



Research article

Evaluating COVID-19 in Portugal: Bootstrap confidence interval

Sofia Tedim¹, Vera Afreixo¹, Miguel Felgueiras², Rui Pedro Leitão³, Sofia J. Pinheiro¹ and Cristiana J. Silva^{4,1,*}

¹ Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

² ESTG, Polytechnic Institute of Leiria and CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal

³ Public Health Unit, Baixo Vouga Primary Care Cluster, Administração Regional de Saúde (ARS) Centro, Av. Dr. Lourenço Peixinho, n 42, 4 andar, 3804-502 Aveiro, Portugal

⁴ Iscte - Instituto Universitário de Lisboa, ISTA, Av. das Forças Armadas, 1649-026 Lisboa, Portugal

* **Correspondence:** Email: cristiana.joao.silva@iscte-iul.pt; cjoaosilva@ua.pt.

Abstract: In this paper, we consider a compartmental model to fit the real data of confirmed active cases with COVID-19 in Portugal, from March 2, 2020 until September 10, 2021 in the Primary Care Cluster in Aveiro region, ACES BV, reported to the Public Health Unit. The model includes a deterministic component based on ordinary differential equations and a stochastic component based on bootstrap methods in regression. The main goal of this work is to take into account the variability underlying the data set and analyse the estimation accuracy of the model using a residual bootstrapped approach in order to compute confidence intervals for the prediction of COVID-19 confirmed active cases. All numerical simulations are performed in R environment (R version. 4.0.5). The proposed algorithm can be used, after a suitable adaptation, in other communicable diseases and outbreaks.


Keywords: COVID-19; bootstrap confidence interval; SAIRP model; centre of Portugal region

Mathematics Subject Classification: 62F40, 62F25, 62P10

1. Introduction

Since the beginning of COVID-19 pandemic, researchers have been using mathematical models to study the dynamics of SARS-CoV-2 spread and the impact of different non pharmacological and vaccination strategies in different regions of the world, see e.g. [1–5]. The mathematical framework and the time scale of epidemic models may depend on the nature of the available real data and on main goals of the scientific study, leading to different types of models, e.g. discrete, hybrid, deterministic,

etc. Deterministic compartmental models, given by systems of ordinary differential equations, have been used in the last decades to model, predict and control the spread of a wide range of infectious diseases helping to understand the transmission dynamics, the impact of preventive and intervention measures and predict outcomes of epidemics [6–13].

In this paper, we focus on a compartmental SAIRP model, first proposed in [14] and studied from the stability point of view in [15] given by a system of five ordinary differential equations, considering piecewise constant parameters, which allows to model the confirmed number of infected individuals with the virus SARS-CoV-2 in Portugal. The choice of this specific model was motivated by the good results when fitting Portuguese data [14]. In [16] the authors provide a Python code that allows to estimate some of the piecewise constant parameters and fit the model to the real data of COVID-19 transmission in Portugal, from March 2, 2020 until April 15, 2021. In our study, we use data from the Primary Care Cluster in Aveiro region, ACES BV, reported to the Public Health Unit (PHU), between March 2, 2020 and September 10, 2021. The main goal of this paper, is to extend the point estimation performed in [16] to interval estimation for the number of active infected individuals with SARS-CoV-2. With this procedure, we can accommodate the variability within the data and therefore provide accurate estimates. Since we are dealing with nonlinear methods, standard regression procedures can not be applied due to the dependence on the underlying distributional assumptions. For this reason, we use bootstrap methodology to obtain confidence intervals. We propose an algorithm, ran in  version 4.0.5 (2021-03-31), which can be used, after a suitable adaptation, to other communicable diseases and outbreaks.

The paper is structured as follows. In Section 2, we introduce the deterministic compartmental SAIRP model. In Section 3, we propose a method for interval estimation for the previsions based on confidence intervals, bootstrap methodology in regression and quantile intervals. The algorithm that implements the proposed methodology is provided. We show that our method allows to construct an adequate bootstrap confidence interval for the data under analysis (the number of active infected cases). We end the paper with Section 4, with some conclusions and challenges for future work.

