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

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Spatio-temporal modeling of COVID-19 prevalence and mortality using artificial neural network algorithms

Nima Kianfar$^a$,*, Mohammad Saadi Mesgari$^a$, Abolfazl Mollalo$^b$, Mehrdad Kaveh$^a$

$^a$ Faculty of Geodesy and Geomatics, K. N. Toosi University of Technology, Tehran 19967-15433, Iran
$^b$ Department of Public Health and Prevention Science, School of Health Sciences, Baldwin Wallace University, Berea, OH 44017, USA

ARTICLE INFO

Keywords:
Artificial neural network
COVID-19
GIS
Spatio-temporal analysis
Variable importance analysis

ABSTRACT

The outbreak of coronavirus disease (COVID-19) has become one of the most challenging global concerns in recent years. Due to inadequate worldwide studies on spatio-temporal modeling of COVID-19, this research aims to examine the relative significance of potential explanatory variables ($n = 75$) concerning COVID-19 prevalence and mortality using multilayer perceptron artificial neural network topology. We utilized ten variable importance analysis methods to identify the relative importance of the explanatory variables. The main findings indicated that several variables were persistently among the most influential variables in all periods. Regarding COVID-19 prevalence, unemployment and population density were among the most influential variables with the highest importance scores. While for COVID-19 mortality, health-related variables such as diabetes prevalence and number of hospital beds were among the most significant variables. The obtained findings from this study might provide general insights for public health policymakers to monitor the spread of disease and support decision-making.

List of Abbreviations

ANN artificial neural network
CW connection weights
FR fatality rate
GA Garson’s algorithm
GIS geographic information system
GR growth rate
PR prevalence rate
PR-IQR prevalence rate in interquartile range
MR mortality rate
MR-IQR mortality rate in interquartile range
MSE mean squared error
PD partial derivatives
RMSEIQR root mean square error in interquartile range
MCW modified connection weights
MS model selection
SLP single-layer perceptron
TMMR trimmed mean mortality rate
VIA variable importance analysis
VIF variance inflation factor

* Corresponding author.
E-mail address: Nimakianfar@email.kntu.ac.ir (N. Kianfar).

https://doi.org/10.1016/j.sste.2021.100471
Received 21 June 2021; Received in revised form 3 October 2021; Accepted 4 November 2021
Available online 11 November 2021
1877-5845/© 2021 Elsevier Ltd. All rights reserved.
Olden et al. (2004) proposed the connection weights approach and argued that this approach is the least biased method that can accurately quantify the contribution of each input variable in predicting the health outcome is difficult (Ripley, 2007).

Previous studies have utilized various ANNs topologies to quantify the contribution of explanatory variables on dependent outcomes. Duh et al. (1998) examined the neural interpretation diagram, Garson’s algorithm, and sensitivity analysis to understand neural network relation weights. They showed that by extending randomization methods to ANNs, the output was evaluated using empirical evidence. Their findings indicated that the most squares method outperformed other methods, which was consistent with the results of multiple linear regressions in terms of partial $R^2$ (Ozesmi and Ozesmi, 1999).

Because of the complexity of interactions between variables, particularly in large datasets, variable importance analysis (VIA) has gained attention in many practical applications (Ferretti et al., 2016). VIA is a critical task in classification or regression problems to improve model interpretability, computational costs, data storage, and ultimately provide a sparse model without sacrificing prediction capacity (Wei et al., 2015). Dealing with various balance scenarios, Duf et al. examined the neural interpretation diagram, Garson’s algorithm, and sensitivity analysis to understand neural network relation weights. They showed that by extending randomization methods to ANNs, the output was evaluated using empirical evidence. Their findings indicated that the most squares method outperformed other methods, which was consistent with the results of multiple linear regressions in terms of partial $R^2$ (Ozesmi and Ozesmi, 1999).

Table 1
Various indicators used as target values for prevalence.

| Indicator                  | Formula                                                                 |
|----------------------------|-------------------------------------------------------------------------|
| Prevalence rate (PR)       | $\frac{\text{Total COVID19 confirmed cases}}{\text{Total population}} \times 10^6$ |
| Prevalence rate in interquartile range (PR-IQR) | $\frac{\text{Total COVID19 confirmed cases (in IQR)}}{\text{Total population}} \times 10^6$ |
| Trimmed mean rate (TMR)    | $\frac{\text{Trimmed mean of COVID19 confirmed cases}}{\text{Total population}} \times 10^6$ |
| Growth rate (GR1)          | $\frac{\text{Cumulative number of confirmed cases at the beginning of the period}}{\text{Total COVID19 confirmed cases}}$ |

Table 2
Various indicators used as target values for mortality.

| Indicator                  | Formula                                                                 |
|----------------------------|-------------------------------------------------------------------------|
| Mortality rate (MR)        | $\frac{\text{Total COVID19 deaths}}{\text{Total population}} \times 10^6$ |
| Mortality rate in interquartile range (MR-IQR) | $\frac{\text{Total COVID19 deaths (in IQR)}}{\text{Total population}} \times 10^6$ |
| Trimmed mean mortality rate (TMMR) | $\frac{\text{Trimmed mean of COVID19 deaths}}{\text{Total population}} \times 10^6$ |
| Growth rate (GR2)          | $\frac{\text{Cumulative number of confirmed deaths at the beginning of the period}}{\text{Total COVID19 deaths}} \times 10^6$ |
| Fatality rate (FR)         | $\frac{\text{Total COVID19 deaths}}{\text{Total COVID19 confirmed cases}} \times 10^6$ |

