



Research article

Fast screening framework for infection control scenario identification

Yohei Kakimoto^{1,*}, Yuto Omae¹, Jun Toyotani¹ and Hirotaka Takahashi²

¹ College of Industrial Technology, Nihon University, 1-2-1 Izumicho, Narashino, Chiba 275-8575, Japan

² Research Center for Space Science, Advanced Research Laboratories, Tokyo City University, 8-15-1 Todoroki, Setagaya, Tokyo 158-0082, Japan

* **Correspondence:** Email: kakimoto.youhei@nihon-u.ac.jp; Tel: +81474742628.

Abstract: Due to the emergence of the novel coronavirus disease, many recent studies have investigated prediction methods for infectious disease transmission. This paper proposes a framework to quickly screen infection control scenarios and identify the most effective scheme for reducing the number of infected individuals. Analytical methods, as typified by the SIR model, can conduct trial-and-error verification with low computational costs; however, they must be reformulated to introduce additional constraints, and thus are inappropriate for case studies considering detailed constraint parameters. In contrast, multi-agent system (MAS) simulators introduce detailed parameters but incur high computation costs per simulation, making them unsuitable for extracting effective measures. Therefore, we propose a framework that implements an MAS for constructing a training dataset, and then trains a support vector regression (SVR) model to obtain effective measure results. The proposed framework overcomes the weaknesses of conventional methods to produce effective control measure recommendations. The constructed SVR model was experimentally verified by comparing its performance on datasets with expected and unexpected outputs. Although datasets producing an unexpected output decreased the prediction accuracy, by removing randomness from the training dataset, the accuracy of the proposed method was still high in these cases. High-precision predictions of the MAS-based simulation output were obtained for both test datasets in under one second of the computational time. Furthermore, the experimental results establish that the proposed framework can obtain intuitively correct outputs for unknown inputs, and produces sufficiently high-precision prediction with lower computation costs than an existing method.

Keywords: fast screening framework; multi-agent system; support vector regression; infection control; identifying effective scheme

1. Introduction

Research on prediction methods for infectious disease transmission has increased with the spread of the coronavirus disease (COVID-19) [1–7]. In a society where infectious disease is widespread, an infection control strategy using techniques such as policies and infection prevention measures is required to contribute to reducing infection rates. Control strategies need to be determined promptly because social conditions change over time.

The SIR model is a well-established analytical method for infectious disease transmission, and uses differential equations to express the change in infection status between infectious and recovered [8]. This model is extended to various models such as the SEIR model, which divides infection status into exposed and infectious, the SEIRD model, which includes “dead” as a status, and the SIRVD model, which considers vaccination [9–11]. These models can be rapidly calculated by solving differential equations. However, as differential equations must be reformulated when a new status or condition is adopted, integrating detailed policies for infection prevention measures is difficult. In addition, the infection rate in SIR or its extended models based on differential equations is represented by a single parameter. Notably, this parameter does not represent the original infectivity of a virus. The parameter is determined by infection control measures, e.g., wearing a mask and maintaining social distance on a micro-scale, and lockdown and stay-at-home orders on a macro-scale, as well as virus infectivity. Consolidating various measures into a single parameter, these analytical models cannot consider detailed infection control measures.

Therefore, Farooqa et al. [11] proposed the framework using an adaptive deep learning model to improve prediction accuracy for the model parameters in the SIRVD model, i.e., an analytical model. The framework simultaneously learns model parameters and updates the model. In addition, Farooqa et al. ensured a more realistic prediction by dividing the infection rate parameter into lockdown existence and non-existence. In contrast, this framework can learn only a unique scenario, i.e., a single combination of various infection control measures that have been implemented in a specific region in the real world. Furthermore, because the infectivity determined by various complex and entangled elements consolidates into one dimension parameters, it is difficult to verify combinations of infection control measures that have not been implemented in the real world by this framework.

A multi-agent system (MAS) simulates the movement of the entire society by arranging multiple agents in a virtual space and creating specific rules to represent agent behavior. In MAS, detailed conditions and agent status are easily given, for example, events due to agent behavior. Furthermore, an MAS is a more flexible analytic method. For these reasons, MAS is often used to verify the effects of infection prevention measures [3,5–7,12–14]. However, the results can be unstable if the simulation outputs (e.g., the total number of infected individuals) are not averaged using multiple random seeds. Moreover, calculation cost rapidly increases for large-scale agents and large virtual space. For example, Omae et al. [7] reported 12 hours average calculation time per scenario using 300 random seeds for an epidemic infection simulation of 45 days and 999 peoples. Thus, the use of MAS for screening many infection control scenarios is not appropriate owing to the calculation time.

Hirose proposed a hybrid method using a combination of multi-agent and differential equations (MADE) [15]. The MADE aims for the early estimation of the total number of infected individuals and simulates infection transition in the early stage by MAS and subsequently performs the infection process using the SEIR model based on the information obtained by MAS. Specifically, the

parameters of the SEIR model, such as the rate of infection, are estimated based on the number of individuals by infection status of the MAS simulation at time t , and the simulation is performed using the SEIR model. As the MAS simulation requires large computation costs in the middle stage owing to increasing infection events, MADE can switch the simulation method from MAS to the SEIR model from the middle to the last stage and thereby reducing the computation costs. However, MADE cannot handle a situation in which a scheme, e.g., lockdown and stay-at-home orders, is implemented in the middle stage of the simulation because the SEIR model that is used after the middle stage is a deterministic method.

To overcome the limitations of the analytical and MAS methods and existing methods, this study proposes a framework for fast screening of MAS-based simulation configurations using a machine learning (ML) model. The ML model dataset is produced by the MAS, has multiple inputs, and outputs the total number of infected individuals. The MAS-based simulator developed by Omae et al. [5, 6] for infectious disease transmission was adopted to create a dataset with multiple scenarios. The framework is then constructed by using the dataset to train a support vector regression (SVR) model. Our framework is called MAS-SVR. SVR is an actively researched type of support vector machine and is able to conduct non-linear regression [16–21]. If the dataset is not too large, SVR can conduct learning in a realistic time and produce high-precision predictions when compared to other regression methods. Exploiting the features of SVR, MAS-SVR can quickly and precisely predict the MAS simulation result while requiring calculation costs in making the dataset and the learning process. Thus, MAS-SVR simplifies screening for many scenarios by preparing the model before the infection prevention measure is required. Note that the MAS-SVR only predicts the total number of infected individuals at the end of the MAS-based simulation, i.e., it does not output any data regarding progress during the simulation runtime. Thus MAS-SVR acts as a screening framework for many MAS scenarios.

