Sains Malaysiana 49(5)(2020): 1165-1174
http://dx.doi.org/10.17576/jsm-2020-4905-22

Imputation Techniques for Incomplete Load Data Based on Seasonality and Orientation of the Missing Values

(Teknik Pengimputan untuk Data Beban tak Lengkap Berdasarkan Kemusiman dan Orientasi Nilai yang Hilang)

NUR ARINA BAZILAH KAMISAN*, MUHAMMAD HISYAM LEE, ABDUL GHAPOR HUSSIN & YONG ZULINA ZUBAIRI

ABSTRACT

In load data, the missing problem always occurs in a set of data. Since it has a seasonal pattern according to days, most of the time, the load usage for the next day is predictable. For this reason, a new model has been developed based on these characteristics. Data containing missing values being divided to its seasonality pattern and for each subdivision, the values from mean, the mean with standard deviation and third quartile are calculated before being rearrange to form a new set of values that will replace the missing values. These three values will be used as imputations for the missing values. To examine the effects of the orientation of the missing values with the choices of imputation, the missing values from the data are divided into three parts: at the front, in the middle and at the end of the data with 5%, 15%, and 25% of missing values. The results from root mean square error and mean absolute error show that the proposed techniques, particularly the mean and the third quartile value, are superior to the other complex methods when dealing with the missing values. The mean imputation is ample when the missing values is presence at the front and in the middle of the data while the third quartile value is superior when the missing values is at the end of the data.

Keywords: Data orientation; missing values; multiple imputation; seasonal load data; seasonality

ABSTRAK

Dalam data beban, masalah kehilangan data selalu berlaku dalam satu set data. Memandangkan ia mempunyai corak bermusim mengikut hari, kebanyakan masa, penggunaan beban untuk hari berikutnya boleh diramal. Atas sebab ini, satu model baru telah dibangunkan berdasarkan ciri-ciri ini. Data yang mengandungi nilai yang hilang yang dibahagikan kepada bentuk pola bermusimnya dan bagi setiap subdata, nilai min, min bersama hasil tambah sisihan piawai dan kuartil ketiga dihitung sebelum disusun semula untuk membentuk satu set nilai baru yang akan menggantikan nilai data yang hilang. Ketiga-tiga nilai ini akan digunakan sebagai pengimputan untuk nilai yang hilang. Untuk mengkaji kesan kedudukan nilai-nilai yang hilang dengan pilihan pengimputan, nilai-nilai yang hilang daripada data dibahagikan kepada tiga bahagian iaitu: di bahagian depan data, di tengah data dan di akhir data dengan 5%, 15% dan 25% nilai yang hilang. Keputusan daripada ralat min punca kuasa dan ralat min mutlak menunjukkan bahawa teknik yang dicadangkan, terutamanya pengimputan nilai min dan kuartil ketiga, memberikan hasil yang lebih bagus daripada kaedah kompleks lain ketika berurusan dengan nilai yang hilang. Pengimputan min adalah bagus apabila nilai-nilai yang hilang berada di hadapan dan di tengah data manakala nilai kuartil ketiga lebih bagus apabila nilai-nilai yang hilang berada pada bahagian akhir data.

Kata kunci: Data beban bermusim; data orientasi; kepelbagaian pengimputan; nilai yang hilang; kemusiman

INTRODUCTION

As mentioned by Winkler and McCarthy (2005), missing data are a very important and serious problem. The observations with missing values are important to show the new outcomes and indicate the absolute fit of a model. Thus, improvement is needed if there is any (Cumming et al. 2007). Missing values may occur due to lack of records, item non-response, machine failure to record observation during an experiment, lost records, and other issues (Kihoro & Athiany 2013). Addressing the issue of missing values is crucial in the process of getting precise and accurate results. As mentioned by Penn (2007), many studies in the literature suggested that how researchers deal with the missing data can influence model estimates and standard errors. The results could lead to biased estimates if the missing data are not treated appropriately. In some instances, the data cannot be analyzed either at the record level or for the overall database. Thus, it is vital to handle missing values properly in all types of analysis (Winkler

& McCarthy 2005). Missing observations in time series data are very common since the data are recorded through time. When one or more observations are missing, it is essential to estimate the missing values to gain a better understanding of the nature of the data. As mentioned by Shukur and Lee (2015), imputing the missing values using an effective method is crucial before performing an analysis.

