International Journal of Engineering & Technology, 7 (3.7) (2018) 51-53

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Treatment of Outlier Using Interpolation Method in Malaysia Tourist Arrivals

Maria Elena Nor[1*], Norsoraya Azurin Wahir [1], G. P. Khuneswari[1], Mohd Saifullah Rusiman[1]

[1]Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia,
Pagoh, Malaysia
**Corresponding author E-mail:mariaelena@uthm.edu.my*

Abstract

The presence of outliers is an example of aberrant data that can have huge negative influence on statistical method under the assumption of normality and it affects the estimation. This paper introduces an alternative method as outlier treatment in time series which is interpolation. It compares two interpolation methods using performance indicator. Assuming outlier as a missing value in the data allows the application of the interpolation method to interpolate the missing value, thus comparing the result using the forecast accuracy. The monthly time series data from January 1998 until December 2015 of Malaysia Tourist Arrivals were used to deal with outliers. The results found that the cubic spline interpolation method gave the best result than the linear interpolation and the improved time series data indicated better performance in forecasting rather than the original time series data of Box-Jenkins model.

Keywords: *Outliers, Box-Jenkins, Linear Interpolation, Cubic Spline Interpolation*

1. Introduction

Outliers is an extreme value which always appears in the time series data and the value is recorded differently from the rest of the data for being either too small or too large. Outliers can be observed as a figure that acts differently from other in the data [3]. The presence of outlier may lead to a poor data analysis, resulting in inaccurate estimation. Research findings also be badly affected especially in a statistical analysis due to the existence of outlier. There are several causes outlier, for example error in data transmission, abnormal spike in time series data, unusual high number of component failure within a given time series, and periodic malfunction of measurement device. Handling or dealing with outlier is a difficult process but in forecasting, it is a critical function because the data used had been passed on, hence affecting the estimation in the results. This had been proven by [1] when the research finding found that the outliers affect various measures including the impacts on the data set, estimation method and skewness coefficient but it did not affect the headcount index. Generally, [8] found that many researchers face problems in handling outliers as a legitimate part of data. [6] stated that the outliers are then removed in order to get the best estimate of population parameters. Unfortunately [7] claimed, there is an argument stating that the removal of outlier without replacing any data may produce invalid and undesirable results. [2] stated the time series data that contain outliers may cause losses in the forecast accuracy and this happens when bias exists during the estimation of model parameter. A common technique in the treatment of outlier is to identify the locations and types of outliers. First, outlier must be detected through suitable methods and the detected outlier needs to be treated. Then, [5] found that [9] considered outlier as a missing data and assumed ARMA model as contaminated series and to be replaced with a new value from the missing data using interpolation method. In this paper, the outlier was detected using fit ARIMA distribution method with SAS software. [4] in their business Tankan surveys, the outliers were regarded as missing values and were treated using different methods when filling the missing values which was cold deck imputation method. Meanwhile, this paper compared two single imputation methods, linear interpolation and cubic spline interpolation, as an outlier treatment in time series data in terms of forecast accuracy. There were also three types of performance indicators being used to compare both methods in describing the goodness fit, namely mean absolute error (MAE), root mean square (RMSE), and coefficient of determination (R^2). Then, the outliers in the data were evaluated and compared before and after the treatment had been applied.

2. Materials and methods

In this study, the Malaysia tourist arrival monthly data were used starting from 1998 until 2015. The data were retrieved from the Ministry of Malaysia Tourism website. The improved time series data is a new set of data which went through the interpolation method. The outliers were then detected using fit ARIMA distribution and got removed. Upon replacing the detected outliers using interpolation method, the actual model of Box-Jenkins time series data and improved time series data were evaluated. Next, both data were compared using forecast accuracy to find the changes before and after interpolation were applied. Subsequently, comparing the goodness of fit and the efficiency between linear interpolation method and cubic spline interpolation method required an evaluation using performance indicators.

All of the statistical analyses were performed using SAS software, S-PLUS software, Minitab software, Microsoft Excel and SRS1Spline software. In this research, interpolation method was used to obtain the new point that was present in the data set. Thus, in order to understand the process of this study, several steps of the outlier treatments were applied to achieve a newly improved time series data:

Step 1: Identify the outliers in the actual time series data using the Box-Jenkins methodology.

Step 2: Remove the outliers in the data depending on the position of detected outlier.