The performance of MAS-SVR was compared with that of the original MAS and an existing method with a similar aim in numerical experiments to establish the superiority of MAS-SVR. In addition, the MAS-SVR model behavior was observed for inputs and outputs not expected in the dataset created by the MAS. Furthermore, the MAS-SVR was validated by sensitivity analysis for the parameters related to stay-at-home orders that are considered effective measures under the novel coronavirus pandemic [22–24]. The experimental results confirmed that the MAS-SVR model can accurately predict the output of the original MAS-based simulation even when the output is outside the range of the training data. Moreover, the results demonstrated the superiority of MAS-SVR compared to an existing method in terms of both estimation accuracy and computation costs. To be predictable, intuitively correct outputs could be generated by the MAS-SVR model, for instance, given unknown inputs such as stay-at-home orders under the coronavirus disease using sensibility analysis. As these results, we believe the dataset should be created with a policy for constructing a high-precision SVR model. By accepting complex inputs and performing fast calculation, MAS-SVR enables fast screening of many scenarios.

2. Combination of multi-agent and differential equations

A combination method, MADE proposed by Hirose [15] has a similar objective as this study in terms of verifying detailed infection control measures with fewer computation costs. Thus, this section

reviews the MADE focusing on its specific methods.

The main idea behind MADE is to simulate the early stage of an epidemic infection spread using MAS which can incorporate detailed infection control measures. The MADE estimates the parameters of the SEIR model represented by differential equations based on the number of infected individuals obtained in the MAS simulation.

The SEIR model is described by the following differential Eqs (2.1)–(2.4):

$$\frac{dS(t)}{dt} = -\lambda S(t)I(t), \quad (2.1)$$

$$\frac{dE(t)}{dt} = \lambda S(t)I(t) - \sigma E(t), \quad (2.2)$$

$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma I(t), \quad (2.3)$$

$$\frac{dR(t)}{dt} = \gamma I(t), \quad (2.4)$$

where S , E , I , and R represent susceptible, exposed, infectious, and recovered, respectively. Then, λ , σ , and γ are the parameters that express infection rate, removal rate, and transmission rate, respectively. The total population is conserved, i.e., $S(t) + E(t) + I(t) + R(t) = \text{const}$, $\forall t$.

The MADE obtains the approximate parameters λ , σ , and γ at time t based on the number of individuals of each status obtained by an MAS simulation at time t and $t + 1$ by following simultaneous differential equations:

$$\lambda(t) = \frac{S(t) - S(t + 1)}{S(t)I(t)}, \quad (2.5)$$

$$\sigma(t) = \frac{\{E(t) - E(t + 1)\} + \{S(t) - S(t + 1)\}}{E(t)}, \quad (2.6)$$

$$\gamma(t) = \frac{R(t + 1) - R(t)}{I(t)}. \quad (2.7)$$

Note that time t in Eqs (2.1)–(2.4) is continuous whereas in Eqs (2.5)–(2.7), time t is regarded as discrete. The time t at where the simulation method is switched from MAS to differential equations is called connecting time.

The MADE can quickly estimate infected individuals by solving the SEIR model (2.1)–(2.4) with the obtained parameters as constant. However, this method has the problem that if some schemes, e.g., lockdown and stay-at-home orders, are implemented after an MAS simulation, its effects on the parameters (2.5)–(2.7) are ignored.

3. Proposed framework

3.1. Outline of the framework

The MAS-SVR framework trains a machine learning model using a dataset comprising input configurations and MAS-based simulation outputs. Once the dataset is generated, this framework performs fast screening of many scenarios, while the MAS component provides the ability to consider detailed constraints.

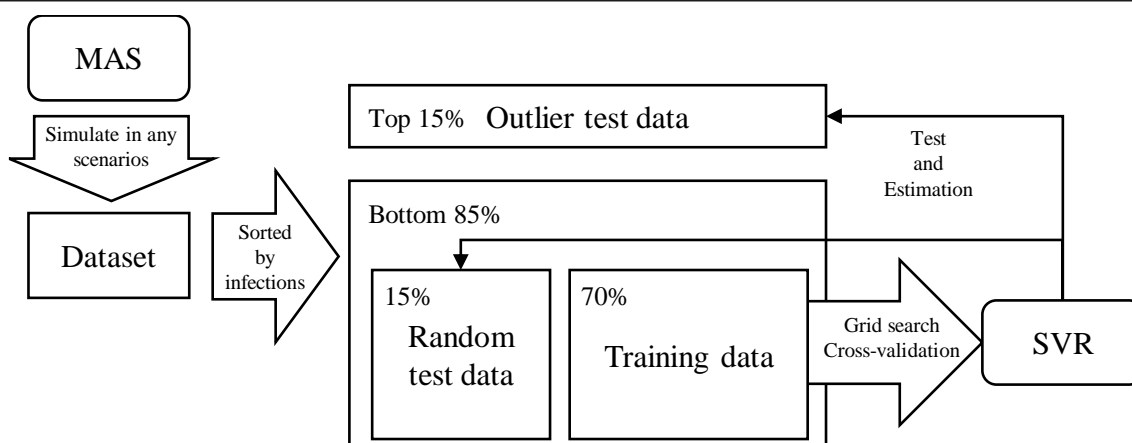


Figure 1. Diagram of the MAS-SVR framework.

Table 1. Advantages and disadvantages of each simulation method.

	Analytical expression	MAS-based simulation	MAS-SVR model
Input flexibility	×	○	○
Computational cost	○	×	○
Detailed output	△	○	×

Note: ○... Strong, △... Normal, ×... Weak.

Firstly, we consider the selection of the machine learning method used to predict the number of infected individuals. Support vector regression (SVR) is a regression method that prevents overfitting by considering a regularization term and an error term is defined as the difference between the measured estimated values. In addition, SVR is able to conduct nonlinear regression using the kernel trick. SVR is not appropriate for handling very large datasets as the $n \times n$ gram matrix must be calculated, where n is the number of training data. In the proposed framework, however, very large datasets will not be used as the MAS computational cost of producing a very large dataset is too high. In addition, since approximating gram matrix and fast learning algorithms are proposed in many studies [25–27], we consider it to be a minor problem. Therefore, this study adopts SVR as the machine-learning model, and refers to the SVR model constructed by this framework as the MAS-SVR model. Figure 1 indicates the method and structure diagram of the MAS-SVR. Note that MAS-SVR is the process framework that generates a dataset using MAS, trains the SVR model on the dataset, and predicts the MAS-based simulation output. However, the use of the SVR model is non-essential, i.e., this framework is applicable to other instances if MAS is available.