2. Deterministic SAIRP model

We considered the deterministic compartmental model for COVID-19 transmission in a homogeneously mixed population with varying total size, proposed in [14, 15]. In this model, it is assumed that the total population $N(t)$, with $t \in [0, T]$ (in days) and $T > 0$, is subdivided into five groups of disjoint individuals: susceptible (S); asymptomatic infected (A); active infected (I); removed (including recovered and COVID-19 induced deaths) (R); and protected (P). Therefore, $N(t) = S(t) + A(t) + I(t) + R(t) + P(t)$, for $t \in [0, T]$. The susceptible sub-population has constant recruitment rate, Λ , and the entire population suffers from natural death, at a rate $\mu > 0$. The susceptible individuals S become infected by contact with active infected I and asymptomatic infected A individuals, at a rate of infection $\beta \frac{\theta(A+I)}{N}$, where θ represents a modification parameter for the infectiousness of the asymptomatic infected individuals A and β represents the transmission rate. Only a fraction q of asymptomatic infected individuals A develop symptoms and are detected, at a rate ν . Active infected individuals I are transferred to the recovered/removed individuals R , at a rate δ , by recovery from the disease or by COVID-19 induced death. A fraction p , with $0 < p < 1$, is protected (without permanent immunity) from infection, and is transferred to the class of protected individuals

P , at a rate ϕ . A fraction m of protected individuals P returns to the susceptible class S , at a rate w . Let $\xi = v q$ and $\omega = w m$. The previous assumptions are described by the following system of ordinary differential equations:

$$\begin{cases} \dot{S}(t) = \Lambda - \beta(1-p)\frac{\theta A(t)+I(t)}{N(t)}S(t) - (\phi p + \mu)S(t) + \omega P(t), \\ \dot{A}(t) = \beta(1-p)\frac{\theta A(t)+I(t)}{N(t)}S(t) - (\xi + \mu)A(t), \\ \dot{I}(t) = \xi A(t) - (\delta + \mu)I(t), \\ \dot{R}(t) = \delta I(t) - \mu R(t), \\ \dot{P}(t) = \phi p S(t) - (\omega + \mu)P(t). \end{cases} \quad (2.1)$$

To systematize and simplify, the parameter's description and notation are resumed in Table 1.

Table 1. Description and notation of the parameters of model (2.1), [14, 15].

Parameter	Description
Λ	Recruitment rate
μ	Natural death rate
β	Transmission rate
θ	Modification parameter
v	Transfer rate from A to I
q	Transfer fraction from A to I
ϕ	Transfer rate from S to P
p	Transfer fraction from S to P
w	Transfer rate from P to S
m	Transfer fraction from P to S
δ	Recovery rate

In this paper, we skip the mathematical analysis of the model, such as, invariant region, equilibrium points and its local and global stability, which was done in [15].

Using the package *deSolve* [17] to find the numerical solution of the system (2.1) and the function *optim* of *stats* package from R version 4.0.5 (2021-03-31), we estimate the parameters vector, using the method *L-BFGS-B* and splitting the time window in $l = 17$ sub-intervals [1, 45, 83, 100, 120, 186, 212, 242, 251, 261, 271, 296, 320, 374, 400, 470, 496, 552], where 1 represents March 2, 2020 and 552 September 10, 2021. The remaining parameters are assumed to take the following fixed values: $N = 363803$, $\lambda = (0.0019 * N)/365$, $\theta = 1$, $\phi = 1/6$, $\mu = 1/(81 * 365)$, $\delta = 1/27$, $v = 1$, $q = 0.15$, $w = 1/35$, with $\xi = v q$ and $\omega = w m$. With these parameter values, we show that model (2.1) fits the number of active cases infected by SARS-CoV-2, from March 2, 2020 until September 10, 2021, from Primary Care Cluster in Aveiro region, ACES BV, see Figure 1. We remark that, since the parameter space is \mathbb{R}_0^+ , we must add this constraint in the *L-BFGS-B* routine.

