IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 13, 2021, accepted August 7, 2021, date of publication August 13, 2021, date of current version August 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104519

Two-Strain COVID-19 Model Using Delayed Dynamic System and Big Data

VASYL MARTSENYUK, MARCIN BERNAS, AND ALEKSANDRA KLOS-WITKOWSKA
Department of Computer Science and Automatics, University of Bielsko-Biala, 43-309 Bielsko-Biala, Poland

Corresponding author: Vasyl Martsenyuk (vmartsenyuk@ath.bielsko.pl)

This work was supported by the European Union's Erasmus + Program for Education under Key Action 2: Strategic Partnerships Grant (Innovations for Big Data in a Real World) under Project 2020-1-PL01-KA203-082197.

ABSTRACT The proposed model is based on COVID-19 Big Data Hub. It enables us to predict pandemics development taking into account multiple virus strains and delays of infectiousness. Two-strain dynamic models with distributed delays have been fitted to the time series retrieved from COVID data hub. The data at the national, regional, and county-level which are seamlessly integrated with World Bank Open Data, Google Mobility Reports, Apple Mobility Reports, have been used. The parameter identification has been fulfilled with the help of COBYLA algorithm. The simulations have been implemented with the help of Julia high-performance computing. The effect of the time delays is analyzed. The considered pipeline utilizes the data from the Hub to generate the COVID model and to produce a reliable prediction.

INDEX TERMS Big data, COVID, delayed dynamic system, distributed delays, epidemiology, high-performance computing, parameter identification, scientific machine learning, virus strain.

I. INTRODUCTION

Nowadays the coronavirus pandemic is considered as a global medical, social, and economic problem. It has affected all branches of human activity, both industry, and science. A lot of researchers are joined to the world-wide projects dealing with the diagnostics, prophylaxis, and treatment of COVID.

One of the current topics, connected with COVID19, is the ability to model its spread and by doing so contain it by incorporating appropriate measures.

A. RELATED WORKS

1) COVID-19 MODELING

All works dealing with the research of COVID-19, are related with the different aspects of prophylaxis, diagnostics, and treatment of the disease. When considering the *prophylaxis*, they are aiming to predict quantitatively and qualitatively (epidemic outbreaks) the epidemic curve. Traditionally the pandemic spread is described by SIR (Susceptible-Infected-Removed) model with the help of non-linear differential equations [1]

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \alpha I, \quad \frac{dR}{dt} = \alpha I, \tag{1}$$

where: S, I, and R denote susceptible, infected, and removed individuals, respectively, while parameters β is the transmission rate, and α is the recovering rate. The classical Susceptible-Infected-Removed model was proposed in [1], which can take under consideration laws and density using aforementioned β and α parameters (as time-variant variables). Several variants of the SIR model were proposed: spatial-temporal dependence [2], quarantine model [3] or stochastic SIR model [4].

Previous study [5] showed the significance of dynamic models with continuously distributed delays when modeling population dynamics in epidemiology. It was numerically evidenced [6] that such type of dynamics is the most adequate for COVID pandemic data.

One of the most effective prophylactic mean for COVID-19 is vaccination. In [7] the authors proposed the vaccination strategy using time-varying linear optimization-based approach to disseminate vaccines among zones. The solution is based on SEIRD model to learn the extent of immunity provided by the vaccines. The model was verified on infection data of New York State. Big data approach provides a means to help contain COVID-19 epidemic. Using AI analytics, big data allows us to easier monitor the process [8]. This approach is presented as a mean to understand and manage the pandemic spread.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

Some attempts were done to apply ML techniques for SIR-modeling. So, in [9] authors proposed various data sources (e.g., dead cases and weather) to predict the risk level for the country. The results showed that models with shallow long short-term memory [10] (e.g., LSTM) proved useful in this case and obtained average accuracy of 78% for over 170 countries.

For the purpose of *diagnostics* and *treatment* of COVID-19 several ML models were proposed [11] to more precisely model a human behaviour. Authors in [12] proposed a useful model to estimate COVID-19 severity using the Harris hawks optimization (HHO) to optimize the Fuzzy K-nearest neighbor (FKNN). The approach provides acceptable and stable solution for group of research patients. Several statistic approaches were also researched as Bayesian non-linear model in [9], Prophet forecasting procedure and autoregressive integrated moving average (ARIMA) model in [10].

2) BIG DATA SOLUTIONS

The last decade success of big data technology, supported by AI in various areas, bring researchers to use this approach also for COVID-19 coronavirus disease distribution modelling and treatment. The growing COVID-19 pandemic changed a local problem to global scale with its high volume, velocity and variety of sources and available methods. The various systems generate massive data volumes (using IT techniques and web technology to) and together with AI support allows researchers making more reliable multiple source big data pipelines [13]. In case of COVID-19 case, useful Big Data can be generated from social media, smartphones, IoT sensors and public data repositories [14].

Authors in [15] is showing that big data opens a possibility for ubiquitous health management and advances in biomedical technologies. To create this big data sets new sensing and data mining technologies should be create to facilities rich physiological and behavioral states of humans. The data format varies and uses text (relation, non-relation and event data), images or videos form. Various data like Computer Tomography (CT) and X-ray was utilised using deep learning techniques to more precisely diagnose each patient [16]. Despite many studies in this area, their analysis [17] showed large uncertainties (methodological flaws and many biases) which have to cope with in order to apply the results in practice.

The paper [18] shows the issue of preparing the data for further analysis - remove private information using Natural Language Processing (NLP) and deep learning techniques. The new proposed approach allows to capture latent syntactic and semantic similarities more efficiently and use data in further processing. Additionally, data format covers both structured or unstructured data. The pandemic is evolving and with each day new data are generated for health monitoring or diagnosis purposes [19]. In [20] the optimized machine learning-based framework using inpatient's facility data was presented, that will give a user-friendly, cost-effective, and time-efficient solution to this pandemic.

The Big Data approach was already successfully applied by the scientist in [21] to select existing medicines, which could be applied for COVID-19 disease treatment. Additionally, real time model for infection prediction across a China was proposed in [22]. The IBM presented cloud-based Big Data research for modelling and drug discovery based on the COVID-19 dataset [23].

The nature language processing using Twitter Big Data and Apache Spark was proposed in [24] and [25]. Using this source only and word analysis using unsupervised Latent Dirichlet Allocation (LDA) machine learning and other methods authors detect 15 government pandemic measures and public concerns and six macro-concerns and found various relationships. In second paper authors used Naive Bayes, Logistic Regression, and multiple feature extraction methods to detect various diseases in the in Saudi Arabia e.g. heart diseases or cancer. Finally, in [26] the technology-driven framework for coordinated and autonomous pandemic management was proposed to propose management plan based on case strategies. The authors stress a need of multiple data source to obtain reliable results.

The high volume of data requires high-performance computing, which allows to process data in efficient and simple way [27]. Unfortunately, machine learning algorithms for Python, MATLAB, and C/C ++ are not optimised for this task. Thus, in this paper the Julia language for high-performance machine learning was utilised as it is offering parallel GPU processing (also in cloud) and is sixteen times faster than corresponding serial version and proved more efficient than implementation in Python [28].

It is crucial when estimating parameters of infinite-dimensional dynamic systems to fit time series, like epidemic curves, to solve the thousands of differential equations during nonlinear optimization. Besides that Julia is compatible with ML tools accessible through R and Python languages, which aggregate data for COVID Big Data Hub [29].

Thus, in this research the high-performance model with delay characteristic was proposed to model the disease evolution in specific areas.

B. RESEARCH GAP, OBJECTIVES, AND CONTRIBUTIONS

For the reasons given, there are many solutions that take into account one source of data, like health data or social data. However, there are very few solutions which utilise multiple Big Data sources. Preferably, they focus on initial stages of Big Data processes, while not taking into consideration complete Big Data pipeline, including collecting, aggregating, processing, and digesting stages.

Also, we see the essential gap of research which offer methods and tools to use mathematical models based on infinite dimensional systems (like the models with continuously distributed delays) to various Big Data sources, considering variety of population dynamics processes, virus strains etc.

One of the most important problems for application such complex dynamic systems is the lack of validated and practically tested techniques of the parameter estimation,

TABLE 1. Notations used through the paper.

$\mathbf{R}_+$	nonnegative real numbers,
$[\cdot]^+$	the positive part of a real or extended real-valued function,
$\tau_M > 0$	the largest value of delays considered,
$\mathtt{E}(\tau)$ and $\mathtt{Var}(\tau)$	mean value and variance of random variable τ,
$\mathbf{C}^1[-\tau_M, 0]$	Banach space of continuously differentiable functions on $[-\tau_M, 0]$,
$\mathbf{C}^+[-\tau_M, 0]$	Banach space of continuous nonnegative functions $x(t)$ for $t \in [-\tau_M, 0)$ and $x(0) > 0$,
$\mathbf{C}^1([-\tau_M, 0], \mathbf{R}^m)$	Banach space of continuously differentiable vector-functions mapping $[-\tau_M, 0]$ onto $\mathbf{R}^n$, $m \in \mathbf{N}$,
x_t or $(x)_t$, $t > 0$	interval of function values $\{x(t+\tau), \tau \in [-\tau_M, 0]\}$ of the function $x(t) \in \mathbf{C}^1[-\tau_M, 0]$.

qualitative behavior, especially in the case of Big Data, known as SciML. On the other hand, parameter estimation for dynamical systems is dealing with NP computational complexity. Unfortunately, we have a lack of research dealing with the possibilities of high-performance computing for such problems.

