Advances in Industrial Engineering, December 2024, 58(2): 291-306
DOI: 10.22059/aie.2024.380927.1906

RESEARCH PAPER

Analyzing and Forecasting of Coronavirus Time-Series Data: Performance Comparison of Machine Learning and Statistical Models

Majid Alimohammadi Ardakani[1], Mohammad Hossein Karimi-Zarchi[2], Davood Shishebori[3*]

[1]Assistant Professor, Department of industrial engineering, faculty of engineering, Ardakan University, Ardakan, Iran.
[2]Ph.D. Candidate, Industrial Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran.
[3]Associate Professor, Industrial Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran.

Received: 15 August 2024, Revised: 26 September 2024, Accepted: 27 September 2024

Abstract

Coronavirus is a respiratory disease caused by coronavirus 2 acute respiratory syndrome. Forecasting the number of new cases and deaths can be an efficient step towards predicting costs and providing timely and sufficient facilities needed in the future. The goal of the current study is to accurately formulate and predict new cases and mortality in the future. Nine prediction models are tested on the Coronavirus data of Yazd province as a case study. Due to the evaluation criteria of root mean square error (RMSE), mean square error (MSE), mean absolute percentage error (MAPE), and mean absolute value of error (MAE), the models are compared. The analysis results emphasize that, according to the mentioned evaluation criteria, the KNN regression model and the BATS model are the best models for predicting the cumulative cases of hospitalization of Coronavirus and the cumulative cases of death, respectively. Moreover, for both hospitalization and death cases, the autoregressive neural network model has the worst performance among other formulations.

Keywords:
Pandemic, Coronavirus, Time Series, Forecasting, Statistical Modeling, Machine Learning Models.

Introduction

Covid-19, or Coronavirus disease 2019, also called as Acute Respiratory Disease 2019-nCoV or commonly Corona, respiratory disease caused by acute respiratory syndrome of coronavirus-2 or SARS-CoV-2. For the first time, the virus was recognized in December 2019, and laboratory research showed that it is a new strain of Coronaviruses. A group of patients infected with this virus showed a new form of viral infection pneumonia. All patients also had a similar history of visiting a beet market in Wuhan, China. The usual symptoms are cough, fever, loss of smell, and shortness of breath. Phlegm production, muscle pain, nausea, sore throat, and red eyes are among its less common symptoms [1]. On March 11, 2020, the World Health Organization (WHO) pronounced the outbreak of the modern Coronavirus (NCOV-2019) as a

[*] Corresponding author: (Davood Shishebori)
Email: Shishebori@yazd.ac.ir

worldwide viral disease pandemic. Since the outbreak of this disease was declared as an epidemic, many countries in the world have been severely affected by the disease of the Coronavirus, and many preventive measures have been considered, including quarantine, wearing masks, rapid tests for the diagnosis of the Corona, keeping distance. Social, and self-quarantine by countries to avoid the spread of the epidemic disease of Coronavirus is being implemented. Despite these measures, Covid-19 is spreading rapidly due to various reasons such as population density, lifestyle, world travel, and the emergence of new strains of this virus. Accordingly, it affects on human wellbeing and the worldwide economy.

Preparing and controlling the spread of infectious respiratory diseases such as Covid-19 requires careful planning and policies. Modeling, estimating, and predicting the spread of viruses and epidemiological characteristics are essential in on time providing the equipment needed to control and reduce their consequences. To have efficient planning, it is very vital to predict the critical indicators in the future. Therefore, formulating and forecasting the number of authenticated instances of Coronavirus, recoveries, as well as deaths, play a critical role in efficient planning to control the spread of Coronavirus in the world.

Various statistical formulations have been presented for predicting some future cases and predicting the behavior of infectious diseases within the close future, as well as predicting some Coronavirus cases in the future [2]. To predict typhoid fever, Zhang et al. [3] applied the seasonal integrated moving average (SARIMA), and three various formulations derived from neural networks, including: radial function neural networks (RBFNN), back propagation neural networks (BPNN), and element recurrent neural networks (ERNN). Moreover, Chen et al. [4] predicted the prevalence of influenza in rural and urban regions of Shenyang by applying the SARIMA. Their research results have been utilized as an authentic report for future influenza abatement and control strategies. Other studies have similarly used autoregressive integrated moving average (ARIMA) models to predict pestiferous diseases such as tuberculosis, dengue fever, and malaria [2-4].

Ceylon [5] applied ARIMA models to forecast the epidemiological behavior of the spread of Coronavirus in Spain, Italy, and Europe. In their research, several ARIMA models with several parameters were selected for countries due to the mean absolute error (MAPE) evaluation criteria. Lotfi and Burko [6] used the ARIMA at the European Center for Control and Prevention (ECDC) to determine the number of authenticated cases and mortality of Covid-19. Tanden et al. [7] applied the ARIMA to predict the cases of Coronavirus in India. Their study represented an upward behavior of Covid-19 instances in the near future days, and time series investigation also showed a non-linear growth in several cases. Based on Italian epidemiological information at the regional and national level, Perron [8] presented an ARIMA to predict the epidemic behavior during the weeks after April 4, 2020 (more than 6 weeks). Its outcomes emphasized that the number of Coronavirus cases in some regions in Italy would level off on the 8th week.

Several studies have provided short-time forecasting of the Coronavirus epidemic by applying machine learning methods in addition SARIMA models. Ghasal et al. [9] applied multiple linear regression and linear regression techniques to forecast the number of casualties in India for 40 days. They report that the death toll in India will double if preventive measures against Covid-19 are not implemented or changed. Parbat and Chakraborty [10] used support vector regression to predict the number of Coronavirus instances in India for 60 days due to time series data reported from 1th March 2020 to 30th April 2020. Their results emphasize that the support vector regression has a precision of about 97% in predicting instances of loss, cumulative confirmed cases, and recovered cases. Antonio Nimoli [11] has used heterogeneous autoregressive (HAR) to predict the cases of Coronavirus in Italy due to reported time series data. Their results show that the HAR model is more precise than the ARIMA models in predicting the instances of Coronavirus. Kibria et al. [12] have applied an ARIMA to predict the expected daily number of Coronavirus instances in Bangladesh due to data between April

20th, 2021, and July 4th, 2021. ARIMA represent the best outcomes among several implemented techniques compared to the autoregressive (AR), moving average (MA), and ARMA models. Khan and Gupta [13] have used a univariate time series technique to estimate the number of Coronavirus infected instances which can be expected in the coming weeks in India. They implemented an ARIMA on data obtained between January 31th, 2020, and March 25 th, 2020, and evaluated it by applying data obtained between March 26th, 2020, and April 4 th, 2020. They utilized this method to predict daily instances of coronavirus for the next 50 days without any extra intermediation. The outcomes represented a growing trend in the forecasted and actual number of Covid-19 instances by about 10000 instances per week, due to data available as of April 4th, 2020. Chowdhury et al. [14] applied the third-wave Coronavirus data set, that includes almost three-month recordings of authenticated instances, by applying short-term memory (LSTM) networks and an adaptive neural fuzzy inference system to forecast new instances of Coronavirus in Bangladesh. The results of both methods were compared, and it was found that LSTM showed more satisfactory results. Al-Assafi et al. [15] proposed a public dataset provided by the ECDC to develop a prediction system for the propagation of the Coronavirus outbreak in Malaysia, Morocco, and Saudi Arabia. For this purpose, some deep learning (DL) was applied to forecast the behavior of the propagation of Coronavirus in these three countries. In a study, Chayoun et al. [16] proposed the ARIMA to analyze the global spread of Coronavirus between January 22th, 2020, and April 7th, 2020. According to several statistical techniques such as machine learning, time series, and ensemble learning, Nair et al. [17] analyze the predict of cumulative authenticated instances of Coronavirus deaths in India. ARIMA model, exponential smoothing, and Holt-Winters in time series, random forest regression in ensemble learning (EL), linear regression (LR) and support vector regression in machine learning (ML) are executed for forecasting. The accuracy of the trained methods is figured out due to criteria such as RMSE, MSE, MAE, and MAPE.

