Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Application of Gaussian process regression as a surrogate modeling method to assess the dynamics of COVID-19 propagation

Alexandra Matveeva\textsuperscript{a, *}, Vasiliy Leonenko\textsuperscript{b}

\textsuperscript{a}Faculty of Digital Transformation, ITMO University, Birzhevaya, 4, Saint Petersburg, 199034, Russian Federation
\textsuperscript{b}National Center for Cognitive Research, ITMO University, Birzhevaya, 4, Saint Petersburg, 199034, Russian Federation

Abstract

In this research, we aimed to assess the possibility of using surrogate modeling methods to replace time-consuming calculations related to modeling of COVID-19 dynamics. The Gaussian process regression (GPR) was used as a surrogate to replace detailed simulations by a COVID-19 multiagent model. Experiments were conducted with various kernels, as a result, in accordance with the quality metrics of the models, kernels were identified in which the surrogate gives the most accurate result (Rational Quadratic kernel and Additive kernel). It was demonstrated that by smoothing the dynamics of COVID-19 propagation, it is possible to achieve greater accuracy in GPR training. The obtained results prove the potential possibility of using surrogate modeling methods to conduct an uncertainty quantification of the multiagent model of COVID-19 propagation.

© 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 11th International Young Scientist Conference on Computational Science.

Keywords: Epidemiology; Mathematical modeling; Multiagent modeling; Surrogate modeling; Gaussian process regression; Uncertainty quantification; COVID-19

1. Introduction

Mathematical modeling proved itself as a useful tool to characterize the dynamics of the spread of diseases such as influenza and COVID-19. Outbreaks of coronavirus infection 2019-nCoV (COVID-19) have had a significant impact...
not only on the global economy, but also on the daily life of each person [1]. Since December 2019, thousands of people around the world have been infected with a new coronavirus called SARS-CoV-2 with a high mortality rate [2]. In this regard, there is a need for tools capable of modeling the spread of infectious diseases.

There are many types of models for predicting and explaining the dynamics of the epidemic, including compartmental and multiagent models. Models based on the compartmental approach have a more simplified structure, which makes it possible to quickly calibrate them to real data, but for a more detailed description of the processes of virus spread, it is necessary to consider parameters such as population heterogeneity and individual characteristics of each agent. To fulfill this requirement, an approach based on so-called multiagent models was created [3].

1.1. Agent-based modeling and uncertainty quantification

Multiagent models, also known as agent-based models (ABM), have some limitations despite all their advantages, in particular, high-level complexity of parameters, long execution time, and complexity of model analysis. As the complexity of agent-based models increases, the number of parameters required to be assessed on real data grows [4]. Due to the presence of stochastic processes for model calibration, as well as a need for uncertainty and sensitivity analysis [5], it is necessary to conduct many simulation launches, which leads to increased time consumption.

Uncertainty quantification (UQ) is an analysis to study the effect of uncertainties in initial or boundary conditions and of other parameters of computational model on their simulated quantities of interest [6]. There are many different uncertainty quantification methods, the main ones are:

- Methods based on Monte Carlo method [7], [8], [9]
- Polynomial chaos expansion [10], [11]
- Stochastic collocation [12]

All the above-mentioned methods require a lot of simulations to provide enough data for the numerical integration of the statistical estimator [13]. One of the solution is to adopt surrogate modeling. Within this approach, a surrogate model (or meta-model) is being developed to approximate the response of the original model at relatively low costs. This surrogate model replaces time-consuming simulations, which makes it possible to conduct the UQ. In this paper, we demonstrate the application of this approach for modeling COVID-19 dynamics.

In this research work an agent-based framework from [14] was used. The basic principle of simulation is as follows: each agent in the population potentially interacts with other agents if they attend the same school (for schoolchildren), workplace (for working age adults), or live in the same household. The infectivity of each agent depends on their day of infection [15]. The modeling step of this model is equal to one day. Agents are randomly selected from the general population and are assigned an infectious status at the beginning of the simulation. Step by step algorithm is described in [15].

