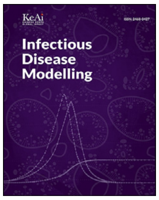
Infectious Disease Modelling 7 (2022) 30–44

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm



Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm

Olga Krivorotko [a, b, *], Mariia Sosnovskaia [b], Ivan Vashchenko [b], Cliff Kerr [c], Daniel Lesnic [d]

[a] *Institute of Computational Mathematics and Mathematical Geophysics Siberian Branch of the Russian Academy of Sciences, 6 Prospect Akademika Lavrentieva Street, Novosibirsk, 630090, Russia*
[b] *Novosibirsk State University, 2 Pirogova Street, Novosibirsk, 630090, Russia*
[c] *Institute for Disease Modeling, Bill & Melinda Gates Foundation, Seattle, USA*
[d] *University of Leeds, LS2 9JT, UK*

ARTICLE INFO

Article history:
Received 23 June 2021
Received in revised form 24 November 2021
Accepted 24 November 2021
Available online 27 November 2021
Handling editor: Dr Lou Yijun

Keywords:
Epidemiology
Agent-based modeling
COVID-19
Interventions analysis
Coronavirus data analysis
Forecasting scenarios
Reproduction number
Optimization
Parameter identification

ABSTRACT

This paper uses Covasim, an agent-based model (ABM) of COVID-19, to evaluate and scenarios of epidemic spread in New York State (USA) and the UK. Epidemiological parameters such as contagiousness (virus transmission rate), initial number of infected people, and probability of being tested depend on the region's demographic and geographical features, the containment measures introduced; they are calibrated to data about COVID-19 spread in the region of interest. At the first stage of our study, epidemiological data (numbers of people tested, diagnoses, critical cases, hospitalizations, and deaths) for each of the mentioned regions were analyzed. The data were characterized in terms of seasonality, stationarity, and dependency spaces, and were extrapolated using machine learning techniques to specify unknown epidemiological parameters of the model. At the second stage, the Optuna optimizer based on the tree Parzen estimation method for objective function minimization was applied to determine the model's unknown parameters. The model was validated with the historical data of 2020. The modeled results of COVID-19 spread in New York State and the UK have demonstrated that if the level of testing and containment measures is preserved, the number of positive cases in New York State remain the same during March of 2021, while in the UK it will reduce.

1. Introduction

COVID-19 was declared pandemic by the WHO on March 11th, 2020. Since that time, more than 219 million people in 192 countries have been infected with the disease, and more than 4.5 million people have died after getting infected. During the last two years mankind have mobilized its resources to fight the pandemic. One of the useful tools in this struggle has been

* Corresponding author. Institute of Computational Mathematics and Mathematical Geophysics Siberian Branch of the Russian Academy of Sciences, 6 Prospect Akademika Lavrentieva Street, Novosibirsk, 630090, Russia.

E-mail addresses: o.krivorotko@g.nsu.ru (O. Krivorotko), m.sosnovskaya@g.nsu.ru (M. Sosnovskaia), i.vashchenko@g.nsu.ru (I. Vashchenko), ckerr@idmod.org (C. Kerr), D.Lesnic@leeds.ac.uk (D. Lesnic).

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2021.11.004

mathematical modeling that uses known historical data to study different scenarios of disease spread (Cooper et al., 2020; Ndairoua et al., 2020).

The models including those for studying coronavirus infections can be divided into two groups: compartmental and agent-based models. In compartmental models, a population is divided into groups sharing similar features and interacting with one another following the mass action law. Agent-based models (ABMs), on the other hand, give each agent (people, social institutes, the state, etc.) a set of features and determines the way the agents interact from random graphs following disease spread principles. In other words, an agent's behavior is determined individually, and their joint behavior is described as the interaction of multiple agents (bottom-top approach). Unlike compartmental models, ABMs are capable of providing a detailed description of an epidemiological situation, especially in a case of inaccurate and insufficient data. ABMs account for the stochastic nature of epidemic spread, makes it possible to estimate the likelihood of different epidemic scenarios, and allows one to evaluate the risks of unfavorable events occurring due to policy changes. The resulted data enable one to make conclusions about the duration, severity and scale of an epidemic, evaluate the efficacy of the preventive and quarantine measures, and assess its economic consequences.

During 2020, many models were developed to predict COVID-19 spread. In (Kai et al., 2020), both compartmental and agent-based models are presented to study mass face-mask wearing and predict its effect on COVID-19 spread. A graph-based ABM is suggested in (Wolfram, 2020). The paper considers a small population of 1000 agents and graphs of different kinds such as fully-connected, Barabasi-Albert, Watts-Strogatz ones, etc. Another approach to agents interactions is demonstrated in (Cuevas, 2020). In this paper, agents are initialized in a 2D space (so-called mesh) to consider a distance between different agents, so such a factor as social distancing can be explicitly accounted for. However, this approach leads to excessive computation complexity, and for that reason, these models consider a small number of agents (up to 500).

Another important direction of modeling has been a comparison of containment measures and projecting the future for different scenarios. The paper published by Chang et al. (Chang et al., 2021) combines an epidemiological SEIR model and the hourly GPS data from the mobile phones of 98 million people in 10 US cities. The model predicted that closing of the most crowded public places such as restaurants and religious establishments would be a sufficient measure to contain the pandemic unlike unilateral measures to limit people's mobility. In (Silva et al., 2020), the authors offer an ABM using SEIR agents to model COVID-19 spread dynamics, where the agents imitate people, businesses and the government. Using the model, they have analyzed seven social-distancing scenarios having different epidemiological and economic effects. The paper has demonstrated that the so-called vertical isolation has no positive effect. In (Mellacher, 2020) macroeconomic epidemiological ABM been presented to study the economic effect COVID-19 would have in different scenarios of pandemic containment such as closing of educational and entertainment facilities. The model was calibrated using the statistical data on country's and business demography, households, employment, profits and wages in Germany.

It is noteworthy that data collecting and processing is a very important step in building an effective COVID-19 spread model. However, in the studies mentioned above data pre-processing for the modeled regions was not performed. Most of them concentrated on building the models and algorithms, whose parameters were considered known either from literature or from experts' estimations, so the issues of identifiability for unknown parameters have remained unresolved, as has the need of devising a regularization algorithm for solving the problem of epidemiological forecasting.

In this study, our focus is on the analysis of data, parameter identification, and regularization algorithms. However, it is a known fact that epidemics develop differently in different locations. To address this issue, in our study, the epidemiological situations in New York State (USA) and the UK were compared and analyzed.

1.1. New York State

Nearly 2 million people were confirmed infected and more than 50 000 people died by April 6, 2021 in NY State (Tracking Coronavirus in N, 2021). In NY State, the pandemic spread rapidly, reaching its peak in March–April of 2020 (see Section 2.1). The healthcare system was overloaded in the very first months of the outbreak. Thanks to the containment measures introduced, the number of infected people reduced to a characteristic plateau that was followed by a second infection wave several months later.

1.2. United Kingdom

Another pandemic scenario was observed in the UK. After the first infection wave had been successfully suppressed by June 2020, the second wave hit the country hard and, due to the B.1.1.7 SARS-CoV-2 variant, the number of infected people increased 10 times compared to the first wave (see Section 2.2). At the time of writing, this has reduced to 7000 cases a day, and the pandemic is on wane.