In order to forecast the number of dead and infected patients in the future, Mokhaires and Alaf [18] used short-term predicting models to predict the number of deaths and infected in the near future. Predicting was done by applying the ARIMA model, combined ARIMA, exponential smoothing, and Holt-Winters with a time value equal to 57. By figuring out the tested models due to MAPE, they found that the exponential smoothing is an efficient prediction algorithm for predicting improved cases with an average absolute error of 2.66% and for predicting confirmed cases with an average absolute error of 1.77%. Holt-Winters, the best algorithm for predicting mortality cases with MAPE equal to 5.33%. Prajapati et al. [19] of predicting algorithms for the total cumulative instances of each country by comparing the predicted value and the reported data and then sorting the models (Prophet, LSTM, Holt-Winters, ARIMA-NARNN, ARIMA) due to MAE, RMSE, and MAPE values. They are done. The combination of Non-linear Autoregressive Neural Network (NARNN) and ARIMA models provided the most efficient result among the selected algorithms with simplified RMSE, which was approximately 33.5% better than one of the most common time series forecasting methods (ARIMA). In another study conducted in 2022, Li et al. [20] chose Britain, South Africa, Brazil, and Nigeria as research samples. Their analyzed data set spanned from March 1, 2020, to January 31, 2021. Their study applied the ARIMA to forecast the propagation of the Coronavirus in the mentioned countries. Chung et al. [21] developed a deep learning algorithm to recognize Coronavirus before appearing symptoms by applying the heart rate (HR) data in a smartwatch. They obtained that their developed deep learning algorithm can simply and literally recognize Coronavirus patients by applying HRs found from a smartwatch before appearing symptoms.

In the current research, it is planned to collect the authenticated cumulative instances and deaths of Coronavirus in Yazd province (in Iran) as a case study from 02/26/2020 to 12/19/2021 (Data obtained from the Ministry of Health and Treatment of the country). Several models are examined, and the best future prediction model is selected. Machine learning and statistical models selected for evaluation include exponential smoothing (ETS), autoregressive integrated moving average (ARIMA), Holt-Winter, KNN regression, Theta, and autoregressive neural network (ARNN) models, cubic spline smoothing (CSS), STL method, and BATS method. Also, the selection evaluation criteria for choosing the most suitable model include mean absolute error (MAE), mean absolute error percentage (MAPE), root mean square error (RMSE), and mean square error (MSE).

The rest of the paper is organized as follows: Section 2 describes selected statistical and machine learning models and their analytical tools and evaluation criteria. In section 3, the model parameter selection and optimization method are brought up, considering time series analysis, cumulative authenticated instances, and cumulative mortality in Yazd province as a case study. Section 4 shows the analysis results of the present work. Finally, Section 5 is related to the conclusion and future suggestions.

Statistical models

To create a 14-day prediction of cumulative instances of Coronavirus for Yazd province, the current study ARIMA, ETS, Holt-Winters, KNN regression, Theta, autoregressive neural network (ARNN), cubic spline smoothing (CSS), the STL, and the BATS statistical models. To have a statistically significant prediction from time series data, 30 observations can be sufficient as minimum required sample size [22].

Time series data can be represented as a sequence of numerical values for which a time label is specified for each numerical value [23]. So, it is a collection of observations arranged by time (or any other quantity). Usually, it is displayed as $X_{t_1}, X_{t_2}, \dots, X_{t_n}$. In classical discussions, time series data are categorized into two classes: non-stationary and stationary data. Non-stationary time series data have patterns known as seasonality. In contrast, stationary time series data have no pattern over time. Accordingly, the mean and variance of non-stationary data are variable over time.

Autoregressive integrated moving average (ARIMA) process

For the first time, in 1976, Box and Jenkins introduced the ARIMA(p,d,q) model [24]. This model is applied to predict non-seasonal stationary data. ARIMA series or autoregressive integrated moving average is ARMA series with the trends. So instead of somehow removing the trend and using the same conventional methods for the ARMA series, we use the ARIMA model simultaneously. Accordingly, a time series that becomes an ARMA(p,q) time series after d differentiation is known as an ARIMA(p,d,q). The general pattern of ARIMA is shown in equation (1):

$$\varphi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B)a_t \tag{1}$$

where AR operator is stationary $\boldsymbol{\varphi_p(B)} = (\mathbf{1} - \boldsymbol{\varphi_1 B} - \cdots - \boldsymbol{\varphi_p B^p})$ and MA operator is invertible $\boldsymbol{\theta_q(B)} = (\mathbf{1} - \boldsymbol{\theta_1 B} - \cdots - \boldsymbol{\theta_q B^q})$. The parameter θ0 can be very different for d>0 and d=0.

Seasonal Integrated Moving Average Autoregressive Process (SARIMA)

The well-known collective seasonal ARIMA model of Box and Jenkins is defined as relations (2) to (7):

$$SARIMA: \phi_P(B^s)\varphi_p(B)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\vartheta_Q(B^s)a_t \tag{2}$$

$$ARIMA: \varphi_p(B)(1-B)^d Z_t = \theta_q(B)a_t \tag{3}$$

$$\phi_P(B^s) = 1 - \phi_1 B^s - \phi_2 B^{2s} - \dots - \phi_P B^{Ps} \tag{4}$$

$$\vartheta_Q(B^s) = 1 - \vartheta_1 B^s - \vartheta_2 B^{2s} - \ldots - \vartheta_Q B^{Qs} \quad (5)$$

$$\varphi_p(B) = 1 - \varphi_1 B^1 - \varphi_2 B^2 - \ldots - \varphi_p B^p \quad (6)$$

$$\theta_q(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \ldots - \theta_q B^q \quad (7)$$

These two terms are polynomials of B_s that do not have a common root. The roots of these polynomials are located outside the unit circle and at it is a purely random process with zero means. For convenience, $\varphi_p(B)$ and $\theta_q(B)$ are autoregressive factors and moving average, respectively. The $\phi_P(B^s)$ and $\vartheta_Q(B^s)$ are called autoregressive factors and seasonal moving average, and often as ARIMA(p,d,q)*(P,D,Q)$_s$* indicate where the index s indicates the seasonal time period.

Exponential smoothing method

This method is one of the most practical and popular predicting methods. Its popularity is due to its flexibility, ease of automation, simple calculations, and favorable performance. The exponential smoothing method, instead of calculating a simple average takes the weighted average of the past values so that the weights, exponentially, reduce the tendency to the past data; It means that it gives more weight to recent data than older data. The exponential smoothing method forecasts series that lack seasonal trends and patterns. Therefore, we must remove the trend and seasonality from the series and then apply this method to the residual series. The forecast that the smoothing method gives at time t+1 is in the form of equation (8):

$$P_{(t+1)} = \alpha z_t + \alpha(1-\alpha) z_{(t-1)} + \alpha(1-\alpha)^2 z_{(t-2)} + \ldots \quad (8)$$

In the above relation, α is called the smoothing constant, and a is the constant value between zero and one. The above relationship shows exponential smoothing using a weighted average of all older observations with exponential descending weights. The smoothing constant α is determined by the user. The default values considered for α and indicating its optimal performance are in the range of 0.1 to 0.2. Also, for choosing α, the trial-and-error method can be a suitable solution. Anyway, a lot of care and obsession should be used in choosing α; Otherwise, it will cause overfitting of the method and decrease the prediction accuracy of the method during the validation period and in the future.