Table 3
The category, name, and source of the variables.

| Category                        | Variable                                | Source                                               |
|---------------------------------|-----------------------------------------|------------------------------------------------------|
| Demographic (25 variables)      | Population, male (% of total population) | World bank (World Bank, February 1, 2021)                   |
|                                 | Population, female (% of total population) | World bank                                                  |
|                                 | Population ages 0-14 (% of total population) | World bank                                                  |
|                                 | Population ages 0-14, male (% of male population) | World bank                                                  |
|                                 | Population ages 0-14, female (% of female population) | World bank                                                  |
|                                 | Population ages 15-64 (% of total population) | World bank                                                  |
|                                 | Population ages 15-64, male (% of male population) | World bank                                                  |
|                                 | Population ages 15-64, female (% of female population) | World bank                                                  |
|                                 | Population ages 65 and above (% of total population) | World bank                                                  |
|                                 | Population ages 65 and above, male (% of male population) | World bank                                                  |
|                                 | Population ages 65 and above, female (% of female population) | World bank                                                  |
|                                 | Population density (people per sq. km of land area) | World bank                                                  |
|                                 | Urban population (% of total population) | World bank                                                  |
|                                 | Urban population growth (annual %) | World bank                                                  |
|                                 | Rural population (% of total population) | World bank                                                  |
|                                 | Rural population growth (annual %) | World bank                                                  |
|                                 | Population in the largest city (% of urban population) | World bank                                                  |
|                                 | Age dependency ratio (% of working-age population) | World bank                                                  |
|                                 | Birth rate, crude (per 1,000 people) | World bank                                                  |
|                                 | Death rate, crude (per 1,000 people) | World bank                                                  |
|                                 | Physicians (per 1,000 people) | World bank                                                  |
|                                 | Nurses and midwives (per 1,000 people) | World bank                                                  |
|                                 | Hospital beds (per 1,000 people) | World bank                                                  |
|                                 | Age dependency ratio, old (% of working-age population) | World bank                                                  |
|                                 | Age dependency ratio, young (% of working-age population) | World bank                                                  |
| Economic (19 variables)         | Labor force participation rate, total | World bank                                                  |
|                                 | Labor force participation rate, male | World bank                                                  |
|                                 | Labor force participation rate, female | World bank                                                  |
|                                 | Employment to population ratio, 15+, total | World bank                                                  |
|                                 | Employers, total (% of total employment) | World bank                                                  |
|                                 | Employers, male (% of male employment) | World bank                                                  |
|                                 | Employers, female (% of female employment) | World bank                                                  |
|                                 | Vulnerable employment, total | World bank                                                  |
|                                 | Unemployment, total | World bank                                                  |
|                                 | Unemployment with advanced education | World bank                                                  |
|                                 | Unemployment, male (% of male labor force) | World bank                                                  |
|                                 | Unemployment, female (% of female labor force) | World bank                                                  |
|                                 | International migrant stock | World bank                                                  |
|                                 | Poverty headcount ratio at national poverty lines | World bank                                                  |
|                                 | Inflation, consumer prices | World bank                                                  |
|                                 | GDP per capita | World bank                                                  |
|                                 | GDP per capita growth | World bank                                                  |
(2020) introduced the nonparametric \( mn - \chi^2 \) variable importance technique, which uses a multivariate continuous response system to select and rank the most influential variables. The method measures the dissimilarities between the distribution of errors caused by the base learner before and after permuting the variable. Casiraghi et al. (2020) used a prediction model, “an explainable machine learning decision system based on additive trees”, which processed clinical, radiological, and laboratory data of COVID-19 patients to predict the risk of severe outcomes. They combined Boruta and random forest in a 10-fold cross-validation scheme to produce variable importance estimates not affected by the presence of surrogates. Pasha et al. (2021) employed multiple linear regression and a nonlinear regression based on 43 socio-economic and meteorological variables of 31 counties in California, United States. They found that the total population, household income, occupation, and transportation are more influential on COVID-19 spread than other variables. Shaffiee Hashghenas et al. (2020) applied ANNs based on particle swarm optimization and differential evolution algorithms to prioritize climatic and urban factors. They found that population density and humidity were the most influential variables to predict the confirmed COVID-19 cases.

In addition to the machine learning algorithms, the geographic information system (GIS) is a robust tool for analysis and visualizing many public health problems (Mollaoğlu et al., 2015, 2018). Recent GIS-based research has shown that several factors such as air quality (Bashir et al., 2020), population flow (Zhang and Schwartz, 2020, Jia et al., 2020), and population density (Ahmadi et al., 2020, Ramirez and Lee, 2020) could contribute to the higher rates of COVID-19 morbidity and mortality. In the Caribbean, Moonsammy et al. (2021) applied spatial lag and linear regression models to identify spatial clusters of COVID-19 and the most influential socio-economic variables. They suggested that COVID-19 cases and deaths in the Caribbean have a spatial connection with mainland countries. They also concluded that population transmission could contribute to higher COVID-19 spread. The consequences of the COVID-19 outbreak on the environment have also been investigated in some studies. For instance, Anibade et al. (2021) examined the levels of three air pollutants, namely particulate matter (PM2.5), Black Carbon (BC), and Polycyclic Aromatic Hydrocarbons (PAHs), in Jamshedpur city, India. Their results indicated that the concentrations of the contaminants were reduced during the lockdown compared to unlock down circumstances and regular days. Gautam (2020) showed that India experienced a large decrease in air pollution concentration during the lockdown, which led to fewer deaths during the outbreak. Gautam (2020) also suggested that lockdowns could help Asian and European countries experience lower levels of NO2. On the other hand, in China, Wang et al. (2020) demonstrated that quarantine actions would not be sufficient to prevent severe air pollution despite reductions in transportation and industrial emissions.