Table 1 summarizes the advantages and disadvantages of the SIR model (analytical expression) MAS-based simulation, and MAS-SVR model.

3.2. MAS-based simulator for the infectious disease transmission

This study assumes multiple scenarios for infectious disease transmission. We generated the dataset of inputs and outputs by conducting infection simulations using the MAS-based simulator proposed in [5, 6] based on multiple scenarios.

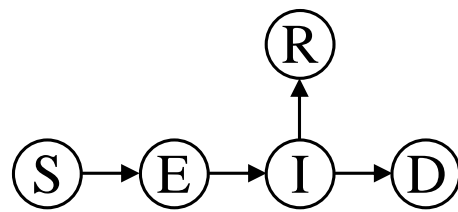


Figure 2. Status transitions in the MAS [5, 6]: Susceptible (S), Exposed (E), Infectious (I), Recovered (R), or Dead (D).

The MAS-based simulator [5, 6] assigns each agent a status of either susceptible, exposed, infectious, recovered, or dead. Figure 2 indicates the transition paths between these statuses. Essentially, it is necessary to discuss whether an exposed person has infectivity toward a susceptible person; however, this study assumes that exposed agents have infectivity based on [5, 6]. Although typically the transition from recovered to susceptible status is considered, we have not included this transition as this study only uses a short simulation period.

Table 2 shows the values of the 24 simulation parameters used in this study. The simulator assumes that three occupation categories of agents, namely, a worker, homemaker, and student, live in each house. Hence, 250 houses would have a population of 750 in the simulated space. Additionally, each agent has a different “departure time”, “trip probability”, and time spent outside home. If the distance between a susceptible person and an infectious or exposed person drops below a threshold, infection events occur. The transition probability from susceptible to exposed is the “infection probability”. Stay-at-home orders are issued when the number of infected individuals reaches the “start condition”. In addition, when the number of infected individuals falls below the “removal condition”, the stay-at-home orders are canceled. The “trip probability” of agents is decreased to “reduced level of trip probability” if stay-at-home orders are issued. Although infection events occur due to the distance between agents, it is unrealistic to make infection events occur in identical coordinates where many agents gather, such as offices, houses, and schools. To resolve this problem, the simulator generates small-scale networks (SSNs) in these identical coordinates and causes infection events to occur in each SSN. Please refer to [5, 6] for further parameter and simulator details.

Making a dataset with some scenarios, it is difficult to target all of the parameters in Table 2 in terms of the computation cost. Therefore, this study focuses on the five parameters indicated in Table 3, which are regarded as the dataset inputs. Combinations of these parameters, that is, $3^5 = 243$ scenarios, are created and the dataset generated by simulating each scenario using the simulator. If a MAS-based simulator includes random events, the result depends on random numbers, meaning a different result is obtained for each simulation. Our simulator instead used 25 random seeds to simulate each scenario 25 times to remove the random number dependency. The results were averaged to produce the output of the dataset. Thus, a total of $243 \cdot 25 = 6075$ simulations were conducted.

3.3. Support vector regression

The regression equation is defined as

$$f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b, \quad (3.1)$$

Table 2. Parameters of the MAS-based simulator for infectious disease transmission.

Parameters	Values
Simulation period	29 [days]
Houses	Refer to Table 3
Initial symptomatic agents	10 [peoples]
Trip prob. (workers)	99.0–100 [%]
Trip prob. (homemakers)	50.0–100 [%]
Trip prob. (students)	99.0–100 [%]
Departure time (workers)	8:30:00 ± 1:30:00
Departure time (homemakers)	10:30:00 ± 1:30:00
Departure time (students)	8:30:00 ± 1:30:00
Stay time outside (workers)	6:00:00–8:00:00
Stay time outside (homemakers)	0:10:00–0:30:00
Stay time outside (students)	5:00:00–6:00:00
Companies	10 [locations]
Stores	10 [locations]
Schools	10 [locations]
Infection prob.	Refer to Table 3
Prob. of hospital visiting	60.0 [%]
Capacity of isolation wards	Refer to Table 3
Fatality rate (Non hospitalization)	10.0 [%]
Fatality rate (hospitalization)	1.0 [%]
Start cond. (stay-at-home)	Refer to Table 3
Removal cond. (stay at home)	10 [peoples]
Reduced level of trip prob.	Refer to Table 3
Maximum SSN size	25 [peoples]

Table 3. Input configuration parameters of the ML training dataset.

Parameters	A	B	C
Houses [unit]	250	500	750
Infection prob. [%]	0.03	0.06	0.09
Reduced trip prob. levels [%]	30	60	90
Capacity of isolation ward [bed]	0	5	10
Start cond. (stay-at-home) [peoples]	75	150	225

where ϕ is a nonlinear map $\phi : X \mapsto \mathfrak{Z}$ from input space X to feature amount space \mathfrak{Z} and $\langle \cdot, \cdot \rangle$ is the inner product. Hereinafter $f(\mathbf{x}; \mathbf{w}, b)$ is abbreviated as $f(\mathbf{x})$. SVR determines parameters \mathbf{w} and b by minimizing the following loss function:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i \in N} \max(0, |y_i - f(\mathbf{x}_i)| - \varepsilon), \quad (3.2)$$

where $N = \{1, 2, \dots, n\}$. The first term of Eq (3.2) is the regularization term to prevent overfitting and the second is an error term (the difference between measured and estimated values). The constants C and ε denote the weights of the error term and the width of the dead zone for errors, respectively. Note that both constants are the SVR hyperparameters. As errors less than hyperparameter ε are ignored, SVR is robust to error.