The main input parameters of the model which are important for uncertainty and sensitivity analysis are introduced in Table 1.

| Parameter name | Description                      |
|----------------|---------------------------------|
| $\alpha$       | Fraction of non-immune individuals in general population |
| $\lambda$      | Infection transmission coefficient |
| $I_0$          | Initial number of infected individuals |

When modeling the dynamics of the spread of COVID-19, building a network of interactions between agents is quite important. This process is completely stochastic and its modeling requires substantial computational costs. Parallelization of calculations leads to a slight acceleration of the model. In order to minimize time-consuming computational processes, it is proposed to apply surrogate modeling methods [16].
1.2. Surrogate modeling

Surrogate modeling is used to approximate expensive models in modern complex tasks. Often surrogate models are parts of evolutionary algorithms as fitness functions or individuals of a population. The main idea of this method is to replace time-consuming calculations with an approximating model [17]. The algorithm for using surrogates is shown in Fig. 1 and consists in the following: the original model (or part of it) is replaced by a surrogate and is presented as a "black box", in which the main problem is finding the function of dependence of the output data on the parameters fed to the input of the model [18]. This model imitates the behaviour of the original computational model and evaluates the corresponding responses of the model based on specific inputs [19].

Fig. 1 – The scheme of using a surrogate model to replace the expensive computer simulations

There are three types of the construction of the surrogate models [20]:

- Simplified models
- Projection-based methods
- Data-fit methods

Examples of simplified models are models of spatial dimensionality reduction, in other words algorithms based on simplifications of the simulated system [21], [22]. The data-fit methods map out latent functions between input and output. Common methods of this type of surrogates are:

- Support vector machines [23]
- Neural networks [24]
- Gaussian processes [25]

In this research, Gaussian Process Regression (GPR) was used as a surrogate to minimize the time costs of the agent-based model, as well as to demonstrate the possibility of estimating the dynamics of COVID-19 propagation with different sets of input parameters.

Gaussian Process Regression is a class of supervised machine learning algorithm, for which it is sufficient to use a small number of parameters to make a prediction. Covariance functions are an important component of GPR models because these functions weigh the contribution of training points to the predicted test targets according to the kernel distance between the observed training points and test points [21]. The following covariance functions were considered in the paper [26]:

- Rational Quadratic (RQ) kernel
- Squared Exponential (SE) / Radial-basis Function (RBF) kernel
Gaussian Process Regression is a class of supervised machine learning algorithm, for which it is sufficient to use a small number of parameters to make a prediction. Covariance functions are an important component of GP with output. Common methods of this type of surrogates are: simplifications of the simulated model, as well as to demonstrate the possibility of estimating the dynamics of COVID-19 propagation. Gaussian processes are parts of evolutionary algorithms as fitness functions or individuals of a population. The main idea of this method is to replace time-consuming calculations with an approximating model. The scheme of using a surrogate model to replace the expensive computer simulations is shown in Fig. 1 and consists in the following: the original model (or part of it) is replaced by a surrogate and is fed to the input of the model (21). The algorithm for using surrogates is as Proper orthogonal decomposition (28), to the model response before regression.

To train a Gaussian process regression model, it is necessary to form a dataset by repeatedly running the original agent-based model of COVID-19 spread (14). To perform this synthetic population of Saint Petersburg, Russia was used.

**I.3. Synthetic population and sampling**

’Synthetic population’ is a synthesized, spatially explicit human agent database representing the population of a city, a region or a country. By its cumulative characteristics, this database is equivalent to the real population but its records are not correspondent to real people (15). The initial synthetic population of St. Petersburg used in the simulation consists of several files described in Table 2.

| File               | Contents                                                                 |
|--------------------|---------------------------------------------------------------------------|
| people_workplaces.txt | Records for each person, along with their age and gender and information about workplaces for each agent |
| households.txt     | The location and descriptive attributes for each household                |
| schools.json       | Records for each school (dictionary)                                     |

There are about five million agents in this synthetic population, modeling on such a large set takes a significant amount of time. In this case, there is a need to conduct research with a sampled population. The figure 2 shows that with a sampled population, modeling is much faster.
To demonstrate the work of the algorithm, it is enough to simulate the spread of COVID-19 not on the entire synthetic population, but only on a certain part. Such a part will demonstrate the spread of the disease not by the example of a large city (in this case, Saint Petersburg), but by the example of a small town with a population of 500,000 people. The scheme of the dependence of the fields of the tables is shown in Fig. 3.

As part of this research, the following sampling algorithm was implemented:

- From the list of all agents “people_workplaces.txt” randomly select the required number of agents
- From the file “households.txt” select the houses corresponding to the agent IDs selected in the first paragraph
- Similarly, from the “schools.json” file, select the schools that the agents attend.

**2. Model development**

The surrogate model was constructed to replace the whole agent-based model to assess the dynamics of COVID-19 propagation. Most GPR implementations model only a single response variable [29], Traditionally, one response variable is treated as a Gaussian process, and multiple responses are modeled independently without considering their correlation [29], but this approach is not ideal.

