Source: Mathematica®, built-in function *GeoGraphics*.

Province	Index	Inhabitants
Leon	1	462 000
Palencia	2	160 000
Burgos	3	355 000
Soria	4	89 000
Segovia	5	154 000
Avila	6	159 000
Salamanca	7	332 000
Zamora	8	173 000
Valladolid	9	520 000

TABLE 1. The nine provinces of Castilla-Leon, their codes and populations.

2. Deterministic model

Given an index $i \in \{1, \ldots, 9\}$ that identifies the province, let $P_i \in [0, 1]$ be the proportion of cumulative infections and $p_i \in [0, 1]$ be the proportion of new infections, scaled from the total population N_i. These proportions depend on the day t: $P_i \equiv P_i(t)$ and $p_i \equiv p_i(t)$, $t \geq 0$. As suggested in [18], the relation $p_i = P_i'$ is assumed (the prime denotes the derivative). The key idea is that the differential equation model is set for $P_1, \ldots, P_9$, while the parameter calibration is conducted for $p_1, \ldots, p_9$ (scaled daily new infections) to avoid serial correlation in errors for cumulative cases and biased parameters.

Within-province dynamics may be governed by a generalized logistic differential equation model of the form

$$P_i' = a_i P_i \left(1 - \left(\frac{P_i}{K_i}\right)^{b_i}\right), \quad i = 1, \ldots, 9. \tag{2.1}$$

The parameters are the growth rate $a_i > 0$, the local asymptotic equilibrium $K_i > 0$, and the flexibility coefficient $b_i > 0$, which are assumed to be time invariant. The saturation effect from K_i implicitly captures public health interventions, without complex mechanistic assumptions about the transmission process. The parameter b_i allows for more flexible S-shaped growth profiles than the classical logistic formulation by Verhulst. It reflects the asymmetry of the curve of daily new infections with respect to the peak (new infections rise quicker than decrease). When $b_i = 1$ and $b_i \to 0$, the logistic and the Gompertz curves are retrieved, respectively. The reader may consult an extensive list of references for the generalized and classical logistic equations, with a variety of applications, in the Introduction section.

Spatial structure, in which individuals interact more intensely with neighbors, may be incorporated as follows. Given two provinces i and j, we write $i \sim j$ whenever they are adjacent. The complete phenomenological model is then the following:

$$P_i' = a_i P_i \left(1 - \left(\frac{P_i}{K_i}\right)^{b_i}\right) + D \sum_{\substack{j \neq i \\ j \sim i}} P_j', \quad i = 1, \ldots, 9. \tag{2.2}$$

To couple the provinces, we have related the daily new infections through a density-independent parameter $D > 0$ that entails positive spatial correlation: when some P_j increases rapidly at t, for $j \sim i$, then P_i augments quicker too. Again, no mechanistic assumptions are made. To construct (2.2), some ideas were taken from the theory of disperse populations in discrete patches [23]. After isolating $P_1', \ldots, P_9'$ in (2.2) symbolically, each P_i' is written as a linear combination of $a_1 P_1(1 - (P_1/K_1)^{b_1}), \ldots, a_9 P_9(1 - (P_9/K_9)^{b_9})$; this is somehow similar to the coupled model investigated a few decades ago in [11] from the system dynamics viewpoint. In contrast to the local model (2.1), which belongs to the class of Bernoulli ordinary differential equations, the coupled system (2.2) does not have a closed-form solution; numerical methods are required for its resolution.

The next section details the calibration of the 28 parameters in the coupled generalized logistic model (2.2).

2.1. Calibration of the deterministic model.

The initial conditions $P_i(0)$ in (2.2) are fixed as the initial data; if any of them is 0, then $P_i(0)$ is set as $1/N_i$ (one infected individual). The 28 parameters are fitted by least-squares optimization for $\{p_i(t)\}_{i=1,\ldots,9,\, t=0,\ldots,113}$ (scaled daily new infections), as recommended by [18]:

$$\min_{\{a_j, b_j, K_j\}_{j=1}^{9},\, D} \sum_{i=1}^{9} \sum_{t=0}^{113} \left(p_i(t) - d_i(t)\right)^2. \tag{2.3}$$

Here $d_i(t)$ denotes the observed datum. This minimum gives a measure of how good the fit is.