Hence the main objectives of the work include:

- developing the general pipeline enabling application of infinite-dimensional dynamic models for the problems of SciML in case of Big Data;
- offering two-strain model based on delay dynamic system;
- implementing high-performing solution of the parameter estimation problem in case of gamma distribution of delays;
- investigation of Big Data solution based on the data from COVID-19 Big Data Hub at different spatial levels of data.

The basic contribution of the paper to the field of knowledge can be stated as proposing the two-strain dynamic models with distributed delays that have been fitted to the time series retrieved from COVID data hub. Moreover, it is included to the general Big Data pipeline. In addition, such solution is based on high-performance computing.

The notations used in the paper are presented in Table 1.

The work is organized as follows. Section II-A describes the Big Data pipeline and the variety of COVID data. Section II-B offers the COVID Delayed Dynamic Systems model and an algorithm SciML. The algorithm for tuning model parameters is presented in detail. Section II-C focuses on particularities of high-performance computing within problem-solving. In Section III we present numerical results in case of different levels of administration level of COVID data. In Conclusion, we are focusing on interpreting the results and some open problems.

II. RESEARCH METHODOLOGY AND DESIGN

The proposed model is based on COVID-19 Big Data Hub [29]. The considered pipeline utilizes the data from the Hub to generate the COVID model and to produce the reliable prediction as it was presented in Fig. 1.

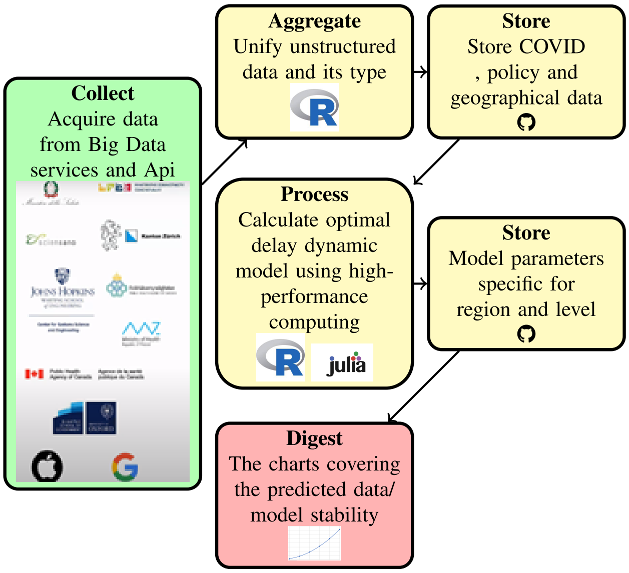

Collect
Acquire data from Big Data services and Api
Aggregate
Unify unstructured data and its type
Store
Store COVID , policy and geographical data
Process
Calculate optimal delay dynamic model using high-performance computing
julia
Store
Model parameters specific for region and level
Digest
The charts covering the predicted data/ model stability

FIGURE 1. Proposed big data pipeline for COVID delayed dynamic systems model.

The method is utilizing the data from various sources and processing them with the help of infinite-dimensional dynamic model with continuously distributed delays.

Each step of the proposed pipeline is presented in the following subsections.

A. COVID BIG DATA DESCRIPTION

The first three elements of a pipeline are provided by COVID-19 Big Data Hub project [29]. The data are collected in various formats (API, batches and streams) from various repositories. In is worth to note that each repository is also the Big data solutions(e.g. [30]). Then collected large volumes of data are processed by the R package aggregation procedures executed on dedicated server every hour. The data are then stored as batches using GitHub repository. Additionally, due to terms of use of some data (World Bank, Google, and Apple) are generated on request from original repository and merged with stored data-set in real-time. During unification procedure, on collecting stage, the data quality is verified by tracking all sources, its documentation and using open-source community. Currently the data are aggregated from 649 different data sources of various types. The data are available at different levels of granularity (countries, regions and cities/counties). Up to date 193 countries, 501 regions and 3908 counties are available. The data covers three major areas:

- population information: deaths number, confirmed cases, tests number, vaccines number, recovered number, hospitalized (with ICU or ventilation) and total number;
- policy level (0: No measures, 1: low restrictions and 2: high restrictions) for: school closing, workplace closing, cancel events, gatherings, transport closing, stay

home restrictions, internal movement restrictions, international movement restrictions, information campaigns, testing policy, contact tracing;

- and localization: longitude and latitude of specific areas.

The selected data are then used to produce the model using next processing script defined using R module and high-computation methods. The models for specific area are then stored for presentation purposes. The contribute elements, with is implementation, were presented in next section.

B. DATA-DRIVEN MODELING

1) THE COMING TO MODELS WITH DELAY

In [31] it was investigated the simplification of the model of coexistence of two virus strains. The model is assigned for describing a spread of various virus strains (for example, pandemic and seasonal influenza). In the model there were made the assumptions: 1) compartments of latent persons are not considered; 2) influenza spread is assumed to be necessarily accompanied by the symptoms, i.e., the absence of compartments of the asymptotically infected; 3) a general size of population N is considered constant.

In [32] it was generalized to a nonstationary system with discrete two discrete temporal delays which describe latent periods for infectiousness. In [5] it was shown experimentally that COVID studying requires using distributed delays due to normal distribution law. So, the given paper applies continuously distributed delays for modeling the COVID pandemic. Pursuing the goal of mapping population processes as well as possible, here we have chosen to use gamma distribution of delays $\tau \geq 0$ as

$$\psi(a, m, \tau_{\min}, \tau) := \begin{cases} 0 & \tau \leq \tau_{\min}, \\ \dfrac{a^{m+1}}{\Gamma(m+1)}(\tau-\tau_{\min})^m e^{-a(\tau-\tau_{\min})} & \tau > \tau_{\min}, \end{cases} \quad (2)$$

where $a, m \geq 0$. The distribution was considered earlier in a different context, for example describing cell maturation times ([33], pp.240-243), where an efficient method of distribution parameter estimation based on experimental population data was offered. Besides that, the non-symmetricity of gamma distribution fits the processes of infectiousness better as compared with the symmetric normal distribution.

Let τ_M be the largest value of the delay considered. Assuming τ is a random variable which is gamma distributed given by ψ, we can estimate its confidence interval with confidence level $c \in (0, 1)$, with the help of applying Chebyshev's inequality to a gamma distribution, resulting in determining the largest value of τ as

$$\tau_M := \mathtt{E}(\tau) + \sqrt{\frac{\mathtt{Var}(\tau)}{1-c}} = \tau_m + \frac{m+1}{a} + \sqrt{\frac{(m+1)}{a^2(1-c)}} \quad (3)$$

2) A TWO-STRAIN MODEL WITH DISTRIBUTED DELAYS

Here we consider two-strain model of COVID growth within the population, which takes into account the following sub-populations: S, sub-population of persons susceptible to both virus strains; I_k, $k = 1, 2$, sub-population of persons which are infected with the kth virus strain; R_k, $k = 1, 2$, sub-population of persons which are recovered after kth virus strain (we suppose that the corresponding immunity has been obtained), but they are susceptible to the jth virus strain, $j = 1, 2$, $j \neq k$; Y_k, $k = 1, 2$, sub-population of persons, which are recovered as a result of jth virus strain, $j \neq k$, currently they are infected with the kth virus strain; R, sub-population of persons which are recovered after being infected with both virus strains. Let N be the total population size. Since $S + \sum_{k=1,2}\left\{I_k + R_k + Y_k\right\} + R = N$, we may omit the sub-population R.

Here we give brief explanations for the parameters used.

- There is a recruitment rate μN into the susceptible class.
- The natural death rate of all sub-populations is μ.
- The rate of having primary infectiousness with the kth virus strain is $\beta_k S I_k$, $k = 1, 2$.
- Susceptible persons are being infected with the kth strain ($k = 1, 2$) of the virus with the gamma-distributed delay τ given by the density
$$\psi_k(\tau) := \psi(a_k, m_k, \tau_{\min,k}, \tau), \quad (4)$$
where a_k, m_k, $\tau_{\min,k}$ are some positive parameters of gamma distribution, $\tau_{\min,k}$ being minimal value of latent period.
- The recovered rates for sub-populations I_k or Y_k are α_k, $k = 1, 2$.
- The rate of having secondary infectiousness with the jth virus strain is $\sigma_j \beta_j R_k I_j$, $k, j = 1, 2$, $k \neq j$.
- The parameter σ_j describes an increase or decrease of susceptibility to strain j as of secondary infection due to possible immune altering as a result of primary infections.
- The delay of secondary infectiousness with the jth virus strain is given by the density $\psi_j(\tau)$, $j = 1, 2$.
- In practice, all rate parameters appeared to be dependent on time. So we consider them as time-varying functions.