Holt-Winters model

The Holt-Winters prediction model is an extension of the exponential smoothing model The Holt-Winters model consists of three sections. The first section is called the mean (fixed value), which shows the general behavior of the model, and the values fluctuate around it. The second section is the behavior (slope of the line) which is constant in terms of time but is considered as a multiple for the variable. The third section, which changes periodically, shows seasonal changes. In such a model, the future value is predicted based on combining these three components. Such a model has several parameters. This group of parameters in this model is known as α, β, γ. The simple form (without trends and seasonal changes) of the Holt-Winters time series model is as equation (9):

$$k_t = \alpha \frac{z_t}{U_{t-1}} + (1-\alpha)(k_{t-1} + k_{b-1}) \quad (9)$$

where z_t is the observation at time *t* and k_t is the smoothed observation at time t. As can be seen, only α parameter is present and each observation is seen as a linear combination of the previous point (k_{t-1}). On the other hand, *U* is called the index of seasonal changes and *Q* is the length of the period of seasonal changes. If the model has a trend, the shape of the model will be in relation (10).

$$b_t = \gamma(k_t - k_{t-1}) + (1-\gamma)(b_{t-1}) \tag{10}$$

As it is known, the parameter γ is also added to the model. Finally, the model presented with seasonal changes will be as follows:

$$U_t = \beta \frac{z_t}{k_t} + (1-\beta)U_{t-Q} \tag{11}$$

In order to predict in the Holt-Winters model, equation (12) is used.

$$P_{t+m} = (k_t + mb_t)U_{t-Q+m} \tag{12}$$

It should be noted that the parameters in this model α, β, γ are smoothing constants and are between zero and one.

Autoregressive neural network (ANN)

In the ANN model, prediction is made in two stages. For the desirable time series, the sequence of the autoregressive model is specified in the first stage. In the second stage, the neural network is invigorated by the training data set regarding the autoregressive sequence. The number of time series or input nodes delays of the neural network is specified from the autoregressive sequence. In this model, the fitted method with a non-seasonal trend includes of two parts h and p, where h represents the number of hidden neurons and p represents the number of input logs. Accordingly, this method is shown as ARNN(p,h). Likewise, the fitted method with seasonal trend is ARNN(p,P,h)$_{(v)}$, which is similar to ARIMA(p,0,0)(P,0,0)$_{(v)}$ with non-linear objectives.

KNN regression

KNN stands for *K*-Nearest Neighbors, which means the name of this algorithm refers to its working method. KNN regression simply takes a set of training samples. The nth training sample consists of a vector, with b features: $(f_1^n, f_2^n, \dots, f_b^n)$، an associated target vector of m features: $(h_1^i, h_2^i, \dots, h_m^i)$. According to a new example, whose characteristics are known $(q_1, \dots, q_b)$; But its purpose is unknown; The new sample features are applied to obtain the k most similar training samples based on the feature vector and distance or similarity. For example, assuming that the distance is Euclidean, the distance between the new sample and the training sample n is calculated as equation (13):

$$\sqrt{\sum_{l=1}^{i}(f_l^{\,n} - q_l)^2} \tag{13}$$

The k training samples closest to the new sample, their k nearest neighbors, or the k most similar samples are considered. KNN is based on learning by analogy. Given a new sample, it appears that the targets of its nearest neighbors are likely to be the same as its unknown target. In this way, the targets of the nearest neighbors are summed to predict the target of the new sample. For example, assuming that targets or k are the nearest neighbors of the vectors: $h^1, \dots, h^u$, they can be combined to predict the target of the new sample as calculated in the average form:

$$\sum_{i=1}^{k} \frac{h^i}{k} \tag{14}$$

Briefly, KNN stores a set of training examples described by n features. Each training sample represents a point in n-dimensional space. Given a new sample, the KNN finds its k closest samples in the n-dimensional space hoping that their targets are similar to its unknown target. In the discussion of time series predicting, the goal of a training sample is a set of data values; And the features that describe the sample are the lagged values of the target. In this way, an autoregressive model is formed [25].

Theta method

This method is according the concept of correcting the local curves of time series data. This behavior is derived from a coefficient named θ (symbolized by the Greek letter Theta) that is directly utilized to the second divergences of the time series:

$$G_{new}''(\theta) = \theta . G_{data}'', where G_{data}'' = G_t - 2G_{(t-1)} - G_{(t-2)} \text{ in time t} \tag{15}$$

If the value of θ decreases gradually, the time series also decreases. The smaller the value of θ coefficient, the lower the degree of the peaks. The coefficient θ can also have negative values. When θ is equals to 0, the time series becomes a linear regression line. Also, if the θ increases (θ is greater than 1), then the time series will widen. The general formula of the Theta method is described in relation (16). The initial time series is split into more theta lines. Each theta line is analogized separately and the projections are indeed combined. For example, when θ equals to 0 and 2, accordingly, we have:

$$Data = \frac{1}{2}\left(L(\theta = 0) + L(\theta = 2)\right) \tag{16}$$

where L(θ = 0) represents Theta line for parameter θ=0. The first theta line (θ equals to 0) is the linear regression line of the time series data, and the second is precisely twice the original time series. The first part $L(\theta = 0)$ represents the time series through a linear behavior. The second part $L(\theta = 2)$, doubles the local flexions and magnifies the short-term trend. The first Theta line is analogized in the common path for a linear behavior. The latter is analogized through ordinary exponential smoothing. An ordinary combination of these two predictions makes the final Theta model prediction for a particular time series.

In general, Theta modeling steps include checking seasonality, deseasonalization, decomposition, extrapolation, and synthesis. An important point is to apply several combinations of Theta lines for each forecast time horizon. Research results show that for longer time horizons, predictions must be more oriented toward long-term trend, while for short-term forecasts, we must consider recent behaviors. This can be simply done by applying several pairs of Theta lines for each forecast time horizon. For example, if the pair θ = 0.00 and θ = 2.50 is applied, more emphasis is placed on the short-term behavior of the time series, while in the case of Theta lines θ = 0.00 and θ = 1.50, the long-term trend becomes more critical [25].

STL method

STL is a filtering method to separate a time series into residual, seasonal, and trend components. Suppose that for the data, the outcome, residual, seasonal, and trend components are indicated by G_r, T_r, S_r, and R_r, respectively, for $r = 1,\ldots,N$. Thus, we have: $G_r = T_r + S_r + R_r$.

STL includes of an order of smoothing operations that use locally weighted or Loess regression. STL has six parameters, as follows, that must be selected by the data analyst:

$j_{(p)}$: number of observations in each seasonal cycle /$j_{(o)}$: number of stable iterations of the outer loop /$j_{(i)}$: number of passes through the inner loop /$j_{(l)}$: smoothing parameter for the low-pass filter /$j_{(s)}$: smoothing parameter for the seasonal component /$j_{(t)}$: smoothing parameter for the trend component.

Choosing the first five is simple. However, the last parameter, j(s), must be carefully tuned for each time series. For more information about this method, refer to the reference [26].

Cubic spline smoothing method

Consider a univariate time series y_j, j = 1,…,n, with a non-linear behavior. It is intended to forecast the series by analogizing the behavior by applying a linear function obtained from the observed time series data. The linear behavior extrapolation approach is broadly applied and accomplishes enough well practically. When the time series has equally spaced data, a cubic spline smoothing is represented as the function $f(j)$ that optimizes overall doubly differentiable functions f in S ($[1,n] \subseteq S \subseteq R$).