Step 3: Replace the detected outliers in the data using linear interpolation and cubic spline interpolation method.

Step 4: Identify the model present in the improved time series data.

Step 5: Evaluate the value of forecast accuracy and performance indicator of the improved time series data.

Step 6:Repeat step 1-5 until the magnitude error perfectly fits the conditions.

3. Results and discussion

Table 1 indicates the magnitude error between actual data of Box-Jenkins model and improved time series data of Box-Jenkins model. The MSE value of actual data between Box-Jenkins, as well improved data of linear interpolation and cubic spline interpolation showed huge differences. In comparison, the improved linear interpolation of second iteration and actual data of Box-Jenkins, there was approximately 3% difference. Next, in the improved linear interpolation method between first and second iterations, the second iteration had been proven to generate the best result compared to the first iteration. As stated in Table 1, the cubic spline interpolation method indicates better results among other errors of magnitude values. First, there were more than 80,000 differences in the MSE values for the first iteration of cubic spline interpolation. Also, less than quarter of MSE value differences were between the actual Box-Jenkins data and improved Box-Jenkins during the second iteration of cubic spline interpolation method. Next, the MAD value of actual Box-Jenkins data indicated a huge difference during the second iteration of cubic spline, in comparison to the first iteration. In testing the MAPE, the first iteration of the improved cubic spline data showed a medium difference from the actual data compared to the second iteration of the improved data that showed 5.2% from 8.4% of value. Thus, the improved cubic spline interpolation of second iteration data the most satisfying result in error magnitude to that of the first iteration. Overall, the result of the error magnitude indicated that there was positive impact during the interpolation method in estimating the missing value which was regarded as outlier in this study. The improved Box-Jenkins data of cubic spline interpolation and the linear interpolation method demonstrated the better results in MSE, MAD, and MAPE values than the actual data of Box-Jenkins. From all of the aforementioned statements, the improved Box-Jenkins of second iteration methods were all proven to acquire the best result each error magnitude. Futhermore, better results of error magnitude could be obtained by applying more iteration.

Table 1: Forecast Accuracy Between Actual Box Jenkins Model and Improved Box-Jenkins for New Time Series

Measurements	Actual Box-Jenkins	Improved Time Series Data			
		Linear Interpolation		Cubic Spline Interpolation	
		1st Itera-tion	2nd Itera-tion	1st Itera-tion	2nd Iteration
MSE	259 693	178 389	305	173 415	107
MAD	175	172	119	120	112
MAPE	8.4 %	8.1%	5.5%	7.6%	5.2%

60 g/cm^3 density
70 g/cm^3 density
80 g/cm^3 density

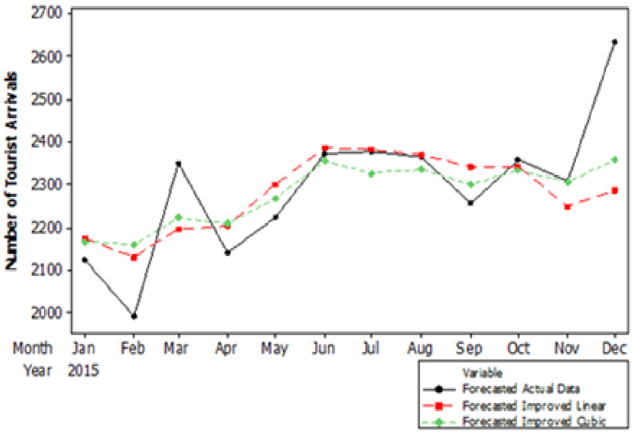

Figure 1: The Time Series Plot Between Forecasted Improved

Table 2 shows the performance indicators between Linear Interpolation and Cubic Spline Interpolation methods using RMSE, R2 and MAE.