COVID-19 transmission is not limited to national borders and geographical territories. The primary focus of many studies that utilized machine learning methods such as ANNs was limited to a specific geographic location and applied pure spatial analysis with few sets of parameters while disregarding the impact of various potential variables over time. Therefore, to bridge the gap, this study investigates the influence of a broad range of explanatory variables (\( n = 75 \)) on disease prevalence and mortality using VIA methods based on ANNs, across the globe. This research optimized ANNs structure using a weighted information criterion (WIC) index to improve modeling accuracy. Moreover, as COVID-19 has shown various behaviors and mutated several times, different indicators were used to estimate mortality and morbidity rates over time. For this purpose, nine targets have been used to study the neural network’s learning process with distinct desires.

2. Materials and methods

2.1. Data

The daily COVID-19 data were obtained from WHO (World Health Organization (WHO) 2021b) from the beginning of March 2020 to the end of February 2021. The data contained new confirmed COVID-19 cases and newly confirmed deaths for all countries. Moreover, nine different indicators were used to study the learning process of further modeling. The formula for each indicator can be found in Table 1 (for prevalence) and Table 2 (for mortality). We divided the COVID-19 data into four equal time intervals (3-month periods): early March 2020 to the end of May 2020 (Period 1), early June 2020 to the end of August 2020 (Period 2), early September 2020 to the end of November 2020 (Period 3), and early December 2020 to the end of February 2021 (Period 4). In addition to COVID-19 data, a set of 75 variables, including demographic, environmental, social, economic, cultural, health, and public transportation variables was compiled at the country level as explanatory variables. The category, name, and source of the variables are presented in Table 3.

2.2. Variables selection

Existence of many correlated explanatory variables (\( n = 75 \)) may cause multicollinearity which can in turn reduce the generalizability of

| Category | Variable | Source |
|----------|----------|--------|
| Environmental | CO2 emissions from transport | World bank |
| Social | Literacy rate, adult total | World bank |
| Health | Prevalence of severe food insecurity in the population | World bank |
| Public transportation | Air transport, passengers carried | World bank |
| Cultural | Religion diversity index | Pew Research Center (April 2014) |
| | Generosity | World happiness report |

| Category | Variable | Source |
|----------|----------|--------|
| Environmental (11 variables) | Gill per capita | World bank |
| Environmental (11 variables) | Gill per capita growth | World bank |
| Environmental (11 variables) | CO2 emissions from transport | World bank |
| Environmental (11 variables) | CO2 emissions from electricity and heat production | World bank |
| Environmental (11 variables) | CO2 emissions from manufacturing industries and construction | World bank |
| Environmental (11 variables) | CO2 emissions from residential buildings and commercial and public services | World bank |
| Environmental (11 variables) | Methane emissions | World bank |
| Environmental (11 variables) | Nitrous oxide emissions | World bank |
| Environmental (11 variables) | PM2.5 air pollution, mean annual exposure | World bank |
| Social (9 variables) | Freedom to make life choices | World happiness report (Helliwell et al., 2018) |
| Social (9 variables) | Happiness | World happiness report |
| Social (9 variables) | Life Ladder | World happiness report |
| Social (9 variables) | Social support | World happiness report |
| Social (9 variables) | Perceptions of corruption | World happiness report |
| Social (9 variables) | Positive affect | World happiness report |
| Social (9 variables) | Negative affect | World happiness report |
| Social (9 variables) | Confidence in national government | World happiness report |
| Health (7 variables) | Life expectancy at birth, total (years) | World bank |
| Health (7 variables) | Prevalence of severe food insecurity in the population | World bank |
| Health (7 variables) | Mortality from CVD, cancer, diabetes or CRD | World bank |
| Health (7 variables) | Incidence of tuberculosis | World bank |
| Health (7 variables) | Diabetis prevalence | World bank |
| Health (7 variables) | Incidence of HIV | World bank |
| Health (7 variables) | Healthy life expectancy at birth | World happiness report |
| Public transportation (2 variables) | Railways, passengers carried | World bank |
| Cultural (2 variables) | Religion diversity index | Pew Research Center (Pew Research Center 4) |
| Cultural (2 variables) | Generosity | World happiness report |

Table 3 (continued)
determined the optimum number of neurons in the hidden layer using and complex interactions between variables (Kayri, 2016). Then we employed. (Yonaba et al., 2010). Therefore, in this study, single-layer perceptron function in the hidden layer and a linear one in the output layer (Fig. 1) (Mollalo et al., 2019). Fig. 1 shows the topology of a single-layer neuron in the hidden layer is connected to the neurons in the output layer and the ones in the hidden layer. Likewise, each neuron in the hidden layer is connected to the neurons in the output layer (Mollalo et al., 2019). The lower model’s WIC index indicates a more efficient model (Egrioglu et al., 2008). Fig. 2 shows the WIC index model selection process.

Different targets were used as the desired value (system output) as COVID-19 has shown various behaviors and mutated several times to estimate mortality and morbidity rates. For this purpose, nine different targets have been used to study the neural network’s learning process with different desires. The accuracy for each of these targets was evaluated by ANNs. A target with highest accuracy suggests a highest suitability for determining the importance of variables and thus was selected as the optimum target for modeling.