There a few methods that can be used to deal with the missing values. For instance, there are simple methods, such as list-wise deletion, pairwise deletion, and mean or mode substitution. A deletion method may be considered since it does not give effects to the analysis. Nevertheless, deletion could result in data loss, which will subsequently lead to a biased outcome and affects the skewness of the distribution (Cokluk & Kayri 2011; Honaker & King 2010). Thus, imputation could help protect the sample size. On the other hand, for mean or mode substitution, the mean value of the remaining observed values is calculated to replace the missing values. This process is considered to be appropriate if the researcher does not have other information (Cokluk & Kayri 2011). It is simple and quick to implement. Nevertheless, the issue of mean substitution is that the value could be unrealistic and even impossible if the value imputed comes from the known value of that particular field (Acock 2005). Thus, this method is not appropriate if the missing value is from marketing databases and surveys. This does not lead the analysis to the desired outcomes due to the poor quality of data (Winkler & McCarthy 2005). Furthermore, there are also other imputation methods, such as single imputation, hot-deck imputation, regression imputation, multiple imputations based on information of the maximum likelihood estimation, and the expectation-maximization (EM) algorithm (Acock 2005). Nevertheless, time series cross-section data often work poorly with these types of imputation methods (Honaker & King 2010). In the real world, to handle the missing values problem, different data require different strategies. Thus, it is necessary to utilize these strategies effectively to obtain the best possible estimates.

In the past, the approach to estimate the missing values for linear time series has involved the use of curve fitting. The details of these approaches can be discovered in many books (Brockwell & Davis 2013; Chatfield 2000; Gerald & Wheatley 2004; Hamilton 1994; Harvey 1990; Janacek & Swift 1993). Subsequently, the advanced models, such as Box-Jenkins model which incorporate space modelling and neural networks as applied to missing values (Damsleth 1980; Kihoro & Athiany 2013). The Box-Jenkins model has been used widely as a technique for dealing with missing values. The earliest study on missing values by using Autoregressive Integrated Moving Average (ARIMA) model written by Damsleth (1980). Damsleth combined forecast and back forecast of ARIMA model to handle the missing values in time series. Today, Box-Jenkins model is considered as a benchmark model for comparison in time series missing values (Ferreiro 1987; Gómez et al. 1992; Kihoro & Athiany 2013; Ruiz & Nieto 2000).

Hybrid models have also become famous recently as a method of imputation for missing values. To overcome the missing values problem, Sorjamaa and Lendasse (2007) combined a nonlinear model named the Self-Organized Maps (SOM) with the linear model Empirical Orthogonal Function (EOF). Furthermore, Honaker and King (2010) proposed multiple imputations to resolve the issue of missing values in time series cross-section data. Kihoro and Athiany (2013) rearranged their data according to their seasonal pattern and the missing values by using simple linear regression. A combination of Autoregressive (AR) and Artificial Neural Network (ANN) models have been used to impute the missing values by studying the pattern and stationarity of the wind speed data first (Shukur & Lee 2015). Zhang et al. (2017) combined three methods which are SOM clustering, the Fruit Fly Optimization Algorithm (FOA) and the Least Squares Support Vector Machine (LSSVM) to impute the missing values in their spatiotemporal data.

Materials and Methods

Load data is considered as a time series data as it is recorded through time. Because it was recorded through time, load data has a cyclic pattern that makes a simple method such as mean substitution, linear interpolation and many more to provide a bad replacement to the missing values. It is relatively complicated to use an advanced model such as seasonal autoregressive integrated moving average (SARIMA) interpolation. By eliminating the cyclic pattern in the data, a simpler interpolation could be used as it converts the seasonal data to a linear data. Moreover, by arranging the data which include the missing value in its seasonal pattern, this could assist in eliminating the cyclic effects.

