Hindawi
Wireless Communications and Mobile Computing
Volume 2019, Article ID 4039758, 13 pages
https://doi.org/10.1155/2019/4039758

Research Article

A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness

Aizaz Chaudhry, Wei Li, Amir Basri, and François Patenaude

Communications Research Centre Canada, Ottawa, ON K2H 8S2, Canada

Correspondence should be addressed to Wei Li; wei.li@canada.ca

Received 25 September 2018; Accepted 7 November 2018; Published 3 January 2019

Academic Editor: Simone Morosi

Imputation of missing data in datasets with high seasonality plays an important role in data analysis and prediction. Failure to appropriately account for missing data may lead to erroneous findings, false conclusions, and inaccurate predictions. The essence of a good imputation method is its missingness-recovery-ability, i.e., the ability to deal with large periods of missing data in the dataset and the ability to extract the right characteristics (e.g., seasonality pattern) buried under the dataset to be analyzed. Univariate imputation is usually incapable of providing a reasonable imputation for a variable when periods of missing values are large. On the other hand, the default multivariate imputation approach cannot provide an accurate imputation for a variable when missing values of other correlated variables used for imputation occur at exactly the same time intervals. To deal with these drawbacks and to provide feasible imputations in such scenarios, we propose a novel method that converts a single variable into a multivariate form by exploiting the high seasonality and random missingness of this variable. After this conversion, multivariate imputation can then be applied. We then test the proposed method on an LTE spectrum dataset for imputing a single variable, such as the average cell throughput. We compare the performance of our proposed method with Kalman filtering and default method for multivariate imputation. The performance evaluation results clearly show that the proposed method significantly outperforms Kalman filtering and default method in terms of imputation and prediction accuracy.

1. Introduction

Cellular data usage over smartphones has resulted in an exponential increase in wireless traffic. To meet the ever-increasing traffic demands, LTE network operators must plan and run networks efficiently. To achieve this goal, it is necessary to measure and predict parameters of the LTE spectrum usage, such as average cell throughput. Forecasting the average cell throughput can play an important role in congestion control and network management. Cell congestion can be anticipated based on the cell throughput prediction and network reconfiguration can be triggered to shift the coverage and capacity to the expectedly affected areas before the users encounter diminished quality of service in those areas. Predicting spectrum usage via throughput can also enable regulators to ensure efficient spectrum management by helping decision makers to better plan spectrum allocations. This work endeavours to develop such capability and is part of the initiative toward the implementation of the Spectrum Environment Awareness (SEA) prototype system at the Communications Research Centre (CRC) Canada [1].

The problem of missing data or missingness is encountered in many types of research [2–5]. Missingness diminishes the quality of the data and adversely impacts the analysis carried out based on such data. This problem has led to extensive research on developing methods for missing data imputation. Data imputation attempts to restore the missing values in a dataset by analyzing the characteristics, rules, relationships, etc., hidden within the dataset. Inappropriate imputation of missing data could lead to incorrect data analysis, false conclusions, and erroneous predictions.

We experience the issue of missing data in LTE spectrum measurements collected in 2016 employing a Rohde & Schwarz R&S®TSMA mobile network scanner [6] placed in the downtown Ottawa area. Parameters such as the number

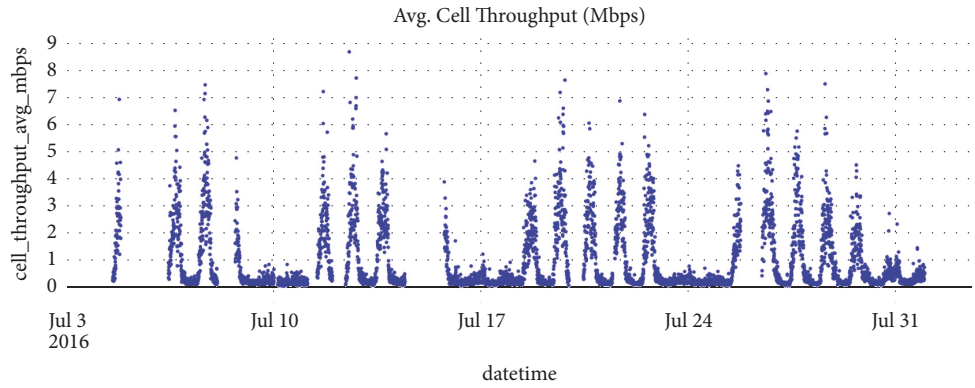

Figure 1: Average cell throughput (Mbps).

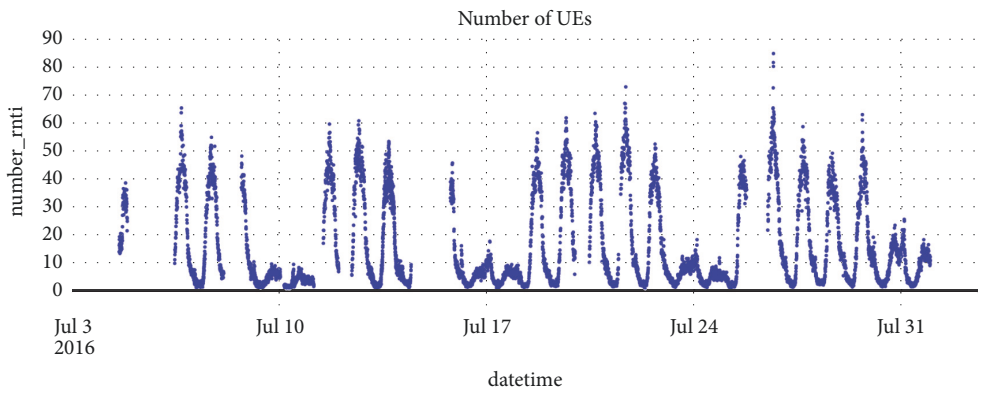

Figure 2: Number of UEs.

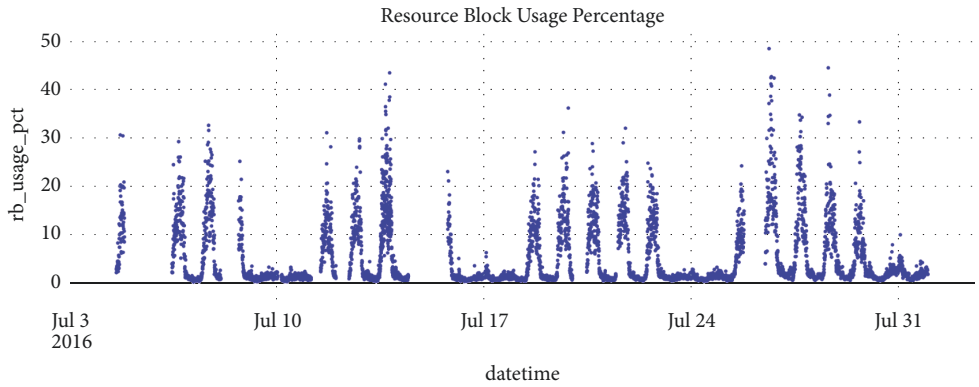

Figure 3: Resource block usage percentage.

of user equipments (UEs), the downlink resource block usage percentage, and the average downlink cell throughput are included in the sensed LTE data among other parameters. Our main aim is to predict the average cell throughput for a week by using its time series values collected from the previous three weeks. To achieve this, we first impute the missing values of the average cell throughput that is shown in Figure 1.

