Thailand Statistician
October 2021; 19(4): 866-879
http://statassoc.or.th
Contributed paper

Imputation Methods in Time Series with a Trend and a Consecutive Missing Value Pattern

Chantha Wongoutong
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand.
Corresponding author; e-mail: fscictw@ku.ac.th

Received: 26 September 2020
Revised: 3 December 2020
Accepted: 3 January 2021

Abstract

Time series with missing values can occur in almost any domain of applied sciences, and ignoring missing values, especially for a large consecutive pattern of missing values, can lead to a loss of efficiency and unreliable results. Applying an appropriate imputation method can replace the missing values with substituted ones and lead to more accurate forecasting. However, the appropriate imputation method depends on the type of time series and the missing data pattern. The focus of this study is on time-series types with a trend when consecutive missing values are apparent. Ten real datasets were used to evaluate the performances of imputation methods with three scenarios of missing artificial data sequences in a time series of 10%, 20%, and 50%. The performances of six approaches for imputing missing values: interpolation, Kalman, moving average (MA), last observation carried forward (LOCF), mean, and linear trend at point (LTP) were compared in terms of root-mean-squared error (RMSE) and mean-absolute-percentage error (MAPE). The performances of the Interpolation, Kalman, and LTP were far superior to the other three imputation methods in the order of 80% on average relative to the Mean imputation method and 30-60% on average relative to the LOCF and MA imputation methods. Hence, the interpolation, Kalman, and LTP methods from this study are appropriate for imputing consecutive missing values for time-series data exhibiting a trend.

Keywords: Missing values, imputation method, consecutive missing values, time series.

1. Introduction

Time-series forecasting is used to predict future values by applying a model based on previous data and assuming that the future data will be similar to the current data (Box and Jenkins 2011). Missing values comprise one of the main problems that frequently occur in data observation or data recording processes as data completeness is vital before applying advanced analyses. Several imputation methods are available to impute missing values that have been utilized in various technologies, including sensors, actuators, mobile devices, and wearable devices (Saunders et al. 2006, Lepot et al. 2017). Characteristics of the time series data and the missing observations can profoundly affect the accuracy of imputation methods.

By including estimates for missing values, a better understanding of the data's nature is possible for more accurate forecasting. When selecting an imputation method, it is important to consider the

time-series pattern: either stationary or non-stationary. Moreover, extensive studies on evaluating methods for statistically handling missing data have been conducted (Grzymala-Busse and Hu 2000, Schefer 2002, Batista and Monard 2003, Honaker and King 2010, Junger and De Leon 2015, Engels and Diehr 2003, Horton et al. 2007, Schlomer et al. 2010). The imputation of missing values for a non-stationary pattern is more complicated than for a stationary one (Walter et al. 2013). Furthermore, some traditional imputation techniques perform well with a trend and no seasonality time-series datasets, while some are appropriate for seasonal time-series datasets, although there is no single univariate imputation technique that fits all types of data patterns (Xu et al. 2016).

Conventional methods such as mean, median, or mode imputation, deletion, etc. are not good enough to handle a complex non-stationary pattern. Schlomer et al. (2010) compared three methods for handling missing data: mean substitution, multiple imputation, and full information maximum likelihood; their results suggest that mean substitution is a poor method for handling missing data for complex time series as in the non-stationary form. Similarly, Bishop (2006) proposed simply substituting the mean or the median of available values for each missing value, although simple algorithms that provide the same results for all missing values have been shown to lead to bias in the results and undervalued standard errors (Crawford et al. 1995, Sterne et al. 2009). However, these studies only focused on filling one isolated, missing value rather than considering a sub-sequence of missing values. Likewise, Moritz and Bartz-Beielstein (2017) compared imputation methods for time-series missing values using the R programming language and suggested that the type of data is important when selecting appropriate imputation methods; their results show that the Kalman and interpolation methods are suitable for time-series data with only the trend component.

Missing patterns can also be classified as either monotone missing or arbitrary missing (Dong and Peng 2013), as shown in Figure 1. In particular, a missing severe pattern can include high rates of missing values over a very long period (Fortino et al. 2015, Wellenzohn et al. 2017). However, only a few studies have endeavored to handle a consecutive range of missing data items and the type of time series simultaneously.

Monotone missing	O	O	O	O	O	O	O	O	O	O	O	O	O	O	X	X	X	X
Arbitrary missing	X	O	O	X	O	O	O	X	O	O	X	O	O	X	X	X	O	X

Figure 1 Missing value pattern classification. O, observed data; X, missing values

A serious missing value pattern problem includes high missing value rates and an extremely long consecutive series of missing values distinct from the existing missing value pattern (the arbitrary missing pattern or the monotone missing pattern). Junninen and Niska (2004) studied missing data imputation in air quality datasets using imputation techniques for univariate time series such as spline interpolation, linear interpolation, and nearest neighbor interpolation. Their results show that univariate methods are dependent on the size of the gap in time series: the long the gap, the less efficient the technique.

A consecutive missing value pattern can occur in time-series data, as shown in Figure 2. However, only limited studies have endeavored to handle a consecutive range of missing data, so there is a strong need for further development. Hence, the focus of this study is on consecutive patterns of missing values in non-stationary time series with a trend by using six well-known imputation methods: interpolation, Kalman, moving average (MA), last observation carried forward (LOCF), mean, and linear trend at point (LTP) available in the imputeTS R package. The performances of the six methods

were compared in terms of mean-absolute-percentage error (MAPE) and root-mean-squared error (RMSE).