The load usage used in this study is recorded from Pusat Bandar Johor Bahru (PBJB). The load usage in PBJB is recorded for a year for every hour in 2010. The data are divided into three sections. The missing values are selected from these three sections. The first section is where the missing values will be taken from the front part of the data (P1). By referring to, the data recorded from 1st January till 30th April is classified as P1. The middle section is where we choose the missing values from the middle part of the data (P2). In this case, the P2 will be the part where the data is lie between 1st May and 31st August. The last section will be the missing values selected from the end of the data (P3) which is from 1st September until 31st December. As can be seen from Figure 1, each section is divided by the red dotted line.

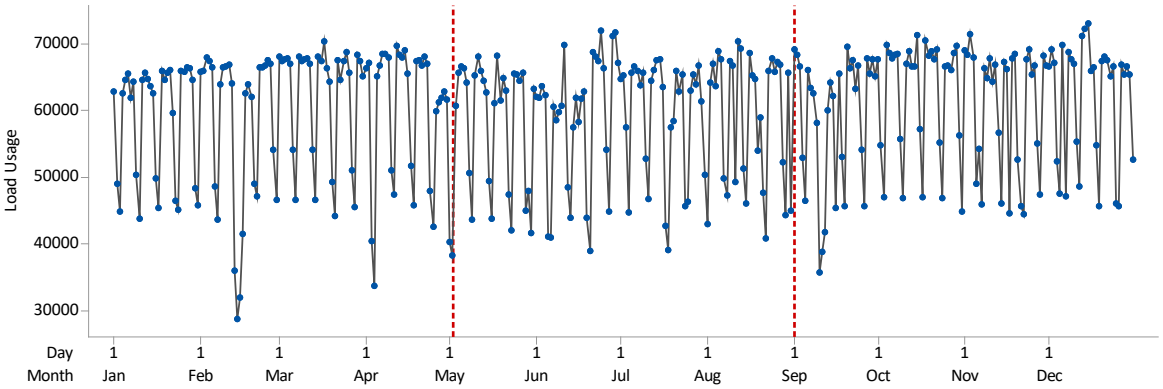

FIGURE 1. The plot on how the data are divided into three categories to observe the effect of the imputation

From the plot in Figure 1, the cyclic pattern of the data is noticeable where it contains 'daily cycle' pattern. To understand the data further a statistical value of the data is generated. From Table 1, the minimum and maximum value in the data is 28751 kW and 73126 kW. Since the minimum and maximum value is quite distant, the standard deviation, gives quite large value which is 9671.63. The mean is at 59872 kW and most of the load values are closer to the maximum value which we can relate with plot in Figure 3 where most of the days the load is above 60000 kW.

TABLE 1. Statistical value for load data PBJB

	Minimum	Maximum	Mean, $\bar{x}$	Standard Deviation, σ	Median, $\hat{x}$	Q1	Q3
Value	28751	73126	59872	9671.63	64683	50841.5	67093.5

Experts do not have any particular agreement on the percentage of the missing values that are suitable (Schlomer et al. 2010). Some suggested a 5% cut-off (Schafer 1999). On the other hand, some suggested 10% as the cut-off (Bennett 2001), whereas the other suggested 20% as the cut-off (Peng et al. 2006). Nevertheless, Schlomer et al. (2010) mentioned that there are two considerations while determining the amount of 'missing values' i.e. *whether the result from the data set is sufficient enough to detect the effect of consequence* and *the pattern of the missing values*. After considering two considerations mentioned by Schlomer et al. (2010), the percentage of missing data being considered is at 5%, 15%, and 25%, respectively. This study considered a year of recorded data and thus 25% missing values is considered sufficient for the highest missing values. By selecting these percentages, it is sufficient enough to observe the effects of the missing values. Since the pattern of the missing values in this study is continuous, these three percentages are suitable for the amount of missing values.

