Treatment of Outliers via Interpolation Method with Neural Network Forecast Performances

To cite this article: N. A. Wahir et al 2018 J. Phys.: Conf. Ser. 995 012025

View the article online for updates and enhancements.
Treatment of Outliers via Interpolation Method with Neural Network Forecast Performances

N. A. Wahir, M. E. Nor, M. S. Rusiman and K. Gopal

Department of Mathematics and Statistics, Faculty of Science, Technology and Human Development, University of Tun Hussein Onn Malaysia, 86400 Johor Malaysia

Abstract. Outliers often lurk in many datasets, especially in real data. Such anomalous data can negatively affect statistical analyses, primarily normality, variance, and estimation aspects. Hence, handling the occurrences of outliers require special attention. Therefore, it is important to determine the suitable ways in treating outliers so as to ensure that the quality of the analyzed data is indeed high. As such, this paper discusses an alternative method to treat outliers via linear interpolation method. In fact, assuming outlier as a missing value in the dataset allows the application of the interpolation method to interpolate the outliers thus, enabling the comparison of data series using forecast accuracy before and after outlier treatment. With that, the monthly time series of Malaysian tourist arrivals from January 1998 until December 2015 had been used to interpolate the new series. The results indicated that the linear interpolation method, which was comprised of improved time series data, displayed better results, when compared to the original time series data in forecasting from both Box-Jenkins and neural network approaches.

1. Introduction

An outlier refers to a certain datum that differs from the others in a dataset, which does not indicate generalizability upon prediction mainly due to inaccurate data. Outliers can also reflect anomalous data from anomalous events that appear to be different from the rest of the data series [1]. With the presence of outliers, the related research findings may be negatively affected, especially during statistical analyses. In fact, accurate data is essential to obtain the accurate models for estimation and forecasting, which can lead to accurate data analysis and unbiased results. Moreover, many factors have been identified to cause the presence of outliers, for example, misinterpretation of data during surveys, malfunction of measurement tools, as well as values reported bizarrely within a time lag or within original samples. Although detecting an outlier may be easier than dealing with it, as for prediction purposes, it is a vital aspect that must be determined due to its negative impact upon the results. Moreover, although many studies have focused on outlier detection [2][3], many have failed to present a solution to handle outliers [4]. Nevertheless, three types of methods have been applied in studies to address issues related to outliers, namely, 1) removal of outlier, 2) replacement of outlier with another suitable value, and 3) the use of a robust method to eliminate outliers. With that, two possible approaches could be adopted when dealing with outliers, which are 1) automated data modification, and 2) manual intervention [5]. Furthermore, the model employed the weighing process that lessened the impact of outliers, such as regression, an example of automated data modification that stays robust with the existence of outlier. In fact, the robust method was applied to treat the outliers in predicting model parameters [6]. Nevertheless, some robust methods have performed well, while some others perform otherwise [7], mainly due to failure in gathering information pertaining to the individual outlier, which could lead to misleading estimation.

On the other hand, manual intervention is applied to replace the original data with forecasted ones in the case of outlier. Generally, many researchers have faced problems in handling outliers as a legitimate part of data [8]. Hence, the outliers are removed [9] so as to obtain the best estimate for
population parameters. Unfortunately, removal of outlier without any replacement may generate invalid and undesirable results [10]. Therefore, removal of outlier should be the last resort of an informed choice and not a routine task. Moreover, handling outliers is not an easy task because this process involves making predictions close to the parameters and offers some trade-offs between variance and bias. Meanwhile, time series data that contain outliers may affect the accuracy in forecasting [11], especially when biasness exists while estimating model parameter. Other than that, the carry-over effect of outliers, which occurs due to misidentified and undetected outliers, could cause inaccuracy in forecast. Thus, another common technique for treating outliers is to identify the locations and the types of outliers. Besides, detecting outliers with the proposed method of intervention model requires iterations between the detection and estimation stages. Although this method is effective during the detection stage, several issues have to be addressed, for instance, biasness in the parameter estimates and its declining competency during detection. Moreover, this particular approach could cause a shift in the types and locations of outliers at varied model estimation iterations, where the worst case is that some outliers cannot be detected or identified. As such, two phases are involved in treating outliers. The first stage refers to detecting outliers using suitable methods, while the second stage is treating the detected outliers. For instance, upon assuming the outliers in the ARMA model as missing data, new values that function as substitutes are definitely needed [13] in [12]. As such, the interpolation method was employed due to its suitability.