The quality of the fit was measured by a weighted coefficient of determination, denoted by

$$R^2 = \sum_{j=1}^l w_j R_j^2. \text{ Each } w_j \text{ represents the weight of a sub-interval (a total of } l = 17 \text{ sub-intervals), and}$$

is defined as the ratio between the number of observations in that interval and the total number of observations (552), with $\sum_{j=1}^{17} w_j = 1$. Each R_j^2 represents the proportion of data variability explained

by our model in that particular interval, and is defined as $R_j^2 = \frac{\sum_{i=1}^{t_j} (\widehat{y}_{ij} - \bar{y}_j)^2}{\sum_{i=1}^{t_j} (\widehat{y}_{ij} - \bar{y}_j)^2 + \sum_{i=1}^{t_j} (y_{ij} - \widehat{y}_{ij})^2}$ where

y_{ij} represents the observed number of covid cases for the i time in the j interval, \widehat{y}_{ij} the corresponding predicted value, \bar{y}_j the corresponding mean and t_j the number of time points for interval j . Therefore, R^2 provides the proportion of data variability explained by our model as a weighted sum of the proportion of data variability explained by each sub-interval. For $l = 17$ sub-intervals, we obtain $R^2 = 0.7862$ and therefore we can conclude that the model explains the majority of the data variance (more than three quarters). We highlight that a better result would be very difficult to achieve due to significant fluctuations in Active Cases, inherent in the data, mainly around day 300.

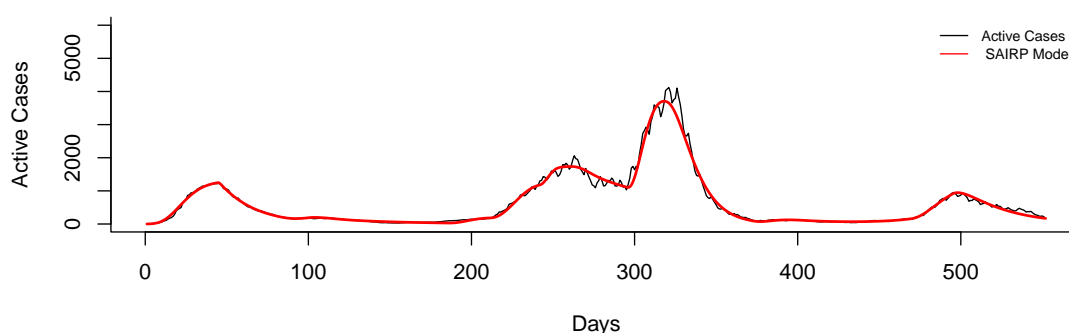


Figure 1. Model (2.1) fit the number of active cases infected by SARS-CoV-2, from March 2, 2020 until September 10, 2021, from Primary Care Cluster in Aveiro region, ACES BV, the observed data and the proposed model fit (SAIRP Model).

3. Confidence intervals for the fit

3.1. Motivation

In the previous section, we presented a deterministic model given by a system of differential equations to fit the considered COVID-19 data set. However, point estimates do not take into account the variability that is underlying the data set and therefore the estimation accuracy is not analysed. To tackle this problem, in this section we introduce confidence intervals for the prediction.

When dealing with linear models, it is usually assumed that the random errors are independent random variables modelled by the Gaussian distribution, and consequently the dependent variable Y is also modelled by the Gaussian distribution. When estimating the error variance, chi-square distribution is used as a consequence of least squares method, and finally the predictions \widehat{Y} are well fitted by the t-Student distribution. Everything works smoothly and, having a well known distribution for \widehat{Y} , inference

procedures are straightforward [18]. The problem arises when least square errors method is used in non linear regression, which is the case presented in this paper. Under these circumstances, the normality hypothesis must be dropped and alternative procedures should be used.

3.2. Bootstrap methodology in regression

The bootstrap is based on resampling, with replacement, an initial data (a large number of times) [19]. In this work, we propose a residual bootstrapped approach in order to compute the confidence interval of the number of COVID-19 active cases.

The bootstrap procedure can be described by the following steps (adapted from [18])

1. Consider the data divided into l time intervals. Each j interval has t_j time points, $j = 1, \dots, l$.
2. For each i time point from j interval, compute the optimum non-linear model using the least square errors method and the corresponding residuals

$$e_{ij} = y_{ij} - \widehat{y}_{ij}$$

with y_{ij} the observed number of covid cases for the i time in the j interval and \widehat{y}_{ij} the corresponding predicted value.