We concentrate our study on the cell possessing physical cell ID (PCI) 65 operating at a 2117.5 MHz center frequency with a 15 MHz channel bandwidth, although the collected LTE data cover several frequency channels. This cell is selected because of the highest number of UEs observed in this cell by the scanner. A high degree of correlation is found between the average cell throughput, number of UEs, and the resource block usage percentage, which can be seen in Figures 1, 2, and 3, respectively. However, the data for these three variables are missing at exactly the same time intervals during the four weeks of July 2016 from July 4th to July 31st, which is clearly visible from Figures 1, 2, and 3.

Multivariate imputation by chained equations (MICE) [7] is a popular approach for imputation of missing data. In

case of three correlated variables X1, X2, and X3, with X1 having missing values, the default multivariate imputation methodology of MICE regresses the observed values of X1 on the other two variables X2 and X3. The imputed values, which are simulated draws from the posterior predictive distribution of X1, then replace the missing values of X1 [8]. Although MICE is able to impute the missing values of the variable, average cell throughput in our case, by regressing its observed values on the other two variables, i.e., the number of UEs and the resource block usage percentage, the resulting imputation is not accurate, as the observed values of the average cell throughput are missing at exactly the same time intervals as the other two variables.

In this work, we propose a novel method of converting a single variable in the LTE spectrum dataset, i.e., the average cell throughput, into a multivariate form in order to successfully apply an existing multivariate imputation approach, such as MICE. To deal with large periods of missing data by exploiting the high seasonality and random missingness of this variable, our method breaks the average cell throughput over twenty-eight days into four separate weeks where each week is considered as a separate variable. These variables are then used as inputs to MICE. Preliminary work in this regard has been presented in [9].

We use our proposed method in combination with MICE and compare its performance in terms of imputation and prediction accuracy with Kalman filtering [10], which is a single imputation approach for single variate data available in the R-package *imputeTS* [11] as R-function *na.kalman*. We also compare the performance of our proposed method with the default multivariate imputation method of MICE, where imputations for average cell throughput are generated using the other two variables, i.e., number of UEs and resource block usage percentage. To compare performance in terms of prediction, we use the times series linear modeling with trend and seasonality components (TSLM) [12]. This prediction approach is available in the R-package *forecast* [13]. The imputation accuracy is measured in terms of the weekly mean of the average cell throughput and the standard error of weekly mean whereas the prediction accuracy is measured in terms of the mean absolute percentage error (MAPE) [14]. The results clearly indicate that the proposed method significantly outperforms the Kalman and default methods in terms of imputation accuracy as well as prediction accuracy.

The rest of the document is organized as follows. Section 2 discusses the related work on imputation and prediction approaches and prediction methods for LTE. Details of data collection, dataset parameters, and missing data analysis in the time period of interest are described in Section 3. The proposed method of converting univariate data into a multivariate form is presented in Section 4. Section 5 gives an in-depth performance evaluation. Section 6 concludes the work and provides some directions for future work.

2. Related Work

Data imputation techniques have long been considered for data analysis and prediction applications in biology [2], electrical power networks [3], city mobility studies [4], and wireless sensor networks [5]. Restoring the important characteristics of a dataset through the addition of plausible values in lieu of the missing data is the objective of data imputation [15]. Missing data mechanisms can be divided into three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR), where most of the missingness scenarios, including our own, belong to MAR.

The missing value is replaced by a value based on the observed subject's characteristics in single imputation. The imputed values are presumed to be the real values that would be there if the data had been complete [16]. A good review of single imputation techniques as well as their pros and cons can be found in [17].

Single imputation may lead to biased estimates of parameters, such as means and correlations. To take imputation uncertainty into consideration, the imputation process is repeated several times to produce multiple imputed datasets; this leads to the concept of multiple imputations. Different datasets are created based on a random draw from different estimated underlying distributions [18]. The uncertainty of the missing values is reflected by the differences between these datasets.

Time is, in fact, an implicit variable although univariate is considered as one column of observations. Single imputation techniques are applicable to univariate imputation. Due to its various tools' development, potential to reduce biased results, and effectiveness in data imputation applications, multivariate multiple imputation has become a very popular approach for missing data analysis as compared to univariate imputation. It maintains the benefits of effective imputation while permitting the statistical uncertainty being carried. Well-known tools that have been developed for multiple imputations include MICE, multiple imputation (mi) [19], Amelia II [20], and Hmisc [21]. These tools function as containers to enable various imputation approaches to be administered separately to each individual variable (represented as one column of observations) aiming at generating plausible replacements for the missing values. Incorporated prevalent imputation approaches include predictive mean matching (PMM) [22], random forests (RF) [23], classification and regression trees (CART) [24], and linear regression prediction method (LRP) [7].

MICE has become known as a trustworthy tool for addressing missing data. In MICE, an imputation approach such as PMM, RF, or LRP is required to be defined for each incomplete variable. In this work, we use PMM for all variables, which is the default imputation approach in MICE. A useful feature of MICE is its capability to identify the set of predictor variables to be considered for each incomplete variable; this is accomplished by setting up the predictor matrix. MICE has been used for missing data imputation in several research areas including research on solar radiation [25], traffic sensing and monitoring [26], and galactic cosmic rays [27].

Prediction finds applications in various circumstances such as weather forecasting, economic trend analysis, dynamic spectrum management, road traffic monitoring, and power consumption estimation. Good imputation helps for a more feasible and accurate prediction. Existing prediction

Table 1: Analysis of weekly data for average cell throughput [9].

Week #	N	Mean	Standard Error of Mean	95% CI for Mean – Lower	95% CI for Mean – Upper	Width of CI
Week 1	1173	0.8562	0.0370	0.7837	0.9287	0.1450
Week 2	1411	0.7878	0.0315	0.7259	0.8496	0.1236
Week 3	1851	0.9079	0.0295	0.8500	0.9659	0.1159
Week 4	1787	0.8634	0.0281	0.8083	0.9185	0.1102

approaches are numerous. Examples include exponential smoothing [28], Poisson process-based forecasting [29], and autoregressive integrated moving average (ARIMA) [30] with seasonal and nonseasonal models. A hybrid prediction approach in [31] models the periodic (or seasonal) component using a trigonometric regression function after the data is decomposed into periodic and residual parts. A probabilistic modeling approach for short-term prediction based on the space-time diurnal (ST-D) method is proposed in [32], which combines the spatial and temporal travel time information for short-term prediction of travel times. TSLM [12] is used to fit linear models to time series including trend and seasonality components. It allows having variables "trend" and "season", which are created from the characteristics of the time series data. The variable "trend" is a simple time trend and "season" is a factor indicating the season, e.g., day, week, month, etc., depending on the frequency of the data. The prediction with TSLM is generated using R-function *forecast* on the fit generated using R-function *tslm*.