The critical parameter λ controls the rate of change between the local variation represented by the integral of the second derivative of f squared and the residual error defined by the sum of the squared residuals. Large values of λ show the function $f(j)$ close to a straight line, while small amounts of λ show a very curved (meandering) function $f(j)$. For more information about this method, refer to the reference [27].

BATS model

The BATS identifier stands for crucial characteristics of the model: Box-Cox transformation, seasonal components, ARMA errors, and trend. To represent the Box–Cox parameter, the adjustment parameter, the seasonal periods ($u_1,\dots,u_T$), and the ARMA parameters p and q are supplemented with the arguments (ω, φ,p,q,u_1,u_2,…,u_T)

The model is the clearest generalization of traditional seasonal innovation trends feasible for multiple seasonal time periods. Nevertheless, it cannot contain a season with an incorrect period, but it is possible to consider multiple seasonal periods for the series. The primary seasonal component alone contains non-zero u_T. For more information about this method, refer to the reference [28].

Evaluation criteria

The evaluation criteria are calculated as follows:

MSE:

$$MSE(Y,\hat{Y}) = 1/s\sum_{c=1}^{s}(Y_c - \hat{Y}_c)^2 = 1/s\sum_{c=1}^{s}{e_c}^2 \tag{17}$$

RMSE:

$$RMSE(Y,\hat{Y}) = \sqrt{MSE(Y,\hat{Y})} = \sqrt{1/s\sum_{c=1}^{s}(Y_c - \hat{Y}_c)^2} \tag{18}$$

MAE:

$$MAE(Y,\hat{Y}) = 1/s\sum_{c=1}^{s}|Y_c - \hat{Y}_c| = 1/s\sum_{c=1}^{s}|e_c| \tag{19}$$

MAPE:

$$MAPE(Y,\hat{Y}) = 100/s\sum_{c=1}^{s}|\frac{Y_c - \hat{Y}_c}{Y_c}| = 100/s\sum_{c=1}^{s}|\frac{e_c}{Y_c}| \tag{20}$$

where $\widehat{Y_c}$ is the forecasted value and Y_c is the real value

Modeling and data analysis

Here, the results of fitting the model and predicting the cumulative values of Corona (cumulative values of hospitalizations and cumulative values of deaths) for Yazd province, according to the models mentioned in the second section, are given.

The used corona data is from 02/26/2020 to 12/19/2021 daily, which was obtained from the Ministry of Health. It should be noted that the available statistics are for people who were admitted and hospitalized in hospitals in Yazd province and do not include people who were admitted to other medical centers or so-called outpatient treatment centers. In this research, the last 14 observed data are considered as test data, and with the rest of the data and models described in the second part, for the next 14 days, cumulative cases of Coronavirus (admissions and deaths) for Yazd province are predicted. Next, the best proposed model (By applying the evaluation criteria mentioned in the Section (2-11)) is selected. It should be noted that the date on which the first case of hospitalization or death was reported is considered as the starting day.

To develop statistical methods due to time series data, the following assumptions are considered:

- Time series data do not include outliers and anomalies.
- The time series data consists of one variable; Data are univariate.
- Data are static and do not require mean and variance over time.
- Over time, the parameters of models and errors do not change and remain constant.

The *R* programming software was used for calculations and coding. According to the data analysis and the results obtained according to Table 1, the best method and model for cumulative cases of hospitalization of Covid-19 in Yazd province based on evaluation criteria or prediction errors, KNN regression model with evaluation criteria MSE=359.185, RMSE = 18.952, MAE = 15.487 and MAPE = 0.045. Accordingly, this method will be applied to forecast the cumulative instances of hospitalization of Coronavirus in Yazd province.

Table 1. Evaluation criteria of models related to cumulative cases of hospitalization of Covid-19 in Yazd province

Model	**MSE**	***RMSE***	**MAE**	**MAPE**
ARIMA	11835.260	108.790	91.740	0.268
Exponential smoothing	5880.527	76.684	64.251	0.188
Holt-Winters	5874.088	76.642	64.215	0.187
ANN	161174	401.465	435.086	1.678
KNN regression	359.185	18.952	15.487	0.045
THETA	111046.800	334.067	298.183	0.873
STL	2210.955	47.020	45.016	0.132
Cubic spline smoothing	10216.460	101.076	84.757	0.248
BATS	8629.444	92.894	77.050	0.225

Based on evaluation criteria or prediction errors, the best method and model for cumulative instances of Coronavirus deaths in Yazd province, the BATS model with evaluation criteria of MSE=17.247, RMSE=4.153, MAE=2.907, and MAPE = 0.067. Therefore, in the future, this method will be applied to forecast cumulative cases of Coronavirus deaths in Yazd province.

Table 2. Evaluation criteria of models related to cumulative cases of covid-19 deaths in Yazd province

Model	**MSE**	***RMSE***	**MAE**	**MAPE**
ARIMA	28.365	5.325	4.553	0.105
Exponential smoothing	28.126	5.303	4.217	0.097
Holt-Winters	28.064	5.297	4.213	0.097
ANN	90.915	13.817	12.339	0.606
KNN regression	19.772	4.446	3.506	0.081
THETA	42.809	6.542	5.902	0.136
STL	23.829	4.881	3.887	0.090
Cubic spline smoothing	20.005	4.472	3.545	0.822
BATS	17.247	4.153	2.907	0.067

Figures 1 to 6 have been drawn to better examine the performance of the nine mentioned forecasting models. According to the examination of Figures 1 to 3, it seems that the STL and KNN regression models for predicting new hospitalization cases of Covid-19 have less error, among which the KNN regression model has the lowest error for forecasting the cumulative instances of hospitalization for Coronavirus. Also, according to the evaluation criteria, the Holt-Winters model and the exponential smoothing have almost the same performance in forecasting cumulative instances of Coronavirus hospitalizations in Yazd province. By examining and analyzing Figures 4 to 6, it can be concluded that the smoothing models of cubic spline, BATS, and STL have less errors than the rest of the models. Meanwhile, the BATS modeling method has the lowest error for predicting the cumulative cases of death. According to the evaluation criteria, the Holt-Winters model and the exponential smoothing have almost the same performance in forecasting cumulative instances of Coronavirus in the province. Nevertheless, what is evident is that the autoregressive neural network method has the worst performance among all the nine mentioned methods, both for hospitalization cases and death cases.