3. For each interval, center the residuals to a mean of 0

$$e_{ij} = e_{ij} - \bar{e}_j$$

with \bar{e}_j the residuals mean in j interval.

4. For each interval j , $n \times t_j$ new samples of residuals were obtained using random sampling with replacement (n bootstrap residuals for each time points). These bootstrap residuals were denoted by e_{ijk}^* , $k = 1, \dots, n$
5. As bootstrap sample to the number of covid cases we consider

$$\{y_{ijk}^* = y_{ij} + e_{ijk}^*, j = 1, \dots, l, i = 1, \dots, t_j\}.$$

6. The model was recalculated for each k and new estimates \widehat{y}_{ijk}^* were obtained. The optimal non-linear model using the least squares errors method was recalculated for each k , and consequently n new fits were obtained for each j interval.

3.3. Quantile intervals

After the step 3 in the previous subsection, we now have n regression models for each j subinterval. Let $\widehat{y}_{(ijk)}^*$ represent, for a fixed i and j , the ordered bootstrap estimate, and suppose that we wish to construct a $(1 - \alpha) 100\%$ confidence interval for y_{ijk} . For a large n , $\widehat{y}_{(lower)}^*$ and $\widehat{y}_{(upper)}^*$, where $lower = \frac{\alpha}{2}n$ and $upper = \left(1 - \frac{\alpha}{2}\right)n$ are non-parametric approaches from \widehat{y}_{ijk}^* quantiles and can be use as bounds for the confidence interval y_{ijk} , that is,

$$IC_{(1-\alpha)} = [\widehat{y}_{(lower)}^*; \widehat{y}_{(upper)}^*].$$

The assumption underlying this procedure is that the errors are independent and identically distributed (IID) among our data.

3.4. Pseudo code algorithm

The implemented algorithm follows the following steps, described in Table 2.

Table 2. Pseudo code algorithm.

<i>Input:</i> Initial conditions, fixed parameters, real data and model equations.
<i>Output:</i> β , m , p estimates, R squared and the active cases curve fit.
1. For each time window sub-interval:
1.1. estimate individually β , m , p minimizing least squares sum;
1.2. based on Step 1.1, estimate 3-uple (β, m, p) ;
1.3. compute prediction solving the model (2.1) using input and Step 1.2;
1.4. calculate ordinary and centered residuals;
1.5. compute the determination coefficient, R squared.
2. Compute the weighted R squared based on the length of each sub-interval.
3. For each time window sub-interval, bootstrap centered residuals.
4. Repeat 1. using real data plus the bootstrap centered residuals, in place of real data.
5. Compute the 95% confidence intervals.

The algorithm includes a deterministic component based on ordinary differential equations and a stochastic component based on bootstrap methods in regression. The novelty of this algorithm is the computation of confidence intervals using a procedure that takes into account the nonlinearity of the data.

3.5. Results

As stated before, the methodology introduced in the previous section does not rely on the normality hypothesis for the residuals. Nevertheless, and since we are re-sampling the residuals and randomly reattach them to the predictions, implicitly we assume that the residuals are identically distributed and that the functional form of the model is correct [18]. Also, a large skewness can be a drawback since some y_{ijk}^* might become quite influential leading to confidence intervals with increased amplitude. When analysing Figure 2, the centered residuals histogram show an even distribution around 0, for the five central classes. Nevertheless, a positive skewness is detected without being of great relevance.

The next step (see subsection 3.2) was to obtain the bootstrap models. Remember that we considered 17 subintervals with a different number of observations in each interval in order to capture the data oscillations, according with several epidemic waves. We remark that, although the time window subdivision is done empirically, it takes into account the public health policies, the adherence of the population to prevention measures and the intensity of the virus transmission. For each interval, 500 new samples were obtained and the corresponding new models were calculated. The consideration of 500 new samples is an acceptable number according to [20], and the time consuming process connected with samples generation.