Consecutive missing pattern																			
	(1)	X	X	X	X	X	X	X	X	O	O	O	O	O	O	O	O	O	O
	(2)	O	X	X	X	X	X	X	X	X	O	O	O	O	O	O	O	O	O
	(3)	O	O	X	X	X	X	X	X	X	X	O	O	O	O	O	O	O	O
	⋮	⋮																	
	(*n*)	O	O	O	O	O	O	O	O	O	O	X	X	X	X	X	X	X	X

Figure 2 Consecutive missing value patterns. O, observed data; X, missing values

The rest of this paper is organized as follows. Section 2 presents material and methods consisting of the imputation procedure used in the study and the data used. The experimental study to compare the methods is described in Section 3. The results of the experimental study and a discussion are given in Section 4. The conclusions are included in Section 5.

2. Material and Methods

2.1 Missing data mechanisms

Three main types of missing data mechanisms under which missing data can occur: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) (Little and Rubin 2019, Rubin 2004). First, the missing data are said to be MCAR when the probability of missing is equal for all cases. Second, the missing data are said to be MAR when the probability of missing is equal only within groups defined in the observed data. Third, the missing data are said to be MNAR when the probability of missing varies for unknown reasons. For univariate time-series imputation, MAR and MCAR are nearly the same. The missing values of MCAR are much easier to estimate, e.g., with the R package MissMech (Jamshidian et al. 2014).

2.2. The real datasets used in the study

Comparing results for real missing data is not possible since the actual values are unknown. Hence, a comparison of imputation methods' performance can only be made for simulated missing data, so data points from a complete dataset are artificially removed. After that, the imputed and real values can be compared. Ten complete real datasets with a trend were obtained from various sources, and a description of these datasets is provided in Table 1.

Table 1 Description of the 10 real datasets used in the study

Dataset	Description	Time Period	Size
T1	Quarterly income from non-industrial agricultural sources in	Q1/1990–Q4/2000	44
T2	Quarterly income of beverage product groups in Thailand	Q1/2007–Q2/2017	42
T3	Quarterly amount of plastics in and synthetic resins in Australia	Q4/1964 – Q4/1974	41
T4	Quarterly financial service charges in Australia	Q1/2002 – Q4/2012	44
T5	Monthly debit card electronic payments in Thailand	Jan 2013 –Dec 2016	48
T6	Monthly number of employed persons in Australia	Jan 1984 –Sep 1987	43
T7	Monthly civilian labor force in Australia: thousands of persons	Feb 1978–Dec 1994	95
T8	Monthly consumer price index in Canada	Jan 1966 –Dec 1969	48
T9	Quarterly Australian govt final consumption expenditure total	Q3/1965–Q4/1980	62
T10	Monthly electronic payment amount in Thailand	Jan 2013 –Dec 2016	48

2.3. Imputation methods evaluated in the study

Imputation is replacing missing data with substituted values (Junnien and Niska 2004). Missing data leads to problems when analyzing the data, and so imputation is used to avoid pitfalls involved with a dataset containing missing values as standard techniques can only be used when the dataset is complete. Thus, imputation is critical to replace missing data items (Little and Rubin 2019), and the R package provides imputation functions (Moritz and Bartz-Beielstein 2017) Six well-known attempts to deal with missing data in a univariate time-series imputation include Interpolation, Kalman, MA, LOCF, mean, and LTP. These six imputation methods in the R package considered in this study are as follows:

(1) Interpolation is a method from imputeTS that replaces missing values with interpolated values using na.interpolation.

(2) Kalman is a method from imputeTS that performs Kalman smoothing using the state space representation of an auto regressive integrated MA model for imputation. The method used is na.kalman with the auto.arima model.

(3) MA is a method from imputeTS with which missing values are replaced by the weighted MA. A semi-adaptive window size is used to ensure that all not available values are replaced. The method used is na.ma where all observations in the window are equally weighted for calculating the mean.

(4) LOCF is a method from imputeTS that replaces each missing value with the most recent non-missing value before it. The algorithm used is na.locf.

(5) Mean from imputeTS fills the missing values with the mean value of a time series using na.mean.

(6) LTP replaces missing values with the linear trend for that point. Regression analysis is applied to the dataset using time as an index variable. After which missing values are imputed with their predicted values.

2.4. The performance metrics

The RMSE and MAPE values for each missing value pattern were respectively calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{miss}} \left(y_i - \hat{y}_i\right)^2}{n_{miss}}}, \; MAPE = \frac{\sum_{i=1}^{n_{miss}} \left(\left|y_i - \hat{y}_i\right| / y_i\right)}{n_{miss}} \times 100,$$

where y_i is the true value, $\hat{y}_i$ is the imputed value, and n_{miss} is the number of missing values.

3. The Experimental Study

Missing values were artificially deleted from the ten complete datasets to produce missing value rates of 10%, 20%, and 50%. For a dataset of n observations and r% missing rate, there are n_{miss} missing values, where $n_{miss} = r \times 0.01 \times n$, and the number of possible patterns is $n - n_{miss} + 1$. For example, for a dataset of 44 observations and a 10% missing value rate, the number of missing values is $10 \times 0.01 \times 44 = 4.4 \approx 5$ and the number of possible patterns in which missing values occur consecutively is 44–5+1 = 40, as summarized in Table 2.