The LTE network produces huge amounts of measurements and diagnosis data. Research attention has been attracted to using existing approaches to evaluate, analyze, and predict the LTE network capacity and performance. A traffic-measurement-based modeling method has been proposed in [33] to look for relationships between LTE network resources and key performance indicators for resource-consumption forecasting based on traffic and service growth. The work in [34] uses ARIMA to forecast the average downlink throughput to anticipate cell congestion in LTE networks. Time series modeling is used in [35] to forecast LTE resource consumption to identify unused resources. Random forests are used in [36] to predict LTE link bandwidth by utilising past throughput and lower layer information. All above works dealt with complete data (without missingness) in parametric predictions and trend forecasts. In [37], a novel technique has been proposed for handling missing values in an LTE-like system for wireless telemetry applications. While this technique works in scenarios with a low percentage of missing data, its effectiveness has not been proven in cases with a high proportion of missing data. Missing data recovery, i.e., imputation, plays a key role in prediction and forecasting accuracy. Our work addresses the high-percentage-missing-data imputation problem and the prediction of LTE spectrum measurements data at the same time with a novel method of improving imputation and prediction accuracy for univariate LTE spectrum data.

3. LTE Spectrum Data

In order to obtain LTE spectrum usage data, we carried out a measurement campaign using a Rohde & Schwarz R&S®TSMA mobile network scanner [6]. This scanner was equipped with the R&S®ROMES4 software [38], to perform measurements over the frequency range of 350 MHz to 4400 MHz. The scanner is capable of detecting LTE cells with their center frequencies and corresponding channel bandwidths as well as measuring the usage of the cells. Following are some of the parameters that are reported as part of our LTE spectrum dataset: MCC (Mobile Country Code), MNC (Mobile Network Code), PCI (Physical Cell ID), number of UEs in the cell (i.e., number of unique RNTI (Radio Network Temporary Identifier) identified in the cell), resource block (RB) usage percentage of the cell, and average cell throughput in Mbps.

Although the LTE spectrum data was collected over several months in 2016, we focus our analysis on July, since it is discovered to be the month with the highest number of observations collected by the scanner. Table 1 exhibits the weekly mean, standard error of the weekly mean, 95% confidence interval (CI) of the weekly mean, and width of the CI, for the variable of interest in this work: average cell throughput of the cell (shown in Figure 1), having PCI 65 operating at 2117.5 MHz center frequency with a 15 MHz channel bandwidth, measured by the scanner during the time period from July 4th to July 31st. In this table, the parameter N indicates the number of observations that were collected in a week for the average cell throughput. The standard error of the weekly mean is computed as $\sigma/\sqrt{N}$, where σ is the standard deviation of the average cell throughput in a week.

An aggregated value for the average cell throughput is obtained every 5 minutes, resulting in 288 total observations per day and 2016 total observations per week. An analysis of the weekly data for the average cell throughput is also displayed in Figure 4. As can be observed from Figure 4(a) as well as from column N of Table 1, Week 1 has the highest while Week 3 has the lowest amount of missing data. The percentage of the available data for the average cell throughput over the four weeks is discovered to be 77.16%; it is computed as the ratio of the sum of number of observations in four weeks in column N to the sum of total number of observations in four weeks and then multiplying this ratio by 100, i.e., ((1173 + 1411 + 1851 + 1787) / (2016 + 2016 + 2016 +2016)) × 100. The percentage of the missing data for the average cell throughput over the four weeks is therefore found to be 22.84%.

Figure 4(b) illustrates the pattern of the weekly data for the average cell throughput, where the blue cells show the availability (or presence) and the yellow cells show the missingness (or absence) of the data. The observed combinations of data presence and absence across the four weeks are displayed. For example, the bottom row represents the percentage of cases where the data was present for all the

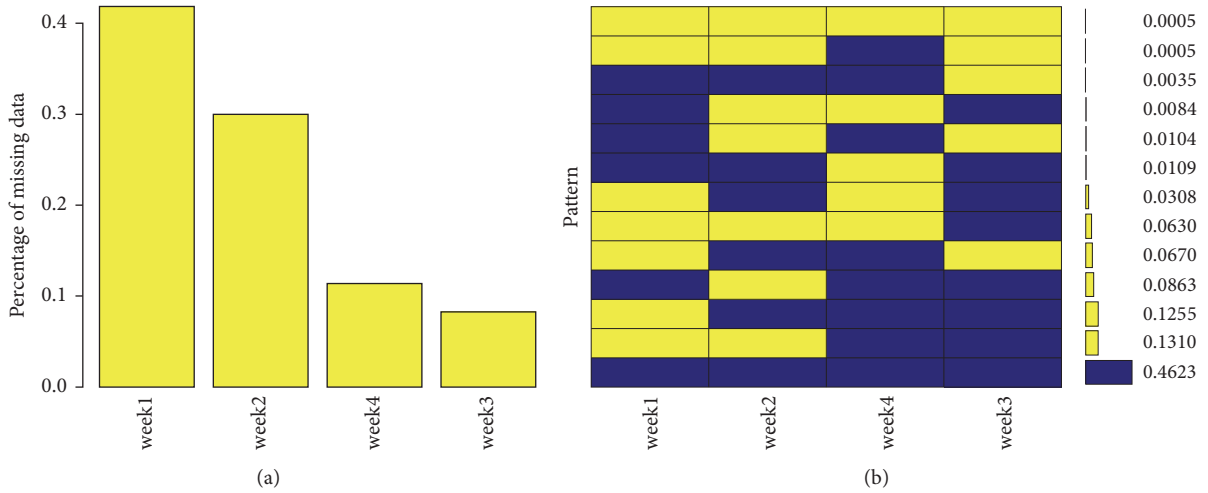

Figure 4: Analysis of weekly data for average cell throughput: (a) percentage of missing data, (b) pattern of weekly data.

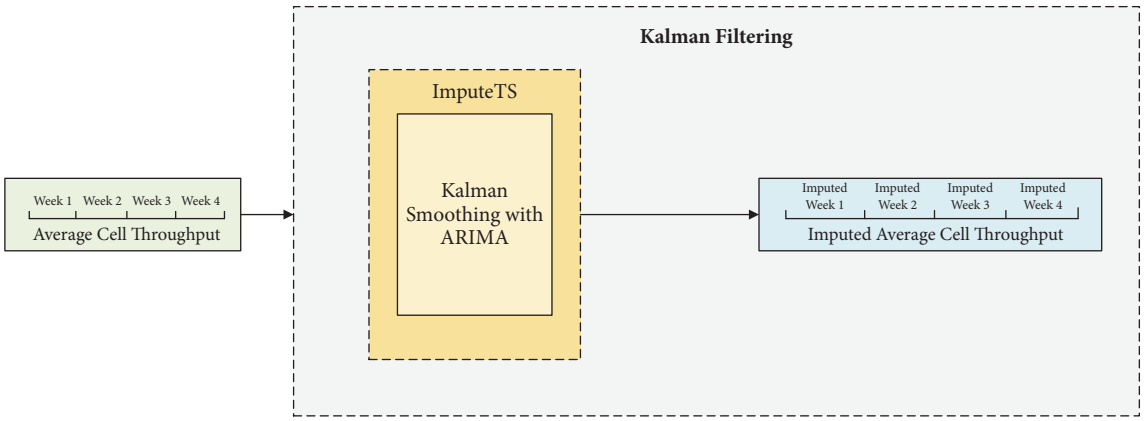

Figure 5: Methodology of Kalman filtering.

weeks simultaneously, which constituted 46.23% of the data while the percentage of the cases when the data was missing for all the weeks simultaneously (represented by the top row) was 0.05%. Note that 0.05% equals to just one observation out of the total number of 2016 per week.

