published by

EQA - International Journal of Environmental Quality
ISSN 2281-4485 - Vol. 53 (2023): 1-10
Journal homepage: https://eqa.unibo.it/

On replacement of outliers and missing values in time series

Loganathan Appaia, Sumithra Palraj*

Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

*Corresponding author E-mail: 350433@msuniv.ac.in

Article info
Received 11/1/2022; received in revised form 21/1/2023; accepted 30/1/2023
DOI: 10.6092/issn.2281-4485/16184

Abstract
Presence of missing values and occurrence of outliers in time series cause many hindrances in the analysis of data. Several methods are proposed for determining estimates to replace the missing values and outliers. Mean, median, the largest order statistic and time series model based forecast values are used as the estimates for replacing missing values and outliers. But, no recommendations have been made so far for selection of the estimation methods. This paper attempts to compare the performance of six such estimation methods. Among them, time series models are fitted applying the autoregressive moving average method, long short-term memory method and Facebook's Prophet method. Models are validated using the test data. Time series of Air Quality Index is used for carrying out for comparative study.

Keywords
Missing Value, Outlier, Autoregressive Moving Average, Facebook's Prophet, Long Short-Term Memory, Mean and Median Imputation

Introduction

Forecasting is one of the main objectives for fitting time series models. Accuracy level of forecast values often decline due to the presence of missing values and occurrence of outliers in the time series. Treating of missing values and outliers is an important task in the time series analysis.

Missing data situation arises for various reasons which include non-response and ambiguity of respondents, data entry errors, system failure, challenges in updating databases. Huang *et al.* (2018) and Rubin (1976) divided the missing data randomness into three categories *viz.*, Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR). When a missing value follows the MCAR mechanism, the missing value for an element does not depend on either the known values or the missing data. At this level of randomness, any missing value treatment approach can be employed without risk of introducing bias into the data. Huang *et al.* (2018) mentioned that under the conditions of MAR and NMAR, missing value observations may depend on the value of that observation. These missing values are considered as significant obstacles in data analysis because they distort the statistical properties of the data and reduce availability of the data. Also, due to the occurrence of missing values in time series, it is difficult to determine correlations with past lag.

In general, time series data may consist of seasonal, trend and random components. Seasonality and trend can be detected easily, but the random component makes time series data difficult to work effectively. Occurrence of small and large values in a time series leads to fluctuations in the time series. But, magnitudes of too large and too small values lead to the occurrence of outliers in the time series. Some of the extreme values may disrupt the pattern of the series and affect the autocorrelation structure in the time series. Tolvi (1998) pointed out that outliers also affect the predicted

DOI: 10.6092/issn.2281-4485/16184

autocorrelation and partial autocorrelation functions. Deutsch *et al.* (1990) found that due to the presence of outlier's autoregressive model can be misidentified either as a moving average or an autoregressive moving average model. According to Ledolter (1989), outlier may increase the estimated variance of the series, making prediction intervals extremely vague.

Tolvi (1998) and Cousineau and Chartier (2010) mentioned that outliers and missing observations can be replaced by appropriate estimates. Several studies reported various aspects of univariate estimation methods. Univariate estimation methods such as mean substitution, median substitution, last observation carried forward, linear interpolation, and seasonal Kalman filter have been compared with respect to imputation accuracy. Imputation accuracy of each approach has been tested based on the difference between the estimated value and actual values. Agbailu *et al.* (2021) and Zeileis *et al.* (2021) suggested that the Kalman filter method is more successful for univariate imputation. Savarimuthu and Karesiddaiah (2021) compared seven missing value estimation strategies for time series analysis. The evaluation metric findings of the TIMIMP and other existing methods reveal that the TIMIMP approach outperforms the other methods.

Kihoro *et al.* (2013) attempted to apply univariate missing value estimation methods using time series forecasting models such as autoregressive integrated moving average model and its variant, seasonal autoregressive integrated moving average for seasonal time series data. They also proposed a univariate imputation method based on direct linear regression. First, linear regression is performed to identify a time series that includes a subsequence which is the most similar to a subsequence before the missing part. Then, the missing part is replaced by the next subsequence of the most similar one.

Recently, some machine learning and deep learning approaches have been proposed to determine estimates for missing values. Cinar *et al.* (2018) attempted to construct recurrent neural networks for estimates to the missing values. They also found that when large scale data is available, deep learning methods can produce estimates with more accuracy in time series analysis.

However, no significant attempt is formed in the literature suggesting the selection of appropriate estimation methods. Main objective of this paper is to compare the performance of six methods for determining estimates *viz.,* mean, median, largest order statistic, autoregressive moving average method, long short-term memory method and Prophet method. Comparative analysis is performed for the time series of air quality index (AQI) of a region. Details of the AQI data and the operating procedure of the six methods are detailed in Section 2. The results of comparative analysis are discussed in Section 3. The conclusions and remarks are presented in Section 4.