As the indicators are not in the same scale, the resulting models have been compared with each other by the normalized root mean square error interquartile index (RMSEIQR) (Li et al., 2019). Compared to the RMSE, which is a scale-dependent index and partly sensitive to outliers and extreme values, RMSEIQR can be used as a practical index for comparing models over various concentration scales (Li et al., 2019). Moreover, RMSEIQR was used as a common tool to assess and measure the uncertainty of the results (Wechsler and Kroll, 2006).

After variable selection, we assessed the relative importance of the selected variables in modeling COVID-19 prevalence and mortality for each period. The following steps explain the process of determining relative importance of variables in each period (Fig. 3):

Step 1: Different target values from COVID-19 data were generated as described in 2.

Step 2: WIC index was used to determine optimum network architecture for modeling each type of target (model selection). Nine of them were chosen from the n * m models (n: number of explanatory variables; m: number of targets) in total.

Step 3: Models were developed based on optimum networks and their RMSEIQR were computed.

Step 4: Two separate models (prevalence and mortality) with the lowest RMSEIQR values for each period were selected.

Step 5: The variables were ranked based on relative importance using VIA methods. Ten different methods were used to perform VIA through the MLP artificial neural network. These ten VIA methods are described in the next section.

2.4. Variable importance analysis (VIA)

The relative importance of input variables refers to each variable’s contribution to predict the dependent variable (Ibrahim, 2013). Ten VIA methods were used to derive the relative importance of variables from these qualified networks: connection weights algorithm, modified

Fig. 1. A single-layer neural network with a non-linear sigmoid transfer function in the hidden layer and a linear function in the output layer.

Fig. 2. WIC index model selection process.
connection weights, most squares, Garson, partial derivatives, stepwise, perturb, Lek’s profile, modified Lek’s profile, and variance-based approaches. The findings of these approaches can be integrated to draw a general inference. For this purpose, the total of the relative weights obtained from various methods (in percent) was calculated for each variable. This was performed individually for each period, for both infected cases and associated deaths. Below, we briefly explained the VIA techniques used in this study to quantify the relative importance of selected variables used in ANNs.

**Connection weights (CW) algorithm**

The main benefit of the CW algorithm is that the relative contribution of each connection weight is preserved for both magnitude and sign (Olden et al., 2004, Ibrahim, 2013). The relative importance of a given input variable can be defined as Eq. (1).

\[ RI = \sum_{j=1}^{m} w_{xy} \times w_{yz} \]

Where \( RI \) is the relative importance of the input layer, \( x \) is the input neuron, \( y \) is the total number of neurons in the hidden layer, and \( z \) is the output neuron. This method estimates the final network weights obtained through network training. The estimates of final weights differ depending on the initial weights used at the beginning of the training phase (Olden et al., 2004).

**Modified connection weights (MCW) algorithm**

Using the same notation as the CW algorithm, after calculating the sum of product of final weights of connections from input neurons to hidden neurons, a correction term (partial correlation) is multiplied by this sum and the absolute value is taken. This absolute value is called the corrected sum. The corrected sum of each input is then divided by the total corrected sum to determine the relative importance of each input in the MCW algorithm, which is calculated as Eqs. (2) and (3) (Ibrahim, 2013).

\[ RI = \sum_{j=1}^{m} w_{xy} \times w_{yz} \times r_{ij} \]

\[ RI = \frac{\sum_{j=1}^{m} w_{xy} \times w_{yz}}{\sum_{j=1}^{m} \sum_{k=1}^{n} w_{yx} \times w_{yz}} \]

![Fig. 2. WIC index for model selection process.](image-url)
(3)

Where $r_{ij,k}$ is the partial correlation of input $i$ with output $j$ after input $k$, which assesses the association degree between two random variables. Moreover, $r_{ij}$ denotes the simple correlation between input $i$ and output $j$.

Most squares

Using the same notation as the CW algorithm, the most squares approach computes the sum of the squared between initial weight ($w_{ij}$) and final weight ($w_{fj}$) for each input. The sum of squared differences for each input is then divided by the total sum of all inputs. Eq. (4) is used to calculate the relative importance of each input (Ibrahim, 2013).

$$RI = \frac{\sum_{j=1}^{n} (w_{ij} - w_{fj})^2}{\sum_{j=1}^{n} \sum_{i=1}^{m} (w_{ij} - w_{fj})^2}$$ (4)

For the COVID-19 data, the steps for determining the relative importance of variables in each period are shown in Fig. 3. The most precise model for COVID-19 confirmed cases modelling in the given period is Arg min of the partial derivatives (PD) method where the output variable in the PD method would decrease when the input variable increases if the PD is negative (Ibrahim, 2013).

Garson’s algorithm (GA)

GA partitions the neural network relative weights and then uses the absolute values of the final correlation weights. Thus, GA does not include the direction of the relationship between the input and output variables (Eq. (5)) (Garson, 1991).

$$RI = \sum_{j=1}^{n} \frac{|w_{ij}w_{fj}|}{\sum_{j=1}^{n} |w_{ij}w_{fj}|}$$ (5)

Partial Derivatives (PD) method

The output variable in the PD method would decrease when the input variable increases if the PD is negative (Ibrahim, 2013).

$$d_{ij} = S_j \sum_{y=1}^{n} w_{iy}I_y (1 - I_y) w_{yh}$$ (6)

$$SSD_i = \sum_{j=1}^{n} (d_{ij})^2$$ (7)
Table 4
Selected models in step 2 and 4.