Therefore, this paper adhered to the method employed in the business Tankan surveys [14], where the outliers were regarded as missing values and treated using the linear interpolation method to fill the missing values. The outliers were detected by using the fit ARIMA distribution method, along with the SAS software program. Furthermore, this method has been widely used to treat outliers since 2004 using datasets derived from five principal survey items of Tankan, namely, profit, sales, net profit, software investment, and fixed investment. As such, this paper compared two approaches, which are Box-Jenkins and neural network approaches, using the similar linear interpolation method in terms of forecast accuracy. The improved series refers to the new series generated after the linear interpolation method was applied as outlier treatment. Next, all the improved series were compared with the original Box-Jenkins time series data and the original neural network data. In addition, several types of forecast accuracy methods were used to evaluate both approaches in describing goodness fit, namely, mean square error (MSE), mean absolute percentage error (MAPE), and mean absolute deviation (MAD). Lastly, the detected outliers in the data series are discussed, inclusive of the comparison made for before and after treatment of outlier.

2. Materials and Methodology

As depicted earlier, the monthly time series data for Malaysian tourist arrivals derived from 1998 until 2015 and retrieved from the official Ministry of Malaysia Tourism website had been employed in this study. [14] also used the same data in forecasting by using several time series approaches. For example, the economic survey carried out in the Bank of Japan regarded the missing values as outliers in the data series [15], in which the fit ARIMA distribution method was employed to detect outliers. Hence, this paper focused on outlier treatment, rather than detection, where the detected outliers were replaced by using the linear interpolation method. After that, the original model of Box-Jenkins time series data and the original neural network model were evaluated. Other than that, data derived from both models were compared with the newly improved series of the same approach using forecast accuracy to identify the changes before and after the treatment. Furthermore, all the statistical analyses were analyzed by using SAS, Minitab, S-PLUS, Microsoft Excel, and SRS1Spline software programs. Besides, this research applied the interpolation method to generate new points for the dataset. Moreover, the interpolation method predicted the values of function \( y(x) \) in any values between \( x_0, \ldots, x_n \) with values of \( y_0, \ldots, y_n \) [16]. Simply put, the interpolation method refers to the process of predicting missing values at \((x, y)\) locations based on their nearest points. With that, several steps were taken in treating the outliers, as listed in the following:
Step 1: Identify the outliers in the original time series data with both Box-Jenkins and neural network.
Step 2: Remove the outliers from the dataset based on the position of those detected.
Step 3: Insert the detected outliers in the dataset using interpolation method.
Step 4: Identify the model present for each improved series.
Step 5: Evaluate the value of forecast accuracy for each improved series.
Step 6: Repeat steps 1-5 until the forecast accuracy perfectly fits the conditions.

2.1. Linear Interpolation
Linear interpolation denotes the simplest form of interpolation that joins two data points with a straight line. Linear interpolation is also a Newton form of polynomial first order. Hence, the missing values can be directly predicted using the linear interpolation equation, as given below:

\[(x) = b_0 + b_1(x - x_0)\]  

Where,
\[x = \text{The independent variable (the time of missing observation)}\]
\[x_0 = \text{A known value of the independent variable (the time point of missing observation)}\]
\[f_1(x) = \text{The value of dependent variable for a value } x \text{ of the independent variable (the missing observation)}\]