4. Proposed Method

In this section, first we discuss the imputation methodology of some well-known imputation approaches with respect to their application to the LTE dataset. These approaches include the univariate Kalman filtering and the classic multivariate default method of MICE. We then introduce our proposed imputation method.

Kalman filtering, defined as Kalman smoothing on the state space representation of an ARIMA model, is usually considered a good approach for imputation of highly seasonal univariate data [39]. The imputation methodology of Kalman filtering is shown in Figure 5. Kalman filtering is not able to provide a reasonable imputation for the variable average cell throughput due to the large periods of the missing values, although this variable has high daily and weekly seasonality as depicted in Figure 1.

On the other hand, the default method shown in Figure 6, which is a well-known multivariate imputation approach used by MICE, cannot provide an accurate imputation for the average cell throughput variable by using the number of UEs and the resource block usage percentage as the missing values of the average cell throughput occur at exactly the same time intervals as the other two variables. This has motivated us to develop a new method that provides better imputations for the average cell throughput when used in combination with the existing multivariate imputation mechanism of MICE. As illustrated in Figure 6, the multivariate multiple-imputation approach of MICE generates multiple different imputations for the average cell throughput.

Periods of missing data in a week are less probable to occur at the same time slots in other weeks considering the randomness of a univariate data such as the average cell throughput. This brings us the intuition of an appropriate method of using the powerful multivariate imputation technique such as MICE while being able to overcome the simultaneous data missing problem in multivariate imputation.

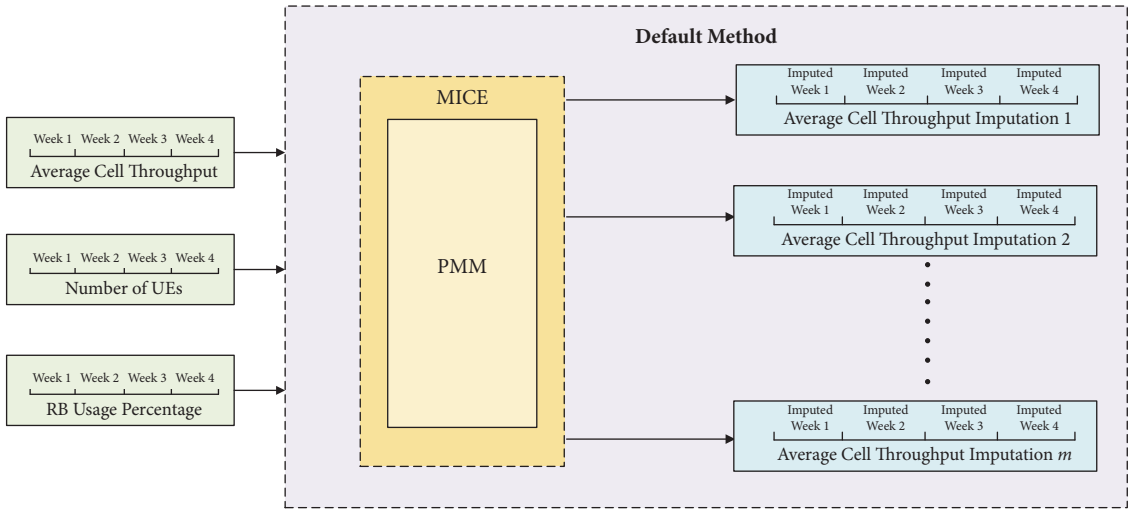

Figure 6: Methodology of default method.

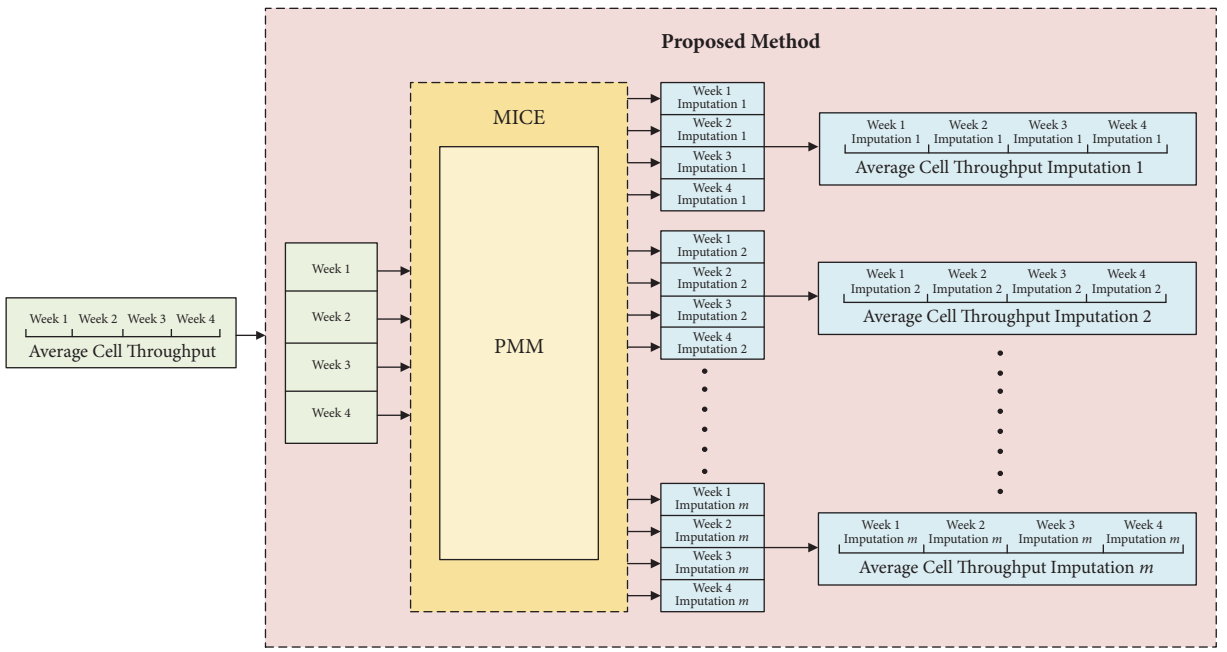

Figure 7: Methodology of proposed method.