After obtaining the bootstrap models, quantile intervals (see subsection 3.3) were computed. These intervals are represented in Figure 3 and for each date the grey area corresponds to the 95% bootstrap confidence interval. It is visually evident that the confidence intervals contain the majority of the active cases (black line). As expected, for the intervals where data fluctuation is intense (basically between

October 2020 and February 2021, and later between June 2021 and September 2021) the confidence intervals reveal a larger range. For example, for the cases record achieved in January 2021, the fitted model (in red) does not apprehend the increase in the active cases, but the 95% confidence interval is able to include these peaks inside its bounds. This example clearly shows the importance of introduce a stochastic component in the predictions.

As we can see in Figure 4, the proportion of points, where the real number of active cases fall within the bootstrap confidence interval ranges from 0.52 (intervals 9,10,11) to 1.0 (interval 6) depending on the interval. At least 50% of the intervals have a proportion higher than (0.80) and 75% have a proportion higher than (0.70), with a mean of 0.80. Considering all the points the general mean for that proportion is 0.79. Weighting by each intervals size the mean is approximately the same. The results show that the real cover rate of the confidence intervals is, in general, below the expected 0.95 rate. Moreover, the obtained bounds are a little optimistic and they should be more far apart to achieve a 0.95 cover rate. One alternative would be to consider in Subsection 3.3 a lower value of α that would lead to an increase in the cover rate.

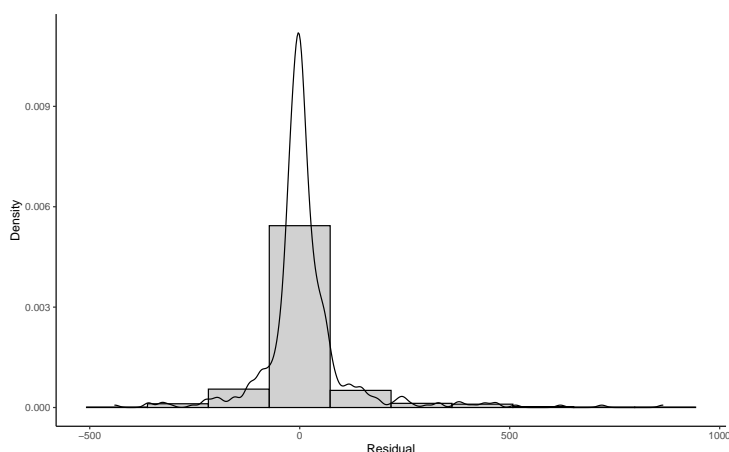


Figure 2. Histogram and density plot for the centered residuals of the fitted model.

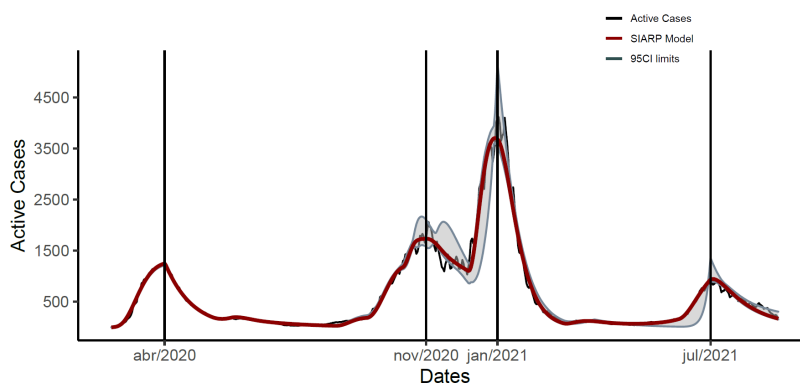


Figure 3. Time series for the number of active cases (in black) throughout the study period. In red is the estimate for the original model and the grey area corresponds to the 95% bootstrap confidence interval. The vertical lines represent the dates of local maximums.