Materials and Methods

The Central Pollution Control Board, India runs three ambient air quality monitoring stations in Chennai at Manali, Alandur and Velachery. These stations maintain the daily records of levels of seven pollutants *viz.,* Particulate Matter 2.5($PM_{2.5}$), Nitrogen oxide (NO), Nitrogen- dioxide (NO_2), NOx, Sulphur-di-oxide (SO_2), Benzene and Toluene. In addition, wind speed, wind direction, relative humidity and temperature are also monitored.

A sub-index is computed every day for each of these pollutants, and the largest among them is determined as the AQI of the respective location for that particular day, as mentioned by the Central Pollution Control Board of India (*https://cpcb.nic.in/National-Air-Quality-Index/*).

The daily AQI determined at Velachery monitoring station for a period of 813 days from January 1, 2018 to March 24, 2020 are considered in this study. The AQI from March 25, 2020 is not considered since a nationwide lockdown was implemented in India due to the COVID-19 pandemic conditions (*https://tnsdma.tn.gov.in/pages/view/covid-19-lockdown-gos*). Record of AQI for six days in the study period, *viz.,* December 5-9, 2019 and December 25, 2019 are not available, and they can be considered as missing observations of the data. Figure 1 display the AQI of Velachery.

Box-Whisker plot of the data in Figure 2 indicates that there are outlying observations in the data. Making decisions from quantitative analysis of data in the presence of outliers may cause severe deviations from the actual information, since magnitudes of the extreme values can affect the calculated measures. Also, applications of time-series analysis methods like model building, require observations recorded at all time points.

It can be noticed that there is no literature defining outliers in quantitative aspects. Many researchers have employed different methods of identifying outliers. Recently, Kolbasi and Unsal, 2019 have prescribed two boundaries for identifying outlying observations, which are determined from the first quartile(Q_1), third quartile(Q_3), and the interquartile range (IQR).

Lower boundary = Q_1 - 3 IQR
Upper boundary = Q_3 + 3 IQR

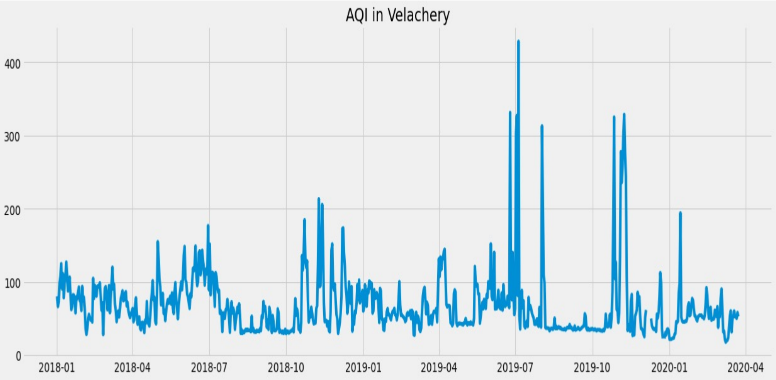

Figure 1
AQI in Velachery station from January 1, 2018 to March 24,2020

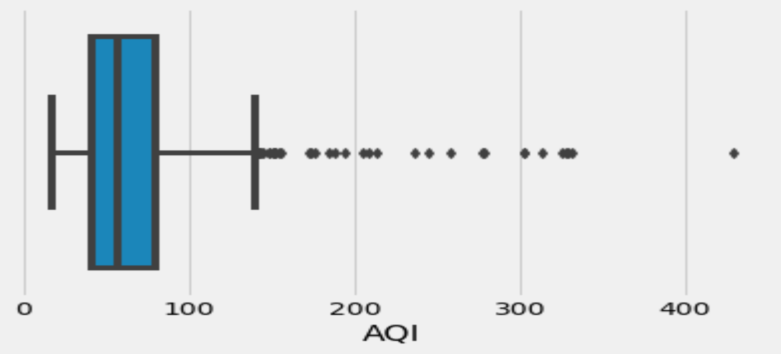

Figure 2
Box-Whisker Plot AQI in Velachery station from January 1, 2018 to March 24,2020

As mentioned by Kolbasi and Unsal, 2019 the observations in the data less than the lower boundary and greater than the upper boundary can be regarded as an outliers.

The Q_1, Q_3 and IQR for AQI recorded at Velachery for the study period are respectively 39.97, 80.38 and 40.41. Hence, the boundaries for identifying outliers for AQI data under study, are -81.26 and 201.62. Since the values of AQI are positive real numbers, none of the values of AQI can be smaller than the lower boundary. However, it is found that 16 observations exceed the upper boundary of 201.62. Thus, there are 16 outlying observations in the data. In total, there are 22 observations in the data which seek serious attention.