In our research we try to apply gaussian process regression to multiple output. The GPR was built using scikit-learn, is an open source machine learning library [30] and trained on the data obtained as a result of repeatedly running of the agent-based model [14] with different values of input parameter $\alpha$ with a range of values $[0.4; 0.93]$. 

![Graph showing execution time](image)
Regression of a Gaussian process with multiple outputs is an area of active study [29]. With an increase in the dimension of the output of the model, it is possible to either use methods of dimensionality reduction, or to formulate of covariance function that describes not only the correlation between data points, but also the correlation between responses [29]. To perform this in future research it is planned to use more flexible tool GPy [31] to build a surrogate.

3. Results

The dataset was divided into training and test sets in the ratio 1:3. The results obtained on the test set are shown in Fig. 4. It can be seen from Fig. 4 a that the Gaussian Process Regression does not work correctly with the Rational-basis kernel. At the same time, surrogate models with Rational quadratic kernel and Additive kernel allow to track the approximate dynamics of propagation with all input parameters, which is demonstrated in Fig. 4 (b, c).

![Fig. 4. The result of the assessment of the COVID-19 spread by Gaussian process regression: (a) RBF kernel; (b) RQ kernel; (c) Additive kernel.](image)

Table 3 demonstrates the metrics of models with different kernels. It can be seen that the best result is given by models with Radial-basis Function kernel and Additive kernel.

| Kernel                      | MAE   | $R^2$ |
|-----------------------------|-------|-------|
| Radial-basis Function (RBF) | 54,633| -0,870|
| Rational Quadratic (RQ)     | 21,833| 0,599 |
| Additive (RBF + RQ)         | 18,001| 0,687 |

Since it is enough to know only the approximate number of infected people to estimate the dynamics, as well as to assess the peak activity of the disease, it is possible to smooth out the graphs of dynamics. It is assumed that this procedure will improve the quality of prediction. The results of the Gaussian process regression model on a test sample of smoothed data are shown in Fig. 5.
Fig. 5. (a) RBF kernel; (b) RQ kernel; (c) RBF+RQ kernel.

Table 4 shows that for a model with RQ kernel and Additive kernel, the values of the $R^2$ and MAE metrics increased by 0.14 and 5.31 on average, respectively.

Table 4. Metrics for GRP models with different kernels.

| Kernel                               | MAE    | $R^2$  |
|--------------------------------------|--------|--------|
| Radial-basis Function (RBF) kernel   | 51,064 | -1,085 |
| Rational Quadratic (RQ) kernel       | 17,013 | 0.710  |
| Additive (RBF + RQ) kernel           | 12,201 | 0.852  |

The obtained results demonstrate the possibility of estimating the dynamics of the spread of COVID-19 at different values of input parameters. According to this result it can be argued that the Gaussian Process Regression can be used for uncertainty and sensitivity analysis of multiagent model.

4. Conclusion

In this research, surrogate model was implemented to replace time-consuming calculations of multiagent models. Four types of kernels for the regression of the Gaussian process were considered. The best results were shown by the Rational quadratic kernel and the Additive kernel (a combination of the kernels of rational quadratic and radial basis functions). As a result of the potential possibility of using surrogate models, namely Gaussian Process Regression, to perform the uncertainty and sensitivity analysis of multi-agent models was demonstrated. To improve the results of predicting dynamics with different input parameters, it was proposed to use the smoothing method.

In future research, it is planned to replace the expensive calculations of the multiagent model [14] with a surrogate to assess the full dynamics (80 days) of the spread of COVID-19. To train this surrogate, it is necessary to form a dataset for a broader range of input parameter values, while paying special attention to values less than 0.3, since in this case the full cycle of infections occurs in about 40 days.

References

[1] van Zyl, Terence L., and Turgay Celik. (2021) "Did we produce more waste during the covid-19 lockdowns? a remote sensing approach to landfill change analysis." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14: 7349-7358.
[2] Paul, Subrata, et al. (2021) "Study of SEIR epidemic model and scenario analysis of COVID-19 pandemic." Ecological Genetics and Genomics 19: 100087.
[3] Keeling, Matt J., and Pejman Rohani. (2011) “Stochastic dynamics.” Modeling infectious diseases in humans and animals. Princeton University Press: 190-231.
[4] Miksch, Florian, et al. “Why should we apply ABM for decision analysis for infectious diseases?—An example for dengue interventions." PloS one 14.8 (2019): e0221564.
References dataset for a broader range of input parameter values, while paying special attention to values less than 0