Equation (3.2) can be rewritten as the following optimization problem by introducing slack variables ξ_i and ξ_i^* :

$$\text{Minimize } \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i \in N} (\xi_i + \xi_i^*), \quad (3.3)$$

$$\text{subject to } y_i - f(\mathbf{x}_i) \leq \varepsilon + \xi_i, \forall i \in N, \quad (3.4)$$

$$-y_i + f(\mathbf{x}_i) \leq \varepsilon + \xi_i^*, \forall i \in N, \quad (3.5)$$

$$\xi_i, \xi_i^* \geq 0, \forall i \in N. \quad (3.6)$$

The regression equation in Eq (3.1) is expressed as the following by arranging the Lagrange function of the optimization problem (3.3)–(3.6) using the Karush-Kuhn-Tucker conditions,

$$f(\mathbf{x}) = \sum_{i \in N} (\alpha_i - \alpha_i^*) \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + b, \quad (3.7)$$

where α_i and α_i^* are the Lagrange multipliers. These multipliers are obtained by solving the dual optimization problem for the problem (3.3)–(3.6). The kernel function determined by $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$ is defined as follows:

$$k(\mathbf{x}_i, \mathbf{x}) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle. \quad (3.8)$$

Then, using Eq (3.8), Equation (3.7) is rewritten as

$$f(\mathbf{x}) = \sum_{i \in N} (\alpha_i - \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}) + b. \quad (3.9)$$

It is unnecessary to replace the inner product of Eq (3.7) with the kernel function Eq (3.8) to directly determine the non-linear map $\phi(\mathbf{x})$. In most cases, the radial basis function (RBF) kernel is used as a kernel function for the SVR, defined as

$$k(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2). \quad (3.10)$$

Equation (3.9) enables the RBF kernel to conduct non-linear regression; this addition greatly improves the performance of the SVR. The constant γ in Eq (3.10) is an SVR hyperparameter. When γ decreases, the influence of one training data point is far reaching.

4. Numerical experiment

4.1. Methods

4.1.1. Verification of MAS-SVR

In this section, we describe the method used to construct and verify the MAS-SVR model, using the dataset obtained by the MAS-based simulation [5,6]. The dataset was created using the five parameters in Table 3, producing 243 scenarios, and was divided into training and test datasets. As the MAS-based simulation [5, 6] includes random event parameters, the results can change even for the same input configuration. Thus, 25 simulations per scenario were conducted using 25 random seeds to reduce the influence of randomness, and the averaged value of their outputs were adopted as the central value of each scenario. Note that the output of MAS, i.e., infected individuals, is the total number of agents excluding susceptible agents at the end of the MAS simulation.

As the effect of random numbers is removed from the dataset, the trained MAS-SVR model predicts the results for the test dataset with high accuracy. Conversely, if the output obtained by a certain input configuration is out of the training dataset range (hereinafter called the training range), it may not apply. Thus, the accuracy of SVR constructed using the MAS-derived dataset may be different depending on if the true output from an input configuration is also in the training range. Therefore, we prepared two test datasets types for this study: a random sampling test dataset in the training range, and an outlier test dataset, containing outputs where the number of infected individuals is larger than the training range upper bound. The SVR is constructed using these two test datasets. Any combination of outputs (infected individuals) of the training data y_{train} , random sampling test data $y_{\text{test}}^{\text{rs}}$, and outlier test data $y_{\text{test}}^{\text{ol}}$ have the relationship

$$y_{\text{train}}, y_{\text{test}}^{\text{rs}} < y_{\text{test}}^{\text{ol}}.$$

The construction and estimation of the MAS-SVR model were conducted in the ratio of 70, 15, and 15%, for the training data, random sampling test data, and outlier test data, respectively. The random sampling test and outlier test datasets were extracted from the bottom 85% and top 15% of outputs respectively. Note that, when extracting the test dataset, a significant difference between the training and random sampling test datasets must not occur. We confirmed via t-test that no significant difference existed between the training and random sampling test datasets ($p \simeq 0.925 > 0.05$).

For estimation of the MAS-SVR models, the determination coefficient

$$R^2 = 1 - \frac{\sum_{i \in N} (y_i - f(\mathbf{x}_i))^2}{\sum_{i \in N} (y_i - \bar{y})^2}, \quad (4.1)$$

is used.

4.1.2. Comparison with MADE

To establish the superiority of MAS-SVR, it was compared with MADE [15]. MAS-SVR and MADE predicted the total number of infected individuals at the end of the MAS simulations using the MAS dataset presented in Section 4.1.1. The prediction results of the two models were estimated by determining the coefficients in Eq (4.1).

MADE's accuracy changed at connecting time t [day] when the simulation method switched from the MAS to the SIER model. Thus, we estimated the prediction accuracy by varying the connecting time t [day] from 0 to 28 [day].

Note that in [15], MADE assumes that a susceptible person is infected only when encountering an infected person. Meanwhile, the MAS [5, 6] used in this study considers an exposed person as an infection source in addition to an infected person. Hence, we reformulate Eqs (2.1) and (2.2) in MADE as follows:

$$\frac{dS(t)}{dt} = -\lambda S(t)I(t) - \lambda S(t)E(t), \quad (4.2)$$

$$\frac{dE(t)}{dt} = \lambda S(t)I(t) + \lambda S(t)E(t) - \sigma E(t). \quad (4.3)$$

Moreover, the parameters in Eqs (2.5)–(2.7) are calculated such that a dead person in the MAS [5, 6] is regarded as a recovered person.

4.1.3. Sensitivity analysis for stay-at-home order

Since the spread of the novel coronavirus disease, numerous researchers have reported that lockdown and stay-at-home orders have contributed to the reduction in the number of infected individuals [22–24]. Several studies have demonstrated that simulating lockdown or stay-at-home orders using MAS is effective [6, 28]. Thus, we performed a sensitivity analysis to verify whether MAS-SVR is a valid screening method for validating stay-at-home orders.

The MAS dataset was divided into train and test data in the experiment described in Section 4.1.1. However, since the aim of this experiment was to confirm the effect of the stay-at-home order by MAS-SVR for unknown inputs, the accuracy of MAS-SVR need not be validated. Accordingly, to construct a model with a higher precision than the one constructed in the experiment described in Section 4.1.1, we utilized the entire MAS dataset as training data for MAS-SVR. Sensitivity analysis for the constructed MAS-SVR was performed using the two parameters indicated as “reduced trip probability levels” and “start condition (stay-at-home)” in Table 3.

4.2. Hyperparameter tuning

The SVR hyperparameters used in this study are C , ε , and γ . The combination method of grid search and k-fold cross-validation (GS+CV) is typically used for hyperparameter tuning, and is summarized in Figure 3.