This study is organized as follows. First, open-source data of COVID-19 spread in New York State and the UK were processed and analyzed using the statistics and machine learning methods to find interdependencies, study seasonality and predict possible future dynamics (see Section 2). Second, we confirmed that the selected ABM met the identifiability condition such as being sensitive to data errors and capable of unambiguous determination of the unknown parameters of COVID-19 spread from additional measurements (Section 3.2). The obtained space of identifiable parameters was specified using the multilevel global-optimization method (Section 4). Finally, scenarios of how the COVID-19 pandemic that could develop in New York State and the UK were assessed concerning available data and certain containment measures (Section 5).

2. Data analysis

Data analysis and data processing are important parts of forecasting modeling. Before data processing begins, one has to understand the character of available data and determine their features. Collecting such daily indicators as the number of tests, diagnosed cases, ventilated COVID patients, etc. (see Table 1) helps to overview the general picture for a considered region, while anomalous time intervals may call for more scrupulous analysis.

2.1. New York State (USA)

To forecast the way the pandemic would develop, the data from the COVID Tracking Project's website (The COVID Tracking Project in USA) were used. The site contains information for each state and for the country as a whole. The feature of the data in question is the method they calculate New Diagnoses: positive cases (confirmed plus probable) summing the total number of confirmed cases and the probable cases of COVID-19 reported by the state or territory, ideally per the “August 5, 2020 CSTE case definition”. Some states are following the older “April 5th, 2020 CSTE case definition” or using their own custom definitions. The latter method of data collection is more suitable for ABMs since it takes into account the percentage of infected people that may have been simply neglected.

We can consider the COVID-19 spread in New York State in more detail, using the epidemic time-series data. Any time series can be decomposed into the following three elements:

$$X(n) = T(n) + S(n) + N(n).$$

Here, $X(n)$ is a time-series value, $T(n)$ is the value of the underlying time-series component; $S(n)$ is the value of a seasonality component, $N(n)$ is the value of a noise component for the n-th day. When analyzing a time series, we found it was most useful to analyze its trend, since this determines an indicator's behavior in time.

The non-smoothed graphs in Fig. 1 demonstrate widely dispersed points of statistics, so these data were smoothed before using them in the model because only the main trends of the curves were necessary for reaching an appropriate result. The graphs demonstrate certain periodicity that is known to be time series seasonality, which is clearly traced during summer in New Diagnoses.

Now, let us consider the average fraction of tests for each day of the week that is calculated as:

$$w_i = \sum\nolimits_{j=0}^{N} \frac{X(i+m\cdot j)}{S(i+m\cdot j)}.$$

Here, i is a week-day number, N is the number of full weeks within a considered time series, $S(i+m\cdot j)$ is the cumulative sum of the New Tests performed within a week corresponding to index j, $m=7$. The results can be seen in Table 2. It is apparent that the number of tests tends to its minimum for Mondays and Tuesdays and reaches its maximum on Fridays and Saturdays, which exemplifies the seasonality (i.e., periodicity) of the considered time series.

COVID-19 pandemic data analysis has shown that there is a dependence between New Tests and New Diagnoses datasets for different regions. In such cases, the behavior of the second indicator is partly determined by that of the first. As a matter of fact, the last day of the week does not mean an abrupt increase in the number of the infected only because it is Friday. However, more people get tested on Friday, so the number of positive tests may increase. For that reason, analyzing the links between the number of infected people and that of new tests becomes crucial. For a proper understanding of the situation, it is not New Diagnoses, but their fraction from the number of New Tests that we want to know. After all, if one tested every person in a region, they would immediately indicate every one infected and their indicator would increase abruptly while their percentage would remain the same. Based on this fact, one can derive a time series that describes the percentage of newly tested people with either positive or potentially positive COVID-19 test.

Fig. 2 shows that the ratio reached its peak in April 2020 and later started to reduce due to an increased number of tests. In autumn, the second pandemic wave began, so the number of daily confirmed cases increased (see Fig. 1), but the ratio between New Diagnoses and New Tests remained at a low level. This implies that testing behavior (i.e., the probability of testing with or without COVID-19 symptoms) did not change.

To confirm change in testing during autumn and winter, the MACD indicator (Bartolucci et al., 2018) was used, whose histogram tracks a function's rise and fall (in our case it was the rise of New Tests). The graph demonstrates that the MACD indicator increased significantly over this period (see Fig. 3).

Table 1
Indicators used in data analyses and their description.

Indicator	Description
New Tests	Number of performed tests
New Diagnoses	Number of diagnosed cases
New Deaths	Number of deaths related to a positive diagnosis

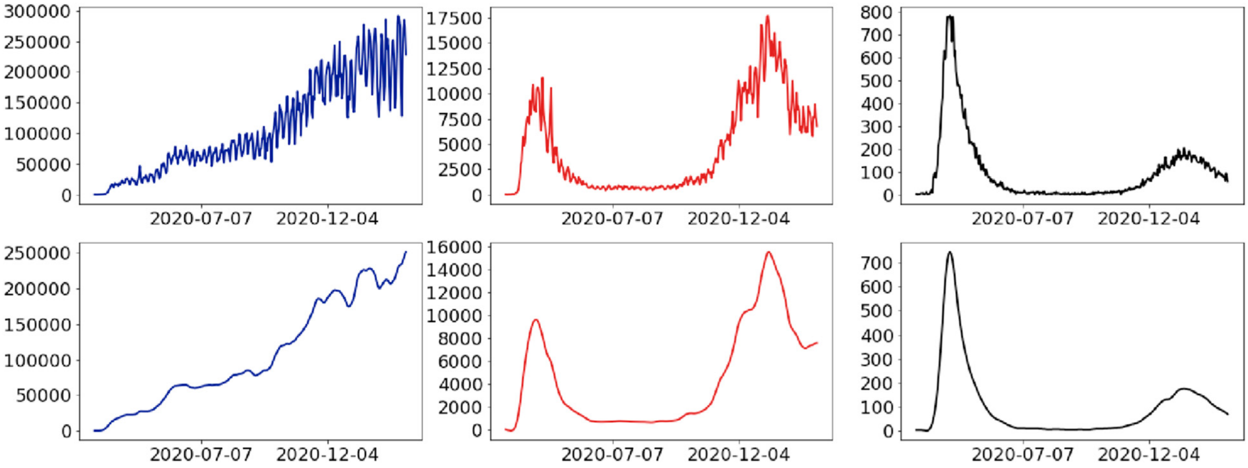


Fig. 1. Graphs of COVID-19 spread in New York State (top line): the number of New Tests (left), New Diagnoses (middle) and New Deaths (right) and their smoothed curves (bottom line).

2.2. Seasonality

The above-mentioned term "seasonality" refers to the periodic fluctuations observed in time series. In other words, if one takes a time-series space and overlaps it against a neighboring space of the same size, their profiles are going to coincide (or differ by a constant). The peak absolute values will correlate with the same time points calculated from the beginning of the space. As for our COVID-19 data, their seasonality can be traced in the New Tests indicator, which has its logic since collecting this statistics involves as a human factor as the features of the healthcare system. The seasonality of a time series makes it possible to forecast the series behavior relative to some average value.

Seasonality is commonly determined with an autocorrelation function (ACF). For a discrete process $X(1)$, $X(2)$, …, $X(n)$ its formula is written as:

$$R(n) = \frac{1}{(M-n)\bullet\sigma^2}\sum\nolimits_{t=1}^{M-n}(X(t)-\mu)(X(t+n)-\mu).$$

Here σ is the standard deviation of a discrete process X, μ is its average value, M and n are positive integers. Besides, to analyze time-series seasonality, a partial correlation (PACF) was applied that removed the linear dependence between shifted time series.