The disaggregation process is an important step considered in this study. A few important methods in time series, such as multiple linear regression, SARIMA and ANN models are being widely used as its consideration of the seasonality contains in the data. This study also put the seasonality into consideration by performing a disaggregation process. In this case, the seasonality is the days. The procedure of how the missing values are conducted as set out listed as follow.

First step The original data are first being assumed to have a missing at random. The percentage of missing value considered in this case is 5%, 15%, and 25%. These percentages are selected.

Second step Data are disaggregated or divided accordingly into days from Monday to Sunday. Considering the study has data with N observations, so this study has $W = (W_1, W_2, W_3, W_4 \ldots, W_n)$ where W represents the weekly period.

$$W_1 = (x_{1,1}, x_{1,2}, \dots, x_{1,6}, x_{1,7})$$
$$W_2 = (x_{2,1}, x_{2,2}, \dots, x_{2,6}, x_{2,7})$$
$$W_3 = (x_{3,1}, x_{3,2}, \dots, x_{3,6}, x_{3,7})$$
$$W_4 = (x_{4,1}, x_{4,2}, \dots, x_{4,6}, x_{4,7})$$
$$\vdots$$
$$W_n = (x_{n,1}, x_{n,2}, \dots, x_{n,6}, x_{n,7})$$

where x represents the day in the week that is selected. The whole idea of the rearranging could be presented in a picture as set out below:

$$\begin{array}{cc} X_t & \\ x_{1,1} & \\ x_{2,1} & \\ x_{3,1} & \\ x_{4,1} & \\ \vdots & \\ x_{n-1,7} & \\ x_{n,7} & \end{array} \rightarrow \begin{array}{ccccccc} Y_1 & Y_2 & \cdots & Y_6 & Y_7 \\ x_{1,1} & x_{1,2} & \cdots & x_{1,6} & x_{1,7} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,6} & x_{2,7} \\ x_{3,1} & x_{3,2} & \cdots & x_{3,6} & x_{3,7} \\ x_{4,1} & x_{4,2} & \cdots & x_{4,6} & x_{4,7} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n-1,1} & x_{n-1,2} & \cdots & x_{n-1,6} & x_{n-1,7} \\ x_{n,1} & x_{n,2} & \cdots & x_{n,6} & x_{n,7} \end{array}$$

where X_t is the real time series data while they are the data after being rearranged by the day.

Third step For each missing value, the mean, mean with standard deviation (mean +σ) and third quartile (Q3) values are calculated from each daily series to substitute the missing point.

Fourth step After the missing points are substituted, the data for each day are rearranged into the original arrangement.

Fifth step The dataset is compared with the original series to observe the performance.

This study considers three imputations. The mean of the data with missing value is considered as the first imputation. Since the seasonality of the data has been removed from the load data, the data become linear. If the data are normal, the mean substitution will be the most suitable imputation for the data.

Other than the mean, this study also considers mean +σ as one of the imputations. This imputation is considered and suitable when the data fluctuate extremely from one point to another. The data occasionally can be extreme due to certain events and holidays. Therefore, by adding the standard deviation to the estimated mean, it provides a greater estimation. Greater estimation is important since the overestimated value is better compared to underestimated value as the load shortage of power supply can cause electricity disruptions. To obtain one, this study only selects the positive standard deviation by adding absolute to the standard deviation.

Other than the two imputations, this study also considers the third quartile (Q3) value imputation. Q3 value is a limit where 75% data are containing below this value. As they are not affected by extreme observations, then Q3 can be a better measure than the mean. Therefore, it can also be a good imputation for a missing value especially if the data are slowly increased over the time.

Since this method involving disaggregation process before being imputed with three different imputation, the method is named as disaggregation-and-imputation (DI) method where DI1 use the mean imputation, DI2 use the mean +σ imputation and DI3 use the Q3 imputation. This method is conducted by using Minitab software. The framework of the study is explained in Figure 2.