Next, from equation (1),

\[b_0 = f(x_0)\] and \[b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}\]

Where,
\[f(x_0) \text{ and } x_0 = \text{coordinates of the initial points for the gap}\]
\[f(x_1) = \text{ending point of the gap}\]

2.2. Box-Jenkins Approach
The Box-Jenkins approach refers to a traditional and well known approach for time series analysis and forecasting. This approach assumes that the data are in a stationary state and suggests variances to achieve stationary. The general Autoregressive Integrated Moving Average Model or ARIMA \((p, d, q)\) is a statistical model employed to analyze data. Besides, the general equation used for the Box-Jenkins approach is given in the following [13] and [17]:

\[\Phi_p(B)\Phi_p(B^s)^{P^d}\theta_q(B)\theta_q(B^s)\delta_t = \theta_q(B)\theta_q(B^s)a_t\]

Where:
\[\Phi_p(B) = \text{order of AR operators}\]
\[s = \text{seasonal length (s=12 for monthly data)}\]
\[\theta_q(B) = \text{order of MA operators}\]
\[B = \text{black-shift operators}\]
\[\delta_t = \text{white noise with normal distribution } N(0, \sigma^2)\]
\[\gamma_t = \text{constant}\]
\[\gamma_t = \text{time series data}\]
2.3. **Neural Network Approach**

Neural network refers to the processing system of patterned software and modeled in smaller scales from the mammalian cerebral cortex. Generally, this approach focuses on solving problems related to text transcription, exploration of data analysis, character recognition, image compression, and prediction in time series. Furthermore, this approach is a non-trivial matter for mathematicians for this approach possesses the ability to produce more accurate and faster results. This approach also used by [18] in time series forecasting.

2.4. **Forecast Accuracy**

In order to measure the forecast accuracy between Box-Jenkins method and Neural Network model, mean absolute deviation (MAD), mean square error (MSE), and mean absolute percentage error (MAPE) were calculated and compared using equations 5, 6, and 7.

\[
\text{MAD} = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t| \quad (5)
\]

\[
\text{MSE} = \frac{\sum (y_t - \hat{y}_t)^2}{n} \quad (6)
\]

\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t} \times 100 \quad (7)
\]

3. **Results and Discussion**

Two approaches were tested in this study to interpolate the outliers embedded within the time series, namely, Box-Jenkins and neural network approaches. The linear interpolation process connects two data points to generate a straight line in the attempt to predict outliers that lurk in the data, which may produce a better fit of smooth graph and data, compared to the original data series. Figure 1 illustrates the time series plot for the original Malaysia tourist arrivals time series data in year 2015, the original Box-Jenkins approach, the original neural network approach, the improved Box-Jenkins, and the improved neural network data series. The original data that reflect tourists exhibited a fluctuating trend for the whole year, while at the end of November 2015, the data suddenly rose to the highest limits, which was similar for both original Box-Jenkins and original neural network data. After applying the interpolation method, the month of December indicated normal observation, similar to the rest of the months. Nevertheless, the original data derived from February and June 2015 depicted the lowest number of tourists, followed by data from the original Box-Jenkins and original neural network data. However, the number of tourists exhibited a more stationary pattern when the linear interpolation was applied to improve the series of both Box-Jenkins and neural network approaches. Thus, all the improved series exemplified better performances after the outlier treatment was applied, in comparison to those before treatment.
Table 1 presents the forecast accuracy between before and after the outlier treatment was carried out, which had been between original data of Box-Jenkins model, original neural network approach, and improved time series data of Box-Jenkins approach. Hence, the mean absolute deviation (MAD), the mean square error (MSE), and the mean absolute percentage error (MAPE) were calculated and compared by using equations (5), (6), and (7) of the original data, inclusive of the improved first and second iteration time series data. The results obtained from the MSE for original data between Box-Jenkins, as well as improved data of linear interpolation for both first and second iterations revealed huge variations. The difference between linear interpolation of the first iteration and the original data of Box-Jenkins was approximately 80000, while the comparison of the original data between Box-Jenkins and improved linear interpolation of the second iteration demonstrated a massive variance, which reached 259 388. In fact, the similar was noted for MAD and MAPE findings. The MAD value for the original data had been 175, displaying only three differences during the first iteration. Nonetheless, as for the linear interpolation of second iteration, the MAD exhibited a trivial value, when compared to the original value of MAD. As for MAPE value, a vast variance was observed between the original data and the improved linear interpolation of first iteration. Upon comparison, the improved linear interpolation of second iteration and the original data of Box-Jenkins displayed approximately 6% of difference. Next, as for the improved linear interpolation method between first and second iterations, the second iteration generated the best result, as compared to the first iteration.