The results confirmed New Tests really had weekly seasonality (7, 14 and 21st days), while the New Diagnoses did not possess this property (see Fig. 4).

2.3. New Tests/New Diagnoses interrelation

We have also considered a percentage change for any current moment in relation to the same moment a week before:

$$pc(n) = \frac{X_{smoothed}(n)}{X_{smoothed}(n-7)} - 1.$$

This statistic demonstrates by how much an average indicator has changed in fractions compared to a previous week. If compared for weekdays, the results become smoother and easier to interpret.

The two time series in Fig. 5 have spaces where their trajectories almost match. In other words, the percent change of one indicator differs from that of the other one by a constant. In terms of tested/infected ratio, such spaces confirm that within this period, the number of infected people grew owing to the increased number of tests and not to a worsening pandemic situation in the region.

For more specific analysis, a window of 28 days was correlated with the previous 28 days for every day in the window (see Fig. 6).

In statistics, a close linear dependence occurs when the absolute correlation value exceeds 0.7. If it exceeds 0.9, it means there is a strong dependence between two indicators (Chaddock, 1925). In NY State, with some deviations, New Diagnoses strongly depended on New Tests. The deviations mean the pandemic develops following its own scenario. For instance, in the

Table 2
Average fraction from the number of tests and its daily distribution for a whole statistics-gathering period (NY State).

Days of the week	Average fraction
Monday	0.113 654
Tuesday	0.127 697
Wednesday	0.134 512
Thursday	0.155 779
Friday	0.162 017
Saturday	0.163 702
Sunday	0.142 640

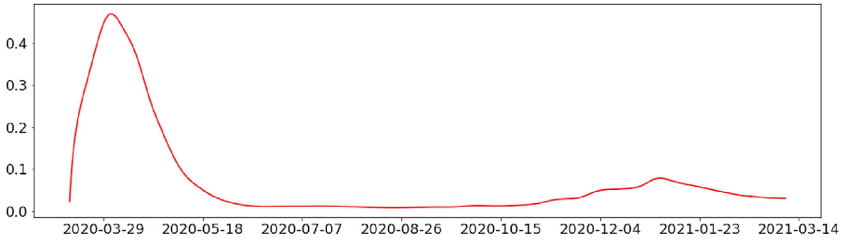


Fig. 2. Fraction of new diagnoses from new tests in NY state.

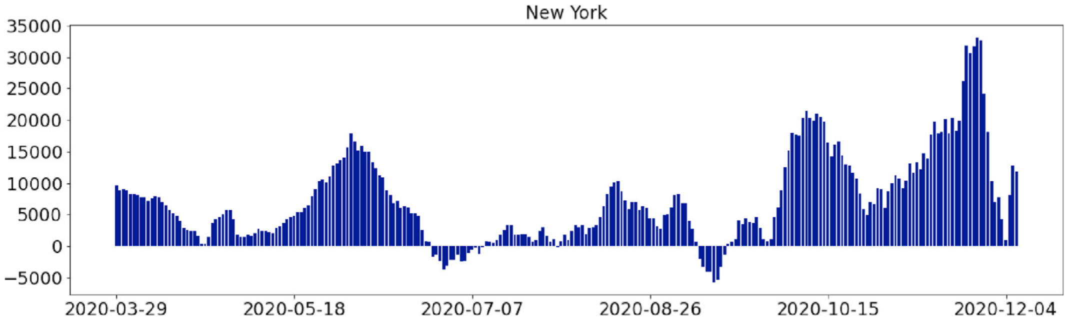


Fig. 3. MACD-indicator of the new tests in NY state.

end of May and the beginning of June, the correlation coefficient exceeded 0.8 within a 28-day backward window when the first pandemic wave in the region was considered defeated and reached the so-called plateau that lasted till October when the second correlation took place and the second pandemic wave began, which means the second wave might have been triggered by the abrupt increase of New Tests. However, the correlation coefficient got back to almost zero by December, indicating that the rise and fall of New Diagnoses did not correlate with New Tests, so the pandemic in the region at that time spread or reduced (depending on graph direction). In February and March, the correlation became very close to one, meaning the number of infected people depended only on a New Tests.

2.4. United Kingdom

To analyze COVID-19 spread in the UK, the data accumulated on an official government web portal were used (Coronavirus (COVID-19) in the UK).

Fig. 7 shows three stages of the infection: the first wave (from 2020.03.09 to 2020.06.01), plateau (from 2020.06.01 to 2020.09.01) and the second wave (from 2020.09.01 to 2021.03.06). The standard deviation of the time series for the second wave was much bigger than for the first one. For that reason, the standard deviation values were considered for two independent time series of New Diagnoses and New Tests (see Table 3). The data for NY State (NY) is put in the table for comparison.

Unfortunately, the data did not allow us to conclude what was the exact reason for the second wave's higher variance. It could have been the week variance of New Tests or something else since, in the UK, they started to register the number of tests only on 2020.04.21. All we know is the standard deviation of New Diagnoses increased with time as well as the variance of New Tests. However, despite the growing variance of New Tests in NY State, the standard deviation value of New Diagnoses remained at the same level.

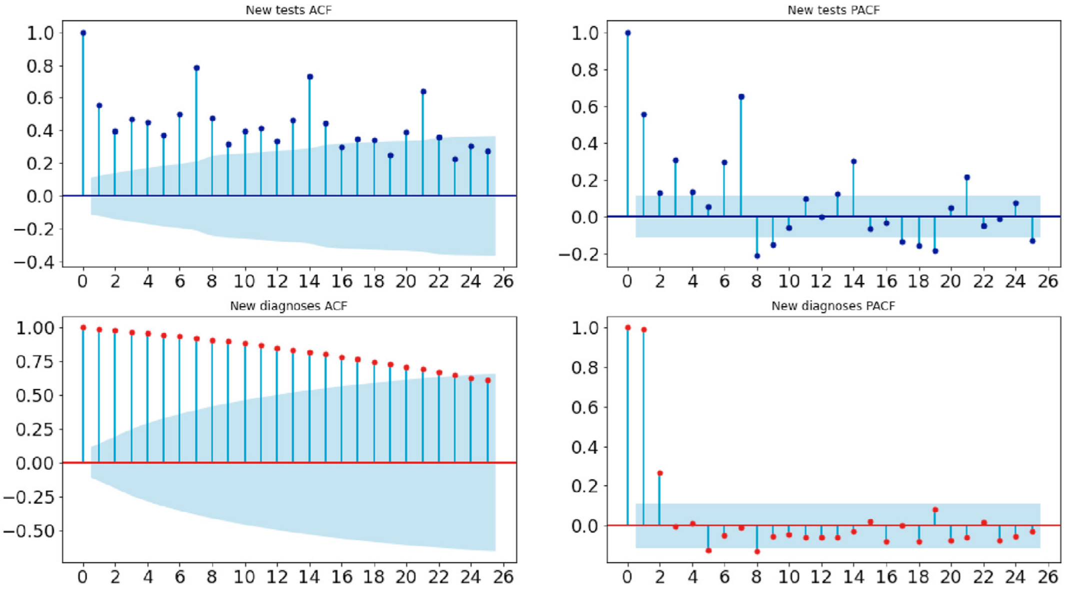


Fig. 4. Graphical results of applying ACF (left) and PACF (right) to New Tests (top) and New Diagnoses (bottom) in NY State.