The GS+CV method incurs high computation costs when comprehensively searching the parameter space, and Kaneko et al. [29] have proposed a more efficient search method. However, we chose to use GS+CV to determine the hyperparameters in this study because the parameter space of the prepared dataset was small. The candidates for each hyperparameter are as follows:

$$\begin{aligned} C &\in \{2^0, 2^1, \dots, 2^{10}\}, \\ \varepsilon &\in \{2^{-10}, 2^{-9}, \dots, 2^{-1}\}, \\ \gamma &\in \{2^{-10}, 2^{-9}, \dots, 2^{10}\}. \end{aligned}$$

The five folds were used in cross-validation (CV), and the validation data were extracted randomly from the training dataset. Moreover, the MAS-SVR model estimation for selected hyperparameters is

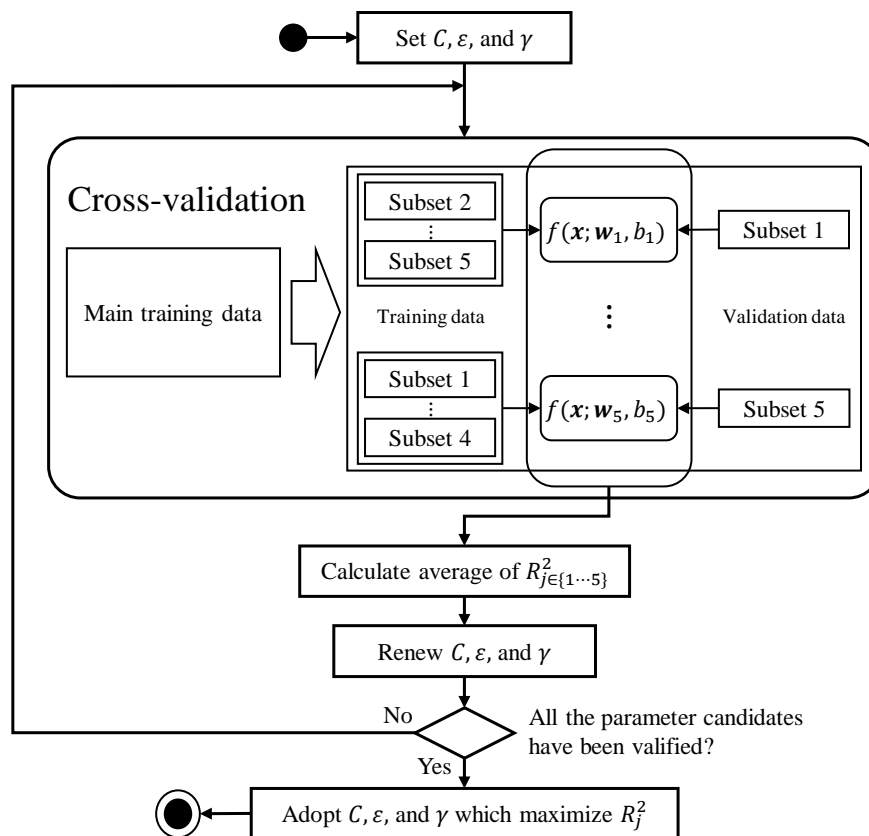


Figure 3. Flowchart of hyperparameter tuning.

conducted using the determination coefficient $R^2_{j \in \{1, \dots, 5\}}$, where j is the fold index. The combination of hyperparameters that obtains the highest R^2_j are adopted.

4.3. Results and discussions: Verification of MAS-SVR

In this section we assess the MAS-SVR model trained using the prepared dataset. The hyperparameter values used were $C = 2^7$, $\varepsilon = 2^{-9}$, and $\gamma = 2^{-3}$, which were obtained by adjustment based on the assumptions outlined in Section 3.2. The total number of infected individuals obtained by the MAS-based simulation was predicted using the MAS-SVR model for each dataset.

In the training dataset, the MAS-SVR model predicted the number of infected individuals with $R^2 \approx 0.9999$. The prepared dataset was created by MAS-based simulation, and the simulator outputs were determined by the MAS parameters. Although the probability events exist in the MAS, the outputs of the dataset are averaged. The training dataset exhibits a clear trend between each input and output; hence, the trained SVR model can predict the output with high precision.

Next, the prediction results were confirmed using the random sampling and outlier test datasets. Figure 4 shows the precision of the prediction results. The vertical and horizontal axes represent the measured and estimated values, i.e., the number of infected individuals obtained by the MAS-based simulation and the SVR model, respectively. The trained MAS-SVR model accurately predicted the outputs of both the random sampling test and training datasets.

For the outlier test dataset, the determination coefficient was high ($R^2 \approx 0.7617$), but lower than

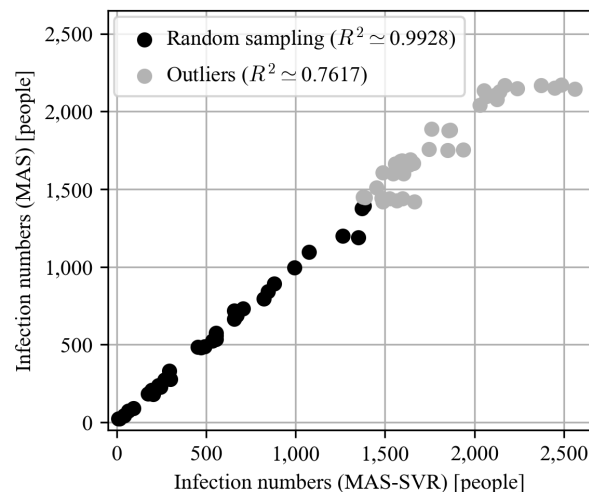


Figure 4. Prediction results for random sampling and outlier test dataset.

that for the random sampling test dataset. Although the outlier test dataset predicted the output range well for $y_{\text{test}}^{\text{ol}} < 2000$, the trained MAS-SVR model overestimated the number of infected individuals for $y_{\text{test}}^{\text{ol}} \geq 2000$, meaning the MAS-SVR model may not correctly predict outputs that lie outside the training range. Therefore, it is important to estimate the expected output range before creating an MAS-derived dataset.

The results in Figure 4 were obtained using the training data averaged from the outputs from 25 random seeds. As each seed result is dependent on a random number, the averaged output from a small number of seeds may have a lower precision than that from a large number of seeds. Figure 5 shows the trend of the determination coefficients with respect to the number of random seeds. The vertical and horizontal axes are the determination coefficient and number of random seeds, respectively, and results are plotted for the training, random sampling test, and outlier test datasets.

High-precision predictions were observed for the training and random sampling test datasets, even for low seed values. In particular, the determination coefficient was almost 1.00 for 10 random seeds. The prediction precision was significantly lower for fewer random seeds for the outlier test dataset, however the determination coefficients were stable and greater than 0.75 for seed numbers exceeding eight. Therefore, if the output range is known in advance, the MAS-SVR model can predict MAS-based simulation results precisely, even for small seed numbers. If the output range is unknown, the preparation of a sufficient number of random seeds and average outputs is required.