The general scheme of the model considered is depicted in Fig. 2. Here $\Psi_k[(x)_t] := \int_{-\tau_{M,k}}^{0} \psi_k(\tau)x(t+\tau)d\tau$ is the distributed delay functional, acting on functions $x_t \in \mathbf{C}^1[-\tau_{M,k}, 0]$, $k = 1, 2$.[1]

For the parameters given above, we lead to the following delayed dynamic system at $t > 0$

$$\begin{aligned} S'(t) &= \mu(t)(N(t) - S(t)) - \sum_{i=1,2} \beta_i(t)S(t)I_k(t), \\ I_k'(t) &= \beta_k(t)\Psi_k[(S(\cdot)I_k(\cdot))_t)] - (\mu(t) + \alpha_k(t))I_k(t), \\ R_k'(t) &= \alpha_k(t)I_k(t) - (\mu(t)R_k(t) + \sigma_j\beta_j(t)I_j(t))R_k(t), \\ Y_k'(t) &= \sigma_k\beta_k(t)\Psi_k[(R_j(\cdot)I_k(\cdot))_t] - (\mu(t) + \alpha_k(t))Y_k(t), \end{aligned} \quad (5)$$

where $k, j = 1, 2$, $k \neq j$.

For the solutions of (5), elements of which are the vector-functions

$$(S, I_1, I_2, R_1, R_2, Y_1, Y_2) \in \mathbf{C}^1([-\tau_{\max}, 0], \mathbf{R}^7),$$

[1] it follows from the properties of $\psi_k(\tau)$ that $\Psi_k[(x)_t] = \int_{-\tau_{M,k}}^{-\tau_{\min,k}} \psi_k(\tau)x(t+\tau)d\tau$

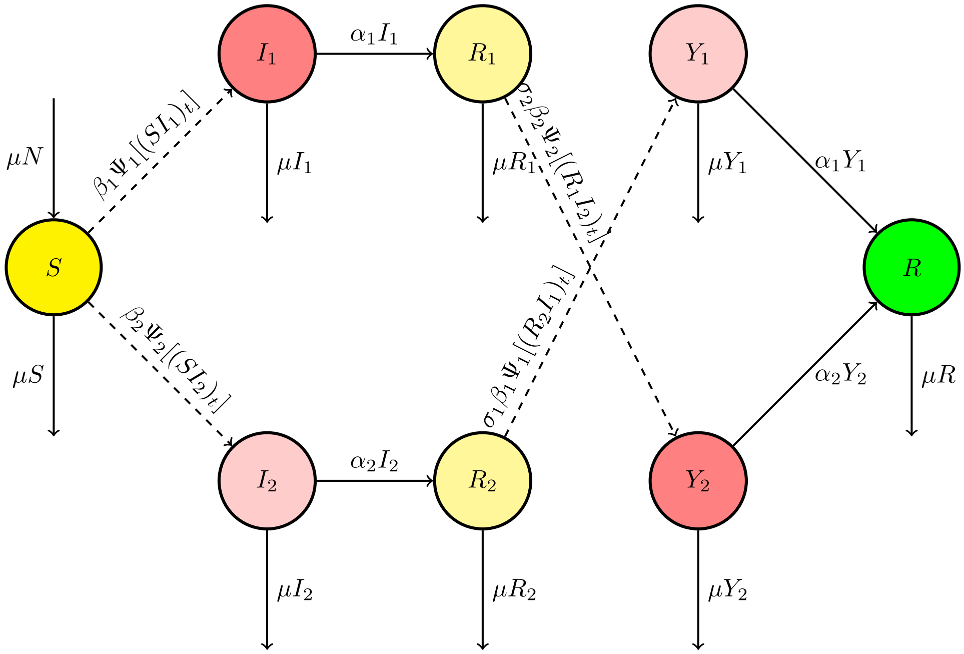


FIGURE 2. Flowchat of the model of coexistence of two COVID strains. The transmissions with delays are marked with dashed arrows. The subpopulations affected by the first strain are at the top, whereas affected by the second one at the bottom.

we consider the initial conditions

$$\begin{aligned} S(t) &= \hat{S}(t) \geq 0, I_k(t) = \hat{I}_k(t) \geq 0, \\ R_k(t) &= \hat{R}_k(t) \geq 0, Y_k(t) = \hat{Y}_k(t) \geq 0, \\ t &\in [-\bar{\tau}_{\max}, 0) \\ S(0) &= \hat{S}^0 > 0, I_k(0) = \hat{I}_k^0 > 0, \\ R_k(0) &= \hat{R}_k^0 > 0, \quad Y_k(0) = \hat{Y}_k^0 > 0, \ k = 1, 2, \end{aligned} \tag{6}$$

where $\tau_{\max} = \max\left\{\tau_{M,k}, k = 1, 2\right\}$.

Given any $\hat{S}(t), \hat{I}_k(t), \hat{R}_k(t), \hat{Y}_k(t) \in C^+[-\tau_{\max}, 0]$, since the right-hand sides of (5) imply Lipschitz condition, there exists a unique trajectory of (5) starting from (6) [34].

3) SCIENTIFIC MACHINE LEARNING

The development of models based on data from COVID data hub can be considered as a problem of scientific machine learning (SciML) [6], [35]. This direction has received currently increased attention since it aims to apply modern machine learning techniques in science, engineering, and medicine, which differs from traditional data-driven machine learning problems in computer science [36].

Let time series $\left\{X_{\exp}(t_j)\right\}_{j=1}^n$ be data from COVID data hub corresponding to instances of time $t_1, t_2, \ldots, t_n$.[2] Fig. 3 presents the COVID modeling as the construction of a machine learning model basing on COVID data hub. Namely, the model constructed can be used for various SciML tasks. We classify the basic of them as a quantitative prediction of the state, investigation of the qualitative behavior of the trajectories, bifurcation analysis, and studying the chaotic dynamics. Nevertheless, the solution of the SciML problem can be presented in the form of a dynamic system with a known closed set of parameters.

We will seek data-driven model (5), (6) in the class of functional-differential equations

$$\frac{dX(t)}{dt} = F(X(\cdot), \Pi), \tag{7}$$

where $\Pi \in \mathcal{P}, F : \mathbf{C}^1[-\tau_{\max}, 0] \to \mathbf{R}^l$ is functional that maps corresponding Banach space of continuously differentiable functions into $\mathbf{R}^l$, $l \in \mathbf{N}$.

We denote the sought data-driven model which is described by (7) as $\mathcal{F}_\Pi$ and its solution as $X(t, \Pi)$ or $X_{\text{pred}}(t)$.[3]

Note that Fig. 3 focuses coping with uncertainties in SciML. Namely, aleatoric uncertainties, which are related with the data, have to be coped with the help of choosing such distribution of experimental data which maximize the loss function of the model $\mathcal{F}_\Pi$. In turn, to cope with epistemic uncertainties, we choose the model designed in a way enabling us minimal values of the loss function for the "worst" distributed data.

4) PARAMETERS IDENTIFICATION

The model (5) is the most affordable for the analysis of COVID pandemic time series. The idea is to use this model for estimating the epidemic curves with the respect to various sub-populations. In turn, these values correspond to the initial conditions of $\hat{S}$, $\hat{I}_1$, and $\hat{R}_1$, which are known from the

[2] Note that $X_{exp}(t)$ includes the raw of COVID data at time t, e.g. $X_{\exp} = \left(S_{\exp}(t), I_{\exp}(t), R_{\exp}(t)\right)^\top$

[3] hereinafter we use subscript 'pred' for the corresponding solutions of (5), (6) or (7)