We propose a novel method of converting a single variable's data in the LTE spectrum dataset, such as the average cell throughput, into a multivariate form by exploiting the high weekly seasonality of this variable before applying multivariate imputation. Our method breaks the average cell throughput of 28 days into four separate weeks in order to make MICE applicable for imputation of this variable. Each week is considered a separate variable; this results in four variables, explicitly Week 1, Week 2, Week 3, and Week 4. These variables (or weeks) are highly correlated due to the high weekly seasonality of the average cell throughput and each of these variables has some missing values. They are then used as input to MICE for multivariate imputation where its default imputation approach, i.e., PMM, is selected for all four variables. The predictor matrix in MICE is set up such that Week 2, Week 3, and Week 4 are utilised to impute the missing values of Week 1; Week 1, Week 3, and Week 4 are utilised to impute the missing values of Week 2, and so on. The resulting imputed variables are combined together in the final step to produce a single imputed variable, culminating in the imputed average cell throughput. The imputation methodology of the proposed method is shown in Figure 7.

MICE incorporates four main steps for multivariate imputation as described in [40]:

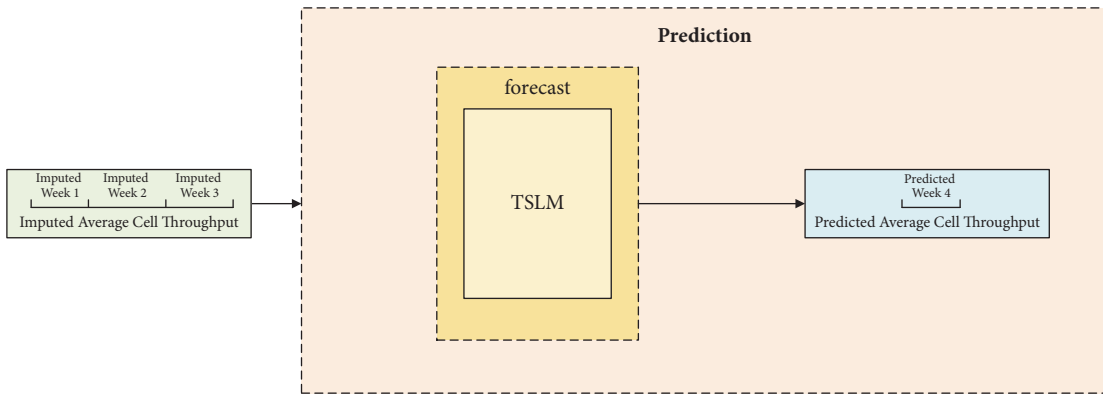

Figure 8: Methodology of prediction.

Step 1. Each week's missing data is first imputed using the mean observed value of that week as a temporary setting of the missing value.

Step 2. The imputed mean values of Week 1 are set back to missing.

Step 3. A linear regression of Week 1 projected by Week 2, Week 3, and Week 4 is acquired using all cases where Week 1 was observed.

Step 4. Imputations of missing Week 1 values are deduced from that regression.

At this stage Week 1 does not have missing values anymore. Steps 2–4 are then reiterated for Week 2. The imputed mean values of Week 2 are set back to missing and a linear regression of Week 2 predicted by Week 1, Week 3, and Week 4 is generated using all cases with observed Week 2. Imputations for missing Week 2 values are obtained from that regression equation. In a similar manner, imputations are obtained for the missing values of Week 3 and Week 4. The observed and the imputed values are finally combined into one complete data set.

5. Performance Evaluation

We compare the performance of our proposed method with Kalman filtering (a univariate imputation approach) and the default method (a classical multivariate imputation approach used by MICE) in terms of imputation and prediction accuracy. The variable in the LTE spectrum dataset that is imputed and predicted for this comparison is the average cell throughput of the cell operating at 2117.5 MHz center frequency as detailed in Section 3.

The proposed and the default methods use the multivariate multiple-imputation approach of MICE and its default imputation approach, i.e., PMM, to generate five different imputations for the average cell throughput. The efficiency of an estimate based on m imputations respective to the one based on an infinite number is calculated as $(1 + \lambda/m)^{-1}$ [41]; here λ denotes the rate of missing information and is 0.2284 in our case as derived in Section 3. An efficiency of 95.6% can be attained with five imputations. In other words, five imputations will generate estimates for mean, standard error of mean, and confidence interval of mean which are 95.6% as efficient (or accurate) as those based on an infinite number of imputations.

To demonstrate the imputation accuracy, we generate plots of the imputed average cell throughput by the three methods as well as tables exhibiting the weekly means of all five imputations of the average cell throughput, standard error of the weekly mean, 95% CI of the weekly mean, and width of the CI. We show plots for one imputation, i.e., imputation #3, for the sake of brevity. Better imputation accuracy is signified by a smaller variation in the weekly means and smaller standard errors of weekly means.

TSLM is the prediction approach that is used to compare the performance of Kalman, the default method, and the proposed method in terms of prediction. The prediction accuracy of the three methods is compared using MAPE, which is calculated as [42]:

$$MAPE = \frac{1}{h}\sum_{t=1}^{h}\left|\frac{A_t - P_t}{A_t}\right| \times 100, \quad (1)$$

where A_t is the actual value, P_t is the predicted value, and h is the number of predicted values. We use three weeks of imputed average cell throughput from the 4th of July, 2016 to the 24th of July, 2016 to predict the fourth week from 25th of July, 2016 to 31st of July, 2016. The prediction methodology is shown in Figure 8. For MAPE calculation, the imputed average cell throughput of the fourth week is used as the actual value, and h is set to 2016, which is the number of observations in a week. The prediction accuracy is also illustrated using plots showing actual (or imputed) vs. predicted average cell throughput.

5.1. Results – Imputation. The imputed average cell throughput generated using Kalman filtering is illustrated in Figure 9. Figure 10 displays one of the five imputations of the average cell throughput when using the default method and Figure 11 shows the same imputation of the average cell throughput when using the proposed method.

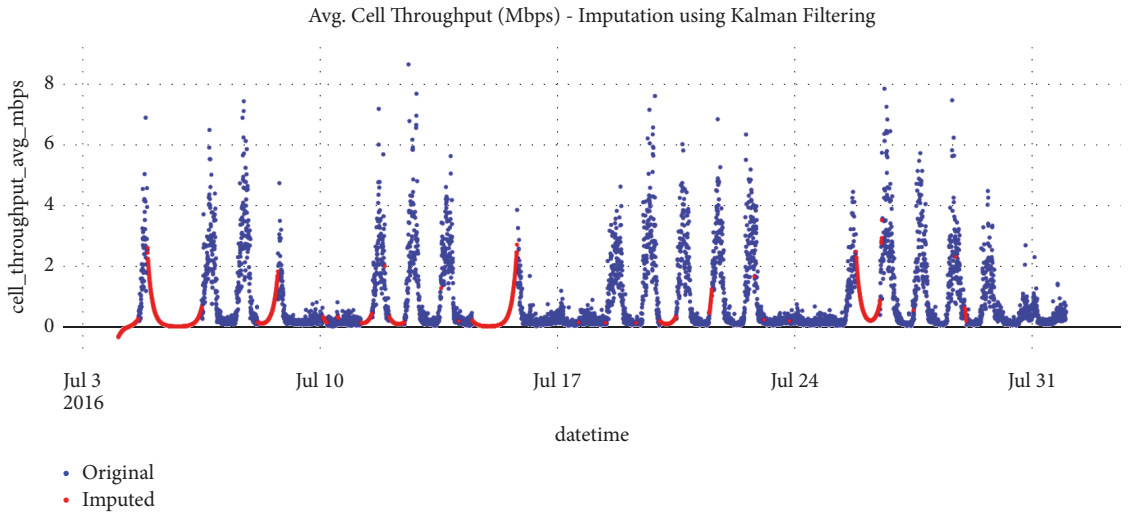

Figure 9: Imputed average cell throughput using Kalman filtering.