| Period | Target type | Optimum number of neurons | RMSEIQR | Selected to perform VIA? |
|--------|-------------|---------------------------|---------|-------------------------|
| 1      | PR-IQR      | 17                        | 0.011   | Yes                     |
|        | TMR         | 25                        | 0.051   | No                      |
|        | GR1         | 28                        | 0.02    | No                      |
|        | MR          | 18                        | 0.085   | Yes                     |
|        | MR-IQR      | 16                        | 0.218   | No                      |
|        | TMMR        | 27                        | 0.512   | No                      |
|        | GR2         | 22                        | 0.245   | No                      |
|        | FR          | 17                        | 0.451   | No                      |
|        | PR          | 24                        | 0.005   | No                      |
| 2      | PR-IQR      | 5                         | 0.003   | Yes                     |
|        | TMR         | 7                         | 0.022   | No                      |
|        | GR1         | 3                         | 0.419   | No                      |
|        | MR          | 13                        | 0.012   | Yes                     |
|        | MR-IQR      | 9                         | 0.021   | No                      |
|        | TMMR        | 21                        | 0.423   | No                      |
|        | GR2         | 16                        | 0.471   | No                      |
|        | FR          | 28                        | 0.474   | No                      |
|        | PR          | 2                         | 0.03    | No                      |
| 3      | PR-IQR      | 5                         | 0.02    | Yes                     |
|        | TMR         | 10                        | 0.165   | No                      |
|        | GR1         | 7                         | 0.421   | No                      |
|        | MR          | 6                         | 0.08    | Yes                     |
|        | MR-IQR      | 8                         | 0.115   | No                      |
|        | TMMR        | 23                        | 0.776   | No                      |
|        | GR2         | 7                         | 0.841   | No                      |
|        | FR          | 26                        | 0.887   | No                      |
|        | PR          | 2                         | 0.057   | No                      |
| 4      | PR-IQR      | 4                         | 0.032   | Yes                     |
|        | TMR         | 8                         | 0.089   | No                      |
|        | GR1         | 5                         | 0.196   | No                      |
|        | MR          | 21                        | 0.015   | Yes                     |
|        | MR-IQR      | 17                        | 0.04    | No                      |
|        | TMMR        | 24                        | 0.426   | No                      |
|        | GR2         | 18                        | 0.359   | No                      |
|        | FR          | 18                        | 0.901   | No                      |

In Eq. (6), $y_i$ is the output with respect to input $x_i$, $N$ denotes the total number of observations in a network with $n_i$ inputs, one hidden layer with $h_0$ neurons, and one output neuron. $S_i$ is the derivative of the output neuron with respect to the corresponding input. $I_{ij}$ is the hidden neuron’s output, and $w_{hi}$ and $w_{ih}$ are the correlation weights between the output neuron and the $h_0$ hidden neuron, and between the $i$th input neuron and the $h_0$ hidden neuron, respectively. In Eq. (7), SS, is the sum of the square partial derivatives.

**Stepwise method**

The stepwise method involves adding or removing one input variable step by step while considering the effect on the output result. Depending on various arguments, the input variables are ranked according to their significance based on the changes in mean squared error (MSE). The largest increases or decreases in MSE due to input deletions are used to classify inputs in order of importance (Sung, 1998).

**Perturb method**

Perturb method aims to measure how minor changes in each input will affect the neural network output. The algorithm modifies one variable’s input values while leaving the others unchanged. The output variable’s responses to each change in the input variable are registered. The input variable with the greatest relative effect on the output is the one with the largest changes. The input variables are classified according to the impact of the small changes (Gevrey et al., 2003).

**Lek’s profile method**

In Lek’s profile method, each input variable is studied while the others are blocked at fixed values. The basic idea behind this method is to create a fictitious matrix that encompasses the entire range of input variables. Each variable is divided into a set of equal intervals between its minimum and maximum values. Except for one, all variables are set to their minimum, first quartile, median, third quartile, and maximum values at the beginning. The median value is subtracted from these five numbers. The output variable’s profile is plotted for the considered values (Gevrey et al., 2003).

**Modified Lek’s profile method**

Despite in Lek’s profile method where the input variables were kept constant at five points, Modified Lek’s profile method selects an input variable and partitions it into 12 parts. Further, a qualified ANN is evaluated for each point of the partitioned variable’s range and is implemented for each fixed values. The average of the outputs for each scale point is determined. This process is repeated until all ANN input variables could be assessed. The resulting curve profile for each input variable is then plotted (do Nascimento et al., 2019).

**Variance based method**

Variance based method computes and updates the variance for given variables. It has the advantage of not requiring the values to be stored for computing the variance at the end. To measure the variance in this method, the sum of squares is updated by previous values according to Eq. (9), and then the variance values are calculated using Eq. (10) (Welford, 1962).

\[
\tau_n = \frac{1}{1} \sum_{i=1}^{n} x_i \quad (8)
\]

\[
SS_n = SS_{n-1} + \left( \frac{n - 1}{n} \right) (x_n - \tau_{n-1})^2 \quad (9)
\]

\[
Var(x_n) = \frac{SS_{n-1} + (x_n - \tau_{n-1})^2}{n} \quad (10)
\]

Where $\tau$ represents the mean of values, $SS_n$ is the corrected sum of squares, and $n$ is the total number of updates.

3. Results

Based on the lowest obtained values for RMSEIQR, we selected prevalence rate in interquartile range (PR-IQR) as the target for modeling the prevalence rates of COVID-19 in each studied period (Table 4). The spatio-temporal variations of prevalence rates in IQRs for each period has been depicted in Fig. 4. According to Fig. 4, the countries in North and South America had a persistent higher prevalence rates in IQR than the rest of the world in all periods. In the period 2, the countries in continental Europe and America showed a relatively increasing trend in COVID-19 prevalence compared to the period 1, as the prevalence rates in IQR values have increased in these areas. The period 3 was the peak of the disease prevalence compared to other periods. During this period, Europe and most countries in north Asia were significantly infected by COVID-19.