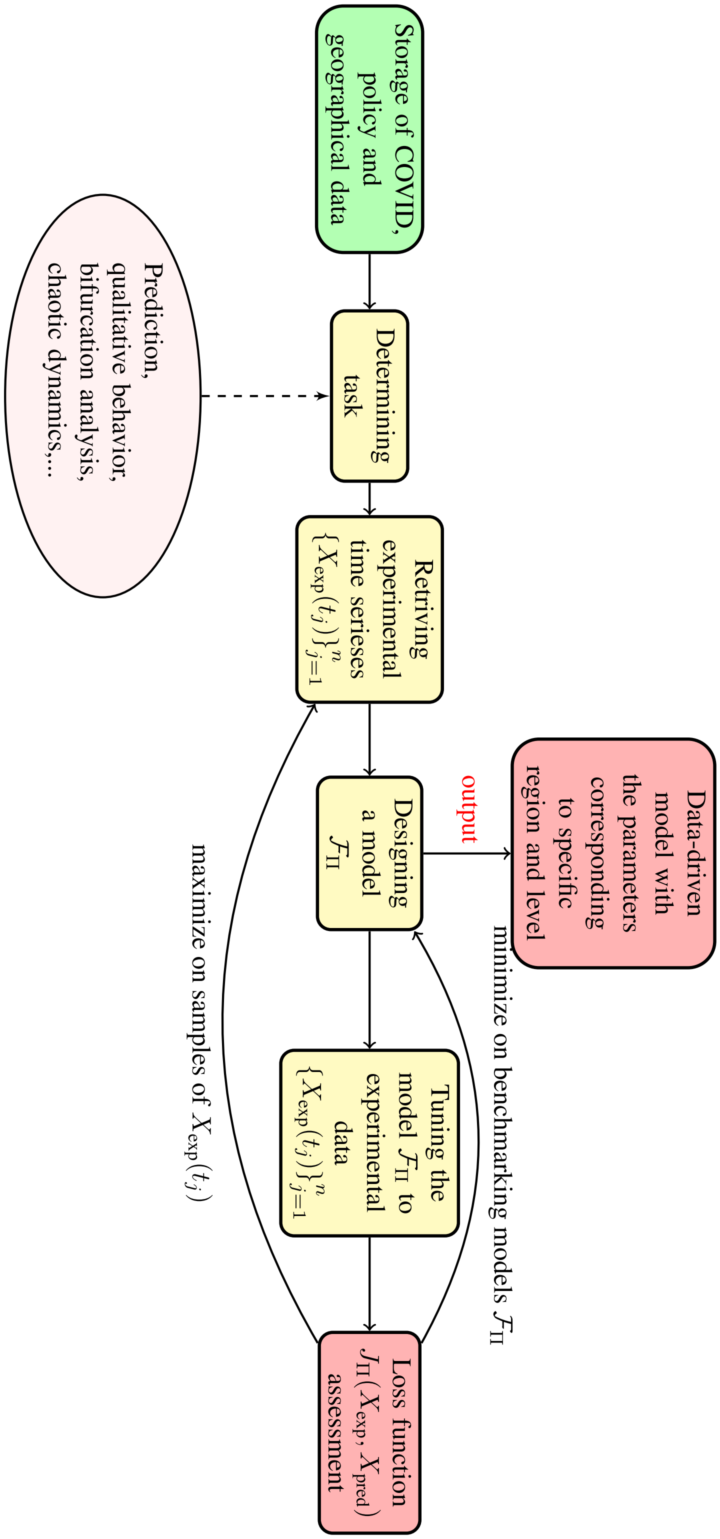


FIGURE 3. Development of machine learning model based on functional differential equation (7).

time series in advance. Estimation should include the rate constants, gamma distribution parameters within ψ function, and unknown initial values contained in the system given by (5), (6).

In order to adjust the predicted data, we need to do them compatible with the data accessible from COVID data hub ("expected data") introducing the denotations for the solutions obtained from (5), (6) as follows

$$\begin{aligned} I_{\text{pred}}(t) &:= \sum_{k=1,2} \left\{ I_{k,\text{pred}}(t) + Y_{k,\text{pred}}(t) \right\}, \\ R_{\text{pred}}(t) &:= N(t) - S_{\text{pred}}(t) - I_{\text{pred}}(t), \end{aligned} \tag{8}$$

Hereinafter we let constant values of $\mu(t)$ and $N(t)$, i.e.,

$$\mu(t) \equiv \mu, N(t) \equiv N, \quad \mu, N \in R_+, \tag{9}$$

whereas $\beta_k(t)$, $\alpha_k(t)$ being periodic functions of the form

$$\begin{aligned} \beta_k(t) &= \beta_k + \tilde{\beta}_k \sin\left(\frac{2\pi}{\omega_{\beta_k}} t\right), \\ \alpha_k(t) &= \alpha_k + \tilde{\alpha}_k \sin\left(\frac{2\pi}{\omega_{\alpha_k}} t\right), \quad k = 1, 2, \end{aligned} \tag{10}$$

where $\beta_k, \tilde{\beta}_k, \omega_{\beta_k}, \alpha_k, \tilde{\alpha}_k, \omega_{\alpha_k} \in R_+$.

Such choice of the form (10) of the rates of infectiousness and recovering is evidenced by the seasonal character of the pandemic growth, which influences the rates significantly and is related to so-called "epidemic waves". Here ω_{β_k}, ω_{α_k}, $k = 1, 2$ are corresponding periods of epidemic processes in days.

Hence the COVID two-strain model (5), (6) under assumptions (9) and (10) depends on 27 unknown parameters; namely

$$\Pi = \left\{ \begin{array}{l} \mu, \beta_1, \tilde{\beta}_1, \omega_{\beta_1}, \alpha_1, \tilde{\alpha}_1, \omega_{\alpha_1}, a_1, m_1, \tau_{\min,1}, \\ \beta_2, \tilde{\beta}_2, \omega_{\beta_2}, \alpha_2, \tilde{\alpha}_2, \omega_{\alpha_2}, a_2, m_2, \tau_{\min,2}, \sigma_1, \sigma_2, \\ \hat{I}_1^0, \hat{R}_1^0, \hat{I}_2^0, \hat{R}_2^0, \hat{Y}_1^0, \hat{Y}_2^0 \end{array} \right\} \in \mathbf{R}_+^{27} \tag{11}$$

In principle, this set of parameters can be estimated from a given time series of the COVID pandemic. For a set of n pointwise experimental data in the time series, say $\left\{S_{\exp}(t_j)\right\}_{j=1}^n$, $\left\{I_{\exp}(t_j)\right\}_{j=1}^n$, $\left\{R_{\exp}(t_j)\right\}_{j=1}^n$ with $t_1, t_2, \ldots, t_n$ being the times of observations, the identification of parameters can be carried out with the following constrained optimization calculations that are expressed in the form

$$\left. \begin{array}{ll} \texttt{minimize} & J(\Pi), \quad \Pi \in R_+^{27} \\ \texttt{subject to} & g_i(\Pi) \geq \Theta, \quad i = 1, 2, \\ & g_j(\Pi) \geq 0, \quad j = \overline{3, 8} \end{array} \right\} \tag{12}$$

Here

$$\begin{aligned} J(\Pi) := \Bigg(\sum_{j=1}^{n} \Big(& (S_{\exp}(t_j) - S_{\text{pred}}(t_j))^2 \\ & + (I_{\exp}(t_j) - I_{\text{pred}}(t_j))^2 \\ & + (R_{\exp}(t_j) - R_{\text{pred}}(t_j))^2 \Big) \Bigg)^{1/2} \end{aligned} \tag{13}$$

is the objective function and

$$\begin{aligned} g_1(\Pi) &= \Pi - \Pi_{\text{lower}} \geq \Theta, \\ g_2(\Pi) &= \Pi_{\text{upper}} - \Pi \geq \Theta, \\ g_3(\Pi) &= N - S_{\exp}(0) - \hat{I}_1^0 - \hat{R}_1^0 - \hat{I}_2^0 - \hat{R}_2^0 - \hat{Y}_1^0 - \hat{Y}_2^0 \geq 0, \\ g_4(\Pi) &= S_{\exp}(0) + \hat{I}_1^0 + \hat{R}_1^0 + \hat{I}_2^0 + \hat{R}_2^0 + \hat{Y}_1^0 + \hat{Y}_2^0 - N \geq 0, \\ g_5(\Pi) &= I_{\exp}(0) - \hat{I}_1^0 - \hat{I}_2^0 \geq 0, \\ g_6(\Pi) &= \hat{I}_1^0 + \hat{I}_2^0 - I_{\exp}(0) \geq 0, \\ g_7(\Pi) &= R_{\exp}(0) - \hat{R}_1^0 - \hat{R}_2^0 \geq 0, \\ g_8(\Pi) &= \hat{R}_1^0 + \hat{R}_2^0 - R_{\exp}(0) \geq 0 \end{aligned} \tag{14}$$

are inequality constraints, where $\Theta \in R^{27}$ is null-vector.

The offered solution of nonlinear optimization problem (12) is based on the COBYLA algorithm [37], which linearly approximates objective function and constraints on 27-simplex $C = C(\Pi_0, \Pi_1, \ldots, \Pi_{27})$ and optimize the simplex on each algorithm iteration. The algorithm transforms problem (12) to the problem without constraints with the help of objective function

$$\Phi(\Pi) := J(\Pi) + \xi [\max\left\{-g_i(\Pi), i = \overline{1, 8}\right\}]^+. \tag{15}$$

We denote its linear approximation on the simplex C as $\hat{\Phi}_C(\Pi)$. An implementation of COBYLA to the problem (12) can be reformulated as the Algorithm 1. Here *stop condition* covers the improvement of objective function, the changes of vertexes and allowed number of iterations.