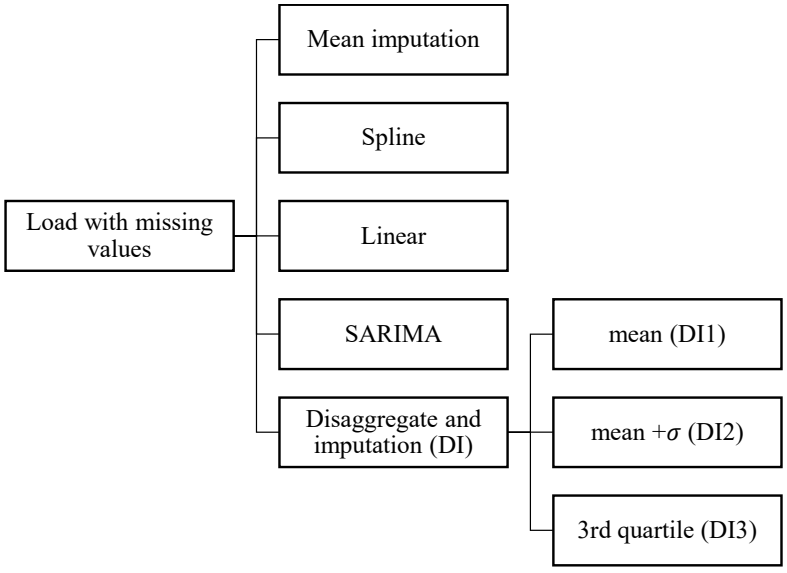

FIGURE 2. Framework on imputation methods used in this study

After data is being disaggregate according to days, the mean, mean+σ and Q3 values are calculated respectively for the 5% missing values, 15% missing values and 25% missing values. Other than that, these values are also calculated for each orientation P1, P2, and P3. After we obtained these values for each of the day, these values will be repositioned into its original sequence to form a set of data that will substitute the missing values. Herewith is the values of the mean, mean+σ and Q3 for selected percentage of missing values according to its position.

TABLE 2. Imputation value for 5%, 15%, and 25% missing values when missing values is at P1

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
				5%			
mean	65009	64843	65506	65651	63704	50069	44398
mean+σ	71543	70650	70258	69870	70448	55059	48106
Q3	68311	67840	67790	68094	67208	54175	46936
				15%			
mean	65681	65279	65461	63738	50554	44654	44654
mean+σ	70455	70270	70436	70802	55306	47708	47708
Q3	68623	68128	67797	67339	54233	46941	46941
				25%			
mean	65339	65054	65155	65318	64049	50518	44455
mean+σ	70304	70357	70405	69937	71003	55541	47617
Q3	68699	68073	70405	68164	67454	54644	46926

TABLE 3. Imputation value for 5%, 15%, and 25% missing values when missing values is at P2

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
				5%			
mean	64734	64709	65417	65439	64001	49996	44251
mean+σ	71275	70532	70153	69663	70340	55016	47897
Q3	68084	67524	67790	67887	67108	54189	46596
				15%			
mean	64419	64477	65102	65206	63835	50201	44324
mean+σ	71237	70571	70037	69587	70465	55288	47754
Q3	67793	67462	67630	67789	67073	54240	46427

				25%			
mean	50824	44899	65358	65012	65100	65047	63748
mean+σ	55387	47243	70155	70239	70357	69661	70755
Q3	54776	46748	68282	67762	67790	67749	67175

TABLE 4. Imputation value for 5%, 15%, and 25% missing values when missing values is at P3

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
				5%			
mean	64734	64612	65408	65429	63613	49837	44509
mean+σ	68363	70364	70194	69583	70345	54668	48138
Q3	67930	67524	67797	67887	67208	53058	46936
				15%			
mean	65199	64898	65667	65657	63432	50383	44629
mean+σ	71861	70806	70508	69902	70526	55066	48307
Q3	68194	67721	67864	68129	67216	53908	47011
				25%			
mean	65116	64838	66281	65764	63800	50497	44737
mean+σ	72161	71087	70597	69800	70580	55440	48531
Q3	68282	68073	67976	68059	67175	54247	47032