In period 4, the prevalence rates slightly decreased compared to period 3. This reduction in changes is more visible in America, maybe due to earlier initiation of vaccination programs. However, the countries of Central and South Africa have had no remarkable differences in prevalence rates (in all periods), except for the southernmost ones, including South Africa and Namibia, which have had the highest prevalence rates (in all periods), except for the southernmost ones.
The changes in MR trends is more visible in America and Europe continents. In period 1, the distribution of MR was almost uniform across the world. Moreover, in the first period, most countries experienced lower MR rates compared to the following periods. In the period 2, South American countries including Brazil, Argentina, Bolivia, Peru, and Colombia experienced higher MRs than other countries. The period 3 shows a relatively significant increase in COVID-19 MRs in continental Europe and North America. Although the highest prevalence rates in IQR were found in period 3 (Fig. 4), period 4 was found to be the peak of mortality rates, especially in the United States, Brazil, South Africa, and some European countries (Fig. 5).

Based on the WIC index, the optimum network architecture for modeling each type of target was identified. Nine models were chosen from a total of n * m (n: number of explanatory variables; m: number of targets) models (step 2). Further, two models with the lowest RMSEIQR were selected for each period, one model for prevalence and the other for mortality (step 4). Table 4 lists the models that were selected in step 2 and 4.

The ANN topologies that were selected to perform VIA are represented as bold rows in Table 4. Fig. 6 to Fig. 9 depicts the twenty most influential explanatory variables on COVID-19 prevalence and mortality for all selected periods, respectively. As can be seen, some of the explanatory variables were among the twenty most important variables across all periods (non-black horizontal bars). Most economic-related variables such as unemployment, gross national income (GNI) per capita, and GNI per capita growth have always been among the most influential explanatory variables on COVID-19 prevalence. In addition, other variables related to public transportation, including rail and air transportation, as well as surface temperature, population density, and urban population were among the most significant variables for cases at all periods. For mortality, diabetes prevalence, the number of hospital beds (per 1000 people), number of nurses and midwives (per 1000 people), negative affect (negative emotions and experiences during life), and air transportation were the most influential explanatory variables for all periods.

In addition, Table 5 lists the two most influential variables for each period based on the median of weights. Fig. 10 depicts the worldwide spatial distribution of PR-IQRs in all periods, along with the most influential variables on the disease prevalence. In addition, Fig. 11 shows the spatial distribution of MRs for all countries, along with the most influential variables in each period.

4. Discussion

The outbreak of COVID-19 has adversely affected many countries around the world. Numerous mutations caused by the SARS-CoV-2 virus
have intensified its spread, making the control of the epidemic even more challenging. Identifying the effective variables and their relationship with disease prevalence and mortality over time can be useful for controlling disease outbreak. ANNs are among the most widely used approaches to model this relationship, particularly as the associated data and computations become more readily available (Augusta et al., 2019).

Since the epidemic of COVID-19, as a contagious disease, is directly related to the geographical concept of an area, GIS can play an essential role in its planning, management, and modeling (Mollalo et al., 2020). GIS has been used in many studies to manage and plan epidemiological issues from spatial perspectives (Meliker and Sloan, 2011, Shrestha et al., 2020). It has also been consistently used to analyze health-related data and can be a valuable tool for analyzing the spread of disease in each region (Meliker and Sloan, 2011). Increasing the power of computers, improving spatial analysis methods, and developing artificial intelligence models have led to the development of advanced and modern GIS applications in disease modeling and prediction (Ghayvat et al., 2021). Therefore, in this study, we utilized GIS technology to develop a spatio-temporal model for COVID-19 prevalence and mortality.

Given that little space-time COVID-19 modeling has been conducted at the global scale, we compiled a geodatabase of potential influential variables on the prevalence and mortality of the disease and ranked relative importance of variables based on VIA methods for four periods of time. Our findings showed that various VIA algorithms yielded varying results. Although the relative importance of variables on prevalence and mortality changed over time, some variables were identified among the top 20 most relevant variables in all periods.

Dealing with complicated interactions among variables, we applied ten different VIA methods to evaluate the influence of potential explanatory variables by optimizing the data storage, advancing the model interpretability, and providing a smaller number of influential variables without losing accuracy. VIA techniques can be implemented to solve the intricacy of interactions among variables on big datasets (Ferretti et al., 2016). For instance, these techniques were used to figure out how well each variable influences the COVID-19 prevalence. Dfuf et al. (2020) implemented a parametric and a nonparametric VIA method and calculated the impact of the 35 companies on the political, economic, and social instability captured by two highly regarded Spanish economic newspapers during the COVID-19 outbreak. The result showed that the nonparametric VIA method outperformed its competitors since it incorporates all the information using the entire distribution errors.

Economic variables have retained their significant impact on higher rates of COVID-19 prevalence over time. Consistent with our findings,
unemployment was found strongly correlated with the increased risk of disease prevalence (Jin et al., 1995). Since unemployment and poverty reduce people’s ability to access health facilities, unemployed people who are infected communicate with others in the society without being treated, which may increase the severity of the disease transmission. Another hypothesis that can explain this association is unemployed individuals and uneducated people are less likely to get vaccinated due to underestimating the positive impacts or overestimating the risks of getting vaccinated, which can cause a higher prevalence of the COVID-19 in a society (Malik et al., 2020, Mollalo and Tatar, 2021). Some other studies, such as (Jin et al., 1995), have shown that unemployment and inadequate social welfare can increase the disease spread.