We now investigate the computation times for the MAS-based simulator and the MAS-SVR model. Figure 6 shows the computation time of the MAS-based simulator for various population scales and infection probability parameters. These two parameters were selected because they strongly affect computation costs. The values in each cell are the average computation time for results obtained by the combinations of 25 random seeds and all other parameters (excluding the two parameters that were being varied). The color depth indicates the length of time, and the values in brackets are the standard deviations.

Figure 6 clearly indicates that both parameters increase computation time in a similar manner. Some parameter combinations resulted in simulator computation times greater than ten hours for a single

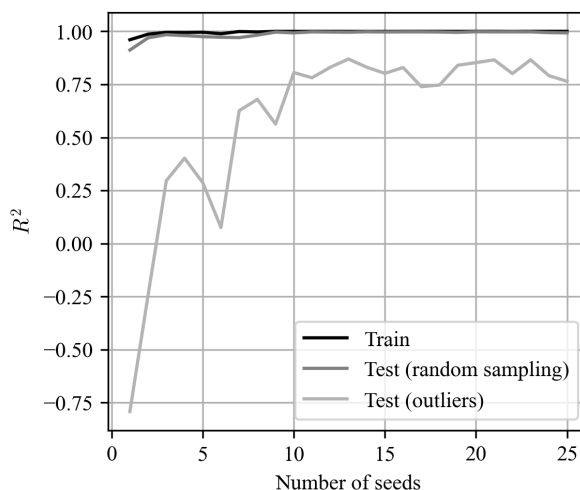


Figure 5. Determination coefficients obtained for different numbers of random seeds. When a large number of random seeds are used for averaging, the randomness is removed and the determination coefficients become stable and have high values.

scenario. This result confirms that the MAS-based simulation is not appropriate for screening many scenarios. In contrast, the computation times for the MAS-SVR model were less than 1 sec for all parameter combinations.

The MAS-SVR framework only predicts the result of the MAS-based simulation, and hence cannot observe the full simulation process. However, by screening scenarios using MAS-SVR, we can determine which scenarios require detailed verification. Therefore, it is important to combine the use of the MAS-SVR framework with MAS.

In this study we constructed and validated a prediction model for an infectious disease transmission case study. However, we believe that MAS-SVR will be applicable to other cases given the generalizability of MAS.

4.4. Results and discussions: Comparison with MADE

The estimation accuracy of MAS-SVR and MADE was compared by predicting the total number of infected individuals using the entire MAS test data, i.e., random sampling and outlier test data.

Figure 7 shows the trend of the determination coefficient R^2 in Eq (4.1) and infection rate λ in Eq (2.5) reflecting Eqs (4.2) and (4.3) obtained by MADE for varying connecting time t . The first vertical, second vertical, and horizontal axes represent R^2 , λ , and connecting time t , respectively. The black-dashed line represents the determination coefficient of MAS-SVR $R^2 \simeq 0.9819$ obtained using the same data.

As shown in Figure 7, the R^2 of MADE is not above the determination coefficient of MAS-SVR until the final stage of MAS simulation ($\simeq 71$ percentile of the simulation period). MAS can appropriate infection control measures such as stay-at-home orders and isolation wards from the middle stage of simulation. However, MADE cannot correctly estimate the infection rate λ unless such time-dependent measures have been reflected in the MAS simulation at the connecting time t . This is because MADE excessively estimates the infection rate λ if the connecting time t is small as shown by Figure 7. If

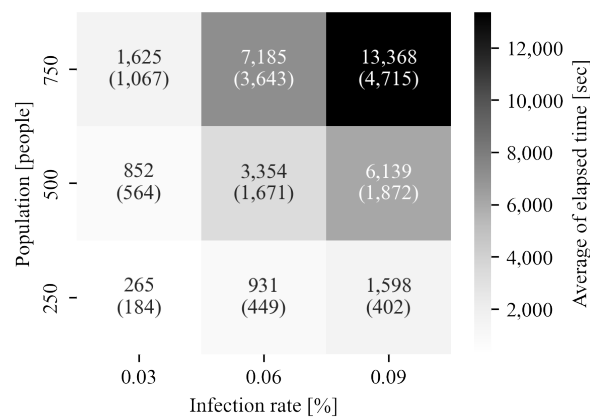


Figure 6. Averages and standard deviations of the MAS-based simulation computation time for varying population scale and infection rate configurations. Color depth indicates time length; standard deviation values are shown in brackets.

the connecting time t is larger than 10 [day], the R^2 of MADE stably increases because the infection control measures in the MAS simulation reflect the infection rate λ ; however, this result does not solve the problem of MAS indicated in Table 1. This shows that MADE has a trade-off relationship between computation costs and prediction accuracy. By contrast, the MAS-SVR proposed in this study can learn outputs such that time-dependent measures are reflected. Therefore, MAS-SVR can predict the total number of infected individuals with sufficiently high accuracy at low computation costs compared with MADE.

4.5. Results and discussions: Sensitivity analysis for stay-at-home order

SVR learns the entire MAS dataset as training data, for constructing high-precision MAS-SVR. The hyperparameters C , ε , and γ are determined by the method described in Section 4.2 with $C = 2^5$, $\varepsilon = 2^{-10}$, and $\gamma = 2^{-3}$.

The vertical and horizontal axes in both Figure 8(a),(b) are the number of infected individuals and the “start condition” of implementing stay-at-home orders. Figure 8(a) plots the number of infected individuals with MAS under initial conditions of 75, 150, and 225 [peoples] by the “reduced trip probability levels.” Figure 8(b) shows the number of infected individuals transitioning by the “reduced trip probability levels” of MAS-SVR.

If the “reduced trip probability level” equals 0.3, the number of infected individuals is approximately constant regardless of the start conditions in Figure 8(a),(b). As the “reduced trip probability level” equals 0.3 and 0.6, the trend and amount of infected individuals are approximately equal; however, they vary significantly between 0.6 and 0.9. To confirm the trend in this range, Figure 8(b) indicates the prediction results of MAS-SVR at 0.7 and 0.8 using dashed lines.

Figure 8(a),(b) reveal that MAS-SVR can predict the number of infected individuals in the same trend as MAS. In particular, MAS-SVR can predict the number of infected individuals who are intuitively correct at the “reduced trip probability levels,” equal to 0.7 and 0.8, which do not exist in the MAS dataset. These results imply that MAS-SVR is valid for screening infection control measures.