For computations, Mathematica® has been used. The system of ordinary differential equations (2.2) is parametrically solved with the built-in function *ParametricNDSolveValue*, with the option *Method* $\to$ {*"EquationSimplification"* $\to$ *"Residual"*} to isolate the derivatives. The minimization is carried out with the routine *FindMinimum*. This function was executed with 800 iterations, for 7.5 hours, and the algorithm converged. The estimates of the parameters that minimize (2.3) are presented in Table 2. The powers b_i are closer to 0 than to 1, so the model for each province is more similar to a Gompertz curve than to a classical logistic curve. The coefficients K_i provide the maximum level of infection at the first wave under no neighboring-provinces effect; for example, in Leon it would have been 1.95% and in Soria 4.39%. The value of (2.3) is 0.0000498. It is observed that $D > 0$, so there is indeed a spatial effect. In fact, if $D = 0$ and local generalized logistic curves are fitted for each province, then the value of (2.3) becomes 0.0000523, that is, 5.06% greater. When $D = 0$, the provinces for which the least-squares error increases are Zamora, Palencia, Avila and Valladolid, in decreasing order of magnitude; this means that these four provinces were the most susceptible to their neighbors during the first wave of the epidemic. Of course, these assertions on D are conditional on the validity of the generalized logistic model (2.1) to describe within-province dynamics, so that any deviation of it is due to spatial factors; in general, the validity of the model is a necessary assumption when performing sensitivity analysis.

Parameter	Estimate	Parameter	Estimate
D	0.130	b_5	0.0979
a_1	0.939	b_6	0.0421
a_2	0.963	b_7	0.0883
a_3	0.922	b_8	0.0262
a_4	1.00	b_9	0.00606
a_5	0.900	K_1	0.0195
a_6	0.874	K_2	0.0243
a_7	0.927	K_3	0.0291
a_8	0.362	K_4	0.0439
a_9	0.411	K_5	0.0381
b_1	0.0724	K_6	0.0347
b_2	0.0330	K_7	0.0284
b_3	0.0213	K_8	0.0347
b_4	0.0492	K_9	0.0429

TABLE 2. Parameter estimates of (2.2) that minimize (2.3).

In Figure 2, the fit of $q_i(t) = p_i(t)N_i$ to the data is plotted. The deterministic model renders a smooth, averaged fit. However, the abrupt variation in consecutive days is not captured, and this is the reason of incorporating stochasticity in the following section.

3. A stochastic model

In order to capture the highly irregular dynamics of the data, stochasticity is incorporated into the coupled generalized logistic model (2.2) [19]. By inspecting the deterministic fit, a Gaussian white noise error is added to the scaled number of daily new cases $p_i(t)$:

$$z_i(t) = p_i(t) + (p_i(t))^{\lambda_i}\,\eta_i(t). \tag{3.1}$$

Here the power $\lambda_i > 0$ is independent of t, the noise $\eta_i(t) \sim \text{Normal}(0, \sigma_i^2)$ is an uncorrelated process, with variance $\sigma_i^2 > 0$ independent of t, and $z_i(t)$ is the new stochastic response. This new response is highly irregular: it is not jointly measurable nor right/left-continuous on any interval. The term $(p_i(t))^{\lambda_i}$ controls the dispersion of the random error; the higher the value of $p_i(t)$, the