Algorithm 1: COBYLA Algorithm Implementation to the Problem (12)

```
Input data: $X_{\exp}$, $\Pi_{\text{lower}}$, $\Pi_{\text{upper}}$, $\Pi_{\text{init}}$
Result: $\Pi_{\text{opt}}$
form the initial simplex $C_{\text{init}}$ with the vertices
  $\Pi_0^{\text{init}}, \Pi_1^{\text{init}}, \ldots, \Pi_{27}^{\text{init}}$;
repeat
    for the current simplex $C$ calculate the values
      $\hat{\Phi}_C(\Pi_i), i = \overline{0, 27}$;
    search the vertex $\Pi_l$ determined by the equation
      $\hat{\Phi}_C(\Pi_l) = \min\left\{\hat{\Phi}_C(\Pi_i), i = \overline{0, 27}\right\}$;
    calculate new vertex as
      $\Pi_{new} := -\theta \Pi_l + (1 + \theta) \frac{1}{27} \sum_{i=\overline{0,27}, i \neq l} \Pi_i$, where
      reflection coefficient $\theta \in (0, 1)$ being chosen as
      small as possible in order $\hat{\Phi}_C(\Pi_l)$ not were the
      least calculated function value so far;
    form modified simplex $C_{new}$ replacing vertex $\Pi_l$
      with $\Pi_{new}$;
    search $\Pi_{opt}$ as a solution of the problem of linear
      optimization
          $\left. \begin{array}{ll} \texttt{minimize} & \hat{\Phi}_{C_{\text{new}}}(\Pi), \\ \texttt{subject to} & \Pi \in C_{\text{new}} \end{array} \right\}$    (16)
until stop condition;
return $\Pi_{opt}$;
```

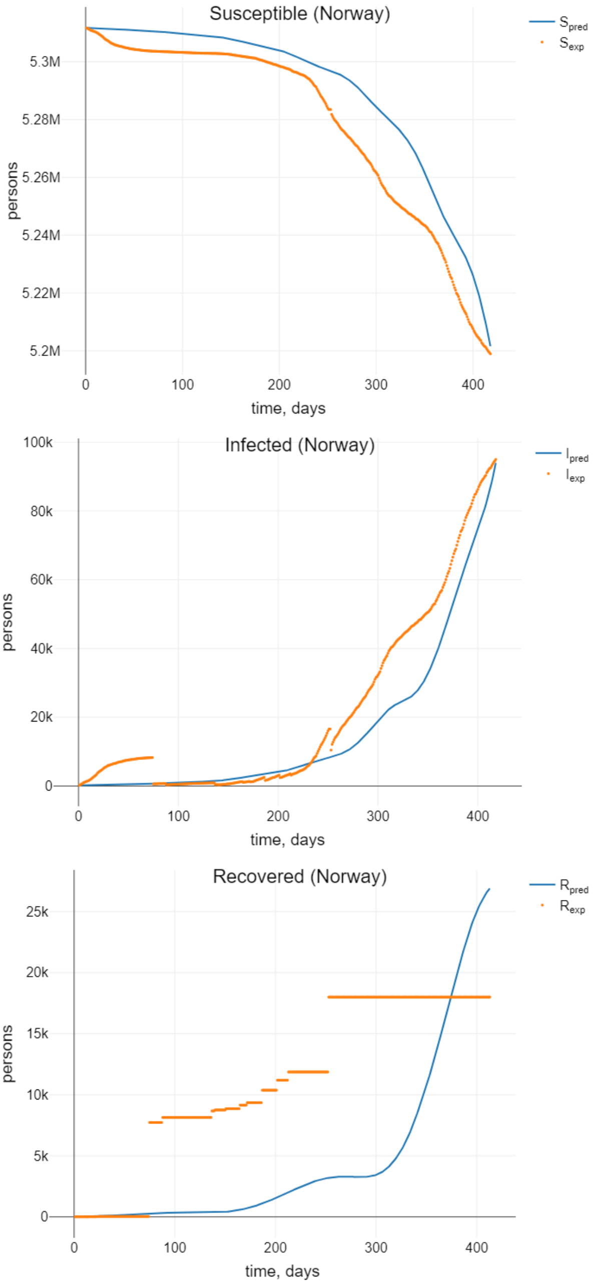


FIGURE 4. Data-driven modeling at country level. Model fitting for Norway.

C. HIGH-PERFORMANCE CALCULATION

In practical development of machine learning model based on COVID data hub data due to Fig. 3, we have used implementation of basic steps of COBYLA related to the forming simplexes and linear approximation within the package `nloptr`, whereas objective function and constraints have to be developed directly.

There are few reasons causing the need for high-performance computation in COVID modeling. On one hand, hereditary models based on differential equations with delay are infinite-dimensional systems that require saving more prehistory states that are used to calculate the next ones. On the other hand, the problem of fitting rate parameters for delayed dynamic systems includes a lot of degrees of freedom, which also leads to the NP-hard problems. Moreover, analysis of qualitative behavior of such type nonlinear models which is traditionally performed with the help of various numerical characteristics results in the numerical integration of the model for a huge number of points in the multidimensional space of the parameters.

The high-performance computations have been implemented in Julia language, which can execute a huge amount of calculations whereas assuring high performance. Since COVID modeling is oriented on cloud computing, Julia facilities supporting clouds and parallel programming are essential. In order to enable high-performance computing through R platform, package `diffeqr` was used for solving delay differential equations [38]. Its core routines are implemented on the basis of suite `DifferentialEquations.jl`, which is developed for SciML enabled simulation and estimation. It enables us high performance solving of equations like (3) directly within R. [39].

The listing implementing the model (5), (6) given parameters (9) and (10) is written in Julia with the help of R package `diffeqr`.

III. EXAMPLES FOR DATA AT DIFFERENT LEVELS

Trying to evaluate the models (5), we were tuning it for training data at various level.

A. COUNTRY LEVEL

Here we use the time series of COVID growth for Norway at $n = 418$ (Fig. 4).

When applying COBYLA algorithm, we let the values of parameters Π needed to form the initial simplex C_{init} as shown in Table 2. The number of evaluations was used as a stop condition. As a result of 200 iterations the value of objective function was 597914.3. Integer code `convergence` $= 5$, which indicates successful completion of the algorithm. The estimated values of the parameters are shown at the column $\Pi_{\text{est}}^{\text{Norway}}$.

B. REGIONAL LEVEL

We investigate time series for Texas (USA) at $n = 343$ (Fig. 5). Initial simplex is constructed based on the same parameters from Table 2. As a result of 600 iterations the value of objective function was 5038572 at successful completion. Column $\Pi_{\text{est}}^{\text{Texas}}$ shows the estimated values of the parameters.

C. COUNTY LEVEL

Time series from SK Speyer (Rheinland-Pfalz, Germany) was investigated (Fig. 6). The value of objective function is 16606.81 for $n = 304$ time instances. The values of parameters Π to construct simplexes and the parameters estimated $\Pi_{\text{est}}^{\text{Speyer}}$, are presented in Table 3. Because of significantly

TABLE 2. The set of parameters used related to the model (5) at the level of country and region.

	Π_{init}	Π_{lower}	Π_{upper}	$\Pi_{\text{est}}^{\text{Norway}}$	$\Pi_{\text{est}}^{\text{Texas}}$
μ	0.05	1e-5	0.1	5.806023e-04	3.240548e-04
β_1	4e-9	1e-9	9e-9	8.620473e-09	2.318532e-09
$\tilde{\beta_1}$	2e-9	1e-9	9e-7	2.624741e-09	2.991082e-09
ω_{β_1}	145	5	195	1.866038e+02	1.917545e+02
α_1	0.02428	0.01428	0.09428	3.092062e-02	1.858831e-02
$\tilde{\alpha_1}$	2e-9	1e-9	9e-7	6.743763e-03	4.781098e-03
ω_{α_1}	145	5	195	9.887024e+01	1.206686e+02
a_1	5	1	6	2.791238e+00	5.567910e+00
m_1	100	6	200	4.867431e+01	5.541286e+01
$\tau_{\min,1}$	6	5	90	1.499519e+01	9.200957e+00
β_2	3e-9	1e-9	9e-9	7.223225e-09	2.170880e-09
$\tilde{\beta_2}$	5e-3	1e-3	9e-2	1.824720e-09	1.996217e-09
ω_{β_2}	160	6	190	6.889193e+01	1.596115e+02
α_2	0.01428	0.00428	0.09428	1.122584e-02	1.427758e-02
$\tilde{\alpha_2}$	5e-3	1e-6	9e-1	4.551662e-03	4.686629e-03
ω_{α_2}	160	6	190	1.024593e+02	1.820862e+02
a_2	4	1	6	5.264204e+00	3.008174e+00
m_2	100	6	200	3.290939e+01	9.765427e+01
$\tau_{\min,2}$	9	1	80	1.879025e+01	1.893898e+01
σ_1	1.5	1.0	95	1.302885e+00	1.500971e+00
σ_2	1.5	1.0	95	1.000000e+00	1.873940e+00
$\hat{I}_1^0$	3000	0	10000	6.748109e-14	7.977431e+02
$\hat{R}_1^0$	0	0	1000	3.452372e-03	1.877187e+01
$\hat{I}_2^0$	3000	0	10000	2.040000e+02	2.229257e+03
$\hat{R}_2^0$	0	0	1000	9.965476e-01	1.922813e+01
$\hat{Y}_1^0$	0	0	0	8.540289e-14	1.636179e-18
$\hat{Y}_2^0$	0	0	0	8.543207e-14	1.323423e-18

other scale of population, both initial values and bounds of the parameters are differed from ones used for both country and regional levels. Besides that, it was evidenced essentially bigger periods for time-varying rates $\beta_k(t)$, $\alpha_k(t)$, $k = 1, 2$, with eventually follows from the smallest size of the county population.

IV. LIMITATIONS OF THE STUDY

Our study trying to offer a comprehensive approach to modeling the COVID pandemic has some limitations. Here we analyze them briefly.

In the given work we have used the deterministic model based on delayed differential equations. When studying epidemic models stochastic approach can be applicable introducing "white" noises and considering stochastic differential equations. Certainly, the form of their solutions fits better to real epidemic curves. However, the deterministic model gives an average approximation in some sense of the solution of the stochastic one.