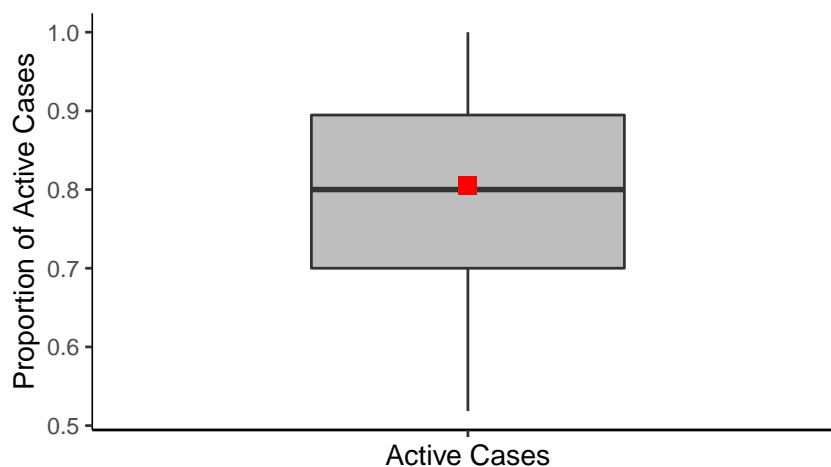


Figure 4. Box Plot for the proportion of points, by interval where the real number of active cases fall within the bootstrap confidence interval. The red dot represents the median.

4. Conclusions and future work

In this paper, we considered a model described by a system of differential equations to fit a COVID-19 real data set. To overcome the fact that deterministic models often use point estimation, which do not take into account the variability that is underlying the data set, we proposed a method for estimating confidence intervals that allowed us to analyse the accuracy of the model prediction. To compute the confidence intervals of the number of COVID-19 active cases, we used a residual bootstrapped approach, and showed that they contain the majority of the COVID-19 active cases. Moreover, the 95% confidence interval is able to include inside its bounds the incidence peaks that occurred during the epidemic period, highlighting the importance of introducing a stochastic component into the predictions. Up to our knowledge, this is the first time that a deterministic compartmental model is enriched with a residual bootstrapped approach in order to compute confidence intervals for the prediction of COVID-19 confirmed active cases.

As future work, robust estimation methods could be employed. Least squares methodology is sensitive to influential observations, that is, observations that individually have significant influence on parameter values. Therefore the class of M-estimators could be a valuable option [21].

Acknowledgement

This work is partially supported by Portuguese funds through CIDMA, The Center for Research and Development in Mathematics and Applications of University of Aveiro, and the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnologia), within project UIDB/04106/2020 (<https://doi.org/10.54499/UIDB/04106/2020>) and project UIDP/04106/2020 (Thematic Line BIOMATH) (<https://doi.org/10.54499/UIDP/04106/2020>). C. J. Silva is also supported by the project “Mathematical Modelling of Multi-scale Control Systems: applications to human diseases (CoSysM3)”, 2022.03091.PTDC, financially

supported by national funds (OE), through FCT/MCTES. M. Felgueiras is supported through FCT project UIDB/00006/2020.

Conflict of interest

Professor Cristiana J. Silva is an editorial board member for AIMS Mathematics and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

References

1. E. Bertuzzo, L. Mari, D. Pasetto, S. Miccoli, R. Casagrandi, M. Gatto, et al., The geography of COVID-19 spread in Italy and implications for the relaxation of confinement measures, *Nat. Commun.*, **11** (2020). <https://doi.org/10.1038/s41467-020-18050-2>
2. B. Machado, L. Antunes, C. Caetano, J. F. Pereira, B. Nunes, P. Patrício, et al., The impact of vaccination on the evolution of COVID-19 in Portugal, *Math. Biosci. Eng.*, **19** (2022), 936–952. <https://doi.org/10.3934/mbe.2022043>
3. S. Moore, E. M. Hill, M. J. Tildesley, L. Dyson, M. J. Keeling, Vaccination and non-pharmaceutical interventions for COVID-19: A mathematical modelling study, *Lancet Infect. Dis.*, **21** (2021), 793–802. [https://doi.org/10.1016/S1473-3099\(21\)00143-2](https://doi.org/10.1016/S1473-3099(21)00143-2)
4. F. Ndairou, I. Area, J. J. Nieto, C. J. Silva, D. F. M. Torres, Fractional model of COVID-19 applied to Galicia, Spain and Portugal, *Chaos Soliton. Fract.*, **144** (2021), 110652. <https://doi.org/10.1016/j.chaos.2021.110652>
5. O. Pinto Neto, D. M. Kennedy, J. C. Reis, Y. Wang, A. C. Brisola Brizzi, G. José Zambrano, et al., Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil, *Nat. Commun.*, **12** (2021), 418. <https://doi.org/10.1038/s41467-020-20687-y>
6. R. M. Anderson, R. M. May, *Infectious diseases of humans: dynamics and control*, Oxford University Press, (1991).
7. R. M. Anderson, R. M. May, M. C. Boily, G. P. Garnett, J. T. Rowley, The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS, *Nature*, **352** (1991), 581–589. <https://doi.org/10.1038/352581a0>
8. N. Bacaër, McKendrick and Kermack on epidemic modelling (1926–1927), *A Short History of Mathematical Population Dynamics*, Springer, (2011). https://doi.org/10.1007/978-0-85729-115-8_16
9. H. W. Hethcote, A thousand and one epidemic models, in *Frontiers in mathematical biology. Lecture notes in Biomathematics* (eds. Simon A. Levin), Springer, (1984), 100, 504–515. https://doi.org/10.1007/978-3-642-50124-1_29