Methodology

Ahn *et al.* (2022), Enders (2010), Jadhav *et al.* (2019), Kihoro *et al.* (2013), Rubin (1976), Tolvi (1998) have suggested to replace the missing values with suitable estimates for obtaining meaningful results. The outlying observations may also be replaced by suitable alternate values as estimates. The following all the methods for estimating the missing values and replacing the outliers with suitable estimates:

Method-1: Estimation by mean (Rubin, 1976; Enders, 2010)

Method-2: Estimation by median (Enders, 2010; Lin and Tsai, 2020)

Method-3: Estimation by the largest order statistic

Method-4: Estimation by forecast values from time series models constructed applying

a) Autoregressive Moving Average (ARMA) method (Deneshkumar, V., Kannan, K.S., 2011; Kihoro *et al.,* 2013; Tolvi, 1998).

b) Long Short-Term Memory (LSTM) method (Chang *et al.,* 2020; Song *et al.,* 2020; Qiu *et al.,* 2020; Janarthanan *et al.,* 2021; Lyu *et al.,* 2021; Mani and Volety, 2021)

c) Facebook's Prophet method (Taylor and Letham, 2018; Shen *et al.,* 2020)

The missing observations and outliers will be commonly referred in the remaining part of the text as "*replaceable observations*".

Suppose that there are *M* number of *replaceable observations* in the given time series. Replacement of these observations with the central measures like mean and median, will not affect the stationary property of the time series. Also, using the largest order statistic as an estimate may not affect the behaviour of the time series. Similarly, a valid time series model can produce forecast values with high precision or lower inaccuracy. Application of the four methods to determine the estimates are described below.

Method-1:

Step-1: If the r_jth observation in the time series is the *jth* *replaceable observation*, compute arithmetic mean of the preceding (r_j - 1) observations.

Step-2: Replace the j^{th} *replaceable observation* with the arithmetic mean.

Step-3: Repeat Step-1 and Step-2 for all the *M replaceable observations.*

Method-2:

Step-1: If the r_j^{th} observation in the time series is the *jth* replaceable observation, compute median of the preceding (r_j-1) observations.

Step-2: Replace the j^{th} replaceable observation with the median value.

Step-3: Repeat Step-1 and Step-2 for all the *M replaceable observations.*

Method-3:

Step-1: If the r_jth observation in the time series is the j^{th} *replaceable observation*, compute the largest order statistic of the preceding (r_j-1) observations.

Step-2: Replace the j^{th} *replaceable observation* with the largest order statistic.

Step-3: Repeat Step-1 and Step-2 for all the *M replaceable observations.*

Method-4:

Step-1: Construct a valid time series model applying the ARMA/ LSTM/ Prophet method based on the (*u*-1) observations preceding the u^{th} *replaceable observation*.

Step-2: Forecast the u^{th} observation from the constructed time series model.

Step-3: Replace the u^{th} *replaceable observation* by the forecast value.

Step-4: If s^{th} ($s>u$) observation in the time series is the next *replaceable observation,* construct a time series model based on the preceding (*s*-1) observations, taking into consideration of the earlier estimates too.

Step-5: Forecast the s^{th} observation applying the time series model of (*s*-1) observations.

Step-6: Replace the s^{th} *replaceable observation* by the forecast value.

Basic theoretical details about the ARMA, LSTM and Prophet method are presented below briefly.

Method-4. (a): Autoregressive Moving Average Method

The ARMA model proposed by Box-Jenkins, 2015 is one of the time series modelling techniques that can be fitted to stationary time series. ARMA(*p,q*) model is a combination both AR and MA of order *p* and *q*, respectively the general form of ARMA model is given by the

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q} \quad [1]$$

where

α is the intercept; β_i is the coefficient of AR component with lag *t-i, i*=1, 2, 3,... *p*; ϕ_j is the coefficient of MA component with lag *t- j, j*=1,2,3,..., *q;* is the error term for the series.

As mentioned by Box and Jenkins, 2015, determination of the co-efficients α, β_i and ϕ_j for the given time series is known as the fitting of the ARMA(*p,q*) model. Here, Python program is used for determining the maximum likelihood estimates of the *p*+*q*+1 model co-efficients. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used for selecting the appropriate model. Accordingly, the model with the lowest values of AIC and BIC is considered as the preferable model.

Method- 4. (b): Long Short-Term Memory Method

LSTM is like a basic recurrent network in that the hidden units are replaced by memory cells. To maintain and update the state of a memory cell, the LSTM model filters the information through the gate structure. The gate structure consists of inputs, forgotten gates, and output gates. There are three sigmoid layers and one tanh layer in each memory cell. which can effectively tackle the problem of gradient disappearance or explosion due to extended periods of information dependencies, learn and predict based on historical data, balance the temporal and nonlinear relationships of the data, and improve prediction outcomes (Chang *et al.,* 2020).