Temporal limitations of experimental data take place. Real experimental data aggregated and stored in the COVID data hub are collected from some limited temporal interval. When using the data at the beginning of the pandemic as training data, we can get parameters of the model which differ significantly from ones obtained basing on the wider period of time. So, what is the reasonable duration of the time series to be used as COVID training data for SciML? Of course, we can observe some seasonal changes for pandemic development. For this purpose we use time-varying parameters α_k, β_k, $k = 1, 2$ in the model. In turn, the temporal limitations of experimental data influence the model significantly.

TABLE 3. The set of parameters related to the model (5) at the level of county with population size $5.0565e + 04$.

	Π_{init}	Π_{lower}	Π_{upper}	$\Pi_{\text{est}}^{\text{Speyer}}$
μ	5e-5	1e-5	0.1	2.042599e-05
β_1	2e-6	1e-6	9e-6	1.552275e-06
$\tilde{\beta_1}$	2e-9	1e-9	9e-7	4.521289e-09
ω_{β_1}	145	5	395	3.536144e+02
α_1	0.2428	0.0428	0.9428	2.573227e-01
$\tilde{\alpha_1}$	2e-9	1e-9	9e-2	9.746765e-03
ω_{α_1}	145	5	195	2.571025e+02
a_1	5	1	7	6.689514e+00
m_1	100	6	200	3.399645e+01
$\tau_{\min,1}$	6	5	90	1.103496e+01
β_2	2e-6	1e-6	9e-6	3.874197e-06
$\tilde{\beta_2}$	5e-3	1e-3	9e-2	4.693135e-09
ω_{β_2}	160	6	390	3.026446e+02
α_2	0.01428	0.00428	0.09428	8.244222e-02
$\tilde{\alpha_2}$	5e-3	1e-6	9e-1	4.104499e-03
ω_{α_2}	160	6	390	2.026553e+02
a_2	4	1	6	4.793328e+00
m_2	100	6	200	1.564635e+02
$\tau_{\min,2}$	9	1	80	6.801479e+00
σ_1	1.5	1.0	95	1.353987e+00
σ_2	1.5	1.0	95	2.164397e+00
$\hat{I}_1^0$	0	0	1000	2.112696e+00
$\hat{R}_1^0$	0	0	100	4.103215e-01
$\hat{I}_2^0$	0	0	1000	2.021346e+00
$\hat{R}_2^0$	0	0	1000	1.787402e-01
$\hat{Y}_1^0$	0	0	0	1.265140e+00
$\hat{Y}_2^0$	0	0	0	6.652395e-02

The real question is how many strains to account for. In the given model we have described the interaction between two COVID strains. The current COVID situation shows the importance of taking into account more strains (alpha, beta, gamma, delta). In turn, it requires considering more subpopulations and a more complicated flowchart of the model of coexistence of COVID strains than in Fig. 2. Moreover, when

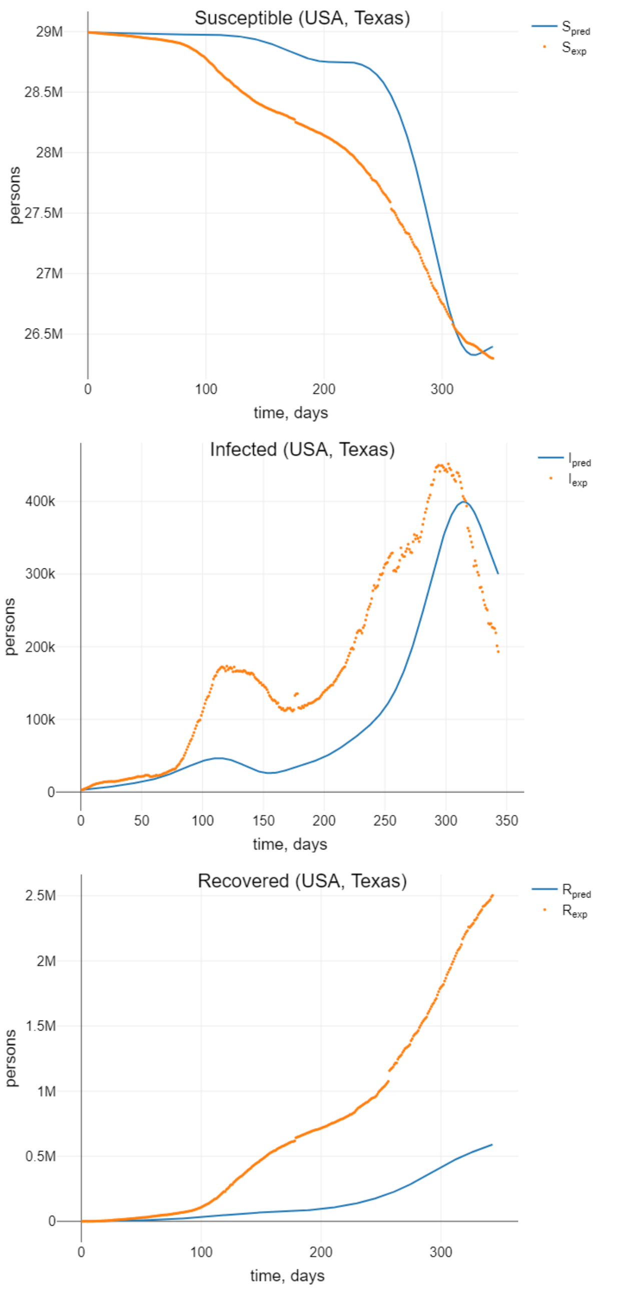


FIGURE 5. **Data-driven modeling at regional level. Model fitting for Texas (USA).**

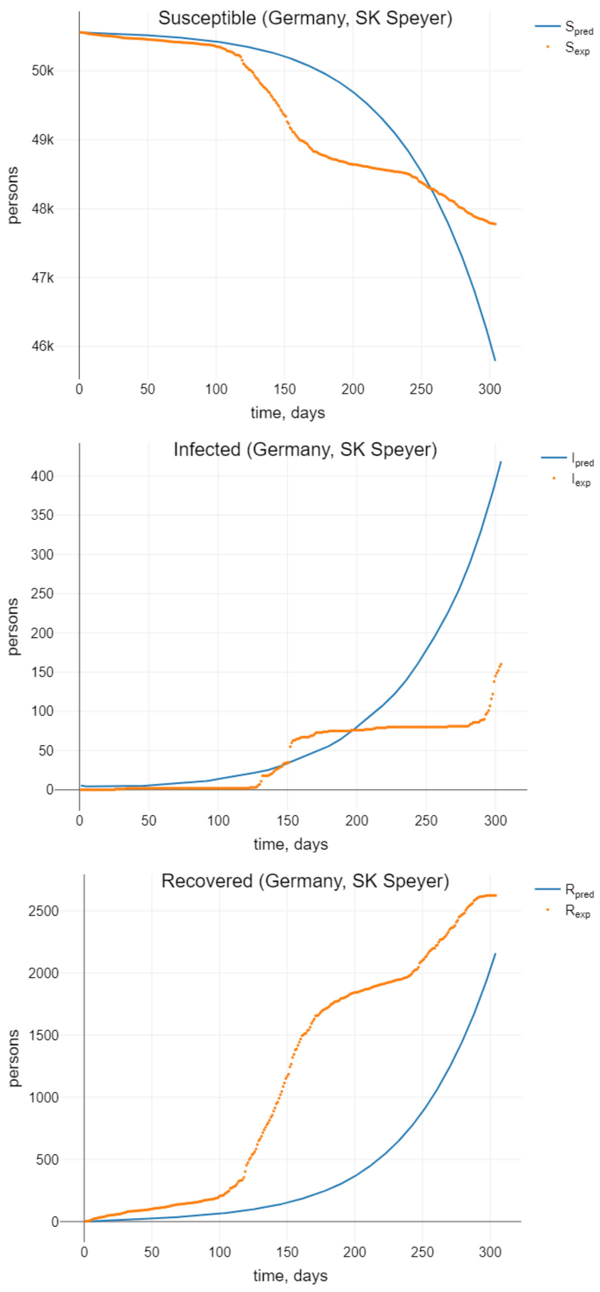


FIGURE 6. **Data-driven modeling at county level. Model fitting for SK Speyer (Germany).**

fitting the model (5) to experimental data we see that the predicted values of infected subpopulations, as a rule, are less than the expected ones. Considering additional strains can explain the shortcoming.

Limitations of biological assumptions can affect the model solutions also. Here we have considered the simplest competitive behavior of the populations basing on the law of mass action. Advanced modeling requires using more sophisticated biological mechanisms like Holling type II and III functional responses.

The model under study (5) does not include vaccination directly, using special terms or variables. On the other hand, here the vaccinated persons are combined with the recovered ones. The advanced study can construct control models and additional subpopulations of vaccinated persons.

Some special case of the model (5) with discrete delays was studied in [32]. It has shown chaotic behavior of the trajectories for some values of delays. Chaotic behavior is likely to hold for the model (5) also, which should be studied deeper.