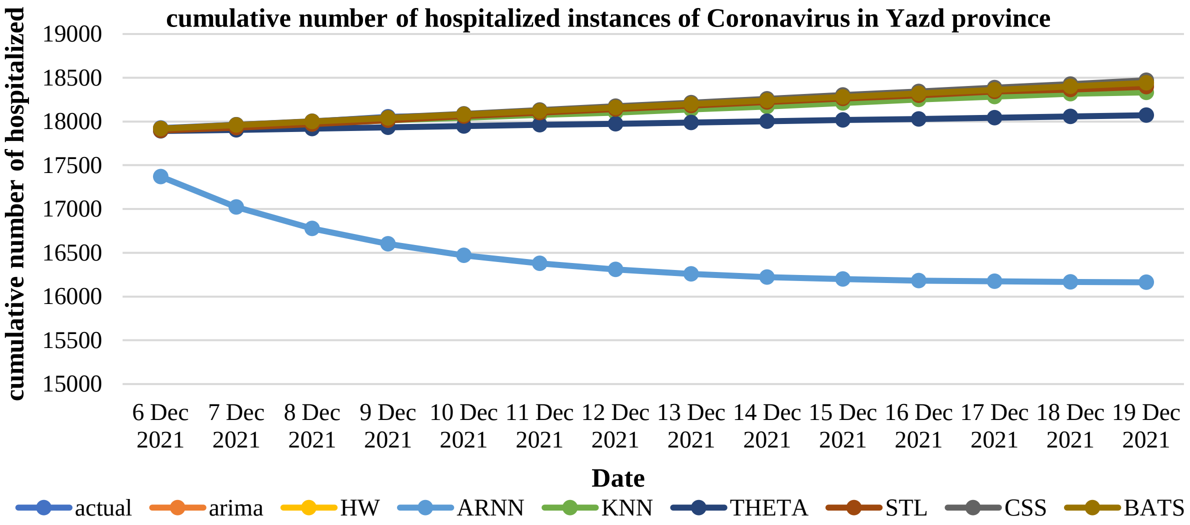


Figure 1. The 14-day prediction chart of the cumulative number of hospitalized instances of Coronavirus in Yazd province

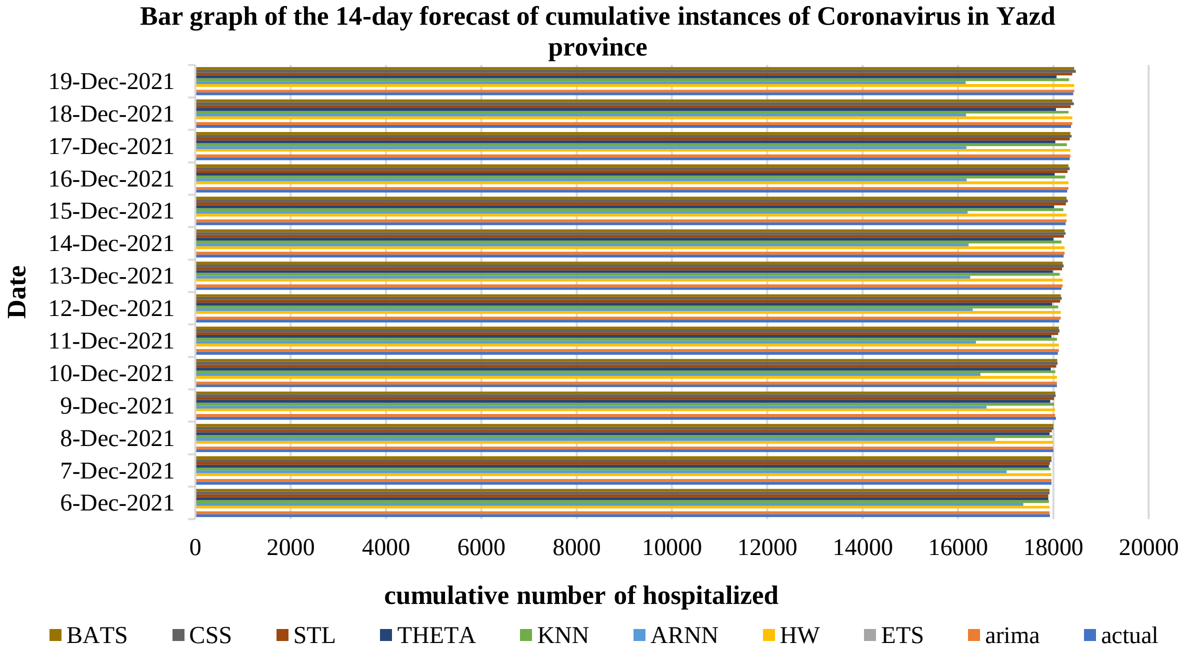
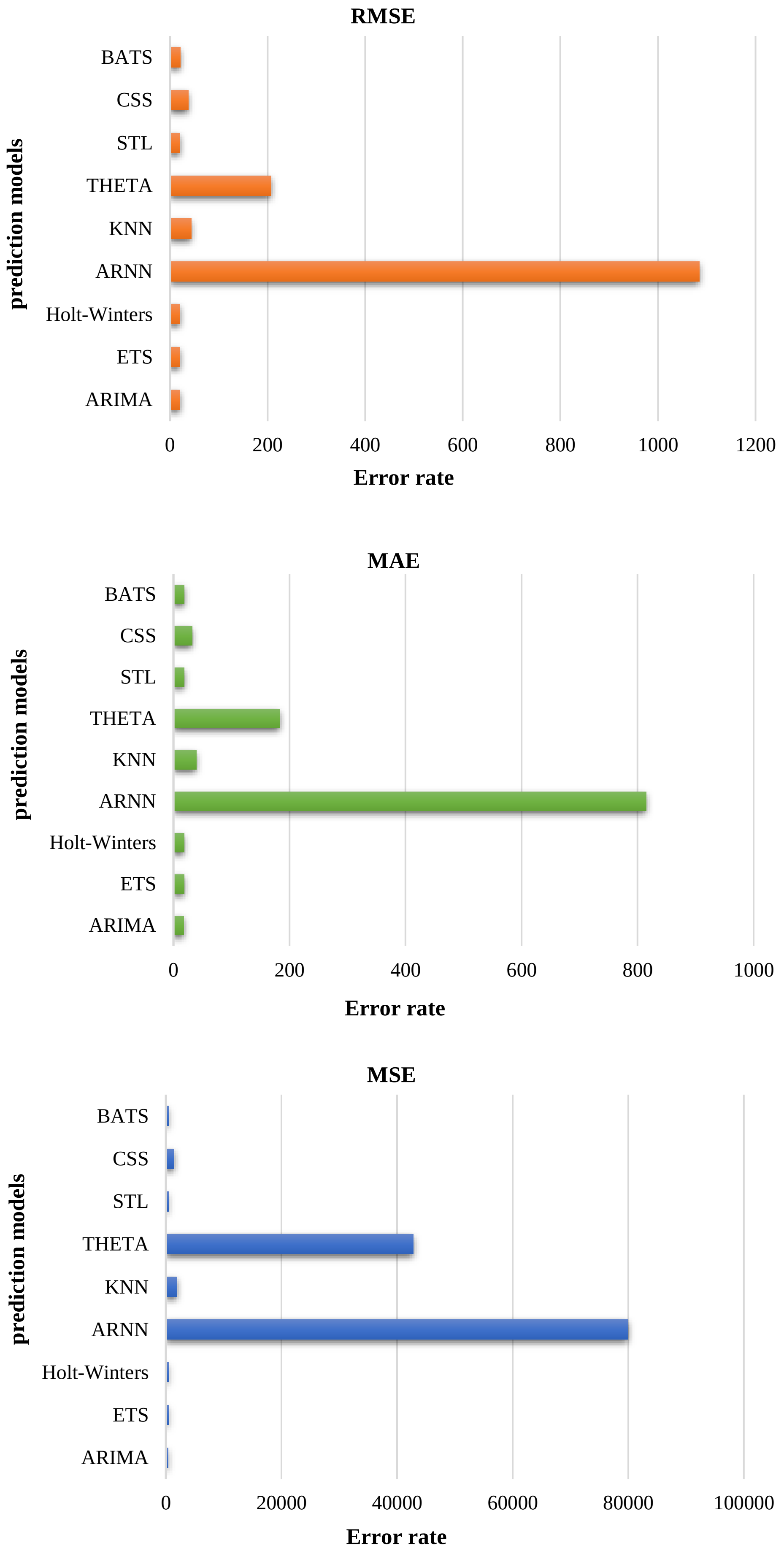


Figure 2. Bar graph of the 14-day forecast of cumulative instances of Coronavirus in Yazd province