Equation (2) indicates that by feeding the previous time

DOI: 10.6092/issn.2281-4485/16184

point's output and the current time point's data into the hidden layer. The forget gate controls which messages are removed from the cell state.

$$f_t= \sigma\left(\omega_f[h_{t-1}, x_t] + b_f\right) \quad [2]$$

Hidden layer and the Sigmoid function by inputting the previous time point's output h_{t-1} as well as the current time point's input message x_t, create a value by using the sigmoid function, and it specifies how many C_t values must be added to the neuron state. By putting the previous time point's output h_{t-1} and the current time point's input message x_t into the hidden layer and the *tanh* function, a campaign is added to the neuron state. Equation (4), update message adds to the neuron state and add f_t to the neuron state C_{t-1} plus, multiply the obtained value i_t and as LSTM at time t. In the neuron state, the input gate selects which new messages are to be remembered.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + bi) \quad [3]$$

$$\widetilde{c_t} = tanh(w_i[h_{t-1}, x_t] + bi) \quad [4]$$

$$C_t=f_t *C_{t-1}+it*\widetilde{c_t} \quad [5]$$

In the neuron state, the output gate decides which messages are to be outputted. It enters the hidden layer and the Sigmoid function by inputting the previous time point's output h_{t-1} and adding the current time point's input message x_t to produce a value O_t using sigmoid activation function. It determines how many neurons in the state must be the output. A *tanh* initially activates the neuron status message before it is amplified by O_t. The multiplied result is the LSTM neural network's output message h_t (Equation [7]) at time t.

$$O_t = \sigma(w_0[h_{t-1}, x_t] + b_0) \quad [6]$$

$$h_t=O_t* \, tanh\, (C_t) \quad [7]$$

Method- 4. (c): Facebook Prophet Method

The Prophet is a method for forecasting time series data, which uses an additive model to determine non-linear trends with yearly, weekly, and daily seasonality and holiday impacts. It works with high seasonal effects and historical data from multiple seasons. Prophet method is robust to missing data and trend shifts, and it handles outliers well, which employs an additive regression strategy to fit the model (Taylor and Letham, 2018). In this case, Tamil Nadu government holidays are considered to decompose the holiday, as the vehicle usage mainly peaks during the festival holidays. The general form of Prophet model is given by the

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad [8]$$

where: is the linear trend (growth); is a seasonality (Periodic change); is a holiday event; is an error term that is supposed to be normally distributed but is not considered by the model.

The piecewise linear function used to solve the trend function, the linear growth model can be written as below,

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma) \quad [9]$$

As per the seasonality, Fourier series approximation can be used for seasonal functions.

$$s(t) = \sum_{n=1}^{n}(a_n cos\left(\frac{2\pi nt}{P}\right) + b_n sin\left(\frac{2\pi nt}{P}\right)) \quad [10]$$

Results and Discussions

Estimates of the *replaceable observations* in the time series of AQI of Velachery station are determined applying the six methods described in Section 2.1. Accordingly, six estimated time series are obtained, and they will be referred correspondingly as mean estimated time series, median estimated time series, largest order statistic estimated time series, ARMA estimated time series, LSTM estimated time series and Prophet estimated time series. Some of the descriptive statistical measures computed for the six time series are presented in Table 1.

It can be noted from Table 1 that application of all the six estimation methods yield the same minimum value, the first quartile, and the maximum value. But, range of the Prophet estimated time series is more than the range of the other five estimated time series. Mean estimated time series, median estimated time series, ARMA estimated time series, and LSTM estimated time series have mean and standard deviation. The mean and standard deviation of the Prophet estimated time series and the largest order statistic estimated time series are relatively larger than the respective measures of other estimated time series.

Table 1. *Descriptive Statistical Measures of Estimated Time Series of AQI of Velachery during January 1, 2018 to March 24, 2020.*

Descriptive Measure	Estimated Time Series					
	Mean	Median	Largest Order Statistic	ARMA	LSTM	Prophet
Minimum	17.5000	17.5000	17.5000	17.5000	17.5000	17.5000
Mean	63.5216	63.3661	66.7343	63.6334	63.9033	64.3624
SD	29.1620	29.1581	35.2849	29.4352	29.5096	30.3942
Quartile 1	41.0000	41.0000	41.0000	41.0000	41.0000	41.0000
Quartile 2	57.4713	57.4713	57.4713	57.0827	57.0827	57.0489
Quartile 3	78.7779	78.7779	81.5151	79.6903	79.6903	80.1972
Maximum	194.0941	194.0941	194.0941	194.0941	194.0941	207.0166

It is interesting to note that difference among the median of all the estimated time series is not significant. The third quartile of the mean estimated time series and median estimated time series are equal. Similarly, the ARMA estimated time series and LSTM estimated time series have equal third quartile. All the six estimated time series have similar characteristics except for mean, standard deviation, and the third quartile. Plot of the six estimated time series is displayed in Fig. 4.