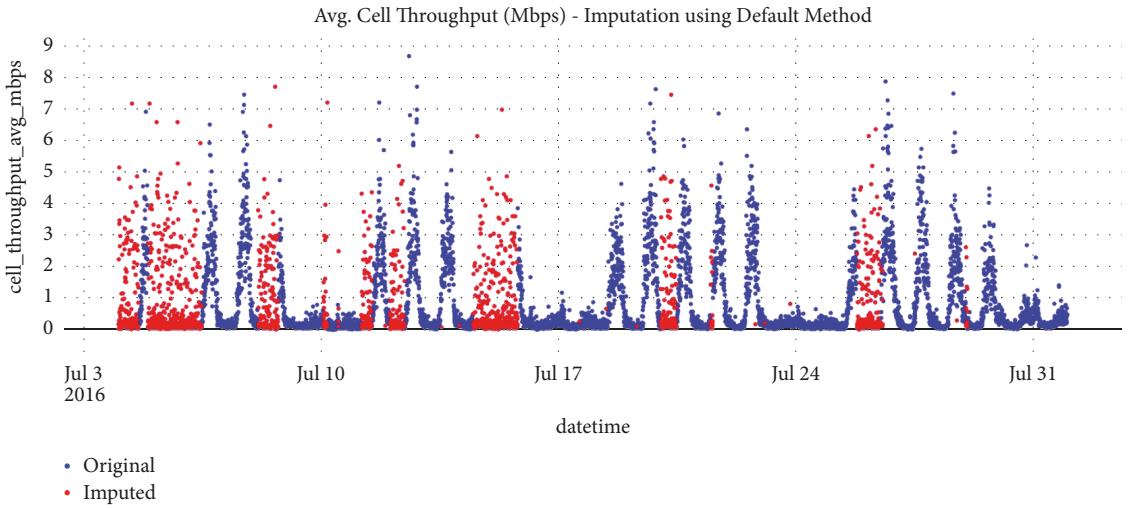

Figure 10: Imputed average cell throughput using default method.

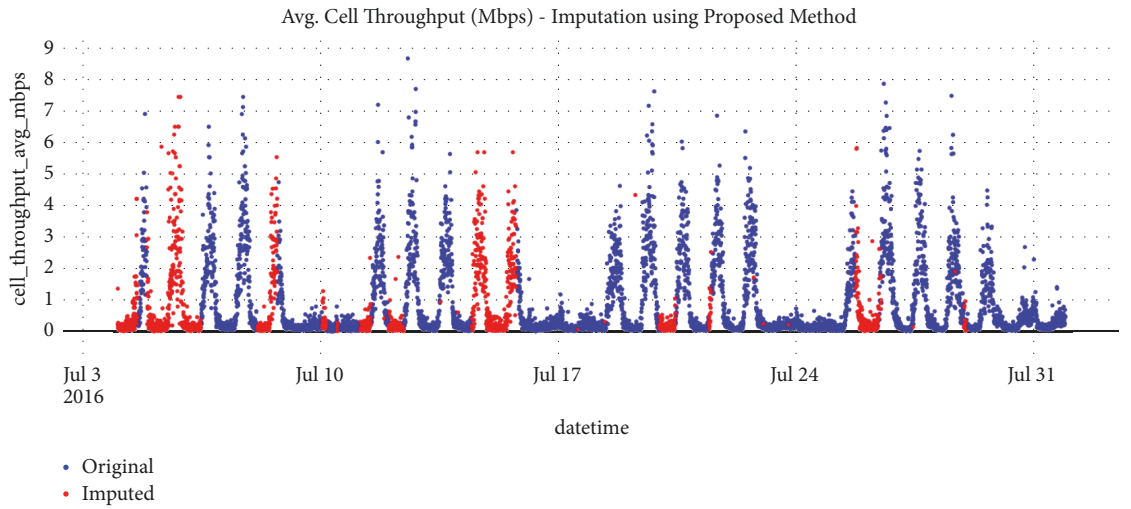

Figure 11: Imputed average cell throughput using proposed method.

Table 2: Analysis of the imputation with Kalman filtering [9].

Week #	N	Mean	Standard Error of Mean	95% CI for Mean – Lower	95% CI for Mean – Upper	Width of CI
Week 1	2016	0.5966	0.0234	0.5508	0.6424	0.0916
Week 2	2016	0.6253	0.0233	0.5795	0.6711	0.0916
Week 3	2016	0.8526	0.0275	0.7987	0.9065	0.1078
Week 4	2016	0.8401	0.0254	0.7903	0.8899	0.0995

Table 3: Analysis of all imputations with default method.

Week #	N	Mean	Standard Error of Mean	95% CI for Mean – Lower	95% CI for Mean – Upper	Width of CI
Week 1	10080	0.8560	0.0124	0.8316	0.8803	0.0487
Week 2	10080	0.8010	0.0117	0.7780	0.8240	0.0460
Week 3	10080	0.9018	0.0126	0.8771	0.9265	0.0494
Week 4	10080	0.8633	0.0119	0.8400	0.8865	0.0465

Table 4: Analysis of all imputations with proposed method [9].

Variable	N	Mean	Standard Error of Mean	95% CI for Mean – Lower	95% CI for Mean – Upper	Width of CI
Week 1	10080	0.8331	0.0127	0.8081	0.8580	0.0499
Week 2	10080	0.8125	0.0118	0.7893	0.8357	0.0464
Week 3	10080	0.8557	0.0123	0.8316	0.8799	0.0483
Week 4	10080	0.8346	0.0115	0.8119	0.8572	0.0453