Table 2 All possible consecutive missing patterns (40 patterns) for a dataset of size 44 with a 10% missing value rate

Data		Time Period												⋯									
		1	2	3	4	5	6	7	8	9	10	11	12	⋯	36	37	38	39	40	41	42	43	44
Complete data		O	O	O	O	O	O	O	O	O	O	O	O	⋯	O	O	O	O	O	O	O	O	O
Consecutive missing value	No.1	X	X	X	X	X	O	O	O	O	O	O	O	⋯	O	O	O	O	O	O	O	O	O
	No.2	O	X	X	X	X	X	O	O	O	O	O	O	⋯	O	O	O	O	O	O	O	O	O
	⋮													⋮									
	No.39	O	O	O	O	O	O	O	O	O	O	O	O	⋯	O	O	O	X	X	X	X	X	O
	No.40	O	O	O	O	O	O	O	O	O	O	O	O	⋯	O	O	O	O	X	X	X	X	X

O, observed data; X, missing values.

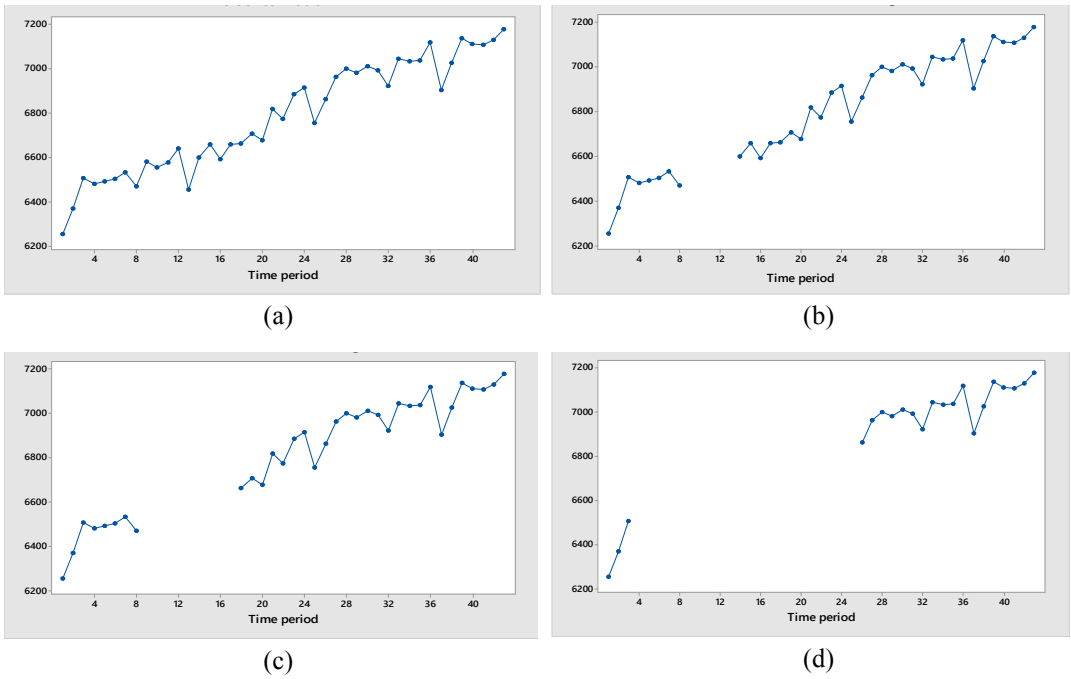

Figure 3 Time-series plot of dataset T1 and examples of consecutive missing value patterns: (a), (b), (c), and (d) refers to complete data and data with missing value rates10%, 20%, 50% with respectively

Table 3 The number of missing values and the number of possible patterns with consecutive missing values for T1 to T10 with varying missing value rates

Dataset	Size	Number of Missing Values			Number of Possible Patterns		
		Missing Rate			Missing Rate		
		10%	20%	50%	10%	20%	50%
T1	44	5	9	22	40	36	23
T2	42	5	9	21	38	34	22
T3	41	5	9	21	37	33	21
T4	44	5	9	22	40	36	23
T5	48	5	10	24	44	39	25
T6	43	5	9	22	39	35	22
T7	95	10	19	48	86	77	48
T8	48	5	10	24	44	39	25
T9	62	7	13	31	56	50	32
T10	48	5	10	24	44	39	25

A flow chart of the steps to evaluate the performance of the six traditional imputation methods is shown in Figure 4.