ARMA models are fitted for the six-time series applying Box-Jenkins procedure for comparing the performance of the six estimation methods. The coefficients of the AR and MA components of each time series are displayed in Table 2 as β-co-efficient and ϕ-co-efficient respectively. There are no moving average components in the ARMA model fitted for the largest order statistic estimated time series, ARMA estimated time series and the Prophet estimated time series. The AIC and BIC values computed for each ARMA model are also displayed in Table 2. It can be observed form Table 2 that *Augmented Dickey-Fuller* test provide evidence that all estimated time series is stationary.

Table 2. *ARMA Models for the Six Estimated Time Series*

Time Series	ADF-Test *p* value	Fitted Model	β-coefficient	ϕ-coefficient	AIC	BIC
Mean Estimated Time Series	0.0022	ARMA (2, 1)	1.5830 and -0.5951	-0.8928	183.148	206.393
Median Estimated Time Series	0.0023	ARMA (2, 1)	1.5738 and -0.5861	-0.8908	187.584	210.829
Largest Order Statistic Estimated Time Series	0.00	ARMA (1, 0)	0.7676	_	355.464	369.411
ARMA Estimated Time Series	0.002	ARMA (2, 0)	0.7530 and 0.0495	_	154.105	172.701
LSTM Estimated Time Series	0.0019	ARMA (2, 1)	1.6407 and -0.6508	-0.9072	135.861	159.105
Prophet Estimated Time Series	0.0	ARMA (2, 0)	0.7547 and 0.0465	_	183.595	202.191

It can be further observed form Table 2 that order of the ARMA model fitted for the mean estimated time series, median estimated time series and LSTM estimated time series are same as (2,1). Similarly, order of the models obtained for ARMA estimated time series and Prophet estimated time series are same as (2,0). Though all the six estimated time series have few similar descriptive properties, there are deviations among the suitability of the ARMA models fitted to them. The fitted models have varied AIC and BIC values. The AIC value of the ARMA model fitted to the LSTM estimated time series is 135.861, which is relatively smaller than the AIC values of the other five models. Similarly, the BIC value of the same model, 159.105 is relatively smaller than the BIC

DOI: 10.6092/issn.2281-4485/16184

values of the other five models. The AIC and BIC values of the model determined from the ARMA estimated time series are moderately larger than such values for LSTM estimated time series, but they are significantly smaller than the values of the remaining four ARMA models. It can be noted from these observations that application of all the six estimation methods yield the estimated time series with similar basic characteristics. However, when it is required to construct time series models for forecasting purposes, LSTM model-based procedure may be adopted for obtaining estimated time series. Thus, estimates of the missing values and outliers may be determined from the forecast values computed from the LSTM model which is fitted based on the observations preceding the missing values and outliers.

Further time series models are fitted to the six estimated time series applying the three methods *viz.,* ARMA method, LSTM method and Prophet method. Mean absolute error (MAEs) and root mean square error (RMSE) are computed for each model for assessing the forecast efficiency of the models, which are displayed in Figure 3.