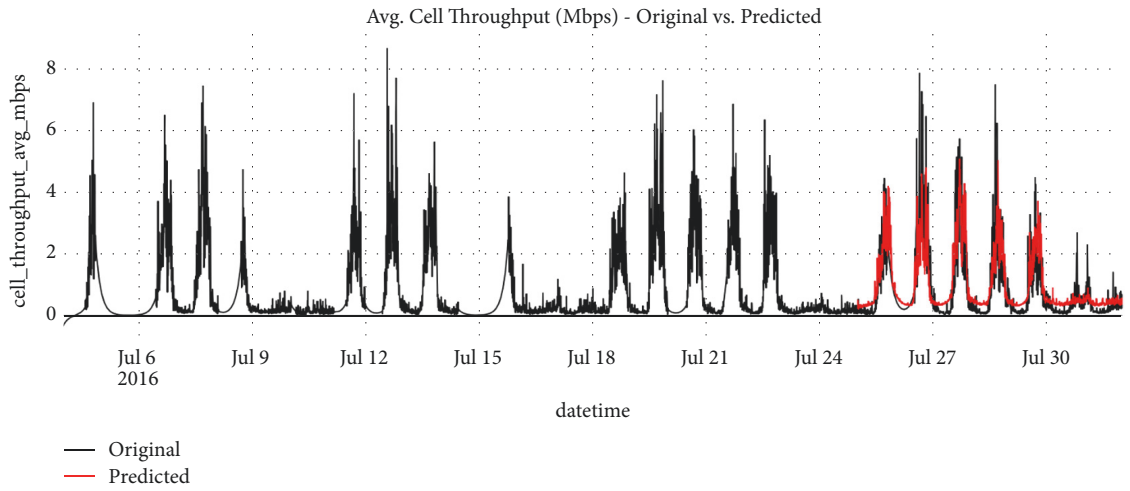

Figure 12: Average cell throughput, original vs. predicted with Kalman filtering.

The weekly analysis of the imputation generated with Kalman filtering is listed in Table 2 in terms of the weekly mean of the imputed average cell throughput, the standard error of the weekly mean, the 95% CI of the weekly mean, as well as the width of the CI.

Table 3 shows this weekly analysis for all five imputations of the average cell throughput when using the default method. To generate this weekly analysis, we sequentially combine the five different imputations of the average cell throughput variable in one column, divide this column into four separate datasets representing four different weeks by selecting the appropriate values for the date and time column in the dataset, and then calculate the mean, the standard error of the mean, the 95% CI of the mean, and the width of the CI for the average cell throughput variable in each week.

Table 4 shows this weekly analysis for all five imputations of the average cell throughput when using the proposed method. For each of the four weekly variables, namely, Week 1, Week 2, Week 3, and Week 4, five different imputations are generated. After these imputations are sequentially combined in a single column for a week, the mean, the standard error of the mean, the 95% CI of the mean, and the width of the CI for the average cell throughput is calculated for that weekly variable.

5.2. Results – Prediction. Figures 12, 13, and 14 show the original vs. predicted average cell throughput when using TSLM with Kalman filtering, TSLM with the default method (DM), and TSLM with the proposed method (PM), respectively. By "original" average cell throughput in these figures, we mean

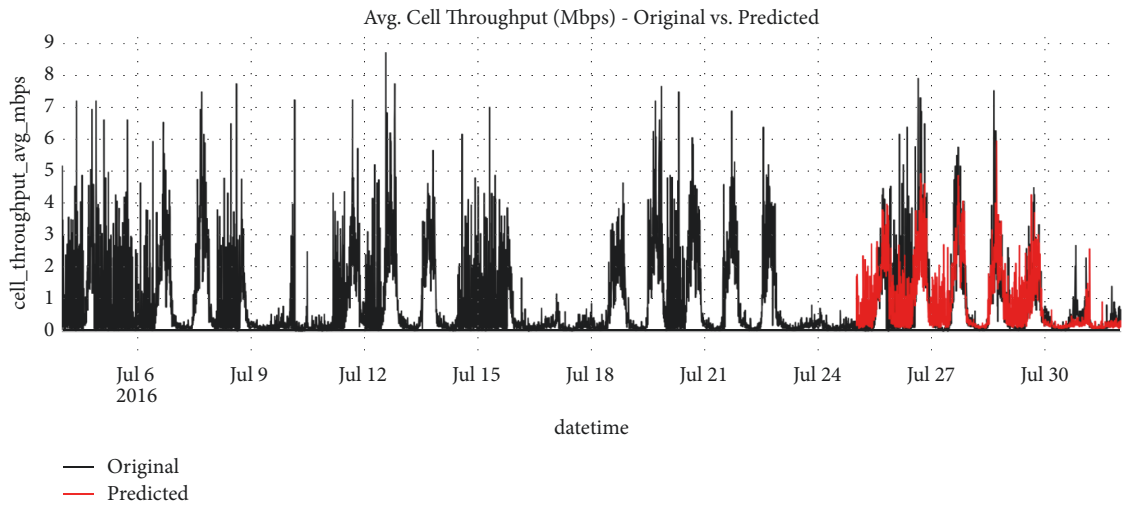

Figure 13: Average cell throughput, original vs. predicted with default method.

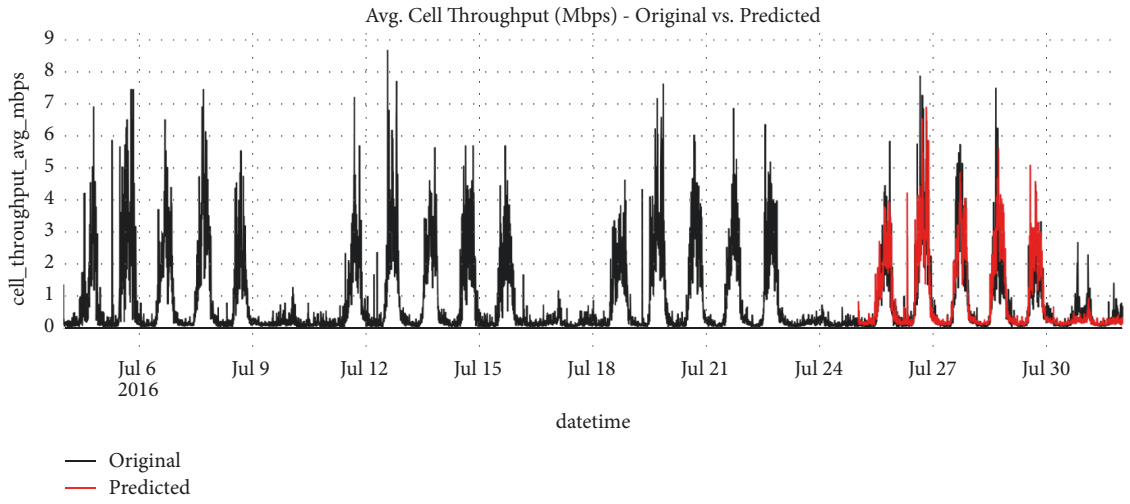

Figure 14: Average cell throughput, original vs. predicted with proposed method.

Table 5: Comparison of prediction accuracy for Kalman filtering, DM, and PM.

Prediction Method	MAPE		
	Kalman Filtering	Default Method	Proposed Method
TSLM	142.01	209.65	63.69

the average cell throughput imputed using Kalman filtering, the default method, and the proposed method, respectively. Three weeks of imputed average cell throughput from 4th to 24th of July, 2016, are used to predict the fourth week from 25th to 31st of July, 2016.

Table 5 shows the comparison of prediction accuracy for the Kalman filtering, default, and proposed methods in terms of MAPE. Note that for the default and proposed methods, the comparison of prediction accuracy was carried out for all five imputations generated. For the sake of brevity, the comparisons shown in Figures 12–14 and Table 5 are based on one imputation.