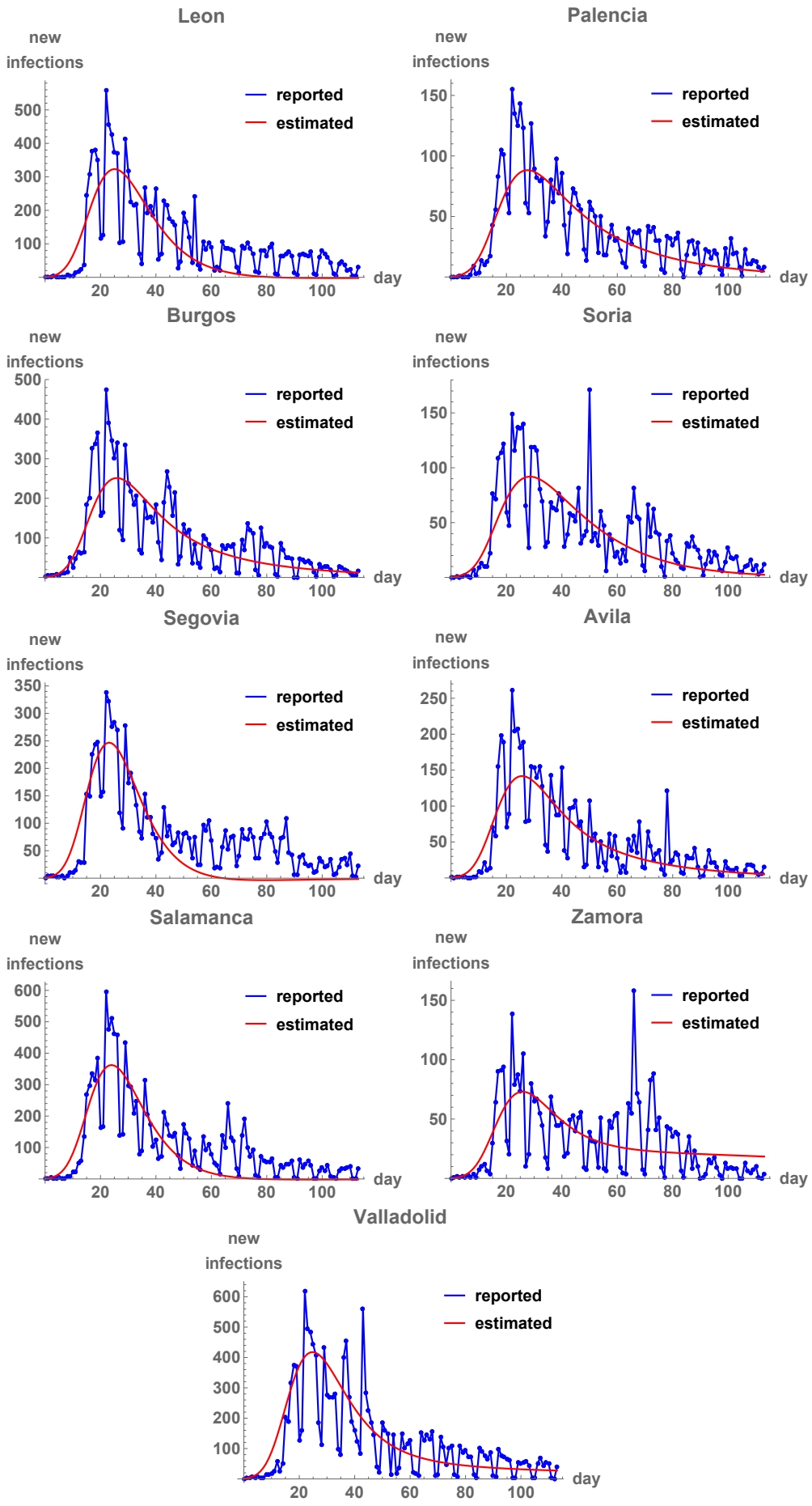

FIGURE 2. Fit of $q_i(t) = p_i(t)N_i$ to the number of daily new cases, by province $i = 1, \ldots, 9$.

larger is the variability the error exhibits. The mean of (3.1) is the output of the deterministic model (2.2).

3.1. **Calibration of the stochastic model.** Given the estimates of D, a_i, b_i and K_i from Table 2, both parameters λ_i and σ_i of the stochastic model (3.1) are calibrated for each province $i \in \{1, \dots, 9\}$ by means of maximum likelihood [16]. The likelihood of the observed time series $d_i = \{d_i(t) : t = 0, \dots, 113\}$ is given by

$$\mathcal{L}(d_i|\lambda_i, \sigma_i) = \prod_{t=0}^{113} \pi_{\text{Normal}(p_i(t),(p_i(t))^{2\lambda_i}\sigma_i^2)}(d_i(t)),$$

where π denotes the probability density function. By maximizing it,

$$\max_{\lambda_i,\sigma_i} \prod_{t=0}^{113} \pi_{\text{Normal}(p_i(t),(p_i(t))^{2\lambda_i}\sigma_i^2)}(d_i(t)),$$