- 186 - 190

13. GPy: A Gaussian process framework in python. [Online].
14. Scikit learn [Online].
15. GPy. GPy: A Gaussian process framework in python. [Online].
16. Uncertainty quantification: theory, implementation, and applications. Vol. 12. Siam, 2013.
17. Ye, Dongwei, et al. (2021) "Uncertainty quantification of a 3D In-Stent Restenosis model with surrogate modelling." arXiv preprint arXiv:2111.06173.
18. Caflisch, Russel E. (1998) "Monte carlo and quasi-monte carlo methods." Acta numerica 7: 1-49.
19. Giles, Michael B. (2015) "Multilevel monte carlo methods." Acta numerica 24: 259-328.
20. Pecherstorfer, Benjamin, Karen Willcox, and Max Gunzburger. (2016) "Optimal model management for multifidelity Monte Carlo estimation." SIAM Journal on Scientific Computing 38(5): A3163-A3194.
21. Crestaux, Thierry, Olivier Le Matte, and Jean-Marc Martinez. (2009) "Polynomial chaos expansion for sensitivity analysis." Reliability Engineering & System Safety 94(7): 1161-1172.
22. Le Gratiet, Loic, Stefano Marelli, and Bruno Sudret. "Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes." Handbook of uncertainty quantification. Springer, Cham, 2017. 1289-1325.
23. Sankaran, Sethuraman, and Alison L. Marsden. (2011) "A stochastic collocation method for uncertainty quantification and propagation in cardiovascular simulations." Journal of biomechanical engineering 133(3).
24. Fang, Kai-Tai, Runze Li, and Agus Sudjianto. Design and modeling for computer experiments. Chapman and Hall/CRC, 2005.
25. Influenza spatial: A spatial model for the spread of influenza [Online]. – URL: https://github.com/vnleonenko/Influenza_spatial (accessed: 15.01.2022).
26. Leonenko, Vasily, Sviatoslav Arzamastsev, and Georgiy Bobashev. (2020) "Contact patterns and influenza outbreaks in Russian cities: A proof-of-concept study via agent-based modeling." Journal of Computational Science 44: 101156.
27. Koeppl, Tobias, et al. (2018) "Numerical modelling of a peripheral arterial stenosis using dimensionally reduced models and kernel methods." International journal for numerical methods in biomedical engineering 34(8): e3095.
28. Burnaev E. V., Prikhodko P. V. (2013). Methodology of construction of surrogate models for approximation of spatially inhomogeneous functions. Proceedings of the Moscow Institute of Physics and Technology (Vol. 5): 122–132.
29. Coccarelli, Alberto, et al. (2021) "A framework for incorporating 3D hyperelastic vascular wall models in 1D blood flow simulations." Biomechanics and Modeling in Mechanobiology 20(4): 1231-1249.
30. Sobester, András, Alexander Forrester, and Andy Keane. Engineering design via surrogate modelling: a practical guide. John Wiley & Sons, 2008.
31. Pecherstorfer, Benjamin, Karen Willcox, and Max Gunzburger. (2018) "Survey of multifidelity methods in uncertainty propagation, inference, and optimization." Siam Review 60(3): 550-591.
32. Zhang, Xinshuai, et al. (2021) "Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization." Computer Methods in Applied Mechanics and Engineering 373: 113485.
33. Sen, Oishik, et al. (2018) "Evaluation of multifidelity surrogate modeling techniques to construct closure laws for drag in shock–particle interactions." Journal of Computational Physics 434-451.
34. Wang, Lipo, ed. Support vector machines: theory and applications. Vol. 177. Springer Science & Business Media, 2005.
35. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
36. Rasmussen, Carl Edward. "Gaussian processes in machine learning." Summer school on machine learning. Springer, Berlin, Heidelberg, 2003
37. The Kernel Cookbook: Advice on Covariance functions [Online]. – URL: https://www.cs.toronto.edu/~duvenaud/cookbook/ (accessed: 20.06.2022).
38. Marrel, Amandine, et al. (2009) "Calculations of Sobol indices for the Gaussian process metamodel." Reliability Engineering & System Safety 94(3): 742-751.
39. Chatterjee, Anindya. "An introduction to the proper orthogonal decomposition." Current science (2000): 808-817.
40. Wang, Bo, and Tao Chen. "Gaussian process regression with multiple response variables." Chemometrics and Intelligent Laboratory Systems 142 (2015): 159-165.
41. Scikit learn [Online]. – URL: https://scikit-learn.org/stable/getting_started.html (accessed: 18.09.2022).
42. GPy. GPy: A Gaussian process framework in python. [Online]. – URL: http://github.com/SheffieldML/GPy (accessed: 18.09.2022).