2.5. Seasonality

The ACF and PACF applied to the New Diagnoses and New Tests time series and the fractions of tests calculated for every day of the week demonstrated the weekly seasonality of New Tests confirmed by ACF (see Fig. 8 and Table 4). However, no significant PACF delays were observed. Results obtained with those from NY State demonstrated the absence of linear dependence did not affect the time series's seasonality. However, a different situation was observed in the UK.

In the next section, an approach to forecasting the New Tests time series will be demonstrated. The approach relies upon several techniques, including SARIMA (Dabral & Murry, 2017), an algorithm requiring a significant correlation for the seasonal delays (7, 14, 21, etc.) of a single parameter.

2.6. Forecasting new tests

To draw forecast curves while modeling, one has to predict a number of certain statistical data sets that are used as input parameters. In our model, such a data set was the New Tests performed in the region since this indicator did not depend on the others and had the highest value of seasonality that determined the seasonality of the other indicators. To extrapolate the New Tests time series, the SARIMA algorithm was used. The result of numerical modeling and forecasting is demonstrated in Section 5.

3. Agent-based mathematical model

This section presents the ABM devised to describe COVID-19 spread and formulates a problem to identify (calibrate) the model's unknown parameters as objective functional minimization. It also presents a scheme for automatic calibration of the parameters for time intervals.

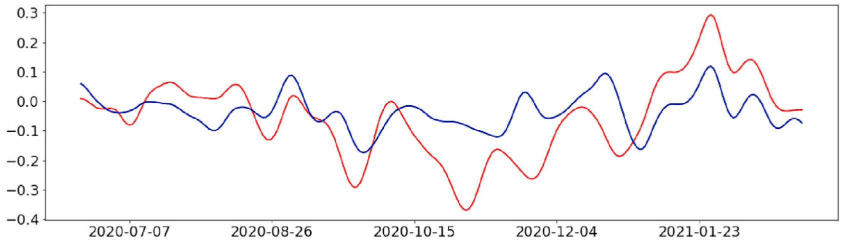


Fig. 5. New Diagnoses percent change (red line) and New Tests percent change (blue line) time-series graphs in NY State.

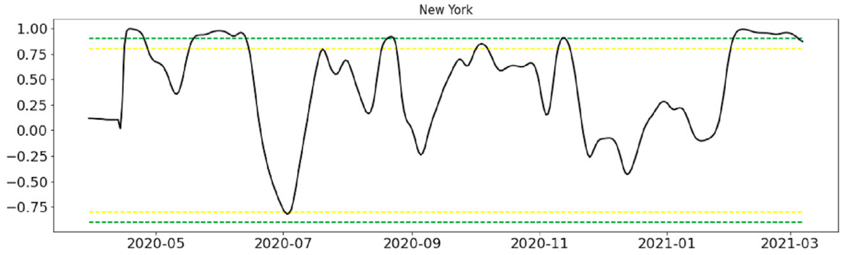


Fig. 6. New Diagnoses percent change and New Tests perecent change correlations within a 28-day backward window with |0.8| and |0.9| correlation's threshold lines for NY state.

3.1. *ABM formulation*

Within the framework of this research, stochastic ABMs for New York State and the United Kingdom were devised. They were built using the Covasim (Kerr et al., 2021). This package had been utilized to predict the number of infected, dead and hospitalized people in the State of Oregon and become one of the tools to make decisions about whether to relax or escalate COVID-19 containment measures. This library is written in Python to study non-trivial COVID-19 dynamics. Its general algorithm is as follows: after all necessary parameters and statistical data are uploaded, the package creates an artificial population with account for age distribution. The model's agent is a person in a particular region. Then, the agents are united into contact networks and the integration loop begins. At every time step (1 day), an agent's status is updated in relation to its contact network and the containment measures relevant for this interval (self-isolation; closed access to public places; wearing face masks, etc.). The agents can interact with one another in particular networks. Depending on the network's structure, both full and random connectivity graphs are built for an agent to Figure out how the infection spreads. The average number of daily contacts is different for every agent and every network. At any time moment, the agents distributed by their age (bins of 0–9, 10–19, …, 90+) are found in their given state (see Fig. 10). More details about the structure of Covasim-based ABMs, their parameters and realization methods can be found in (Covasim documentation).

Every agent has their set of properties and characteristics that can be divided into 2 groups: constant (belong to each particular agent and do not change while modeling) and time-dependent.

3.2. *Time–independent agent characteristics*

- Age ($t*$). All the agents are subdivided into age groups of 10 years (0–9 years, 10–19, …, 90 + . The age distribution depends on the demographic situation in a studied region.
- Social status (determined by an agent's age $t*$). Depending on their age, agents contact one another in contact networks. All agents have contacts in households and public places. Agents of 6–21 years old can also have contacts in educational institutions with agents of their age. Agents of 22–65 years old contact at work (see Fig. 9). Depending on a contact's structure, the transmission parameter β is multiplied by corresponding constant w_β ($w_\beta = 3$ for households, 0.6 – for educational institutions, 0.3 – for public places), i.e. the likelihood of virus transfer is different for every network.
- Likelihood of disease progression (determined by an agent's age $t*$). These parameters characterize disease progression (see Fig. 10). Their description is given in Table 5.

3.3. *Time-dependent agent characteristics*

- Agent's epidemiological status. Each agent may have one of the 9 stages of the disease $\vec{X} = (S, E, A, Y, M, H, C, R, D)$ (Fig. 10).
- Agent's chance to be tested for COVID-19 ($\tilde{p}(X(t))$) that is determined by the agent's epidemiological status. The agents are tested daily, the number of tests corresponds to the statistical data obtained in a region. At every modeling step, the tests are distributed across the population, the agents whose status is marked with an orange frame in a Fig. 10 can be given a positive result. The agents whose test is recognized as positive are marked as “confirmed” and included in New Diagnoses. The model assumes that the likelihood for an agent to be tested as a symptomatic carrier is higher and this chance ratio is controlled by parameter $\tilde{p}(X(t))$ at is restored from solving the inverse problem (see Section 3.2).

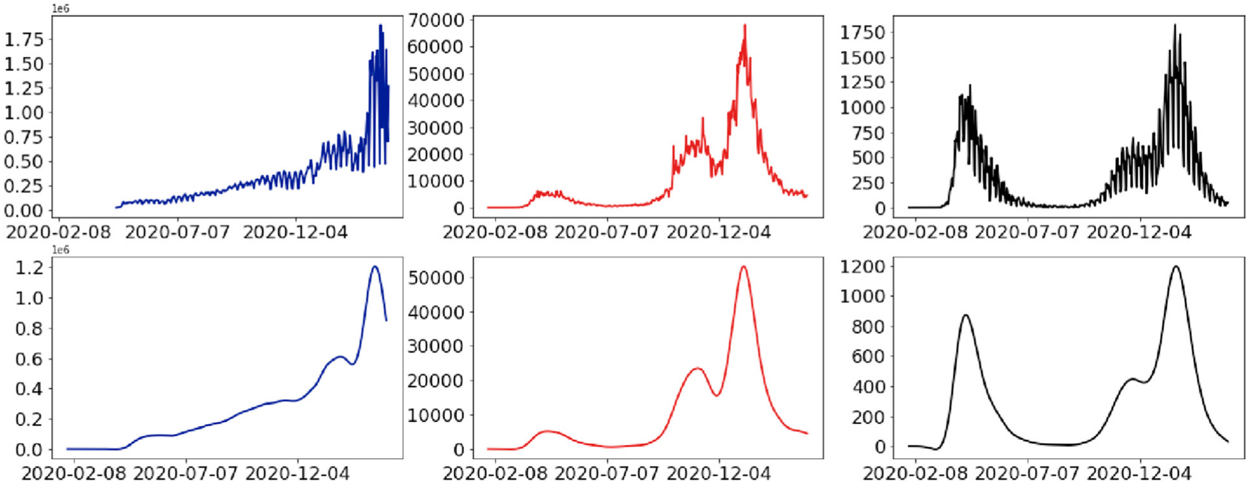


Fig. 7. COVID-19 spread graphs in the UK (top line): New Tests (left), New Diagnoses (middle), New Deaths (right) and their smoothed curves (bottom line).