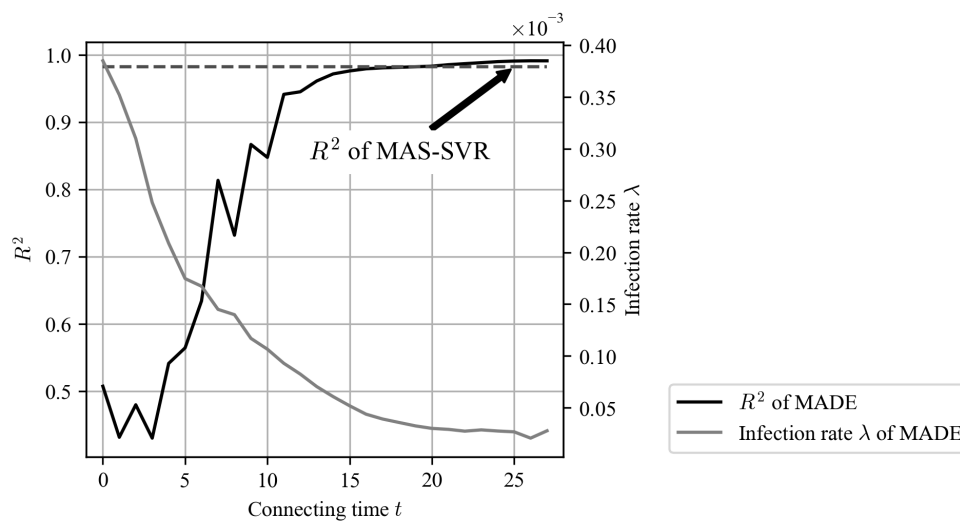


Figure 7. Determination coefficient R^2 transition of MADE for connecting time t . If the simulation method is switched at the early time t , MADE excessively estimates the infection rate λ and the prediction accuracy decreases. As the computation costs by MAS increase if time t is large, this trend indicates that MADE has a trade-off structure.

5. Conclusions

A fast-screening framework for infection prevention measures is proposed comprising of a MAS-derived dataset and an SVR. We used MAS-SVR to predict the training range and outlier test datasets. The results confirmed that MAS-SVR could predict the MAS-based simulation output for the training-range test dataset with remarkably high precision ($R^2 \approx 0.9928$). On the outlier test dataset, while the prediction precision decreased compared with that of the training range, the MAS-SVR model obtained high-precision prediction results ($R^2 \approx 0.7617$). Moreover, MAS-SVR can calculate prediction results with lower computational costs than the MAS-based simulator; for example, MAS-SVR calculated the result for one scenario within 1 sec, whereas the MAS-based simulator took over 10 hours to simulate one scenario. Thus, MAS-SVR is appropriate for screening many input configurations and hence many scenarios. In addition, the superiority of MAS-SVR in terms of computation costs and prediction precision was confirmed when compared to the existing method MADE [15] by revealing that MADE has a trade-off relationship between their two factors. Furthermore, we showed that MAS-SVR would be able to obtain intuitively correct output for unknown inputs using the sensitivity analysis.

As MAS-SVR only predicts the output of the MAS-based simulation, it can be effectively combined with MAS by screening the input configurations using MAS-SVR, and then verifying the obtained configuration with MAS. While this study used infectious disease simulation as the simulation case study, MAS-SVR is applicable to other cases where MAS is applicable; therefore, MAS-SVR may have implications in many fields.

The scale of the dataset obtained by MAS in this study is not large; however, if the scale increases, MAS-SVR may incur heavy computation costs because the efficiently learning methods of SVR have not been developed. Since the approximate algorithms for gram matrix and fast learning algorithms

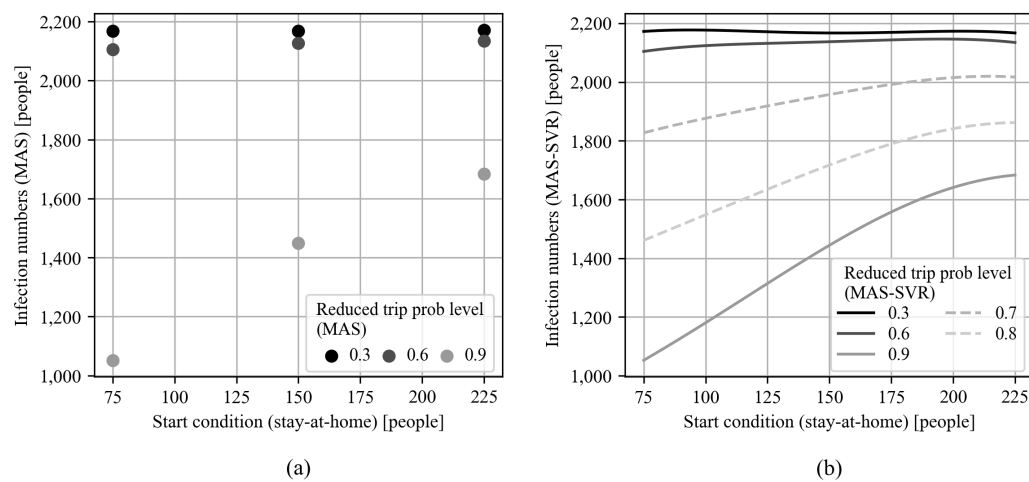


Figure 8. Transition of the number of infected individuals by “reduced trip probability levels” under the “start condition” of stay-at-home. Prediction by (a) MAS and (b) MAS-SVR. This implies that MAS-SVR can predict the number of infected individuals who are intuitively correct for inputs that do not exist in the train dataset.

have been proposed by various studies [25–27], we will consider the new MAS-SVR that is aimed at reducing computation costs in the future works.

Acknowledgments

This work was supported by JSPS Grant-in-Aid for Scientific Research (C) (Grant No.21K04535 to J.Toyotani).

Conflict of interest

The authors have no conflict of interest.