the infinitesimal probability around d_i,

$$\Pr\{\text{datum} \in [d_i(t), d_i(t) + \delta d_i(t)],\, t = 0, \dots, 113\} = \mathcal{L}(d_i|\lambda_i, \sigma_i)\, \delta d_i(0) \cdots \delta d_i(113),$$

is also maximized. After applying $-\log$, the maximization problem is more conveniently given as

$$\min_{\lambda_i,\sigma_i} \left[\lambda_i \sum_{t=0}^{113} \log p_i(t) + 114 \log \sigma_i + \sum_{t=0}^{113} \frac{(\text{dato}_i(t) - p_i(t))^2}{2\,(p_i(t))^{2\lambda_i}\,\sigma_i^2} \right].$$

We use Mathematica® with the built-in instruction *NMinimize*, in the region $\lambda_i \in (0, 1)$ and $\sigma_i \in (0, 1)$. If some $p_i(t)$ is a small negative number or zero, it is changed to 10^{-8}. In Table 3, the optimal values of λ_i and σ_i are reported. These are then plugged in (3.1).

Province i	1	2	3	4	5
λ_i	0.0701	0.549	0.757	0.539	0.00168
σ_i	0.000365	0.0127	0.0871	0.0216	0.000345
Province i	6	7	8	9	
λ_i	0.645	0.0863	0.837	0.764	
σ_i	0.0399	0.000586	0.223	0.0943	

TABLE 3. Optimal values of λ_i and σ_i for the stochastic model.

In Figure 3, the fit of the stochastic model is illustrated. We have taken the stochastic process $\max\{z_i(t)N_i, 0\}$, whose statistics are determined with Monte Carlo simulation. We show the mean (which is approximately equal to the deterministic fit) and probabilistic intervals, as well as an example of a randomly realizable path. One appreciates the similarity in pattern of the realizable path and the data, which justifies the need of stochasticity.

4. Conclusions

As shown in this paper, a phenomenological model may be useful to capture faithfully the dynamics of an epidemic. In the case study of Castilla-Leon (Spain) and the first wave of COVID-19, we have used a coupled system of generalized logistic differential equations. The coupling comes from the spatial effect due to neighboring provinces of Castilla-Leon. The calibration of the 28 parameters is based on the daily new infections through the derivative of the system. The process is computationally intensive, but optimal parameters can be obtained at the end. The model yields a smooth, averaged curve that follows the pattern of the data. However, stochasticity is needed to obtain realizable paths that resemble the abrupt changes of the data. It is incorporated into the model via a random noise error, whose parameters are determined by maximum likelihood estimation.