10. K. J. B. Villasin, E. M. Rodriguez, A. R. Lao, A Deterministic Compartmental Modeling Framework for Disease Transmission, in *Computational Methods in Synthetic Biology. Methods in Molecular Biology* (eds M.A. Marchisio), Humana, **2189** (2021), 157–167. https://doi.org/10.1007/978-1-0716-0822-7_12
11. Y. Guo, T. Li, Modeling and dynamic analysis of Novel Coronavirus Pneumonia(COVID-19) in China, *J. Appl. Math. Comput.*, **68** (2022), 2641–2666. <https://doi.org/10.1007/s12190-021-01611-z>
12. T. Li, Y. Guo, Modeling and optimal control of mutated COVID-19 (Delta strain) with imperfect vaccination, *Chaos Soliton. Fract.*, **156** (2022), 111825. <https://doi.org/10.1016/j.chaos.2022.111825>
13. T. Li, Y. Guo, Optimal control and cost-effectiveness analysis of a new COVID-19 model for Omicron strain, *Physica A.*, **606** (2022), 128134. <https://doi.org/10.1016/j.physa.2022.128134>
14. C. J. Silva, C. Cruz, D. F. M. Torres, A. P. Muñuzuri, A. Carballosa, I. Area, et al. Optimal control of the COVID-19 pandemic: controlled sanitary deconfinement in Portugal, *Sci. Rep.*, **11** (2021), 3451. <https://doi.org/10.1038/s41598-021-83075-6>
15. C. J. Silva, G. Cantin, C. Cruz, R. Fonseca-Pinto, R. Fonseca, E. S. Santos, et al., Complex network model for COVID-19: human behavior, pseudo-periodic solutions and multiple epidemic waves, *J. Math. Anal. Appl.*, in press. <https://doi.org/10.1016/j.jmaa.2021.125171>
16. Z. Abreu, G. Cantin, C. J. Silva, Analysis of a COVID-19 compartmental model: a mathematical and computational approach, *Math. Biosci. Eng.*, **18** (1992), 7979–7998. <https://doi.org/10.3934/mbe.2021396>
17. K. Soetaert, T. Petzoldt, R. W. Setzer, Solving Differential Equations in R: Package deSolve, *J. Stat. Softw.*, **33** (2010), 1–25. <https://doi.org/10.18637/jss.v033.i09>
18. J. Fox, *Applied Regression Analysis and Generalized Models*, Sage, Los Angeles, (2016).
19. D. A. Freedman, Bootstrapping Regression Models, *Ann. Stat.*, **9** (1981), 1218–1228. <https://doi.org/10.1214/aos/1176345638>
20. R. Davidson, J. MacKinnon, Bootstrap Tests: How many bootstraps?, *Economet. Rev.*, **19** (2000), 55–68. <https://doi.org/10.1080/07474930008800459>
21. D. Q. F. de Menezes, D. M. Prata, A. R. Secchi, J. C. Pinto, A review on robust M-estimators for regression analysis, *Comput. Chem. Eng.*, **147** (2021), 107254. <https://doi.org/10.1016/j.compchemeng.2021.107254>



AIMS Press

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)