RMSE
prediction models
BATS
CSS
STL
THETA
KNN
ARNN
Holt-Winters
ETS
ARIMA
0
200
400
600
800
1000
1200
Error rate
MAE
prediction models
BATS
CSS
STL
THETA
KNN
ARNN
Holt-Winters
ETS
ARIMA
0
200
400
600
800
1000
Error rate
MSE
prediction models
BATS
CSS
STL
THETA
KNN
ARNN
Holt-Winters
ETS
ARIMA
0
20000
40000
60000
80000
100000
Error rate

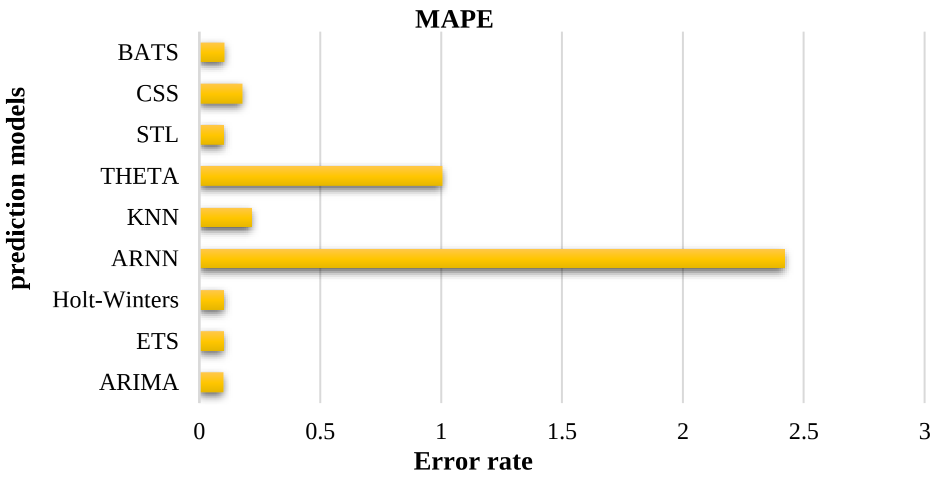


Figure 3. Chart of evaluation criteria for cumulative instances of hospitalization of Coronavirus in Yazd province

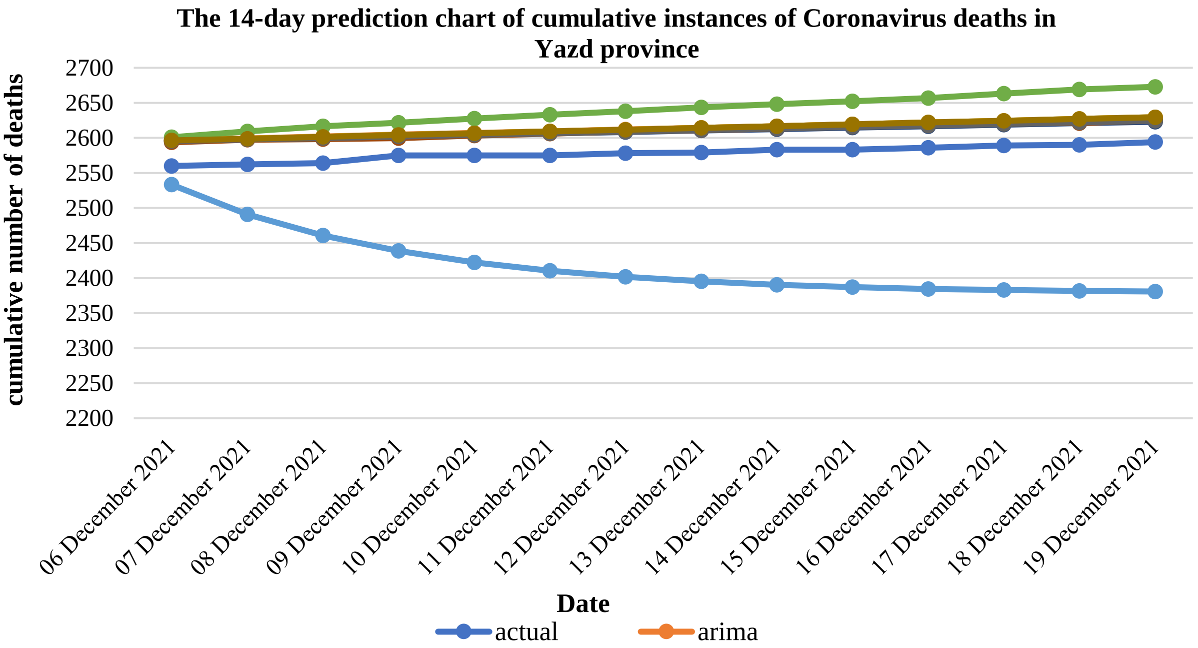


Figure 4. The 14-day prediction chart of cumulative instances of Coronavirus deaths in Yazd province