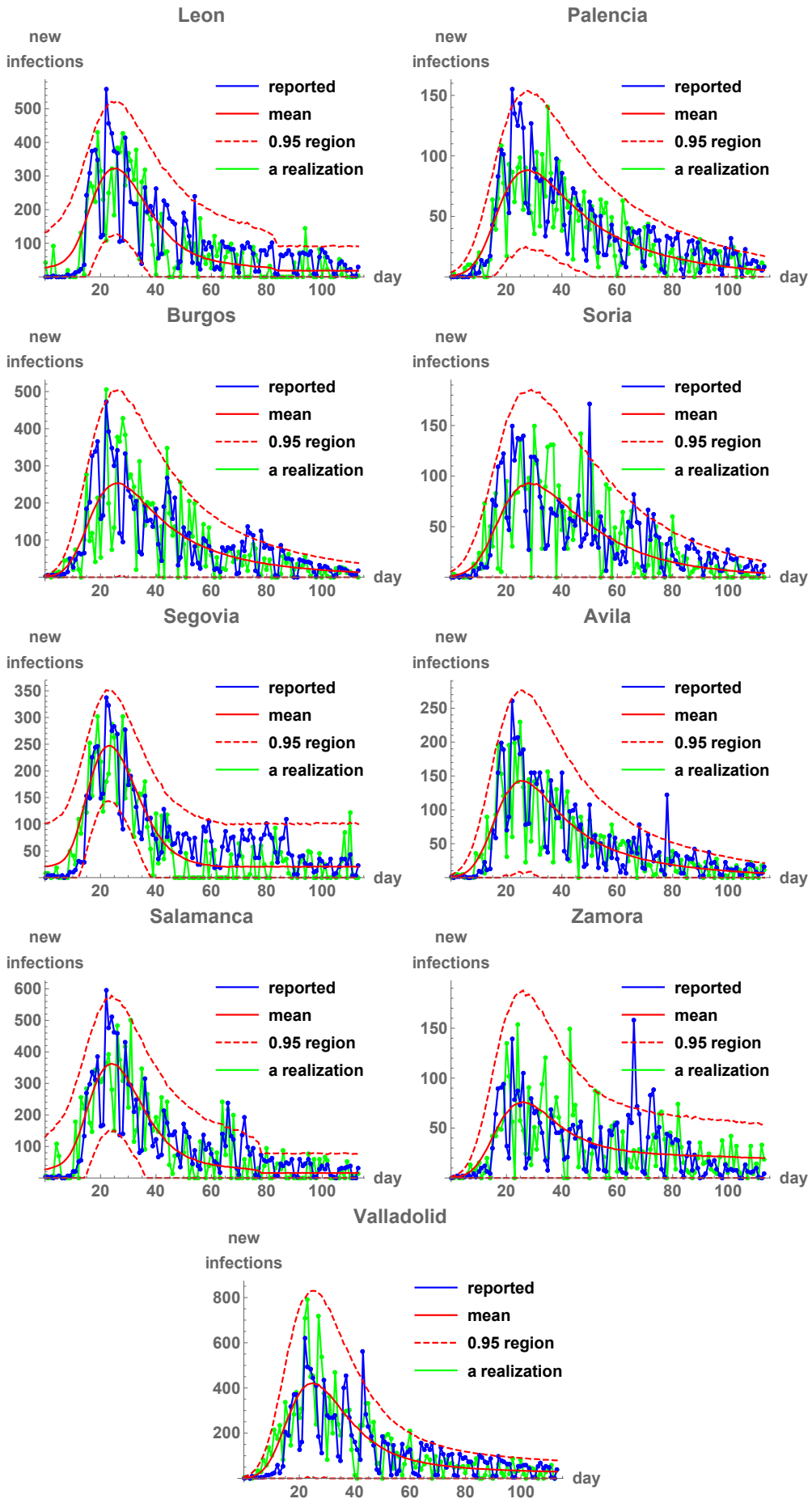

FIGURE 3. Fit of the stochastic model, for each province $i = 1, \ldots, 9$. The mean and probabilistic intervals are shown, as well as an example of a randomly realizable path.

The main limitation when fitting models of coupled differential equations rigorously is the time to execute the optimization procedure. Once the optimal parameters are available, it is simple to incorporate a random noise and to estimate the parameters of dispersion. Several extensions of the present paper may be devised, but constrained to optimizing parameters of coupled differential equations with higher efficiency. This is not easy, due to the well-known curse of dimensionality. Future works could be based on dealing with several waves of infection at once (through a sum of logistic responses), on estimating the deterministic and random error parameters at once (through an intensive likelihood maximization procedure), on dividing the space into finer subregions, or on adding mechanistic processes of infection. Nonetheless, it is important to emphasize that, sometimes, rather than augmenting the complexity of a simple but satisfactory model with mechanistic considerations, it might be better to treat the error as random, and to apply a stochastic fit.

Funding

Julia Calatayud has been supported by a postdoctoral contract from Universitat Jaume I, Spain (Acció 3.2 del Pla de Promoció de la Investigació de la Universitat Jaume I per a l'any 2021). Jorge Mateu has been supported by the grant PID2019-107392RB-I00 from Spanish Ministry of Science and the grant AICO/2019/198 from Generalitat Valenciana.

Conflict of Interest Statement

The authors declare that there is no conflict of interests regarding the publication of this article.

Data Availability Statement

The infection cases have been retrieved from the open data portal of Castilla-Leon: `https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html`

References

[1] Aviv-Sharon, E. and A. Aharoni (2020), Generalized logistic growth modeling of the COVID-19 pandemic in Asia, *Infectious Disease Modelling*, *5*, 502–509.

[2] Benítez, D., G. Montero, E. Rodríguez, D. Greiner, A. Oliver, L. González and R. Montenegro (2020), A phenomenological epidemic model based on the spatio-temporal evolution of a Gaussian probability density function, *Mathematics*, *8*(11), 2000.