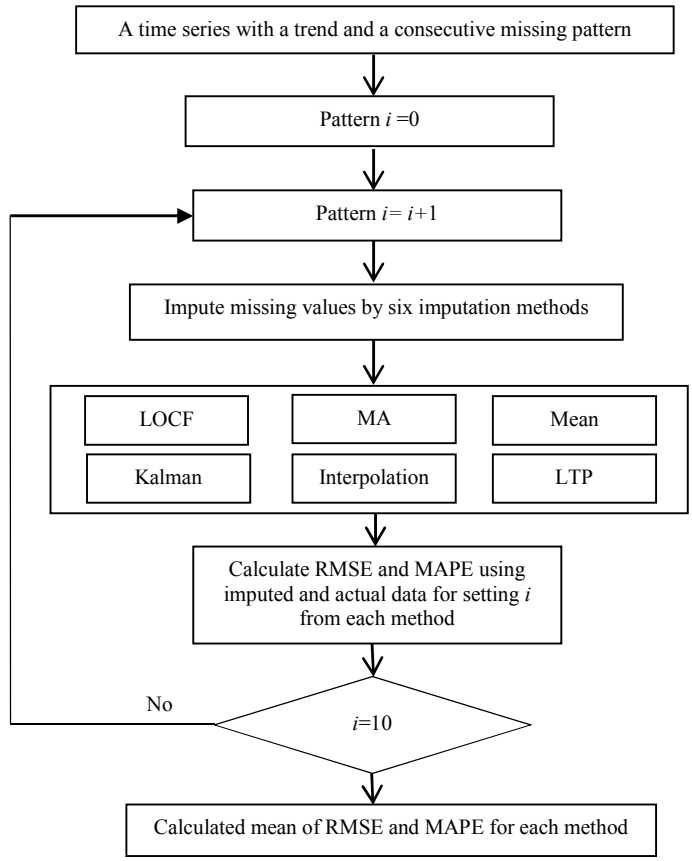

Figure 4 A flowchart of the steps to evaluate the performance of the six imputation

The number of missing values and all possible patterns of consecutive missing values at missing rates of 10%, 20%, and 50% for T1 to T10 are reported in Table 3. For each of the missing rates in each dataset, 10 patterns were randomly selected from all of the possibilities, and the missing values were replaced using the imputed values generated by the six well-known imputation methods. Afterward, the average RMSE and MAPE for the 10 missing value patterns were used to evaluate the performances' imputation methods.

4. Results and Discussion

The performance of the imputation methods was considered via their average MAPE and RMSE values for 10 missing value patterns: the lower the metric's value, the better the imputation method's performance. The average MAPE and RMSE results for the six well-known imputation methods with Interpolation, Kalman, MA, LOCF, Mean, and LTP for datasets T1 to T10 with missing value rates 10%, 20%, and 50% are reported in Tables 4 and 5, respectively.

As an example from the average MAPE results in Table 4, for dataset T1 at a missing value rate of 10%, the average MAPE values for the interpolation, Kalman, MA, LOCF, mean, and LTP methods were 1.6502, 1.3973, 2.5592, 3.5896, 16.8153, and 1.7347, respectively, and at a missing value rate of 20%, they were 1.9093, 1.8396, 5.7669, 6.5967, 17.0914, and 1.6384, respectively. This trend was the same for missing value rates of 50%: 1.8652, 2.1624, 8.0273, 12.691, 18.5656, and 1.8529, respectively. Furthermore, the average RMSE values in Table 5 show the same trend for all three missing value rates.

These results support that the three imputation methods: interpolation, Kalman, and LTP, outperformed the others at imputing the missing values. Indeed, the same conclusion can be drawn from the plots of the combined average MAPE and RMSE for T1 to T10 according to the missing value rate. Furthermore, the combined average MAPE for T1 to T10 according to the missing value rate, as shown in Figure 5.

Table 4 Average MAPE values for T1–T10 with various missing values rates

Missing Rate	Dataset	Imputation Method					
		Interpolation	Kalman	MA	LOCF	Mean	LTP
10%	T1	1.6502	1.3973*	2.5592	3.5896	16.8153	1.7347
	T2	8.3743	7.1286*	9.5886	12.4716	51.1384	14.5044
	T3	6.8736	5.6012*	10.336	12.5007	52.5131	15.3625
	T4	1.7424	1.4003*	2.2188	3.3947	21.3211	4.1956
	T5	4.5778	3.9177	9.4899	7.5399	32.2652	4.8701
	T6	1.0073	0.8469	1.2205	1.6718	3.8810	0.7774*
	T7	0.6990*	0.5925	0.9470	0.9973	5.8691	0.7527
	T8	0.2044	0.2018*	0.7203	0.7795	4.4564	0.4402
	T9	5.0246	4.0756*	6.7953	5.1377	17.0993	4.5686
	T10	4.0119	4.4711	4.6204	6.882	16.5435	3.9480*

*The best performance in terms of MAPE.

Table 4 (Continued)

Missing Rate	Dataset	Imputation Method					
		Interpolation	Kalman	MA	LOCF	Mean	LTP
20%	T1	1.9093	1.8396	5.7669	6.5967	17.0914	1.6384*
	T2	8.0073	6.4090*	23.8735	17.9807	63.7979	13.8224
	T3	8.2222	8.0353*	22.5845	15.5834	65.3907	23.3849
	T4	1.4185	1.2919*	5.2132	6.4018	14.2734	2.9142
	T5	5.0425	4.3065*	9.9625	8.1962	22.6708	5.9962
	T6	0.8484	0.7738*	1.1794	1.6082	3.1909	0.8323
	T7	0.6218	0.5385*	1.2235	1.0990	4.3918	0.8525
	T8	0.3392	0.3374*	1.3387	1.2977	4.4436	0.5179
	T9	5.8639	4.347*	7.8095	7.9584	24.0899	4.8325
	T10	3.5857	3.4727*	6.3675	6.5763	13.5272	4.6029
50%	T1	1.8652	2.1624	8.0273	12.691	18.5656	1.8529*
	T2	8.8227	8.4363*	32.9942	36.7473	69.6177	11.5988
	T3	14.0791	12.9905*	27.9627	34.5174	66.865	26.0955
	T4	2.9800	2.1167*	8.0958	12.9763	21.3792	4.9201
	T5	7.1999	5.6871*	13.6902	17.9445	20.1693	9.0261
	T6	1.5965	1.3482	2.4615	3.3831	3.3583	0.9680*
	T7	0.7872*	0.8489	1.8310	3.0092	7.4985	0.9296
	T8	0.8097	0.7300	2.7168	3.5129	3.5335	0.5925*
	T9	5.9569	4.9769	9.2640	14.3581	23.4198	4.8683*
	T10	5.1031*	5.7225	8.6439	12.7671	22.5772	7.0087

*The best performance in terms of MAPE.