As can be seen from Table 2 to 4, we could see for each percentage of missing values and orientation, the value of mean will be the smallest value followed by the value of Q3 and the largest value will be the mean +σ value. This is expected from the plot in Figure 2 where the minimum and maximum values gap show how these values will be arranged to substitute the missing value, we give an example of how to impute the missing values for 5% missing data that missing at P2 by using this method. By selecting the mean +σ as the imputation from Table 3, the new set of the missing values imputation will be as below, depending on the day the missing value starts and ends:

..., 64734, 64709, 65506, 65651, 63704, 50069, 44398, 64734, 64709, 65506, ...

RESULTS AND DISCUSSION

Tables 5 and 6 are the error measurement outcomes of the findings. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are selected as the error

measurements to observe the goodness-of-fit for these missing values study as it is the most widely used and theoretical relevance in statistical modelling (Hyndman & Koehler 2006). Apart from that, the RMSE is also more sensitive than other measures occasionally whilst the MAE is slightly smaller than RMSE and it is an easier statistic to understand compared to RMSE. It was also recommended by Willmott and Matsuura (Willmott & Matsuura 2005) that MAE is an explicit measure of average error magnitude.

Beside their assists in showing the robustness of the models, these tests could show the significant of the findings if the outcome support each other result. To observe the performance of the model approach, four models widely used in handling missing values of time series are compared. These include the linear model, cubic spline model, mean imputation and SARIMA forward and back propagation model.

TABLE 5. RMSE for selected models for 5%, 15%, and 25% missing values

RMSE		Linear	Spline	Mean	SARIMA	DI1	DI2	DI3
5%	P1	2584	2580	3051	2187	**354**	1526	967
	P2	2837	2896	1586	2104	**1109**	1461	1159
	P3	3675	3769	2249	2219	885	903	**674**
15%	P1	1748	1748	1003	1516	**938**	1311	1133
	P2	1508	1564	1574	1317	**546**	752	574
	P3	1856	1904	1534	1243	781	875	**753**
25%	P1	1165	1185	1597	1072	**597**	850	734
	P2	1287	1301	1381	1118	**626**	839	726
	P3	1042	1091	1281	979	598	579	**518**

TABLE 6. MAE for selected models for 5%, 15%, and 25% missing values

MAE		Linear	Spline	Mean	SARIMA	DI1	DI2	DI3
5%	P1	32802	38694	32802	45765	**5312**	14502	22883
	P2	53897	55026	39985	22024	**21070**	27761	39985
	P3	66149	67850	39940	40473	15935	16250	**12124**

15%	P1	89130	89131	77329	66843	**47818**	57799	51162
	P2	82955	86043	72444	538406	**30043**	41378	31554
	P3	102076	104713	84364	68344	42964	48118	**41440**
25%	P1	102535	104251	140513	94357	**52499**	74772	64609
	P2	118362	119726	127073	102834	**57595**	66799	77214
	P3	89044	99266	116609	84833	54455	52678	**47123**

From Tables 5 and 6, we could see that the results of RMSE and MAE are match. When RMSE gives the smallest value at DI1, MAE also shows a smallest value at DI1 and vice versa. And because RMSE gives the value after being square root, the error values from RMSE are smaller compare to MAE.

The DI1 provides a good substitution for missing values located at P1 and P2 whilst the DI3 gives a good imputation for missing values located at P3. This is acceptable for all percentages of missing values if refer to the placement of the missing values. Since the missing values in P1 and P2 are at the beginning and in the middle of the data, mean imputation is best to substitute the misses. The data do not increase rapidly through the time thus by taking the mean of the remaining data, it will provide rational values which are closer to the missing values.