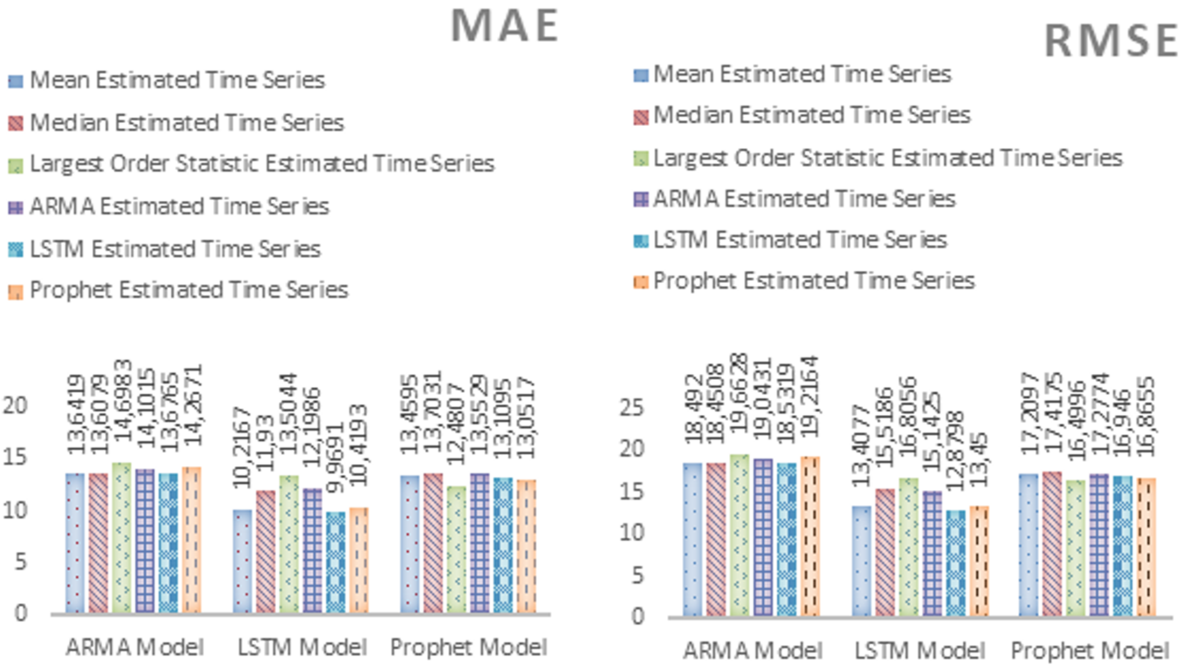

Figure 3. *MAE and RMSE of the Time Series Models of the Estimated Time Series*

Application of the three model fitting methods to the six estimated time series yield varied MAE and RMSE values. The ARMA method provides relatively better fit to the median estimated time series with the smallest MAE and RMSE respectively 13.6079 and 18.4508. The ARMA method provides poor fit to the largest order statistic estimated time series with larger MAE of 14.6983 and RMSE of 19.6628. Interestingly, the ARMA method provides a moderate fit to the ARMA estimated time series with MAE of 14.1015 and RMSE of 19.0431, which are significantly larger than the corresponding values for median estimated time series

The LSTM method provides relatively better fit to the LSTM estimated time series with the smallest MAE and RMSE respectively 9.9691 and 12.8798. The LSTM method provides poor fit to the largest order statistic estimated time series with larger MAE of 13.5044 and RMSE of 16.8056. Prophet model provides relatively better fit to the largest order statistic estimated time series with the smallest MAE and RMSE respectively 12.4807 and 16.4996. The Prophet method provides poor fit to the median estimated time series with larger MAE of 13.7031 and RMSE of 17.4175. Interestingly, the Prophet method provides moderate fit to the Prophet estimated time series with MAE of 13.0517 and RMSE of 16.8655, which are significantly larger than the corresponding values for the largest order statistic estimated time series.

It can be noted from these discussions that the median method of determining estimates to the *replaceable observations* in a time series may be followed, when ARMA model is fitted to the given time series.