Table 5 Average RMSE values for T1–T10 with various missing value rates

Missing Rate	Dataset	Imputation Method					
		Interpolation	Kalman	MA	LOCF	Mean	LTP
10%	T1	15.350	16.435	22.477	34.029	143.459	14.742*
	T2	568.007	467.530*	768.909	968.847	3811.519	796.363
	T3	2923.137	2535.194*	4403.978	5135.350	17477.630	4991.240
	T4	14.565	11.470*	18.696	29.914	147.409	30.927
	T5	248.606	231.428*	435.570	389.224	1024.490	273.448
	T6	79.278	67.034	94.966	123.119	232.219	61.361*
	T7	52.946	44.282*	70.529	54.109	271.136	54.202
	T8	0.272	0.269*	0.944	1.057	4.188	0.543
	T9	388.561	292.794*	427.301	346.020	1462.845	316.779
	T10	2882.189*	3180.847	3313.476	4667.407	12956.030	2883.215

*The best performance in terms of RMSE.

Table 5 (Continued)

Missing Rate	Dataset	Imputation Method					
		Interpolation	Kalman	MA	LOCF	Mean	LTP
20%	T1	18.010	17.692	47.799	59.480	133.007	15.084*
	T2	603.289	489.972*	1464.515	1652.336	4114.717	877.933
	T3	3192.769	2981.958*	7130.571	6708.752	17985.270	7081.749
	T4	12.940	11.873*	49.759	59.279	110.386	24.226
	T5	267.512	245.698*	484.456	474.148	1217.167	315.944
	T6	70.326	63.988*	93.306	122.853	230.694	69.779
	T7	51.056	42.332*	84.841	76.708	290.364	57.845
	T8	0.456	0.455*	1.751	1.841	5.235	0.658
	T9	352.423	309.228*	598.463	553.078	1438.380	333.265
	T10	3011.917	2839.857*	5483.550	5469.922	9826.430	3613.053
50%	T1	20.665	21.937	77.050	126.976	153.524	18.453*
	T2	741.869	717.626*	2524.699	3933.271	3904.998	912.523
	T3	5515.782	5314.645*	11600.140	16453.420	18681.100	9015.737
	T4	25.875	18.920*	78.873	130.601	169.701	39.650
	T5	396.438	326.707*	761.012	1018.855	1447.916	456.717
	T6	123.218	105.761	190.536	264.795	286.000	76.764*
	T7	63.692	68.222	106.294	159.994	309.018	62.526*
	T8	1.070	0.995	3.535	4.805	5.778	0.775*
	T9	453.720	385.902	740.066	1221.644	1349.981	377.344*
	T10	4177.992*	4616.736	7763.198	11086.690	13027.250	5638.134

*The best performance in terms of RMSE.

Furthermore, heatmaps were used to visualize average MAPE clusters for the various missing value rates and imputation methods simultaneously, as shown in Figure 6. The first rows and the columns of the expression matrix were achieved by hierarchical clustering. In this study, the clustering algorithm used Euclidean distance, and the complete linkage method using the hclust function in the R statistics package version 4.0.3 was applied to finding similar group patterns in datasets T1–T10. Next, the dendrograms clustering the algorithm branches were rotated so that the blocks of high and low expression values were nearby in the expression matrix. Finally, visualization was realized by applying a color scheme to display the expression matrix. The tree branches were rotated to create blocks in which the individual values were the closest in both directions. These are color-coded by expression values.

The heatmaps clearly show the patterns picked out by the clustering algorithm as three clustering groups for the MAPE average (worst to best). The first is classified as the Mean imputation method, the second as the MA and LOCF imputation methods, and the third as the LTP, Kalman, and interpolation imputation methods. These results support that the interpolation, Kalman, and LTP methods outperformed the others at imputing the missing values, thus supporting the same trend shown in Figure 6.