3.4. Parameter identification problem

The ABM developed was also characterized by unknown parameters vector $\vec{q} = (E(0),\beta,\beta_d,\beta_c,\tilde{p}(X))$. To specify the model's parameters a variational inverse problem formulation was performed to minimize the misfit function:

$$J(\vec{q}) = \sum\nolimits_{t_i=1}^{T} \frac{|Y_d(t_i) - Y_m(t_i,\vec{q})|}{M_{diag}} + \frac{|D_d(t_i) - D_m(t_i,\vec{q})|}{M_{death}}. \quad (1)$$

Here, $Y_d(t_i), Y_m(t_i,\vec{q})$ are smoothed daily statistical and model data of New Diagnoses; $D_d(t_i), D_m(t_i,\vec{q})$ are smoothed daily statistical and model data of New Deaths; T – the number of modeled days, M_{diag}, M_{death} are normalising terms. Parameters β_d and β_c determine days and values changes of piece-wise parameter β (see details in Section 3.3). In paper (Krivorotko, Kabanikhin, Sosnovskaya, & Andornaya, 2021) the sensitivity-based identifiability analysis of the COVID-19 pandemic agent model in the Novosibirsk region was shown that the transmission parameter \beta is a more sensitive to measurements and needs to be accurate calibration.

3.5. Automatic parameter calibration

The model assumed that parameter β was a piece-wise constant. The longer was a considered time interval, the more unknown parameters it included. Since every launch of the model's calibration algorithm was rather time-consuming, the time interval in question was divided into periods of 1 month. For example, for NY State the first period was 2020.03-02 - 2020.04.01, the second – 2020.04.02–2020.05.03; for the UK – 2020.02.07–2020.03.08 and 2020.03.09–2020.04.07, etc.

Each period was sequentially calibrated, so the parameters restored at a previous step were used in the following iteration of the optimization algorithm. Thus, for the initial period considered the unknown parameter vector was

$$\vec{q}_1 = (E(0),\beta,\beta_d(1),\beta_c(1),\tilde{p}(X)),$$

where $E(0)$ is the initial number of infected agents, β is a contagiousness parameter value, $\beta_d(1)$ is the day parameter β changes, $\beta_c(1)$ is the value by which parameter β changes on day β_d, $\tilde{p}(X)$ is a test level parameter in relation to statistical data. For all the following periods (second, third, etc.):

$$\vec{q}_i = (\beta_d(i),\beta_c(i)).$$

Table 3
Standard deviation values for particular New Diagnoses and New Tests time series during different waves of COVID-19.

Indicator	The 1st stage	The 2nd stage	The 3rd stage
(UK) New Tests std	10 626.7	16 613.8	77 382.5
(UK) New Diagnoses std	455.8	201.1	5727.2
(NY) New Tests std	5636.6	9543.9	31 127
(NY) New Diagnoses std	1137	114.2	1144

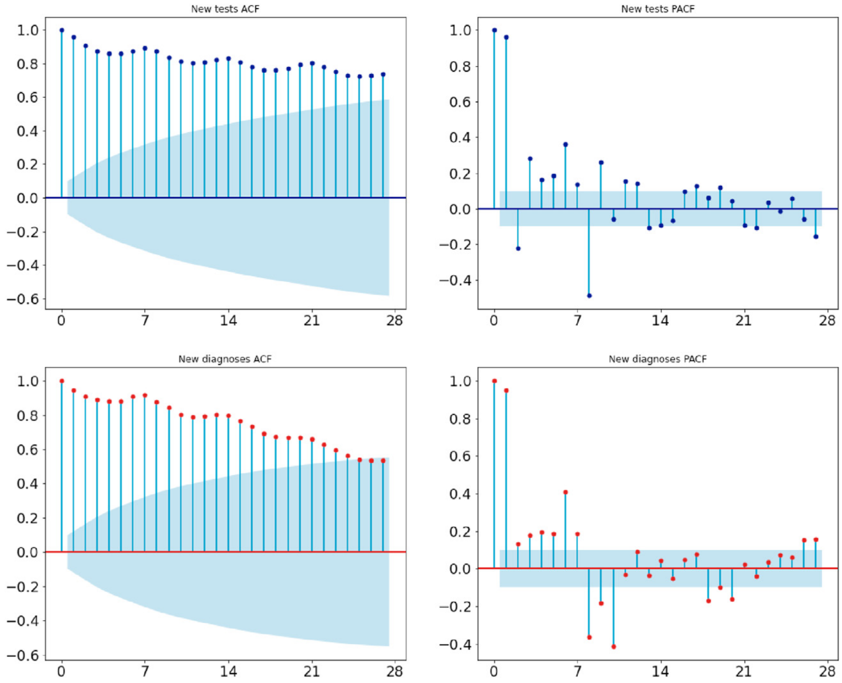


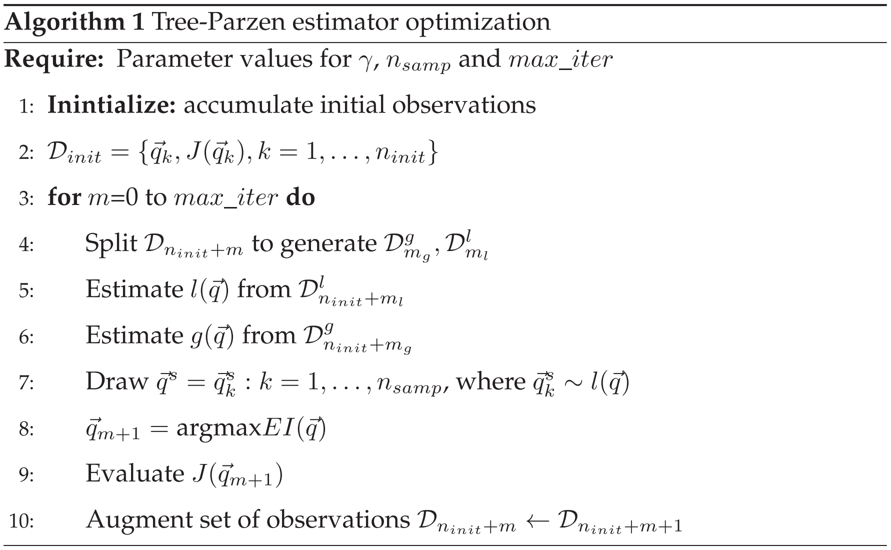
Fig. 8. Graphical results of applying ACF (left) and PACF (right) to New Tests (top) and New Diagnoses (bottom) for the UK.

4. Methods and approaches

Identifying (calibrating) the parameters of an ABM, so that its outputs match observed data, is quite a complex and computation-intensive task due to the large number of the parameters involved. There are different approaches to the problem (see (Kabanikhin & Krivorotko, 2020) and references therein). In most cases, the parameters are selected manually or one uses averaged experimental results neglecting the specific features of a studied region are used. Analysis of the papers describing epidemiological ABMs has demonstrated that no algorithm could be considered superior for identification of model parameters (Hazelbag et al., 2020). According to the paper "… it appears that calibrating individual-based models in epidemiological studies of HIV, malaria and TB transmission dynamics remains more of an art than a science."