Table 2: Performance Indicators between the Linear Interpolation and the Cubic Spline Interpolation

Measurements	Types of Interpolation			
	Linear Interpolation		Cubic Spline Interpolation	
	1st Iteration	2nd Iteration	1st Iteration	2nd Iteration
RMSE	422.36	17.45	416.43	10.36
R^2	0.92	0.93	0.92	0.93
MAE	172.45	119.12	162.56	112.21

From the previous result, the forecast accuracies for all values of MSE, MAD, and MAPE of linear interpolation and cubic spline interpolation showed better results compared to the values of the actual data of Box-Jenkins. In order to know the best interpolation method for this time series data, several tests were ran to find the best method. Thus, the performance indicators were used on the linear interpolation method and cubic spline interpolation method based on the improved tourist arrival data of the Box-Jenkins model. Table 2 indicates the results' performance for both methods. Both methods fit the data very well and cubic spline interpolation method obviously generated the best results in comparison to linear interpolation method. There were several digit differences in the RMSE and MAE values. Although their R^2 tests had the same value during the first iteration and second interpolation of linear interpolation and cubic spline interpolation which was 0.92, the cubic spline interpolation method gave smaller error value in RMSE compared to the linear interpolation method which was approximately to six points in the first iteration. Then, in the second iteration between linear interpolation and cubic spline interpolation, there were 7.09 differences in point. For MAE value, the cubic spline interpolation indicated a much lower value in the second iteration than the first iteration. This was the same result for the linear interpolation which had a lower value during the second iteration. The cubic spline interpolation showed a much lower value in MAE compared to the linear interpolation method. Obviously, all values for the performance indicators showed better results for cubic spline interpolation, making the cubic spline interpolation as the best fit method to interpolate the missing values in this time series data rather than in the linear interpolation method.

4. Conclusion

Time series data usually contained many unexpected outliers. Discarding the outlier observations during the analyzing and forecasting may produce bias results. Perceiving the outliers as missing values and subsequently make treatment of outlier using interpolation method, the missing value can therefore be replaced using this method. In this study, two types of interpolation methods were used to interpolate the missing value in the tourist arrival time series data and later both methods were evaluated using three performance indicators, namely MAE, RMSE, and R^2 to obtain the best method. The first and second iterations of improved data using both methods were applied in order to indicate the performance indicators and the best result to discard the outliers. Afterwards, the actual time series data of the Box-Jenkins model were compared with improved linear interpolation data and improved cubic spline interpolation data using forecast accuracy. The results indicated that the improved time series data of the Box-Jenkins appeared to have smaller error in the forecast accuracy which were MAPE, MSE, and MAD than the actual time series data. The observation upon the performance indicators between two interpolation methods concluded that both methods fit the data very well. The best method in terms of the degree of complexities was the cubic spline interpolation method not the linear interpolation method.

Acknowledgement

The authors would like to express our gratitude to Malaysia Tourism Development Cooperation (MTDC) for providing tourist arrivals data and University Tun Hussein Onn Malaysia (UTHM) for providing facilities and financial supports. This research was carried out under UTHM Short Term Grant. Hence, the authors acknowledge and welcome the constructive comments and suggestions provided by reviewers.

References

[1] Álvarez, E., et al., The Effect of Outliers on the Economic and Social survey on Income and Living Conditions. World Academy of Engineering and Technology, International Journal of Social, Behavioral, Education, Economic, Business and Industrial, Engineering, 2014. 8(10). p. 3268- 3272.

[2] Chen, C., & Liu, L. M. Forecasting Time Series With Outliers. Journal of Forecasting, 1993. 12(1). p. 13-35.

[3] Grubbs, F. E. Procedures for detecting outlying observations in samples. Technometrics, 1969.11(1) p.1-21.

[4] Ishikawa, A., Endo, S., et al., Treatment of Outliers in Business Surveys: The Case of Short-term Economic Survey of Enterprises in Japan (Tankan) Bank of Japan, 2010. (No. 10-E-8).

[5] Ismail, S. O. Accommodation of Outliers In Time Series Data: a Case Study. Asian Journal of Mathematics and Statistics, 2008. 1(1). p. 24-33.

[6] Judd, C. M., & McClelland, G. H. Data analysis: A Model Comparison Approach. San Diego, CA.: Harcourt Brace Jovanovich, 1989.

[7] Orr, J. M., Sackett, et al., Outlier Detection Treatment in I/O Psychology: A Survey Of Researcher Beliefs and an Empirical Illustration. Personnel Psychology, 1991.44. p.473-486.

[8] Pigott, T. D., A Review of Methods for Missing Data. Educational Research And Evaluation, 2001,7(4): 353-383.

[9] Xie, Z., Case study IX in Time Series Analysis: Miscellaneous Cases Study: World Scientific Publishing. Plc. London. UK, 1993, 250-271.