Limitations of available data according to level on data hub (lack of some data) take place. Currently, not all data are accessible at all three levels under consideration. Eventually, there are a few reasons for the lack, including incompatibility

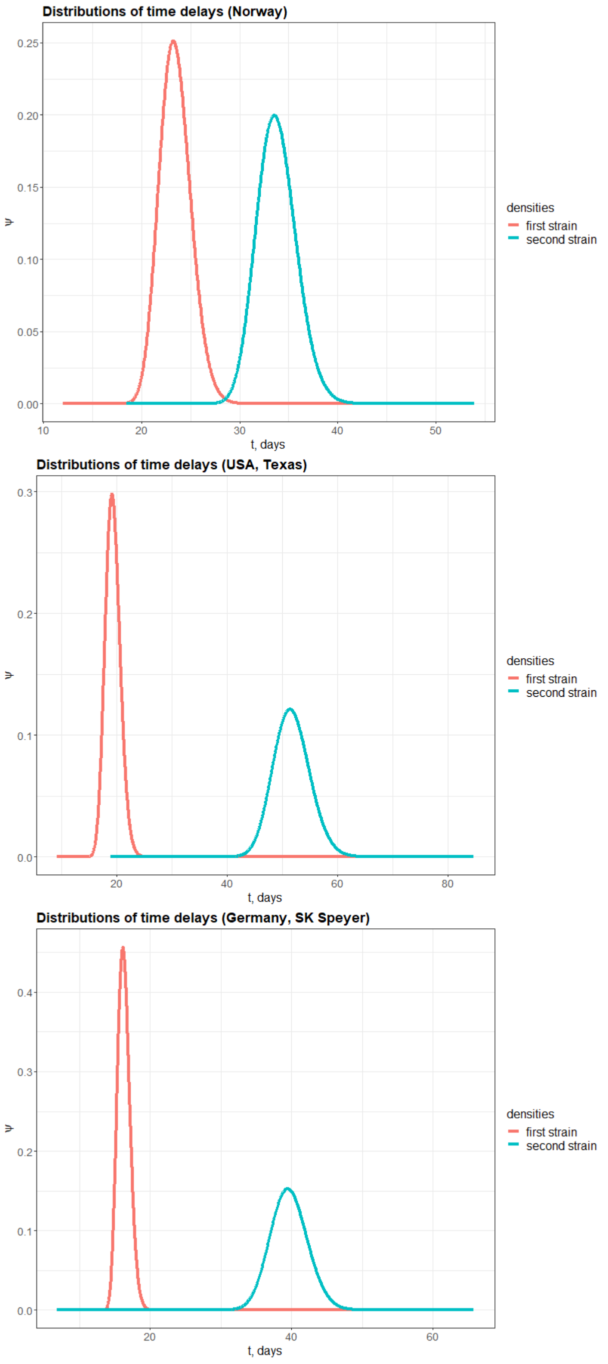


FIGURE 7. **Density distributions for data-driven models at various levels.**

of data storage of some countries, real lack of the data (primarily at level 2 and the most often at the level 3), access restrictions to public health data for some countries, etc. Currently, the situation with the lack of COVID data is really improving all over the world.

V. CONCLUSION

Processing Big Data stored on COVID data hub with the purpose of SciML has shown the importance of high-performance computing, mostly within the getting of model solutions and the tuning system parameters. For example, in the case of the basic operations of the delayed differential equations integration, using Julia parallel computing in R through diffeqr package enables us 14 times faster calculations.

Note, that the parameters fitting due to Algorithm 1 lie in most on initial guess Π_{init}. When working with a real COVID data hub we have chosen an initial guess in dependence on the size of the population. In turn, the first few steps of the bisection method would give us a suitable starting point.

Fig. 7 presents plots of delay densities at a different level of locations, being actually latent periods. Really they contain worth epidemiological indicators describing virus strains that can be extracted. $\tau_{\min}$, being left-side bound for non-zero density values indicates the minimal possible latent period for the virus strain given. Peaks of the plots show us the most probable values of latent periods. Nevertheless the level of pandemic data, we see similar distribution parameters describing the latent stage. In turn, it evidences that we observe the extension of the same virus strains.

When comparing latent periods estimated on Fig. 7 with ones of the experimental epidemiological study [40], we conclude that in population models like (5) constructed on the basis of real pandemic data latent periods will be somewhat bigger. It is caused by the fact that in practice latent period is clearly identified for symptomatically infected patients only. On the other hand, time delays in (5) are not truly latent periods. It is most likely the period to the next pathogen exposure, which, in turn, is dependent in most on the prophylaxes' actions with the respect to coronavirus extension (like quarantine). So, advancing the model (5) should take into account quarantine effects, data on which can be ingested from COVID data hub.

When analyzing plots at Fig. 4, 5, 6, we see admissible following to general tendencies of population growth, whereas numerical approximation is appeared worse. Note that fitting experimental data is better for Big Data at the country level, whereas it is worse for regional and county levels. The eventual reason is the large-scaling nature of population dynamics processes, which is more suitable at the macro level of population and is not considering individuals or small groups. Hence, the important benefits of the models (5) are seen to predict qualitative behavior of pandemic development, mainly outbreaks, at different levels.

On the other hand, better fitting to experimental data requires considering additional virus strains supported by last microbiological research.

The paper presents a Big Data good practice related to epidemiology analytics basing on modeling with the help of delayed dynamical systems. It was fulfilled within the framework of the project iBigWorld [41] which is devoted to gaining skills for developing innovative solutions based on Big Data in real-world applications. Working with COVID Big Data would be a good opportunity to get competencies when using cutting-edge technologies based on Big Data and machine learning.

ACKNOWLEDGMENT

The authors are grateful to reviewers for valuable comments that allowed to improve the paper significantly. The paper presents a Big Data good practice related to epidemiology analytics basing on modeling with the help of delayed dynamical systems. It was fulfilled within the framework of the stage "Intellectual Output 1" of the Erasmus+ project iBigWorld [41], which was devoted to the comprehensive investigation of the good practices in Big Data, including the COVID data hub. In particular, financing of management, dissemination activities, meetings, and intellectual output, which follow from this research, was conducted from the project budget. On the other hand, all activities related to preparing and publishing the paper were financed by the University of Bielsko-Biala within the framework of the discipline "Technical Informatics and Telecommunications".

REFERENCES

[1] Y. Chen, P. Lu, and C. Chang, "A time-dependent SIR model for COVID-19 with undetectable infected persons," 2020, *arXiv:2003.00122*. [Online]. Available: https://arxiv.org/abs/2003.00122

[2] K. Biswas and P. Sen, "Space-time dependence of corona virus (COVID-19) outbreak," 2020, *arXiv:2003.03149*. [Online]. Available: http://arxiv.org/abs/2003.03149

[3] N. Crokidakis, "Modeling the early evolution of the COVID-19 in Brazil: Results from a Susceptible-Infectious-Quarantined-Recovered (SIQR) model," 2020, *arXiv:2003.12150*. [Online]. Available: https://arxiv.org/abs/2003.12150

[4] G. Gaeta, "A simple SIR model with a large set of asymptomatic infectives," 2020, *arXiv:2003.08720*. [Online]. Available: http://arxiv.org/abs/2003.08720

[5] X. Liu, X. Zheng, and B. Balachandran, "COVID-19: Data-driven dynamics, statistical and distributed delay models, and observations," *Nonlinear Dyn.*, vol. 101, no. 3, pp. 1527–1543, Aug. 2020, doi: 10.1007/s11071-020-05863-5.

[6] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[7] S. Roy, R. Dutta, and P. Ghosh, "Optimal time-varying vaccine allocation amid pandemics with uncertain immunity ratios," *IEEE Access*, vol. 9, pp. 15110–15121, 2021.

[8] *Understanding the Covid-19 Pandemic as a Big Data Analytics*. Accessed: Feb. 15, 2013. [Online]. Available: https://healthitanalytics.com/news/understanding-the-fCOVID-19g-pandemic-as-a-big-data-analytics-issue

[9] C. Bayes, V. S. Rosas, and L. Valdivieso, "Modelling death rates due to COVID-19: A Bayesian approach," 2020, *arXiv:2004.02386*. [Online]. Available: http://arxiv.org/abs/2004.02386

[10] B. Mbaye Ndiaye, L. Tendeng, and D. Seck, "Analysis of the COVID-19 pandemic by SIR model and machine learning technics for forecasting," 2020, *arXiv:2004.01574*. [Online]. Available: http://arxiv.org/abs/2004.01574

[11] R. Dandekar and G. Barbastathis, "Neural network aided quarantine control model estimation of COVID spread in Wuhan, China," 2020, *arXiv:2003.09403*. [Online]. Available: http://arxiv.org/abs/2003.09403

[12] H. Ye, P. Wu, T. Zhu, Z. Xiao, X. Zhang, L. Zheng, R. Zheng, Y. Sun, W. Zhou, Q. Fu, X. Ye, A. Chen, S. Zheng, A. A. Heidari, M. Wang, J. Zhu, H. Chen, and J. Li, "Diagnosing coronavirus disease 2019 (COVID-19): Efficient Harris Hawks-inspired fuzzy K-nearest neighbor prediction methods," *IEEE Access*, vol. 9, pp. 17787–17802, 2021, doi: 10.1109/ACCESS.2021.3052835.