In our model, the vector of unknown parameters q was calibrated using the Optuna hyperparameter optimization software (OPTUNA) to be one of the latest optimizers designed to adjust hyperparameters in machine learning algorithms and neural networks. The optimizer is based on the tree-structured Parzen estimator (TPE) that in many ways is similar to the Bayesian optimizer (Rasmussen & Williams, 2006). However, unlike the Bayesian optimizer that calculates $p(J(q)|q)$, TPE calculates $p(q|J(q))$ and $p(J(q))$ to determine the parameters domain to minimize functional J by performing Parzen window density estimation, to generate two separate distributions specifying the high - and low-quality regions of the input-space respectively. For this, $l(q)$ and $g(q)$ probability distributions are introduced. $l(q)$ is interpreted as representing the probability of a region in the input space yielding a high-quality observation while similarly, $g(q)$ represents low-quality regions. A full TPE optimization procedure is described in Algorithm 1 (Bergstra et al., 2011).

Algorithm 1. Tree-Parzen estimator optimization



```
Algorithm 1 Tree-Parzen estimator optimization
Require: Parameter values for $\gamma$, $n_{samp}$ and $max\_iter$
Inintialize: accumulate initial observations
$\mathcal{D}_{init} = \{\vec{q}_k, J(\vec{q}_k), k = 1, \ldots, n_{init}\}$
for $m$=0 to $max\_iter$ do
    Split $\mathcal{D}_{n_{init}+m}$ to generate $\mathcal{D}^g_{m_g}, \mathcal{D}^l_{m_l}$
    Estimate $l(\vec{q})$ from $\mathcal{D}^l_{n_{init}+m_l}$
    Estimate $g(\vec{q})$ from $\mathcal{D}^g_{n_{init}+m_g}$
    Draw $\vec{q}^s = \vec{q}^s_k : k = 1, \ldots, n_{samp}$, where $\vec{q}^s_k \sim l(\vec{q})$
    $\vec{q}_{m+1} = \text{argmax} EI(\vec{q})$
    Evaluate $J(\vec{q}_{m+1})$
    Augment set of observations $\mathcal{D}_{n_{init}+m} \leftarrow \mathcal{D}_{n_{init}+m+1}$
```

5. Modeling and forecasting

In this section, we consider mathematical models and scenarios of COVID-19 spread in NY State and (Section 5.2), the UK (Section 5.3) as well as how do different interventions affect the effective reproduction number (Section 5.4).

5.1. Initial datasets

To build and analyze the ABMs in the two considered regions, the following data were used:

1. Information on population's age distribution according to the local government statistics;
2. Information on the average family size according to the UN data (Household Size and 2019, 2019);
3. Statistical data on the people infected with COVID-19, who recovered and died including the number of tests performed that were collected from:
   - The COVID Tracking Project (New York State): https://covidtracking.com/data;
   - The official UK Government website for data and insights on Coronavirus (United Kingdom): https://coronavirus.data.gov.uk/.

For every region, the modeling results for the New Diagnoses and New Deaths datasets were analyzed. In Sections 5.2 and 5.3 one can find the graphs of 45-day forecasts validated with historical data. The forecasts have an 80% confidence interval to characterize 10% and 90% quantiles.

The forecast scenarios of COVID-19 spread for each region were considered for four intervention types in Section 5.4 (Fig. 13):

- Baseline (black curve);
- Return to offices 30% of workers from remote work (blue curve);
- Closure of 20% of public places and transition 20% of workers to the remote work (green curve);
- Increasing number of people in public places by 60% (red curve).

5.2. COVID-19 spread simulation in NY state

After identification of the parameters by minimization of the misfit function (1), the model was devised and validated with historical data. Fig. 11 presents the results of a 45-day forecast with restored vector of unknown parameters for NY State,

Table 4
Average fraction from the number of tests and its daily distribution for a whole statistics-gathering period in the UK.

Days of the week	Average fraction
Monday	0.125 222
Tuesday	0.135 905
Wednesday	0.159 167
Thursday	0.166 208
Friday	0.157 837
Saturday	0.131 992
Sunday	0.123 668

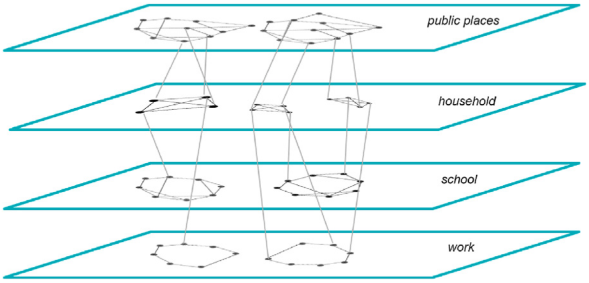


Fig. 9. Agent contacts layers and their interactions in the COVID-19 spread ABM.

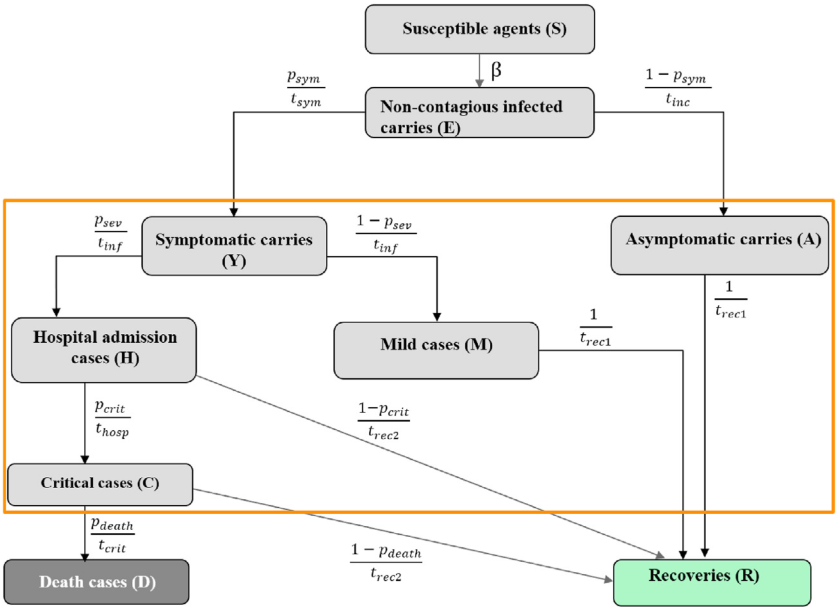

Susceptible agents (S)
β
$\frac{p_{sym}}{t_{sym}}$
Non-contagious infected carries (E)
$\frac{1-p_{sym}}{t_{inc}}$
$\frac{p_{sev}}{t_{inf}}$
Symptomatic carries (Y)
$\frac{1-p_{sev}}{t_{inf}}$
Asymptomatic carries (A)
$\frac{1}{t_{rec1}}$
Hospital admission cases (H)
Mild cases (M)
$\frac{1}{t_{rec1}}$
$\frac{p_{crit}}{t_{hosp}}$
$\frac{1-p_{crit}}{t_{rec2}}$
Critical cases (C)
$\frac{p_{death}}{t_{crit}}$
$\frac{1-p_{death}}{t_{rec2}}$
Death cases (D)
Recoveries (R)

Fig. 10. Agent state transition diagram in Covasim that is based on a SEIR-type compartment model. The orange frame marks those agent states that can give positive COVID-19 tests.