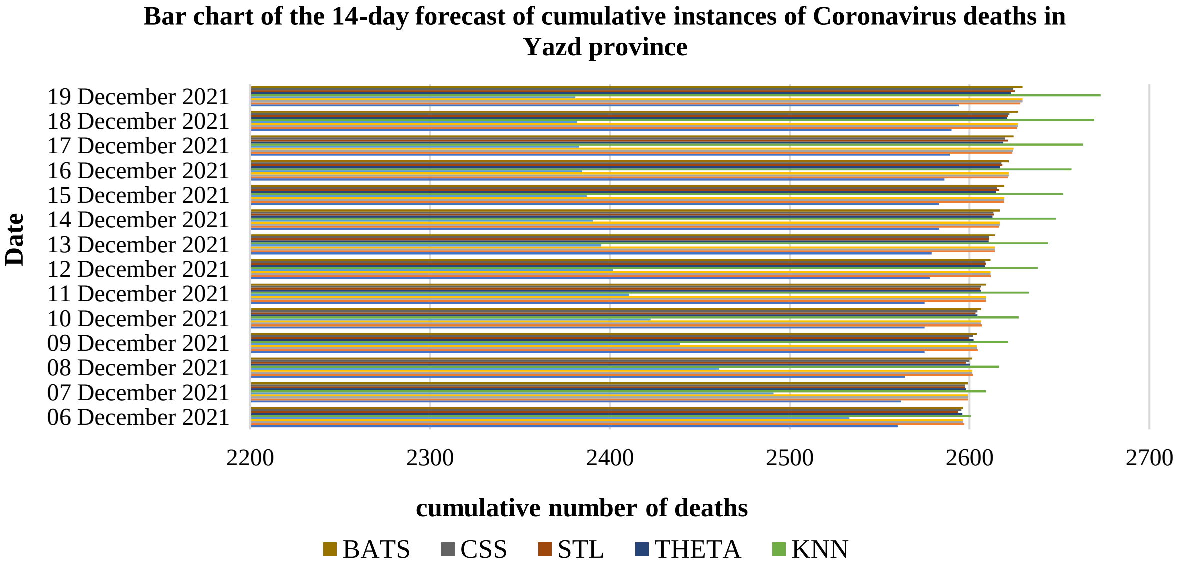
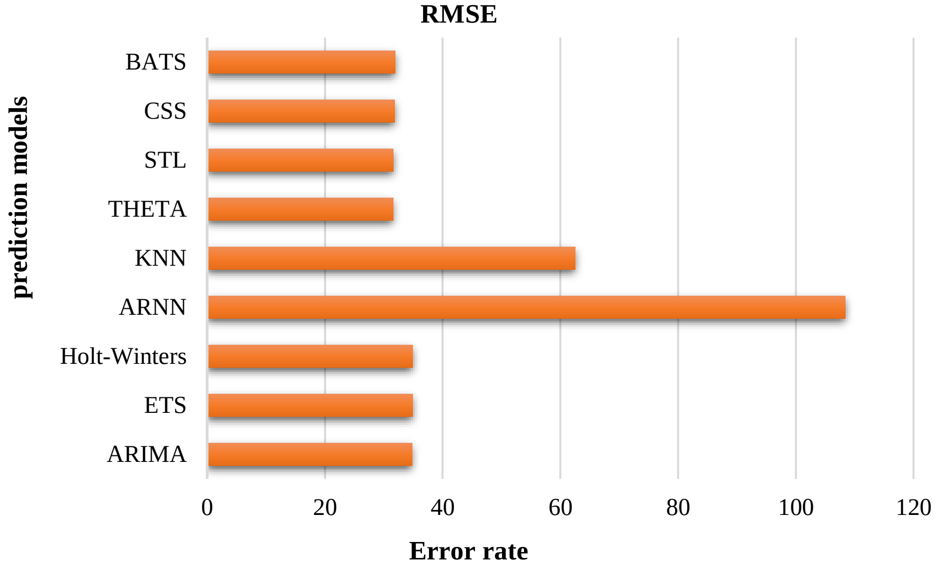
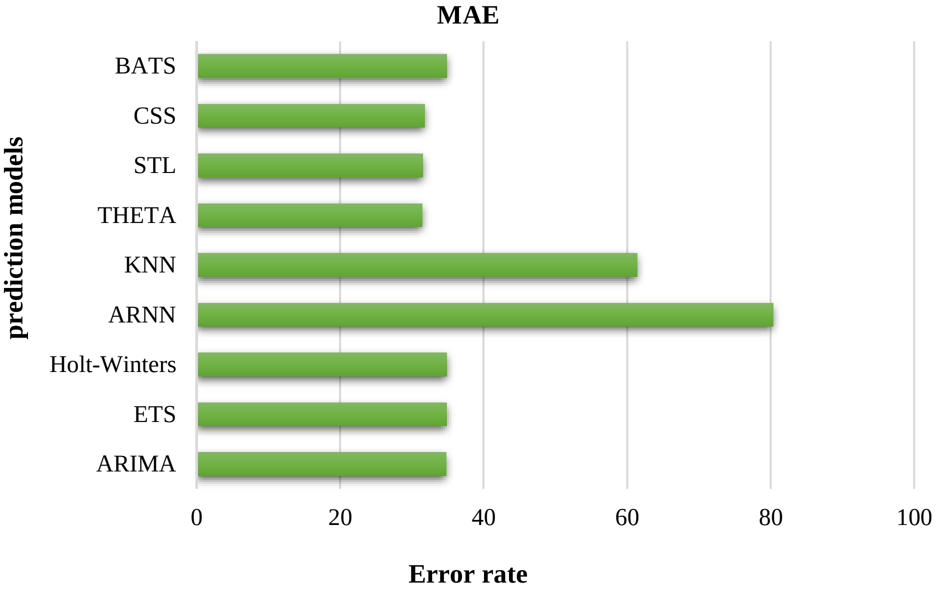
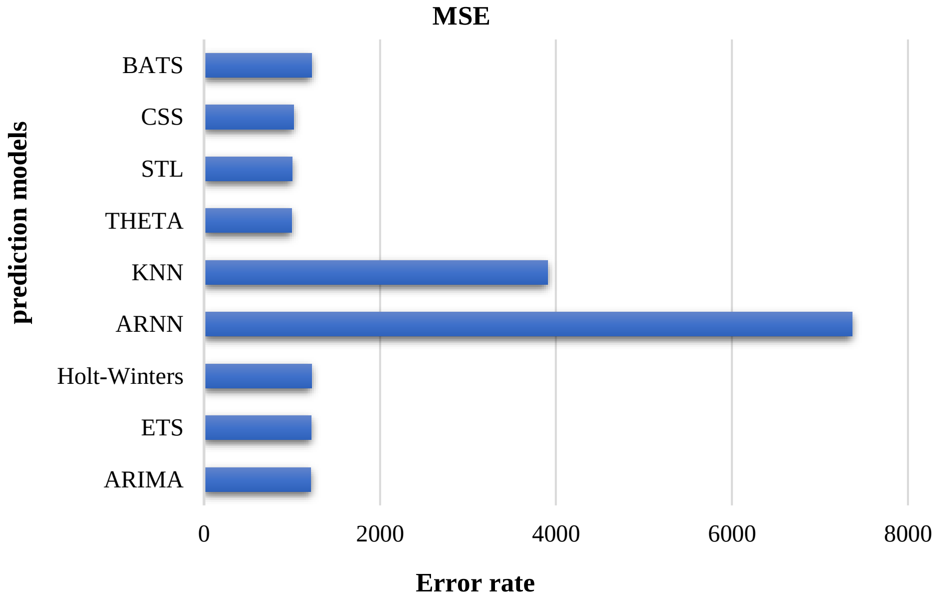


Figure 5. Bar chart of the 14-day forecast of cumulative instances of Coronavirus deaths in Yazd province


RMSE
prediction models
BATS
CSS
STL
THETA
KNN
ARNN
Holt-Winters
ETS
ARIMA
0
20
40
60
80
100
120
Error rate


MAE
prediction models
BATS
CSS
STL
THETA
KNN
ARNN
Holt-Winters
ETS
ARIMA
0
20
40
60
80
100
Error rate


MSE
prediction models
BATS
CSS
STL
THETA
KNN
ARNN
Holt-Winters
ETS
ARIMA
0
2000
4000
6000
8000
Error rate

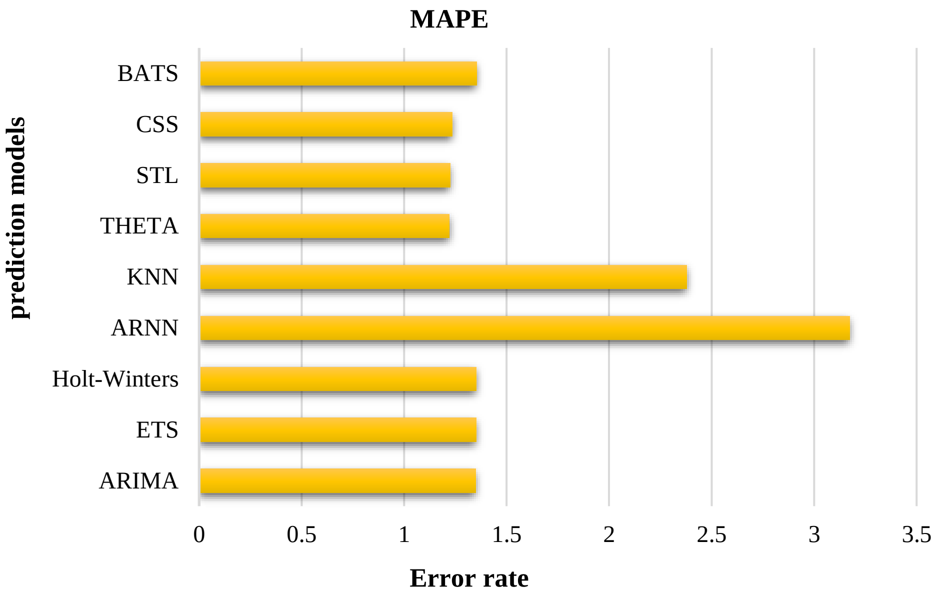


Figure 6. Chart of evaluation criteria for cumulative instances of Coronavirus deaths in Yazd province

The practical results and managerial perspectives from the present study are as follows:

- ✓ The results of the analysis represented that the best model according to the mentioned evaluation criteria for forecasting the cumulative instances of hospitalization of Coronavirus is the KNN regression model and for the cumulative cases of death is the BATS model. Also, the autoregressive neural network model performs worst among all nine models, both for hospitalization and death cases.
- ✓ The most critical point that can be taken from the analysis of the results is that it is impossible to consider a single model for cases of hospitalization and death. Depending on the special conditions of each region as well as the index required for prediction, different prediction models should be tested and the most efficient fitted model should be selected.
- ✓ In the context of the subject of the current research, by using the models that were introduced many years ago (the so-called classical models), better results can be obtained in practice; While the results obtained from newer models, such as the autoregressive neural network model, indicate inappropriate performance in predicting Covid-19. Therefore, it can be concluded that using new models is not always effective in practice, and according to the type of data, classic models may have better results and efficiency in forecasting.
- ✓ The obtained results can be used to predict and control other viral diseases that may emerge as a pandemic or even epidemic and endemic in the future.