Referring to the result in Tables 5 and 6, DI3 provides a good estimation when the missing values occur at the end of the data for all percentages of missing value. This is adequate with the pattern of the data where it has shown an increasing trend along the time which can be referred in Figure 1. A larger estimation from mean value will be appropriate for the misses at P3. And from Table 2 until Table 4, Q3 has a value which is higher than mean, therefore, Q3 will be appropriate to substitute the missing values in P3.

Other than that, the Q3 is more suitable to impute the missing value compare to mean+σ because if we refer to Table 2 until Table 4, mean+σ has larger difference with mean compare to Q3 and since the increment in the data trending is very little, Q3 is more appropriate for imputation compare to mean+σ.

Based on the results from the percentage of misses at 5%, 15%, and 25% for P1, the difference between the errors with other methods is rather large. Therefore, this study moved the misses to the middle of the data and observed that the difference of error between the methods, especially the proposed methods becomes smaller and closer to other methods. This is true as the missing values percentage become larger, the error will become larger too because the data used to impute the missing values will be smaller.

From the results in Tables 5 and 6, linear, spline, mean imputation, and SARIMA provide larger values of RMSE and MAE than the proposed technique. The reason for that is because methods such as linear and cubic spline interpolation are imputed from equations of certain models which assume that the data follow a certain pattern. Nevertheless, this only provides good imputation if the data follow the pattern. As for mean imputation, if the data fluctuate around the mean, the mean imputation provides reasonable imputation to the missing values.

For the SARIMA model, the process consists of identifying the order, parameter estimation and model checking. The process is lengthy, time-consuming and becomes more complex, as it occasionally involves forecasting and back forecasting particularly when the missing values are at the middle of the data. For this reason, imputation by SARIMA is less preferable than other methods. Furthermore, from the RMSE and MAE value in Tables 5 and 6, this method is occasionally inferior compared to mean imputation, a much simpler method.

Conclusion

Missing values are a common issue in the actual database. Thus, a number of common techniques have been developed to deal with missing values. Most importantly, one must choose the appropriate method so that it is able to impute the missing values and the estimates are the closest to the actual one. The appropriate method may primarily depend on the type of data. Time series data often contain a certain pattern which is predictable. Imputing time series data can be relatively challenging due to the seasonality of the time series data. This study used load data to test the

imputation technique to deal with missing values in a seasonal time series data.

The first step is by eliminated the seasonal period by rearranging the data into days. By removing the seasonality, the complexity of the data could be reduced and a simpler technique could be applied. Three imputations are selected for this purpose which is the mean, mean+σ and Q3 value. Each of these imputations is selected for its own benefit to encounter the missing values. These imputations have different value but close to each other. From mean, mean+σ to Q3 values, we can see how these values increased and gives the result of why DI1 and DI3 are suitable for specific orientation.

The orientation of the missing values in the data is also important when considering the imputation of the missing values. If the data shows a trend, then its imputation should also consider the trend factor. As can be concluded from this study, DI1 which used mean imputation provides the best estimation of the missing values if the missing values are at the beginning or in the middle of the data. However, DI3 which used the Q3 is proper for a missing value at the end of the data. This shows that the location of the missing values should be taken into consideration before imputing the missing values.

In conclusion, it is important for one to understand their data before applying a method. Although this method has longer steps, but it gives a lot of improvement in substituting the missing data with the imputations. By understand the data, the complexity of the data such as the seasonality and trend effects could be distinguished and a simpler method could be used to overcome the problem of missing values. On the other hand, the orientation of where the missing value lies in is also important. By recognising the placement of the missing values could help one to choose an appropriate method and could improve the imputations better.

ACKNOWLEDGEMENTS

This work was supported by the Universiti Teknologi Malaysia under Grant Q.J130000.2654.17J90.

REFERENCES

Acock, A.C. 2005. Working with missing values. *Journal of Marriage and Family* 67(4): 1012-1028.

Bennett, D.A. 2001. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health* 25(5): 464-469.

Brockwell, P.J. & Davis, R.A. 2013. *Time Series: Theory and Methods*. New York: Springer Science & Business Media.

Chatfield, C. 2000. *Time-Series Forecasting*. Boca Raton: Chapman & Hall/CRC.