Table 5
Parameters of disease progression probability.

Parameter	Description
$p_{sym}(t^*)$	Probability of developing symptoms
$p_{sev}(t^*)$	Probability of developing severe symptoms (requires hospitalization)
$p_{crit}(t^*)$	Probability of critical condition (ICU)
$p_{death}(t^*)$	Probability of death

where the dots mark the real data accumulated over one year from 2020.03.02 to 2021.03.06 (for New Diagnoses red dots and for New Death blue dots represent modeling data which was used during the solution of inverse problem, while black ones represent testing data which was used for validation). The forecast for New Diagnoses assumed the rate of daily tests remained unchanged.

According to the graphs, despite the size of the confidence interval that started to increase from the October 2020, the increment rate of New Diagnoses was closed to the test data. In terms of New Deaths, the modeling results for this statistics were less accurate than for New Diagnoses. This can be explained by the unweighted type of misfit function (1). The problem of determining the weight coefficients needs further consideration.

5.3. COVID-19 spread simulation in the UK

In the same way as for New York State, the parameters that minimize mismatch function (1) were identified for the United Kingdom. Fig. 12 presents the results of a 45-day forecast for the UK with restored vector of unknown parameters q, where the dots mark the real data accumulated from 2020.02.08 to 2021.03.06 (for New Diagnoses red dots represent modeling data which was used during the solution of inverse problem, while black ones represent testing data which was used for validation) The forecast for New Diagnoses assumed the rate of daily tests remained unchanged.

Although modeling results for New Diagnoses differed from real data by about 20%, the prognosis for validation data (marked by black dotes) was accurate (with relative error less than 5%)

5.4. Effective reproduction number

Additionally, effective reproduction number was calculated for every region. This number indicate how many persons an infectious agent infects on average during the time it has been infectious and calculated as:

$$\mathcal{R}(t) = \frac{I_N(t) \cdot f}{I_C(t)}, \tag{2}$$

where $I_N(t)$ is the number of new infections on day t, $I_C(t)$ is the number of actively infectious people on day t and f is the average duration of infectiousness. If $\mathcal{R}(t) < 1$, the pandemic is considered to stop spreading and keeps spreading otherwise.

Effective reproduction number was calculated considering different scenarios, presented in Section 5.1. The numerical results for both New York State and the United Kingdom are illustrated at Fig. 13.

The graph for New York State demonstrate that under baseline condition (black line) epidemic is under control during the forecasting period from 2021.03.07 to 2021.04.22 (because the value of $\mathcal{R}(t)$ is around 1). Increasing number of people at workplace (blue line) and in public places (red line) provoke the epidemic growth in the region, while closure 20% of public places and workplaces helps to stop propagation (green line).

In the United Kingdom the epidemic spread was decreasing in February and March 2021 (the value of $\mathcal{R}(t)$ was under 1). As a result, under all considered scenarios COVID-19 propagation was expected to decrease because of relatively small number of infected people in the population during the forecast period.

6. Conclusions and discussions

Mathematical models are effective tools to deal with the time evolution and patterns of disease outbreaks. They provide us with useful predictions in the context of the impact of intervention in decreasing the number of infections and deaths.

This paper provides a detailed analysis of the statistical data, including the number of tested, positive, mortality cases on COVID-19 spread in NY State and the UK and presents a solution to the problem of identification of unknown epidemiological parameters (transmissibility; the initial number of infected individuals; probability of being tested, etc.) in an ABM. The problem has been considered as the minimization of a target functional in relation to daily numbers of tested, positive and mortality cases in the studied regions and become an important modification of the Covasim package (Household Size and 2019, 2019). The minimization problem has been solved using the Optuna hyperparameter optimization software and the Parzen estimation method. Normal gradient descent methods do not work with Covasim or other agent-based models, due to the stochastic variability between model runs that makes the landscape very “bumpy” (i.e., many transient local minima). One way of getting around this is to perform many different runs and take the average. However, averaging over many runs is computationally expensive, since running N simulations of ABM will only reduce the noise by $\sqrt{N}$ (Covasim documentation: ht).

The results of data analysis in every studied region showed the weekly seasonality of New Tests, which helped us forecast the future values of this time series. It worth noting that some countries have their own rules of statistical analysis that have to be accounted for when carrying out modeling, e.g. in the USA, the New Diagnoses indicator contains a certain percentage of probable cases, while New Deaths in the UK accounts for all the deaths that have occurred within 28 days since a positive COVID-19 test, even if such death has not been provoked by the virus.

Due to the high sensitivity of the transmissibility parameter, the accuracy of its identification becomes crucial for uncovering the pattern of COVID-19 spread in an investigated region (Krivorotko et al., 2021). However, not all containment

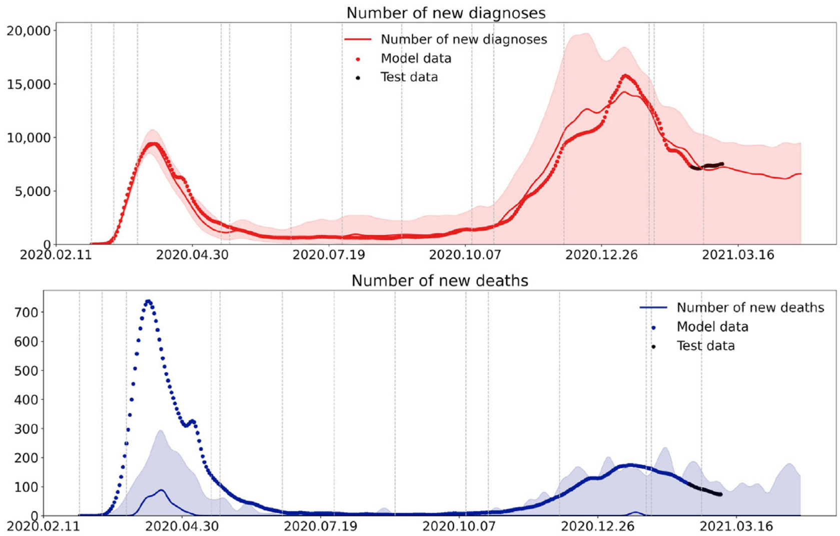


Fig. 11. Model calibration results for 10 simulations + a 45-day forecast for New Diagnoses (the top graph) and New Deaths (the bottom graph) in NY State. The shaded areas are 10% and 90% quantiles, the solid line – the median value of modeling result, and dots – real data. The dashed vertical lines are COVID-19 containment measures.