Moreover, as presented in Table I, the neural network method showed better results among the other values. First, more than 20,000 differences were recorded for MSE results for the first iteration of linear interpolation. Besides, less than a quarter of MSE values differed between the original neural network and the improved series during the second iteration. Next, the MAD values of the original neural network data indicated vast differences during the second iteration of improved series, in comparison to the first iteration that revealed more than 30,000 variances. As for MAPE results, the first iteration of the improved series showed a huge gap between the first and the second iterations, when compared to the Box-Jenkins approach. Therefore, the improved series of neural network approach of the second iteration data offered the most satisfying results for forecast accuracy. Overall, the result of this forecast accuracy indicated a positive impact after outlier treatment using the linear interpolation
method in predicting outliers. In addition, the improved series of Box-Jenkins and neural network approaches after outlier treatment provided better results in MSE, MAD, and MAPE values, in comparison to the original data of both approaches. As such, the improved series from the second iteration of neural network had been proven to offer the best results for each forecast accuracy, and this is followed by the second iteration of Box-Jenkins approach. Therefore, better results of forecast accuracy could be acquired by applying more iteration.

Table 1. Forecast Accuracy between Original Box-Jenkins Approach, Original Neural Network Approach, Improved Box-Jenkins and Improved Neural Network for Time Series Data.

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Before Outlier Treatment</th>
<th>After Outlier Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original Box-Jenkins Approach</td>
<td>Original Neural Network Approach</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>259.693</td>
<td>32.516</td>
</tr>
<tr>
<td>MAD</td>
<td>175</td>
<td>174</td>
</tr>
<tr>
<td>MAPE</td>
<td>8.45%</td>
<td>7.94%</td>
</tr>
</tbody>
</table>

4. Conclusion
In conclusion, the presence of one or more outliers in time series data possesses the potential to cause biasness in estimation model, hence greatly affecting the estimate variance. Thus, removing any aberrant observation and applying data that contain outliers may affect variance and generate inaccurate results. Therefore, the outliers in this study had been regarded as missing values and treated by using the linear interpolation method to treat outliers. Besides, the original data series of two approaches from the Box-Jenkins and the neural network had been examined to interpolate the outliers embedded in the Malaysia tourist arrival data. After that, both approaches from the original series were evaluated and compared with the newly improved series to determine the variances before and after the outlier treatment. Furthermore, the findings obtained from all the improved series of Box-Jenkins and neural network approaches displayed smaller values for MAPE, MSE, and MAD, as compared to the original series. As a conclusion, the forecast accuracy offered better results after carrying out the outlier treatment using the linear interpolation method, in comparison to that before the treatment.

Acknowledgments
The authors would like to gratefully thank the Malaysian Tourism Development Cooperation (MTDC) for providing tourist arrivals data and University Tun Hussein Onn Malaysia (UTHM) for both their technical and financial support. This research was carried out under the UTHM Short Term Grant. Hence, the authors acknowledge and welcome any constructive comments and suggestions provided by the respected reviewers.

References
[1] Battaglias F 2006 On outlier detection in multivariate time series An Unpublished Paper Presented at a Seminar in the University of Liverpool, United Kingdom