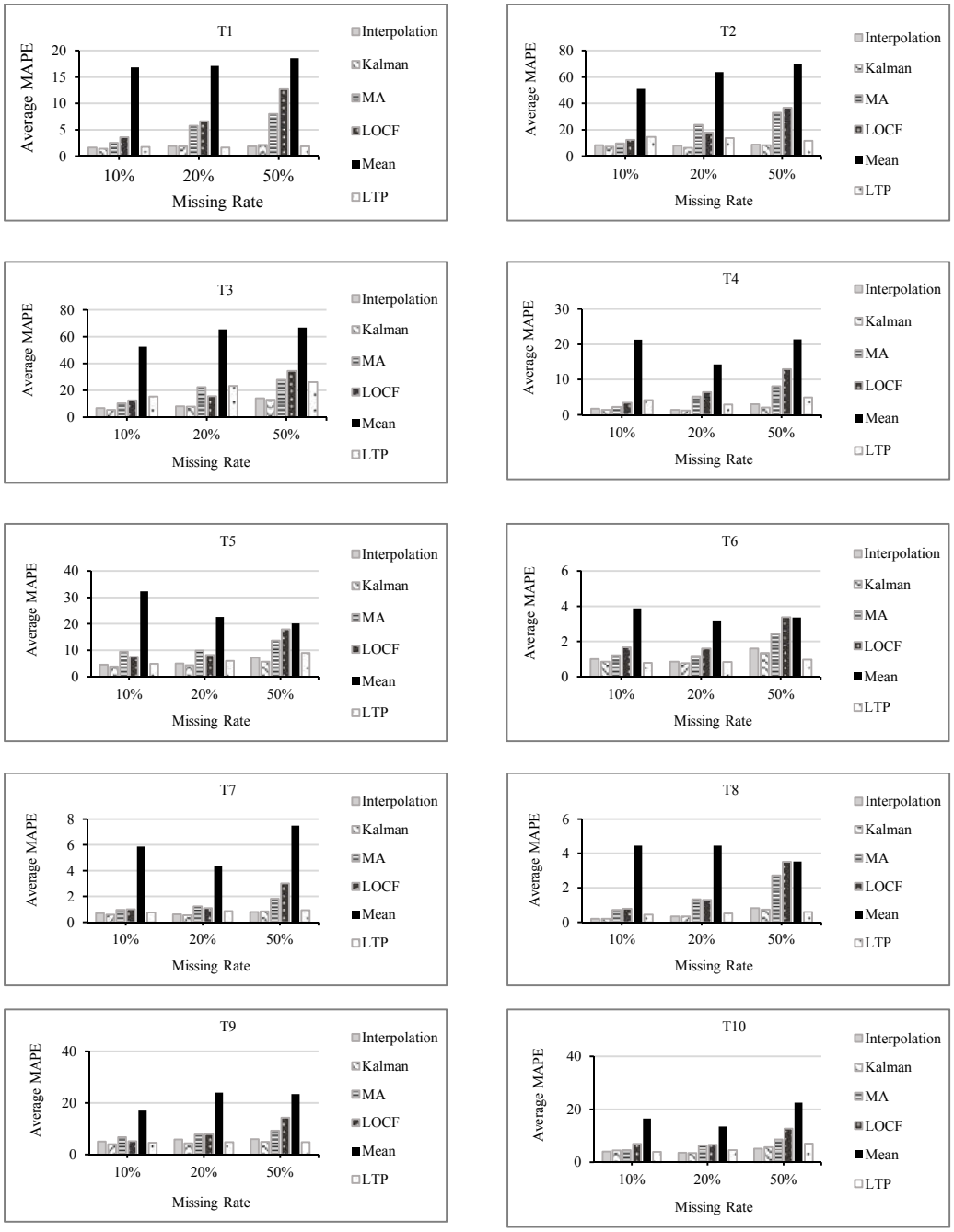

Figure 5 Performance of the imputation methods according to missing value rate: combined average MAPE for dataset of T1 to T10