DOI: 10.6092/issn.2281-4485/16184

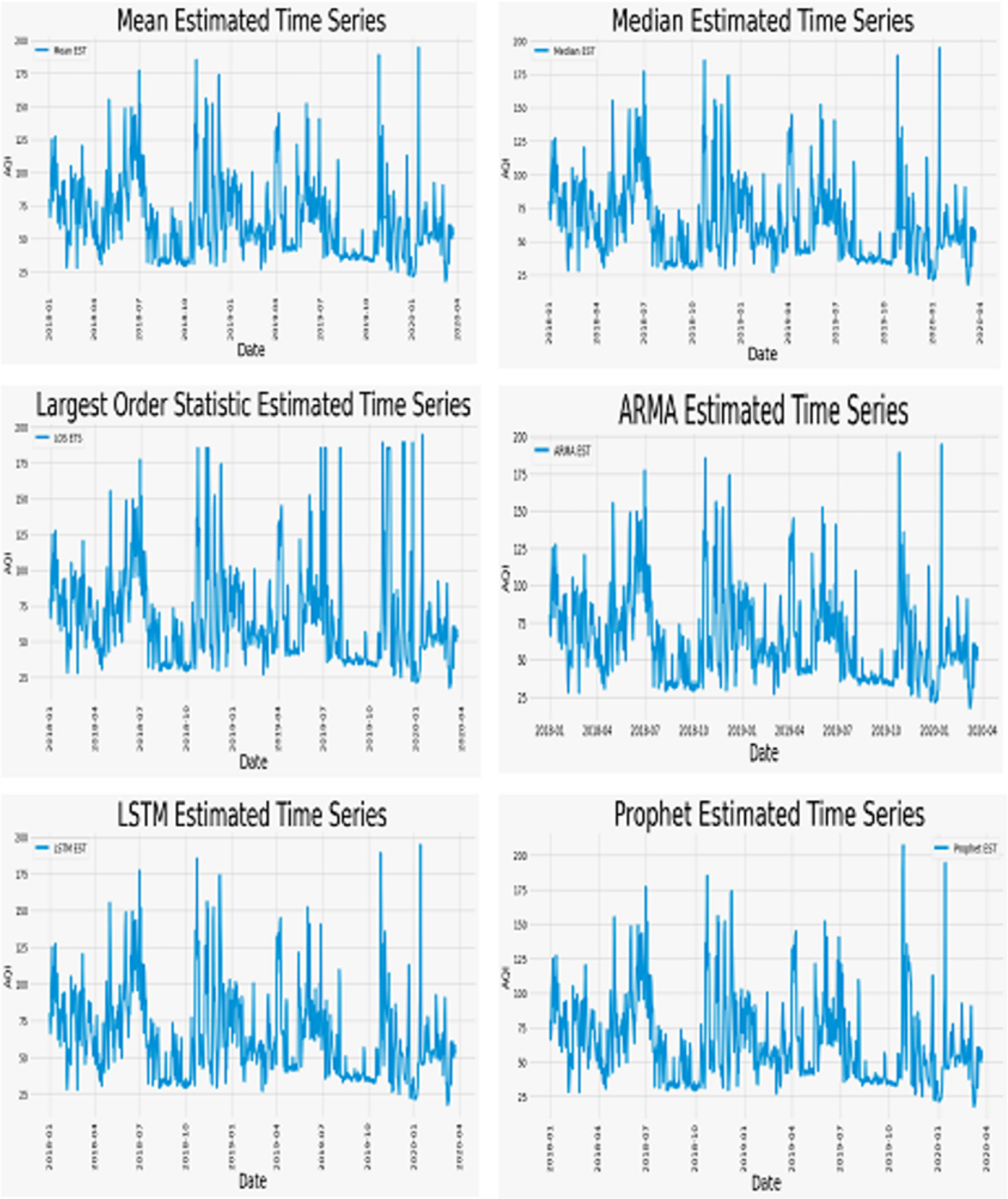

Figura 4. *Estimated Time Series of AQI*

DOI: 10.6092/issn.2281-4485/16184

The LSTM method of determining estimates may be more appropriate when LSTM method is considered for fitting time series model. The largest order statistic method of estimation may be suitable, when Prophet method is employed for determining the time series model.It can also be observed that difference between the MAEs of ARMA models fitted to median estimated time series, mean estimated time series and LSTM estimated time series are very moderate. Similarly, there is a marginal difference among the MAEs of the Prophet models fitted to the largest order statistic estimated time series, Prophet estimated time series and LSTM estimated time series. These observations also hold to the comparison of respective RMSEs. If magnitudes of such differences of MAEs and RMSEs are in admissible range, LSTM method may be considered for determining the estimates for the *replaceable observations* in a given time series. Thus, application of the LSTM method of finding estimates can provide estimated time series for which the MAE and RMSE of ARMA model, LSTM model and Prophet model will be relatively small.

Concluding Remarks

Presence of missing values and outliers in time series cause hindrances in conducting statistical analysis. Discordance of such observations will affect the analysis of time series, when successive observations in the time series are autocorrelated. This study considered six different methods of finding estimates for replacing the missing values and outliers *viz.,* mean, median, largest order statistic, forecast values from ARMA model, LSTM model and Prophet model. Application of these methods do not affect the stationary property and behavioural patterns of the time series. Performance of these estimation methods was discussed with the time series models constructed for the estimated time series obtained from the six methods. The time series models were constructed applying the conventional method (ARMA), a deep learning method (LSTM) and a machine learning method (Facebook Prophet).

Comparison of the MAEs and RMSEs of the time series models reveals that the smallest MAE and RMSE can be obtained, when the missing values and outliers in the time series are replaced by the estimates evaluated from the forecast values of the LSTM model for the preceding observations.

References

AHN H., SUN K., KIM K.P. (2022) Comparison of missing data imputation methods in time series forecasting. Computers. Materials and Continua, 70(1):767-779. https://doi.org/10.32604/cmc.2022.019369

AGBAILU A.O., SENO A., CLEMENT O.O. (2020) Kalman filter algorithm versus other methods of estimating missing values: time series evidence. African Journal of Mathematics and Statistics Studies, 4(2):1-9. https://doi.org/10.52589/AJMSS-VFVNMQLX

BOX G.E.P., JENKINS G.M., REINSEL G., LJUNG G.M. (2015). Time series analysis: forecasting and control. John Wiley & Sons. ISSN 978-1-118-67502-1

CINAR Y.G., MIRISAEE H., GOSWAMI P., GAUSSIER E., AIT-BACHIR A. (2018) Period-aware content attention RNNs for time series forecasting with missing values. Neurocomputing, 312:177-186. https://doi.org/10.1016/j.neucom.2018.05.090

CHANG Y.S., CHIAO H.T., ABIMANNAN S., HUANG Y.P., TSAI Y.T., LIN K.M. (2020) An LSTM-based aggregated model for air pollution forecasting. Atmospheric Pollution Research, 11(8):1451-1463. https://doi.org/10.1016/j.apr.2020.05.015

DENESHKUMAR V., KANNAN K.S. (2011) Outliers in time series data. Int. J. Agricult. Stat. Sci, 7(2):685-691. ISBN 0973-1903.