Demographic variables were other influential variables affecting the COVID-19 spread. Due to the contagious nature of COVID-19, the higher population density and overcrowding in an area are associated with the greater likelihood of disease occurring (Sigler et al., 2021, Sirkeci and

Fig. 6. The 20 most influential explanatory variables on COVID-19 a) prevalence b) mortality in the period 1.
Yucesahin, 2020). On the contrary, countries with a lower population density showed lower prevalence rates of COVID-19 in all periods, such as Australia and Russia. Consistent with our findings, a recent study (Mansour et al., 2021) shows that the higher population density rates in Oman could result in a higher prevalence of COVID-19. A research by Ahmadi et al. (2020) suggests that population density and intra-provincial movement are directly associated with the spread of the coronavirus in Iran. Other studies confirm that higher population density increases the chance of transmission of the virus (Coşkun et al., 2021, Rocklov and Sjödin, 2020) and can alter the prevalence and mortality rates (Bhadra et al., 2021).

The use of public transit was persistently found significant on COVID-19 prevalence in all periods. A possible explanation might be that many people in public transportation stand together for a long time in a closed environment especially transportation by plane and train. As a result, the contagious virus can rapidly be transmitted from infected people.
individuals to other passengers, causing the disease to spread more severely. Zheng et al. (2020) showed that the infected individuals during the incubation period brought the disease from Wuhan, China to other cities and nations by using public transportation such as flights, trains and buses. In New York, Cordes and Castro (2020) suggested that people who rely on public means of transportation might be at higher risks of COVID-19 due to contact with other infected passengers, consistent with our findings.

Regarding COVID-19 mortality, diabetes prevalence was found to be a significant variable in all periods. Inadequate and poor immunological responses to viral infections may be among the leading cause of mortality in COVID-19 patients with diabetes (Critchley et al., 2018).
increased blood sugar level in a person with diabetes can severely damage the beneficial intracellular bacteria, which in turn increases the viral binding affinity and reduces the virus removal (Muniyappa and Gubbi, 2020, Gazzaz, 2021). Exploring the spatial variations of COVID-19 in the Caribbean, Moonsammy et al. (2021) found that the higher prevalence of diabetes in the Caribbean could increase COVID-19 deaths. A meta-analysis on more than 16,000 patients also found that diabetes in patients with COVID-19 doubled the risk of death (Kumar Fig. 9. The 20 most influential explanatory variables on COVID-19 a) prevalence b) mortality in the period 4.
Consistent with our results, other researchers have shown a strong relationship between diabetes prevalence and COVID-19 mortality (Huang et al., 2020, Guo et al., 2020).

There were several caveats and limitations in this study that should be acknowledged. First, due to the worldwide distribution of this study, it is most likely that some countries have not provided accurate statistics about COVID-19 prevalence and deaths, which may bias the results. Another limitation of this study was associated with different lockdown policies and stay-at-home restrictions for each country. Some countries quickly began quarantine policies after the pandemic was announced than others that did not make any specific lockdown policy. Although we tried to find the most influential factors related to COVID-19 prevalence and mortality for all countries at the same time, a study on a higher spatial resolution (sub-country level) can provide more reliable results. Despite above-mentioned limitations, the findings may help policymakers to track the spread of disease over time based on the most significant variables identified by the employed models.

5. Conclusions

In summary, we examined ten different VIA methods to estimate the relative importance of potential explanatory variables on COVID-19 prevalence and mortality at a global scale. Due to the numerous mutations of the virus, various targets were considered for modeling to

| Period   | Variable              | Median of Weights | Variable                | Median of Weights |
|----------|-----------------------|-------------------|-------------------------|-------------------|
| Period 1 | Population density    | 1.778             | Diabetes prevalence     | 1.755             |
|          | GNI per capita        | 1.775             | Hospital beds           | 1.675             |
|          | Unemployment          | 2.11              | Diabetes prevalence     | 1.995             |
|          | Population density    | 1.973             | Nurses and midwives     | 1.778             |
| Period 2 | Population density    | 1.775             | Hospital beds           | 1.684             |
|          | Air transport, passengers carried | 1.645 | Negative affect        | 1.648             |
|          | GNI per capita        | 1.721             | Diabetes prevalence     | 1.764             |
|          | Unemployment          | 1.673             | Hospital beds           | 1.688             |

Fig. 10. Spatio-temporal distribution of the most influential variables on PR-IQRs for each period.
enhance the accuracy of the results. Our findings indicated that the extracted relative importance from different models by VIA methods varies over time. However, several variables were persistently among the most influential variables on the prevalence and mortality of the disease in all periods. Unemployment, population density, air and rail transportation, urban population, GNI per capita, GNI per capita growth, and surface air temperature were among the most significant variables on disease prevalence in all periods. Regarding COVID-19 mortality, diabetes, air transportation, number of hospital beds, number of nurses, and negative affect were among the most influential variables. Better spatial resolution can improve the validity of the results in future studies. Policymakers and epidemiologists can use spatio-temporal analysis to monitor and evaluate COVID-19 prevalence and mortality concerning significant variables.

CRediT authorship contribution statement

**Nima Kianfar:** Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. **Mohammad Saadi Mesgari:** Validation, Supervision, Investigation, Writing – review & editing. **Abolfazl Mollalo:** Methodology, Validation, Writing – review & editing. **Mehrdad Kaveh:** Investigation, Writing – review & editing.