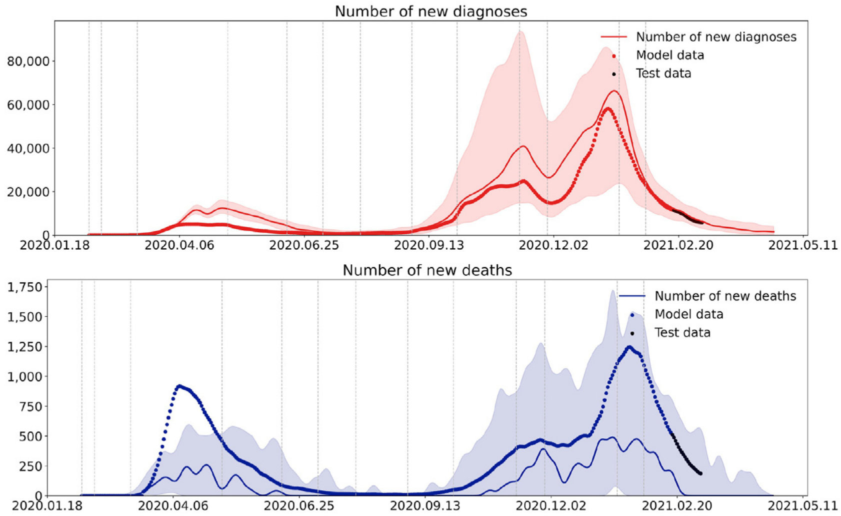


Fig. 12. Model calibration results for 10 simulations + a 45-day forecast for New Diagnoses (the top graph) and New Deaths (the bottom graph) in the UK. The shaded areas are 10% and 90% quantiles, the solid line – the median value of modeling result, and dots – real data. The dashed vertical lines are COVID-19 containment measures.

measures affect the pattern. For that reason, when developing an ABM calibration algorithm based on epidemiological data, we paid special attention to transmissibility and the times this parameter changed while modeling. These characteristics were determined from solving the minimization problem as a piecewise-constant function, while solving the inverse problem restored the parameter together with its times of change.

The devised ABM has been validated with historical data. The modeling results for the two regions in question have demonstrated that preserving the introduced containment measures would have sustained New Diagnoses in NY State during March 2021 and would have reduced them in the UK.

The proposed agent-based model has the following limitations: we do not fix population number changing during model year, consider waning of immunity to coronavirus, or the possibility of re-infection. We also use more simplistic contact structures than in real life.

Our future plans are investigation of model identifiability to real data and sensitivity analysis. It will also be necessary to investigate the influence of vaccination on COVID-19 propagation.

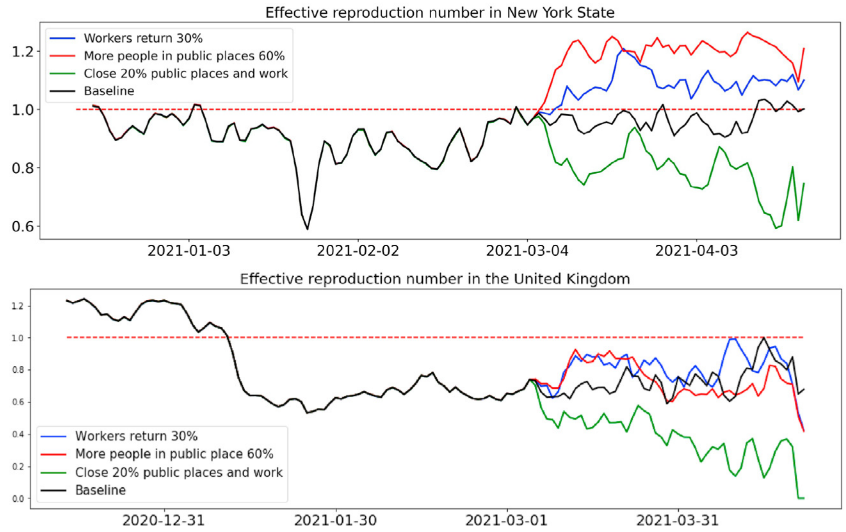


Fig. 13. Effective reproduction number in New York State (top) and in the United Kingdom (bottom) for four considered scenarios. The red dashed line represents $\mathcal{R}(t) = 1$.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The data analysis part (section 2) is supported by the Russian Foundation for Basic Research and Royal Society (project no. 21-51-10 003). The agent-based mathematical model construction and analysis of numerical results (sections 3, 4, 5) was supported by the Russian Science Foundation (project no. 18-71-10 044) and the Royal Society IEC∖R2∖202 020 − International Exchanges 2020 Cost Share between UK and Russia. Authors would like to thank Professor Sergey Kabanikhin and Doctor Alexey Romanyukha for the problem analysis and fruitful discussions. Also we would like to thank Yan Reznichenko for language help.

References

Bartolucci, F., Cardinali, A., & Pennoni, F. (2018). A generalized moving average convergence/divergence for testing semi-strong market efficiency. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, 101–105. https://doi.org/10.1007/978-3-319-89824-7_18

Bergstra, J., Bardenet, R., Bengio, Y., et al. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, 24.

Chaddock, R. E. (1925). Principles and methods of statistics. *The Annals of the American Academy of Political and Social Science, 123*(1), 229–229.

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, *589*, 82–87. https://doi.org/10.1038/s41586-020-2923-3

Cooper, I., Mondal, A., & Antonopoulos, C. G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals, 139*(110057). https://doi.org/10.1016/j.chaos.2020.110057

Coronavirus (COVID-19) in the UK: https://coronavirus.data.gov.uk/.

Covasim documentation: https://docs.idmod.org/projects/covasim/en/latest/index.html.

The COVID Tracking Project in USA: https://covidtracking.com/data.

Cuevas, E. (2020). An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Computers in Biology and Medicine, 121*(103827). https://doi.org/10.1016/j.compbiomed.2020.103827

Dabral, P. P., & Murry, M. Z. (2017). Modelling and forecasting of rainfall time series using SARIMA. *Environmental Processes, 4*, 399–419.

Hazelbag, C. M., Dushoff, J., Dominic, E. M., Mthombothi, Z. E., & Delva, W. (2020). Calibration of individual-based models to epidemiological data: A systematic review. *PLoS Computational Biology, 16*(5), Article e1007893. https://doi.org/10.1371/journal.pcbi.1007893

Household size. UN https://population.un.org/Household/#/countries/840, (2019).

Kabanikhin, S., I., & Krivorotko, O., I. (2020). Mathematical modeling of the Wuhan COVID-2019 epidemic and inverse problems. *Computational Mathematics and Mathematical Physics, 60*(11), 1889–1899. https://doi.org/10.1134/S0965542520110068

Kai, D., Goldstein, G. F., Morgunov, A., Nangalia, V., & Rotkirch, A. (2020). *Universal masking is urgent in the COVID-19 pandemic: SEIR and agent-based models, empirical validation, policy recommendations*. arXiv preprint arXiv:2004.13553. https://doi.org/10.13140/RG.2.2.21662.08001

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., et al. (2021). Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLoS Computational Biology, 17*(7), Article e1009149. https://doi.org/10.1371/journal.pcbi.1009149

Krivorotko, O., I., Kabanikhin, S., I., Sosnovskaya, M., I., & Andornaya, D., V. (2021). Sensitivity and identifiability analysis of COVID-19 pandemic models. *Vavilov Journal of Genetics and Breeding, 25*(1), 82–91. https://doi.org/10.18699/VJ21.010

Mellacher, P. (2020). *COVID-Town: An integrated economic-epidemiological agent-based model*. Munich Personal RePEc Archive Paper. No. 103661.

Ndairoua, F., Area, I., Nieto, J. J., & Torresa, D. F. M. (2020). Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. *Chaos, Solitons & Fractals, 135*(109846). https://doi.org/10.1016/j.chaos.2020.109846

OPTUNA: hyperparameter optimization framework: https://optuna.org/.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press, 026218253X.

Silva, P. C. L., Batista, P. V. C., Lima, H. S., Alves, M. A., Guimarães, F. G., & Silva, R. C. P. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals, 139*(110088). https://doi.org/10.1016/j.chaos.2020.110088

Tracking Coronavirus in New York: Latest Map and Case Count https://www.nytimes.com/interactive/2021/us/new-york-covid-cases.html.

Wolfram, C. (2020). An agent-based model of COVID-19. *Complex Systems, 29*(1), 87–105. https://doi.org/10.25088/ComplexSystems.29.1.87