[13] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: A survey," *J. Big Data*, vol. 2, no. 1, pp. 1–21, Dec. 2015.

[14] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. Treister. *Transforming Health Care Through Big Data Strategies for Leveraging Big Data in the Health Care Industry*. Accessed: Feb. 15, 2013. [Online]. Available: http://ihealthtran.com/big-data-in-healthcare

[15] Q. Zhang, V. Piuri, E. A. Clancy, D. Zhou, T. Penzel, W. W. Hu, and H. Zheng, "IEEE access special section editorial: Smart health sensing and computational intelligence: From big data to big impacts," *IEEE Access*, vol. 9, pp. 30452–30455, 2021.

[16] M. M. Islam, F. Karray, R. Alhajj, and J. Zeng, "A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19)," *IEEE Access*, vol. 9, pp. 30551–30572, 2021.

[17] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. Aviles-Rivero, C. Etmann, C. McCague, and, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nat Mach Intell*, vol. 3, pp. 199–217, Mar. 2021.

[18] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "A novel COVID-19 data set and an effective deep learning approach for the de-identification of Italian medical records," *IEEE Access*, vol. 9, pp. 19097–19110, 2021.

[19] M. Eisenstein, "Infection forecasts powered by big data," *Nature*, vol. 555, no. 7695, pp. 2–4, 2018.

[20] M. A. Awal, M. Masud, M. S. Hossain, A. A.-M. Bulbul, S. M. H. Mahmud, and A. K. Bairagi, "A novel Bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data," *IEEE Access*, vol. 9, pp. 10263–10281, 2021.

[21] H. Li, S. Liu, X. Yu, S. Tang, and C.-K. Tang, "Coronavirus disease 2019 (COVID-19): Current status and future perspectives," *Int. J. Antimicrob Agents*, vol. 55, no. 5, 2020, Art. no. 105951.

[22] B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 784–790, 2020.

[23] *IBM Releases Novel Ai-Powered Technologies to Help Health and Research Community Accelerate the Discovery of Medical Insights and Treatments for COVID-19*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.ibm.com/blogs/research/2020/04/ai-powered-technologies-accelerate-discovery-covid-19/

[24] E. Alomari, I. Katib, A. Albeshri, and R. Mehmood, "COVID-19: Detecting government pandemic measures and public concerns from Twitter Arabic data using distributed machine learning," *Int. J. Environ. Res. Public Health*, vol. 18, no. 1, p. 282, Jan. 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/1/282

[25] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using Twitter, apache spark, and machine learning," *Appl. Sci.*, vol. 10, no. 4, p. 1398, Feb. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/4/1398

[26] F. Alam, A. Almaghthawi, I. Katib, A. Albeshri, and R. Mehmood, "IResponse: An AI and IoT-enabled framework for autonomous COVID-19 pandemic management," *Sustainability*, vol. 13, no. 7, p. 3797, Mar. 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/7/3797

[27] K. Gao, G. Mei, F. Piccialli, S. Cuomo, J. Tu, and Z. Huo, "Julia language in machine learning: Algorithms, applications, and open issues," *Comput. Sci. Rev.*, vol. 37, Aug. 2020, Art. no. 100254. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157401372030071X

[28] Z. Huo, G. Mei, G. Casolla, and F. Giampaolo, "Designing an efficient parallel spectral clustering algorithm on multi-core processors in Julia," *J. Parallel Distrib. Comput.*, vol. 138, pp. 211–221, Apr. 2020. https://www.sciencedirect.com/science/article/pii/S0743731519308743

[29] E. Guidotti and D. Ardia, "COVID-19 data hub," *J. Open Source Softw.*, vol. 5, no. 51, p. 2376, Jul. 2020.

[30] O. Wahltinez. (2020). *COVID-19 Open-Data: Curating a Fine-Grained, Global-Scale Data Repository for SARS-COV-2*. [Online]. Available: https://goo.gle/covid-19-open-data

[31] V. P. Martsenyuk, I. E. Andrushchak, and A. M. Kuchvara, "On conditions of asymptotic stability in SIR-models of mathematical epidemiology," *J. Autom. Inf. Sci.*, vol. 43, no. 12, pp. 59–68, 2011, doi: 10.1615/jautomatinfscien.v43.i12.70.

[32] V. Martsenyuk, K. Augustynek, and A. Urbas, "On qualitative analysis of the nonstationary delayed model of coexistence of two-strain virus: Stability, bifurcation, and transition to chaos," *Int. J. Non-Linear Mech.*, vol. 128, Jan. 2021, Art. no. 103630, doi: 10.1016/j.ijnonlinmec.2020.103630.

[33] A. Beuter, L. Glass, M. C. Mackey, and M. S. Titcombe, *Nonlinear Dynamics in Physiology and Medicine*. New York, NY, USA: Springer , 2003, doi: 10.1007/978-0-387-21640-9.

[34] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations* (Applied Mathematical Sciences), vol. 99. New York, NY, USA: Springer, 1993. [Online]. Available: https://books.google.ch/books?hl=vi&lr=&id=KZTaBwAAQBAJ&oi=fnd&pg=PA1&ots=LM-ikP7W6C&sig=mQKlgvbxJaaaErkFT2pMuFC48HQ#v=onepage&q&f=false

[35] N. Baker, F. Alexander, T. Bremer, A. Hagberg, Y. Kevrekidis, H. Najm, M. Parashar, A. Patra, J. Sethian, S. Wild, and K. Willcox, "Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence," USDOE Office of Science (SC), Washington, DC, USA, Tech. Rep. 1478744, 2019. [Online]. Available: https://www.osti.gov/biblio/1478744-workshop-report-basic-research-needs-scientific-machine-learning-core-technologies-artificial-intelligence

[36] *What is Scientific Machine Learning*. Accessed: Feb. 15, 2021. [Online]. Available: https://www.oden.utexas.edu/research/centers-groups/scientific-machine-learning/

[37] M. J. D. Powell, "A direct search optimization method that models the objective and constraint functions by linear interpolation," in *Advances in Optimization and Numerical Analysis*. Amsterdam, The Netherlands: Springer, 1994, pp. 51–67, doi: 10.1007/978-94-015-8330-5_4.

[38] *Cran—Package Diffeqr*. Accessed: Jan. 5, 2021. [Online]. Available: https://cran.r-project.org/web/packages/diffeqr/index.html

[39] *Differential Equations: Scientific Machine Learning (SCIML) Enabled Simulation and Estimation*. Accessed: Jan. 5, 2021. [Online]. Available: https://diffeq.sciml.ai/dev/

[40] W. Dhouib, J. Maatoug, I. Ayouni, N. Zammit, R. Ghammem, S. B. Fredj, and H. Ghannem, "The incubation period during the pandemic of COVID-19: A systematic review and meta-analysis," *Systematic Rev.*, vol. 10, no. 1, Apr. 2021, doi: 10.1186/s13643-021-01648-y.

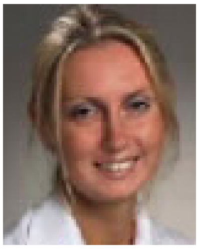
[41] *IBIG World Erasmus+Project—Ibigworld*. Accessed: Jan. 29, 2021. [Online]. Available: http://ibigworld.ni.ac.rs/

VASYL MARTSENYUK received the master's degree in applied math, the Ph.D. degree in systems analysis and decision making, and the D.Sc. degree in systems analysis and decision making from the Taras Shevchenko National University of Kyiv, Ukraine, in 1993, 1996, and 2005, respectively, and the Dr.Hab. degree, in 2015.

From 1997 to 2015, he was working as a Professor, the Chair of the Medical Informatics Department, and the Vice-Rector of Ternopil State Medical University, Ukraine. He received the title of a Professor of technical sciences in Poland, in 2015. He joined the Department of Computer Science and Automation, University of Bielsko-Biala, Poland, as a Professor, in 2015. His research interests include machine learning, decision making, big data, computer graphics, web-programming, medical informatics, biosensors, dynamic systems, and population dynamics.

MARCIN BERNAS received the master's degree in computer science from the University of Silesia in Katowice, and the Ph.D. degree in computer science from the Silesian University of Technology.

He is currently working as an Associate Professor with the Department of Computer Science and Automation, University of Bielsko-Biala. His current research interests include sensor networks, traffic management, machine learning (neural networks), blockchain, time series, big data, and classification.



ALEKSANDRA KLOS-WITKOWSKA received the master's degree in physics from the University of Silesia in Katowice, in 2001, and the Ph.D. degree in physics, in 2007. The Ph.D. project was carried out with the Medical Physics Department, University of Silesia in Katowice.

She is a Laureate of prestigious Maria Curie's scholarship. She completed scientific internships: Max Planck Institut für Biophysikalische Chemie, Germany, University of Ioannina, Greece, and University of Helsinki, Finland. She is currently working as an Associate Professor with the Department of Computer Science and Automation, University of Bielsko-Biala. Her research interests include medical and biological processes modeling, and biosensors design.