Cokluk, O. & Kayri, M. 2011. The effects of methods of imputation for missing values on the validity and reliability of scales. *Educational Sciences: Theory and Practice* 11(1): 303-309.

Cumming, G., Fidler, F. & Vaux, D.L. 2007. Error bars in experimental biology. *The Journal of Cell Biology* 177(1): 7-11.

Damsleth, E. 1980. Interpolating missing values in a time series. *Scandinavian Journal of Statistics* 7(1): 33-39.

Ferreiro, O. 1987. Methodologies for the estimation of missing observations in time series. *Statistics & Probability Letters* 5(1): 65-69.

Gerald, C.F. & Wheatley, P.O. 2004. *Applied Numerical Analysis with MAPLE.* Boston: Addison-Wesley.

Gómez, V., Maravall, A. & Peña, D. 1992. Computing missing values in time series. *Computational Statistics* 1: 283-296.

Hamilton, J.D. 1994. *Time Series Analysis*. Volume 2. New Jersey: Princeton University Press.

Harvey, A.C. 1990. *Forecasting, Structural Time Series Models and The Kalman Filter*. Cambridge: Cambridge University Press.

Honaker, J. & King, G. 2010. What to do about missing values in time-series cross-section data. *American Journal of Political Science* 54(2): 561-581.

Hyndman, R.J. & Koehler, A.B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* 22(4): 679-688.

Janacek, G.J. & Swift, L. 1993. *Time Series: Forecasting, Simulation, Applications*. New York: Ellis Horwood.

Kihoro, J. & Athiany, K. 2013. Imputation of incomplete non-stationary seasonal time series data. *Mathematical Theory and Modeling* 3(12): 142-154.

Peng, C.Y.J., Harwell, M., Liou, S.M. & Ehman, L.H. 2006. Advances in missing data methods and implications for educational research. In *Real Data Analysis,* edited by Sawilowsky, S.S. North Carolina: IAP. pp. 31-78.

Penn, D.A. 2007. Estimating missing values from the general social survey: An application of multiple imputation. *Social Science Quarterly* 88(2): 573-584.

Ruiz, E. & Nieto, F.H. 2000. A note on linear combination of predictors. *Statistics & Probability Letters* 47(4): 351-356.

Schafer, J.L. 1999. Multiple imputation: A primer. *Statistical Methods in Medical Research* 8(1): 3-15.

Schlomer, G.L., Bauman, S. & Card, N.A. 2010. Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology* 57(1): 1-10.

Shukur, O.B. & Lee, M.H. 2015. Imputation of missing values in daily wind speed data using hybrid AR-ANN method. *Modern Applied Science* 9(11): 1-11.

Sorjamaa, A. & Lendasse, A. 2007. Time series prediction as a problem of missing values: Application to ESTSP2007 and NN3 competition benchmarks. Paper presented at the, *International Joint Conference on Neural Networks* 2007 (IJCNN 2007).

Willmott, C.J. & Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30(1): 79-82.

Winkler, A. & McCarthy, P. 2005. Maximising the value of missing data. *Journal of Targeting, Measurement and Analysis for Marketing* 13(2): 168-178.

Zhang, Z., Yang, X., Li, H., Li, W., Yan, H. & Shi, F. 2017. Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level. *Journal of Hydrology* 553: 384-397.

Nur Arina Bazilah Kamisan* & Muhammad Hisyam Lee
Mathematics Department
Faculty of Science
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor Darul Takzim
Malaysia

Abdul Ghapor Hussin
Faculty of Science and Defence Technology
Universiti Pertahanan Nasional Malaysia
50300 Kuala Lumpur, Federal Territory
Malaysia

Yong Zulina Zubairi
Pusat Asasi Sains Universiti Malaya
Universiti Malaya
50300 Kuala Lumpur, Federal Territory
Malaysia

*Corresponding author; email: nurarinabazilah@utm.my

Received: 12 August 2019
Accepted: 24 January 2020