Conclusions and future directions

In this study, nine statistical and machine learning models, including ARIMA, Theta, exponential smoothing, Holt-Winters, autoregressive neural network, KNN regression, BATS method, cubic spline smoothing method, and STL method, were applied and comprehensively analyzed to forecast the cumulative number of hospitalized instances and mortality of Coronavirus in a case study. They were used and compared with each other using MSE, RMSE, MAE, and MAPE evaluation criteria. Meanwhile, the last 14 data were considered as test data. Summarizing the results represented that the best model for forecasting cumulative cases of hospitalization and mortality of Covid-19 in Yazd province as a case study is different. Regarding evaluation criteria, for cumulative cases hospitalized in Yazd province, the best model is KNN regression; While for cumulative cases of death, the BATS method was chosen as the best model. Due to the results of the present study, the following are suggested to future researchers:

- ✓ Prediction of Covid-19 by considering the effect of wearing a mask and observing social distance
- ✓ Prediction of Coronavirus according to the effect of vaccination against Coronavirus,
- ✓ Prediction of Coronavirus according to the combination of methods such as the Theta method and STL analysis method,
- ✓ Applying the nine proposed modeling methods for predicting other infectious or pandemic diseases,
- ✓ Modeling of proposed modeling methods for other geographical regions under different weather conditions.

References:

[1] "Q&A on coronaviruses (COVID-19)". *World Health Organization*. Retrieved 11 March 2020.

[2] Tran, T.T., Pham, L.T. and Ngo, Q.X., 2020. Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (Case study: Iran). *Global Journal of Environmental Science and Management*, *6*(Special Issue (Covid-19)), pp.1-10.

[3] Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A.A. and Li, X., 2013. Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PloS one*, *8*(5), p.e63116.

[4] Chen, Y., Leng, K., Lu, Y., Wen, L., Qi, Y., Gao, W., Chen, H., Bai, L., An, X., Sun, B. and Wang, P., 2020. Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010–2018. *Epidemiology & Infection*, *148*.

[5] Ceylan, Z., 2020. Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of The Total Environment*, *729*, p.138817.

[6] Bayyurt, L. and Bayyurt, B., 2020. Forecasting of COVID-19 cases and deaths using ARIMA models. *medrxiv*, pp.2020-04.

[7] Tandon, H., Ranjan, P., Chakraborty, T. and Suhag, V., 2022. Coronavirus (COVID-19): ARIMA-based Time-series Analysis to Forecast near Future and the Effect of School Reopening in India. *Journal of Health Management*, *24*(3), pp.373-388.

[8] Perone, G., 2020. An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy. *MedRxiv*, pp.2020-04.

[9] Ghosal, S., Sengupta, S., Majumder, M. and Sinha, B., 2020. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases-March 14th 2020). *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), pp.311-315.

[10] Parbat, D. and Chakraborty, M., 2020. A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons & Fractals*, *138*, p.109942.

[11] Naimoli, A., 2022. Modelling the persistence of Covid-19 positivity rate in Italy. *Socio-Economic Planning Sciences*, *82*, p.101225.

[12] Kibria, H.B., Jyoti, O. and Matin, A., 2022. Forecasting the spread of the third wave of COVID-19 pandemic using time series analysis in Bangladesh. *Informatics in medicine unlocked*, *28*, p.100815.

[13] Khan, F.M. and Gupta, R., 2020. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety Science and Resilience*, *1*(1), pp.12-18.

[14] Chowdhury, A.A., Hasan, K.T. and Hoque, K.K.S., 2021. Analysis and prediction of COVID-19 pandemic in Bangladesh by using ANFIS and LSTM network. *Cognitive Computation*, *13*, pp.761-770.

[15] Alassafi, M.O., Jarrah, M. and Alotaibi, R., 2022. Time series predicting of COVID-19 based on deep learning. *Neurocomputing*, *468*, pp.335-344.

[16] Chyon, F.A., Suman, M.N.H., Fahim, M.R.I. and Ahmmed, M.S., 2022. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *Journal of Virological Methods*, *301*, p.114433.

[17] Nair, S., Ckm, G., Varsha, R., Ghosal, S., Vergin, M. and Anbarasi, L.J., 2022. Intelligent Forecasting Strategy for COVID-19 Pandemic Trend in India: A Statistical Approach. In *Artificial Intelligence and Technologies: Select Proceedings of ICRTAC-AIT 2020* (pp. 553-560). Springer Singapore.

[18] Mukhairez, H H, & Alaff, A J, 2022. Short-term Forecasting of COVID-19. In *Computational Intelligence for COVID-19 and Future Pandemics*, (pp. 257-266). Springer, Singapore.

[19] Prajapati, S., Swaraj, A., Lalwani, R., Narwal, A. and Verma, K., 2021. Comparison of traditional and hybrid time series models for forecasting COVID-19 cases. *arXiv preprint arXiv:2105.03266*.

[20] Li, C., Sampene, A.K., Agyeman, F.O., Robert, B. and Ayisi, A.L., 2022. Forecasting the severity of COVID-19 pandemic amidst the emerging SARS-CoV-2 variants: adoption of ARIMA model. *Computational and*

Mathematical Methods in Medicine, *2022*.

[21] Chung, H., Ko, H., Lee, H., Yon, D.K., Lee, W.H., Kim, T.S., Kim, K.W. and Lee, J., 2023. Development and validation of a deep learning model to diagnose COVID-19 using time-series heart rate values before the onset of symptoms. *Journal of Medical Virology*.

[22] Yaffee RA, McGee M. 2000, An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®. *Elsevier*.

[23] Sardar, I., Akbar, M.A., Leiva, V., Alsanad, A. and Mishra, P., 2023. Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: Methodology, evaluation, and case study in SAARC countries. *Stochastic Environmental Research and Risk Assessment*, *37*(1), pp.345-359.

[24] Kufel, T., 2020. ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, *15*(2), pp.181-204.

[25] Assimakopoulos, V. and Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. *International journal of forecasting*, *16*(4), pp.521-530.

[26] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. 1990. STL: A seasonal-trend decomposition. *J. Off. Stat*. 6(1):3-73.

[27] Hyndman, R.J., King, M.L., Pitrun, I. and Billah, B., 2005. Local linear forecasts using cubic smoothing splines. *Australian & New Zealand Journal of Statistics*, *47*(1), pp.87-99.

[28] De Livera, A.M., Hyndman, R.J. and Snyder, R.D., 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American statistical association*, *106*(496), pp.1513-1527.