Declarations of Competing Interest

None.

Acknowledgments

We would like to thank anonymous reviewers for taking the time and effort to review the manuscript.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

World Health Organization (WHO), Archived: WHO Timeline- COVID-19. 2020ab;
Available from: www.who.int/news/item/27-04-2020-who-timeline-covid-19.

World Health Organization (WHO), WHO Coronavirus (COVID-19) Dashboard. 2021b;
Available from: https://covid19.who.int/.

Wang, C., et al., 2020. A novel coronavirus outbreak of global health concern. Lancet 395 (10223), 470–473.

Mansour, S., et al., 2021. Sociodemographic determinants of COVID-19 incidence rates in Oman: geospatial modelling using multiscale geographically weighted regression (MGWR). Sustain. Cities Soc. 65, 102627.

Mollalo, A., Rivera, K.M., Vahedi, B., 2020. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. Int. J. Environ. Res. Public Health 17 (12), 4204.

Mollalo, A., et al., 2019. A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States. Int. J. Environ. Res. Public Health 16 (1), 157.
Duh, M.-S., Walker, A.M., Ayanian, J.Z., 1998. Epidemiologic interpretation of artificial neural networks. Am. J. Epidemiol. 147 (12), 1112–1122.

Olden, J.D., Jackson, D.A., 2002. Illuminating the ‘black box’: a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Modell. 154 (1-2), 135–150.

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Modell. 178 (3-4), 389–397.

Ibrahim, O., 2013. A comparison of methods for assessing the relative importance of input variables in artificial neural networks. J. Appl. Sci. Res. 9 (11), 5692-5700.

Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecol. Modell. 116 (1), 15–31.

Ferretti, F., Saltelli, A., Tarantola, S., 2016. Trends in sensitivity analysis practice in the last decade. Sci. Total Environ. 568, 666–670.

Wei, P., Lu, Z., Song, J., 2015. Variable importance analysis: a comprehensive review. Reliab. Eng. Syst. Saf. 142, 399–432.

Dfuf, I.A., et al., 2020. Variable importance analysis in imbalanced datasets: A new approach. IEEE Access 8, 127404–127430.

Casiraghi, E., et al., 2020. Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments. IEEE Access 8, 196299-196325.

Patha, D.F., et al., 2021. An analysis to identify the important variables for the spread of COVID-19 using numerical techniques and data science. Case Stud. Chem. Environ. Eng. 3, 100067.

Shaffiee Haghhebous, S., et al., 2020. Prioritizing and analyzing the role of climate and urban parameters in the confirmed cases of COVID-19 based on artificial intelligence applications. Int. J. Environ. Res. Public Health 17 (10), 3730.

Mollalo, A., et al., 2015. Geographical information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. Zoonoses Public Health 62 (1), 18–28.

Mollalo, A., et al., 2018. Machine learning approaches in GIS-based ecological modeling of the sand fly Phlebotomus papatasi, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. Acta Trop. 188, 187–194.

Bashir, M.F., et al., 2020. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci. Total Environ. 728, 138835.

Zhang, C.H., Schwartz, G.G., 2020. Spatial disparities in coronavirus incidence and mortality in the United States: an ecological analysis as of May 2020. J. Rural Health 36 (3), 433–445.

Jia, J.S., et al., 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. Nature 582 (7812), 389–394.

Ahmadi, M., et al., 2020. Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci. Total Environ. 729, 138705.

Ramirez, L.J., Lee, J., 2020. COVID-19 emergence and social and health determinants in Colorado: a rapid spatial analysis. Int. J. Environ. Res. Public Health 17 (11), 3856.

Moonnammy, S., et al., 2021. COVID-19 modelling in the Caribbean: spatial and statistical assessments. Spatial Stat. Temporal. Epidemiol. 37, 100416.

Ambade, B., et al., 2019. COVID-19 lockdowns reduce the Black carbon and polycyclic aromatic hydrocarbons of the Asian atmosphere: source apportionment and health hazard evaluation. Environ. Develop. Sustain. 1–20.

Gautam, S., 2020. The influence of COVID-19 on air quality in India: a boon or inutile. Bull. Environ. Contam. Toxicol. 104 (6), 724–726.

Gautam, S., 2020. COVID-19: an air pollution remains low as people stay at home. Air Quality Atmos. Health 13, 853–857.

Wang, P., et al., 2020. Severe air pollution events not avoided by reduced anthropogenic activity. Sci. Total Environ. 751, 141663.

Shrestha, S., et al., 2020. Spatial epidemiology: an empirical framework for syndemics research. Soc. Sci. Med. 113S32.

Chaves, S.E., 2020. Recognizing suspects and predicting the spread of contagion based on mobile phone location data (counteract): a system of identifying covid-19 infectious and hazardous sites, detecting disease outbreaks based on the internet of things, edge computing, and artificial intelligence. Sustain. Cities Soc. 69, 102706.

Gazi, K., 2020. COVID-19 mortality and severity of disease in COVID-19 pneumonia—a systematic review, meta-analysis, and meta-regression. Diabetes Metab. Syndr. Clin. Res. Rev. 14 (4), 395–403.

Guo, W., et al., 2020. Diabetes is a risk factor for the progression and prognosis of COVID-19. Diabetes Metab. Res. Rev. 36 (7), e3315.

Giovanni, Greenbelt. NASA/GSFC, MD,USA. NASA goddard earth sciences data and information services center (GES DISCO) (2021). Available from: https://giovanni.gsfc.nasa.gov/.