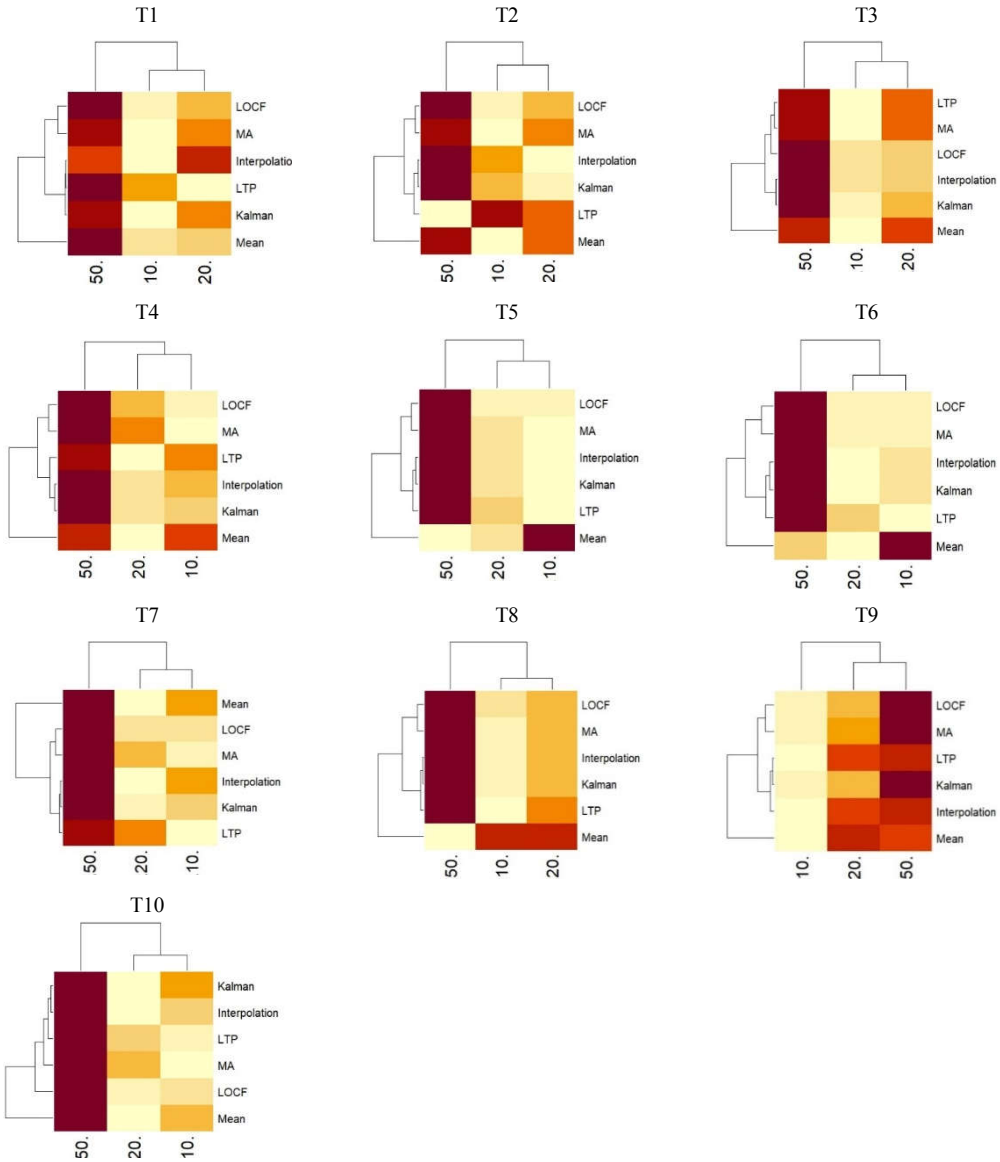

Figure 6 Heatmaps of the clustering of the imputation methods according to the missing value rate using the average MAPE for all of the datasets (T1–T10)

The results in Table 6 indicate better MAPE and RMSE values for the interpolation, Kalman, and LTP methods calculated relative to the other three methods. For the missing value rates of 10%, 20%, and 50%, the overall average of percentage improvement in MAPE over MA, LOCF, and mean were 46.61%, 53.77%, and 81.74% for the interpolation method; 52.46%, 59.11%, and 83.74% for the Kalman method; and 29.74%, 38.47%, and 77.85% for the LTP method, respectively. Similarly, the overall average of percentage improvement in RMSE over MA, LOCF, and Mean were 45.08%, 53.13%, and 79.02% for the Interpolation method; 50.12%, 57.85%, and 80.87% for the Kalman method; and 34.32%, 44.31%, and 75.88% for the LTP method, respectively.

A clear illustration of these findings is given in Figure 7(a)–(f) along with the heatmap results in Figure 6. These results clearly show that the interpolation, Kalman, and LTP imputation methods considerably outperformed the three traditional ones in terms of these metrics.

Table 6 The overall average of percentage improvement in MAPE and RMSE by the Interpolation, Kalman, and LTP imputation methods over the MA, LOCF, and Mean imputation methods for T1–T10

Missing rate	Improvement % in MAPE by Interpolation			Improvement % in RMSE by Interpolation		
	MA	LOCF	Mean	MA	LOCF	Mean
10%	30.94	40.72	84.23	29.12	36.48	81.46
20%	53.96	52.64	83.23	51.97	54.00	80.84
50%	54.93	67.96	77.76	54.14	68.92	74.75
Overall Average	46.61	53.77	81.74	45.08	53.13	79.02
Missing rate	Improvement % in MAPE by Kalman			Improvement % in RMSE by Kalman		
	MA	LOCF	Mean	MA	LOCF	Mean
10%	39.58	48.56	86.11	36.76	44.18	83.52
20%	59.26	58.16	85.10	56.52	58.45	82.59
50%	58.54	70.60	80.00	57.09	70.91	76.49
Overall Average	52.46	59.11	83.74	50.12	57.85	80.87
Missing rate	Improvement % in MAPE by LTP			Improvement % in RMSE by LTP		
	MA	LOCF	Mean	MA	LOCF	Mean
10%	3.49	19.96	80.40	12.92	24.89	78.93
20%	38.10	32.96	77.49	39.32	41.42	75.74
50%	47.62	62.48	75.67	50.73	66.63	72.97
Overall Average	29.74	38.47	77.85	34.32	44.31	75.88

5. Conclusions

The aim of the study was to impute consecutive missing values in time series with a trend. The performances of the six imputation methods: interpolation, Kalman, MA, LOCF, mean, and LTP were compared in terms of MAPE and RMSE values using ten real datasets with missing rates of 10%, 20%, and 50%. The performances of Interpolation, Kalman, and LTP were similar for all cases and notably superior to the other three imputation methods (MA, LOCF, and mean). Increasing the missing value rate decreased the performances of all of the imputation methods. The results show that the percentage improvement in MAPE and RMSE by interpolation, Kalman, and LTP were 80% on overall average relative to the mean imputation method and 30-60% on overall average relative to the LOCF and MA imputation methods. Hence, the interpolation, Kalman, and LTP imputation methods from this study are appropriate for imputing consecutive missing values for time-series data exhibiting a trend. In future work, the approach will be applied to other missing value patterns, such as arbitrary missing.