[3] Berihuete, A., M. Sánchez-Sánchez and A. Suárez-Llorens (2021), A Bayesian model of COVID-19 cases based on the Gompertz curve, *Mathematics*, *9*(3), 228.

[4] Birch, C.P. (1999), A new generalized logistic sigmoid growth equation compared with the Richards growth equation, *Annals of Botany*, 83(6), 713–723.

[5] Chowell, G., D. Hincapie-Palacio, J.F. Ospina, B. Pell, A. Tariq, S. Dahal, S.M. Moghadas, A. Smirnova, L. Simonsen and C. Viboud (2016), Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics, PLoS Currents, *8*.

[6] Chowell, G., L. Simonsen, C. Viboud and Y. Kuang (2014), Is west Africa approaching a catastrophic phase or is the Ebola epidemic slowing down? Different models yield different answers for Liberia, PLoS Currents, *6*.

[7] Hsieh, Y.H. (2009), *Richards model: a simple procedure for real-time prediction of outbreak severity*, in: Z. Ma, Y. Zhou and J. Wu (eds.), Modeling and Dynamics of Infectious Diseases, Series in Contemporary Applied Mathematics, World Scientific, Singapore, 216–236.

[8] Hsieh, Y.H. (2010), Pandemic influenza A (H1N1) during winter influenza season in the southern hemisphere, *Influenza and Other Respiratory Viruses*, *4*(4), 187–197.

[9] Hsieh, Y.H., J.Y. Lee and H.L. Chang (2004), SARS epidemiology modeling, *Emerging infectious diseases*, *10*(6), 1165.

[10] Hsieh, Y.H. and S. Ma (2009), Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005, *The American Journal of Tropical Medicine and Hygiene*, *80*(1), 66–71.

[11] Kendall, B.E. and G.A. Fox (1998), Spatial structure, environmental heterogeneity, and population dynamics: analysis of the coupled logistic map, *Theoretical Population Biology*, *54*(1), 11–37.

[12] Lee, S.Y., B. Lei and B. Mallick (2020), Estimation of COVID-19 spread curves integrating global data and borrowing information, *PloS One*, *15*(7), e0236860.

[13] Marusic, M., Z. Bajzer, S. Vuk-Pavlovic and J.P. Freyer (1994), Tumor growth in vivo and as multicellular spheroids compared by mathematical models, *Bulletin of Mathematical Biology*, *56*, 617–631.

[14] Pelinovsky, E., A. Kurkin, O. Kurkina, M. Kokoulina and A. Epifanova (2020), Logistic equation and COVID-19, *Chaos, Solitons & Fractals*, *140*: 110241.

[15] Pell, B., Y. Kuang, C. Viboud and G. Chowell (2018), Using phenomenological models for forecasting the 2015 Ebola challenge, *Epidemics*, *22*, 62–70.

[16] Rossi, R.J. (2018), *Mathematical Statistics: An Introduction to Likelihood Based Inference*, John Wiley & Sons, New York.

[17] Sachs, R.K., L.R. Hlatky and P. Hahnfeldt (2001), Simple ODE models of tumor growth and anti-angiogenic or radiation treatment, *Mathematical and Computer Modelling*, *33*(12–13), 1297–1305.

[18] Shen, C.Y. (2020), Logistic growth modelling of COVID-19 proliferation in China and its international implications, *International Journal of Infectious Diseases*, *96*, 582–589.

[19] Smith, R.C. (2013), *Uncertainty Quantification: Theory, Implementation, and Applications*, SIAM, Philadelphia.

[20] Spratt, J.S., J.S. Meyer and J.A. Spratt (1996), Rates of growth of human neoplasms: Part II, *Journal of Surgical Oncology*, *61*(1), 68–83.

[21] Wang, P., X. Zheng, J. Li and B. Zhu (2020), Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics, *Chaos Solitons & Fractals*, *139*, 110058.

[22] Wu, K., D. Darcet, Q. Wang and D. Sornette (2020), Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world, *Nonlinear Dynamics*, *101*(3), 1561–1581.

[23] Zhang, B., D.L. DeAngelis and W.M. Ni (2021), Carrying capacity of spatially distributed metapopulations, *Trends in Ecology & Evolution*, *36*(2), 164–173.