References

1. J. M. Carcione, J. E. Santos, C. Bagaini, J. Ba, A simulation of a COVID-19 epidemic based on a deterministic SEIR model, *Front. Public Health*, **8** (2020), PMC7270399. <https://doi.org/10.3389/fpubh.2020.00230>
2. G. Barwolff, Mathematical modeling and simulation of the COVID-19 pandemic, *Systems*, **8** (2020), 24. <https://doi.org/10.3390/systems8030024>
3. P. C. Silva, P. V. Batista, H. S. Lima, M. A. Alves, F. G. Guimaraes, R. C. Silva, COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions, *Chaos, Solitons Fractals*, **139** (2020), 110088. <https://doi.org/10.1016/j.chaos.2020.110088>

4. Y. Wei, J. Wang, W. Song, C. Xiu, L. Ma, T. Pei, Spread of COVID-19 in China: analysis from a city-based epidemic and mobility model, *Cities*, **110** (2021), 103010. <https://doi.org/10.1016/j.cities.2020.103010>
5. Y. Omae, Y. Kakimoto, J. Toyotani, K. Hara, Y. Gon, H. Takahashi, Reliability of multi-agent based infection simulator with parameters of isolation wards, *ICIC Express Lett., Part B Appl.*, **12** (2021), 577–586.
6. Y. Omae, Y. Kakimoto, J. Toyotani, K. Hara, Y. Gon, H. Takahashi, Impact of removal strategies of stay-at-home orders on the number of COVID-19 infectors and people leaving their homes, *Int. J. Innovative Comput. Inf. Control*, **17** (2021), 1055–1065. <https://doi.org/10.24507/ijicic.17.03.1055>
7. Y. Omae, J. Toyotani, K. Hara, Y. Gon, H. Takahashi, Effectiveness of the COVID-19 contact-confirming application (COCOA) based on multi-agent simulation, *J. Adv. Comput. Intell. Intell. Inf.*, **25** (2021), 931–943. <https://doi.org/10.20965/jaciii.2021.p0931>
8. W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, in *Proceedings of the Royal of London. Society A, Containing Papers of a Mathematical and Physical Character*, Seasonality and period-doubling bifurcations in an epidemic model, Royal Society, **115** (1927), 700–721. <https://doi.org/10.1098/rspa.1927.0118>
9. J. L. Aron, I. B. Schwartz, Seasonality and period-doubling bifurcations in an epidemic model, *J. Theor. Biol.*, **110** (1984), 665–679. [https://doi.org/10.1016/s0022-5193\(84\)80150-2](https://doi.org/10.1016/s0022-5193(84)80150-2)
10. I. Korolev, Identification and estimation of the SEIRD epidemic model for COVID-19, *J. Econom.*, **220** (2021), 63–85. <https://doi.org/10.1016/j.jeconom.2020.07.038>
11. J. Farooqa, M. A. Bazaz, A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies, *Chaos, Solitons Fractals*, **138** (2020), 110148. <https://doi.org/10.1016/j.chaos.2020.110148>
12. A. Sharma, S. Bahl, A. K. Bagha, M. Javaid, D. K. Shukla, A. Haleem, Multi-agent system applications to fight COVID-19 pandemic, *Apollo Med.*, **17** (2020), 41–43. https://doi.org/10.4103/am.am_54_20
13. A. Badica, C. Badica, M. Ganzha, M. Ivanovic, M. Paprzycki, Multi-agent simulation of core spatial SIR models for epidemics spread in a population, in *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, IEEE, Jaipur, (2020), 1–7. <https://doi.org/10.1109/ICRAIE51050.2020.9358293>
14. Y. Vyklyuk, M. Manylich, M. Skoda, M. M. Radovanovic, M. D. Petrovic, Modeling and analysis of different scenarios for the spread of COVID-19 by using the modified multi-agent systems—Evidence from the selected countries, *Results Phys.*, **20** (2021), 103662. <https://doi.org/10.1016/j.rinp.2020.103662>
15. H. Hirose, Pandemic simulations by made: A combination of multi-agent and differential equations, with novel influenza A (H1N1) case, *Information*, **16** (2013), 5365–5390.
16. B. E. Boser, I. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, ACM, Pittsburgh, (1992), 144–152. <https://doi.org/10.1145/130385.130401>

17. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297. <https://doi.org/10.1007/BF00994018>
18. D. Basak, S. Pal, D. C. Patranabis, Support vector regression, *Neural Inf. Process. Lett. Rev.*, **11** (2007), 203–224.
19. M. Awad, K. Rahul, *Efficient Learning Machines*, 1st edition, Apress, Berkeley, 2015.
20. B. Scholkopf, A. J. Smola, R. C. Williamson, P. L. Bartlett, New support vector algorithms, *Neural Comput.*, **12** (2000), 1207–1245. <https://doi.org/10.1162/089976600300015565>
21. J. Feng, L. Liu, D. Wu, G. Li, M. Beer, W. Gao, Dynamic reliability analysis using the extended support vector regression (X-SVR), *Mech. Syst. Sig. Process.*, **126** (2019), 368–391. <https://doi.org/10.1016/j.ymsp.2019.02.027>
22. S. Liu, T. Yamamoto, Role of stay-at-home requests and travel restrictions in preventing the spread of COVID-19 in Japan, *Transp. Res. Part A Policy Pract.*, **159** (2022), 1–16. <https://doi.org/10.1016/j.tra.2022.03.009>
23. L. Silva, D. F. Filho, A. Fernandes, The effect of lockdown on the COVID-19 epidemic in Brazil: evidence from an interrupted time series design, *Cad. Saude Publica*, **36** (2020), <https://doi.org/10.1590/0102-311x00213920>
24. J. H. Fowler, S. J. Hill, R. Levin, N. Obradovich, The effect of stay-at-home orders on COVID-19 cases and fatalities in the United States, preprint, medRxiv, 2020.04.13.20063628. <https://doi.org/10.1101/2020.04.13.20063628>
25. P. Drineas, M. W. Mahoney, Approximating a gram matrix for improved kernel-based learning, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **3559** (2005), 323–337. https://doi.org/10.1007/11503415_22
26. M. Tohme, R. Lengelle, F-SVR: A new learning algorithm for support vector regression, in *IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings (ICASSP)*, 2008. <https://doi.org/10.1109/ICASSP.2008.4518032>
27. P. Y. Hao, Pair-v-SVR: A novel and efficient pairing nu-support vector regression algorithm, *IEEE Trans. Neural Networks Learn. Syst.*, **28** (2017), 2503–2515. <https://doi.org/10.1109/TNNLS.2016.2598182>
28. P. M. Dunuwila, R. A. Rajapakse, Evaluating optimal lockdown and testing strategies for COVID-19 using multi-agent social simulation, in *2020-2nd International Conference on Advancements in Computing (ICAC)*, (2020), 240–245. <https://doi.org/10.1109/ICAC51239.2020.9357132>
29. H. Kaneko, K. Funatsu, Fast optimization of hyperparameters for support vector regression models with highly predictive ability, *Chemom. Intell. Lab. Syst.*, **142** (2015), 64–69. <https://doi.org/10.1016/j.chemolab.2015.01.001>



AIMS Press

©2022 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)