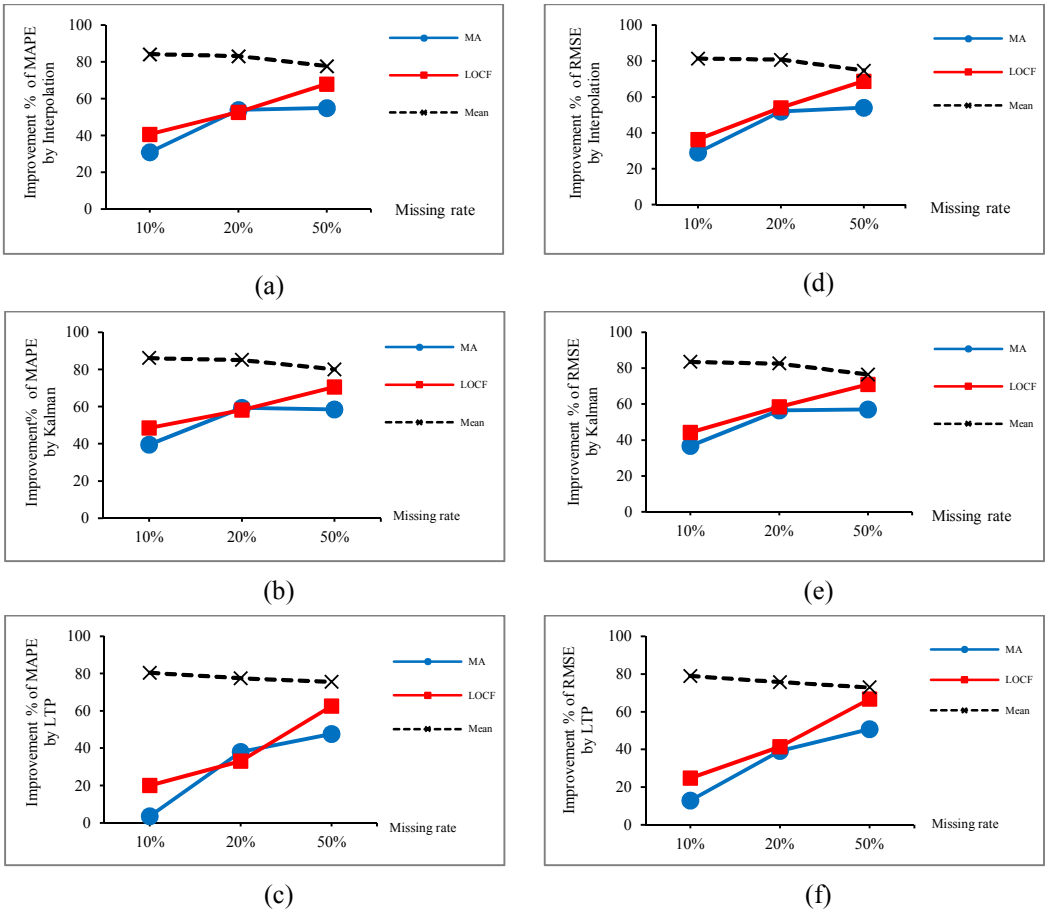

Figure 7 The average percentage improvement: (a), (b), and (c) refers to improvement in MAPE by Interpolation, Kalman and LTP and the average percentage improvement: (d), (e), and (f) refers to improvement in RMSE by Interpolation, Kalman and LTP over the MA, LOCF, and Mean imputation methods for T1-T10 by varying the missing value rates of 10%, 20%, and 50%

Acknowledgments

The author would like to thank Kasetsart University and as well as International SciKU Branding (ISB), Faculty of Science Kasetsart University for providing the facilities to conduct the research.

References

Batista GE, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell. 2003; 17(5-6): 519-533.

Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

Box GE, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control. New York: John Wiley & Sons; 2011.

Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol. 1995; 48(2): 209-219.

Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus 2013; 2(1): 222.

Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol. 2003; 56(10): 968-976.

Fortino G, Galzarano S, Gravina R, Li W. A framework for collaborative computing and multi-sensor data fusion in body sensor networks. Inf Fusion. 2015; 22: 50-70.

Grzymala-Busse JW, Hu M. A comparison of several approaches to missing attribute values in data mining. In: Ziarko W, Yao YY, editors. RSCTC 2000: International Conference on Rough Sets and Current Trends in Computing; 2000 Oct 16; Berlin, Heidelberg. Springer; 2000. pp. 378-385.

Honaker J, King G. What to do about missing values in time-series cross-section data. Am J Pol Sci. 2010; 54(2): 561-581.

Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007; 61(1): 79-90.

Jamshidian M, Jalal SJ, Jansen C. MissMech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). J Stat Softw. 2014; 56(6): 1-31.

Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmos Environ. 2004; 38(18): 2895-2907.

Junger WL, De Leon AP. Imputation of missing data in time series for air pollutants. Atmos Environ. 2015; 102: 96-104.

Lepot M, Aubin JB, Clemens FH. Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment. Water Int. 2017; 9 (10): 1-20.

Little RJ, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 2019.

Moritz S, Bartz-Beielstein T. imputeTS: time series missing value imputation in R. R J 2017; 9(1): 207.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. [cited 2020 Sep 30], Available from: http://www.R-project.org/.

Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 2004.

Saunders JA, Morrow-Howell N, Spitznagel E, Doré P, Proctor EK, Pescarino R. Imputing missing data: a comparison of methods for social work researchers. Soc Work. 2006; 30(1): 19-31.

Scheffer J. Dealing with missing data. Res Lett Inf and Math Sci. 2002; 3: 153-160.

Schlomer GL, Bauman S, Card NA. Best practices for missing data management in counseling psychology. J Couns psychol. 2010; 57(1): 1.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338.

Walter YO, Kihoro JM, Athiany KH, Kibunja HW. Imputation of incomplete non-stationary seasonal time series data. Math Theor Model. 2013; 3: 142-154.

Wellenzohn K, Böhlen MH, Dignös A, Gamper J, Mitterer H. Continuous imputation of missing values in streams of pattern-determining time series. In: Markl V, Orlando S, Mitschang B, Andritsos P, Sattler KU, Breß S, editors. EDBT 2017: Proceedings of the 20th International Conference on Extending Database Technology; 2017 March 21-24; Italy. 2017. OpenProceedings.org; 2017. pp. 330-341.

Xu J, Li Y, Zhang Y, Mahmood A. Wsn missing data imputing based on multiple time granularity. Int J Future Gener Commun Netw. 2016; 9(6): 263-274.