COUSINEAU D., CHARTIER S. (2010) Outliers detection and treatment: a review. International Journal of Psychological Research, 3(1):58-67. https://doi.org/10.21500/20112084.844

DEUTSCH S.J., RICHARDS J.E., SWAIN J.J. (1990) Effects of a single outlier on ARMA identification. Communications in Statistics-Theory and Methods, 19(6):2207-2227. https://doi.org/10.1080/03610929008830316

ENDERS C.K. (2010) Applied Missing Data Analysis, Guilford press. ISSN 978-1-60623-639-0

DOI: 10.6092/issn.2281-4485/16184

HUANG M.W., LIN W.C., TSAI C.F. (2018) Outlier removal in model-based missing value imputation for medical datasets. Journal of healthcare engineering, 2018:1-9. https://doi.org/10.1155/2018/1817479

JADHAV A., PRAMOD D., RAMANATHAN K. (2019) Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138

JANARTHANAN R., PARTHEEBAN P., SOMASUNDARAM K., NAVIN ELAMPARITHI P. (2021) A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society, 67:102720-102731. https://doi.org/10.1016/j.scs.2021.102720

KIHORO J.M., ATHIANY H., WALTER O.Y., W K H (2013) Imputation of incomplete non-stationary seasonal time series data. Mathematical Theory and Model, 3(12):142-154. ISBN 2225-0522

KOLBASI A., UNSAL A. (2019) A Comparison of the Outlier Detecting Methods: An Application on Turkish Foreign Trade Data. Journal of Mathematics and Statistical Science, 5:213-234. ISBN 2411-2518

LEDOLTER J. (1989) The effect of additive outliers on the forecasts from ARIMA models. International Journal of Forecasting, 5(2):231-240. https://doi.org/10.1016/0169-2070(89)90090-3

LIN W.C., TSAI C.F. (2020) Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53(2). 1487-1509. https://doi.org/10.1007/s10462-019-09709-4

LYU J., YIN S., SHANG C.C., MA Y., SUN N., SHEN G., LIU C. (2021) Sensitivity analysis of isoprene and aerosol emission in a suburban plantation using long short-term memory model. Urban Forestry and Urban Greening, 64:127303- 127310. https://doi.org/10.1016/j.ufug.2021.127303

MANI G., VOLETY R. (2021) A comparative analysis of LSTM and ARIMA for enhanced real-time air pollutant levels forecasting using sensor fusion with ground station data. Cogent Engineering, 8(1):1936886-1936912. https://doi.org/10.1080/23311916.2021.1936886

QIU J., WANG B., ZHOU C (2020) Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLoS ONE, 15(1):1-15. https://doi.org/10.1371/journal.pone.0227222

RUBIN D.B. (1976) Inference and missing data. Biometrika 63(3):581-592. https://doi.org/10.1093/biomet/63.3.581

SAVARIMUTHU N., KARESIDDAIAH S. (2021) An unsupervised neural network approach for imputation of missing values in univariate time series data. Concurrency and Computation: Practice and experience, 33(9):1–16. https://doi.org/10.1002/cpe.6156

SHEN J., VALAGOLAM D., McCALLA S. (2020) Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O_3, NO_2, SO_2, CO) in Seoul, South Korea. PeerJ, 8:1-18. https://doi.org/10.7717/peerj.9961

SONG X., LIU Y., XUE L., WANG J., ZHANG J., WANG J., JIANG L., CHENG Z. (2020) Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. Journal of Petroleum Science and Engineering, 186:106682-106700. https://doi.org/10.1016/j.petrol.2019.106682

TAYLOR S.J., LETHAM B. (2018) Forecasting at scale. American statistician, 72(1):37-45. https://doi.org/10.1080/00031305.2017.1380080

TOLVI J. (1998) Outliers in time series: A review. University of Turku. Department of Eco-nomics, Research reports, 76:1-30.

ZEILEIS A., GROTHENDIECK G. (2005) Zoo: S3 infrastructure for regular and irregular time series. Journal of Statistical Software, 14(6):1-27. https://doi.org/10.18637/jss.v014.i06