5.3. Discussion. Imputation using neither Kalman filtering nor the default method is able to recreate the daily seasonality of the average cell throughput and cannot recreate the missing daily peaks, as can be seen from Figures 9 and 10. Kalman filtering uses neighboring values of the missing values while the default method uses other variables for imputation. The proposed method performs significantly better than the Kalman and default methods as can be seen from Figure 11. By exploiting the weekly seasonality of the average cell throughput, it recovers the daily seasonal patterns and imputes the missing daily peaks.

Reflecting poor imputation accuracy, Table 2 exhibits a high variation in the weekly means of the average cell

throughput with Kalman imputation. A smaller variation of the weekly means is observed with the default method as can be seen from Table 3. The smallest variation of weekly means is achieved with the proposed method as shown in Table 4, which indicates that the proposed method clearly outperforms Kalman filtering and the default methods in terms of imputation accuracy. Lower values for the standard error of weekly mean are observed for the default method as well as the proposed method in comparison with Kalman, as can be seen from the corresponding values in Tables 3 and 4.

When Kalman filtering is used for imputation, the prediction accuracy of the prediction approach, namely, TSLM, is poor when compared with the proposed method, as can be seen from the higher value of MAPE for Kalman filtering in Table 5. TSLM is not able to provide a good fit for the predicted average cell throughput, which can be seen from Figure 12 due to poor imputation of the average cell throughput with Kalman filtering shown as "original" in this figure.

Compared to the proposed method, the prediction accuracy observed with the default method is very low, which is reflected by the high MAPE value of the default method in Table 5 and its poor fit with TSLM illustrated by the predicted average cell throughput in Figure 13. The proposed method provides the best prediction accuracy when compared with Kalman filtering and the default method, which is reflected by the lower MAPE value of the proposed method in Table 5. As can be seen from Figure 14, TSLM is able to provide a much better prediction with the proposed method, as compared to those with the Kalman and default methods.

6. Conclusions

We proposed a novel method for improving imputation and prediction accuracy that converts univariate data into a multivariate form applicable for multivariate imputation. The proposed method has been tested for imputing a single variable (having high seasonality and large periods of missingness) in an LTE spectrum dataset, i.e., the average cell throughput. The high weekly seasonality of this variable is leveraged by the proposed method as well as the fact that missing data is less likely to occur in different weeks at the same time. The results show that the proposed method significantly outperforms Kalman filtering and the default multivariate method in terms of imputation and prediction accuracy. This method can be useful for imputing any univariate data that is highly seasonal and has large periods of missingness. It provides an efficient way for automatic data imputation in big data analytics.

Imputation using neither Kalman filtering nor the default method is able to recreate the missing daily peaks whereas the proposed method reinstates these peaks by exploiting the weekly seasonality of the average cell throughput. The smallest variation of weekly means is achieved with the proposed method when compared with the Kalman and default methods, which highlights the imputation accuracy of the proposed method. The proposed method also provides the best prediction accuracy when compared with the Kalman and default methods, which is reflected by the lower MAPE value of the proposed method.

For future work, we plan to evaluate the prediction accuracy of our proposed method in the presence of other prediction approaches, such as, seasonal decomposition of time series by LOcally wEighted regreSsion Smoother (LOESS), a.k.a. STL [43]; trigonometric series for seasonality with Box-Cox transformation, autoregressive moving-average errors, trend and seasonal components (TBATS) [44]; and Holt-Winters filtering [45].

The proposed and the default methods use the multiple-imputation approach of MICE to generate five different imputations for the average cell throughput. For the sake of brevity, we showed plots of one of the five imputations to compare performance. Similarly, the results shown for comparison of prediction accuracy were based on the same imputation. Future work includes a plan to develop a method for imputation selection that can pick that imputation from among multiple imputations that may deliver the best prediction accuracy.

Data Availability

The LTE measurement data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] W. Li, M. Wang, Y. Qiu, Y. Ge, D. Kidston, and A. Chaudhry, "Integration of intelligent tasking subsystem in spectrum environment awareness applications," in *Proceedings of the 12th IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '17)*, pp. 1–5, Cagliari, Italy, June 2017.

[2] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[3] C. C. Turrado, F. S. Lasheras, J. L. Calvo-Rollé, and A. J. Piñón-Pazos, "A new missing data imputation algorithm applied to electrical data loggers," *Sensors*, vol. 15, no. 12, pp. 31069–31082, 2015.

[4] S. Nikfalazar, C.-H. Yeh, S. Bedingfield, and H. A. Khorshidi, "A new iterative fuzzy clustering algorithm for multiple imputation of missing data," in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ '17)*, Naples, Italy, July 2017.

[5] J. Q. Lin, H. C. Wu, and S. C. Chan, "A new regularized recursive dynamic factor analysis with variable forgetting factor for wireless sensor networks with missing data," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '17)*, pp. 1–4, Baltimore, MD, USA, May 2017.

[6] Rohde & Schwarz (R&S), "TSMA Autonomous Mobile Network Scanner – Walk and Drive Testing with Flexible Connectivity," https://cdn.rohde-schwarz.com/au/TSMA_bro_en_3607-1513-12_v0900.pdf.

[7] S. van Buuren and K. Groothuis-Oudshoorn, "Mice: multivariate imputation by chained equations in R," *Journal of Statistical Software* , vol. 45, no. 3, pp. 1–67, 2011.

[8] D. S. Bouhlila and F. Sellaouti, "Multiple imputation using chained equations for missing data in timss: a case study," *Large-scale Assessments in Education*, vol. 1, no. 4, pp. 1–33, 2013.

[9] A. Chaudhry, W. Li, A. Basri, and F. Patenaude, "On improving imputation accuracy of LTE spectrum measurements data," in *Proceedings of the Wireless Telecommunications Symposium (WTS '18)*, pp. 1–7, Phoenix, AZ, USA, April 2018.

[10] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, UK, 1989.

[11] S. Moritz, "Package 'imputeTS' – Time Series Missing Value Imputation," https://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf.

[12] A. I. McLeod, H. Yu, and E. Mahdi, "Time Series Analysis with R," *Handbook of Statistics–Time Series Analysis: Methods and Applications*, vol. 30, pp. 661–712, 2012.

[13] R. J. Hyndman et al., "Package 'forecast'," https://cran.r-project.org/web/packages/forecast/forecast.pdf.

[14] J. Boudoukh, M. Richardson, and R. F. Whitelaw, "The best of both worlds: a hybrid approach to calculating value at risk," *RISK*, vol. 11, no. 4, pp. 64–67, 1998.

[15] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo, *Missing Data: A Gentle Introduction*, Guildford Press, New York, NY, USA, 2007.

[16] M. N. Norazian Ramli, A. S. Yahaya, N. A. Ramli, N. F. F. M. Yusof, and M. M. A. Abdullah, "Roles of imputation methods for filling the missing values: A review," *Advances in Environmental Biology*, vol. 7, no. 12, pp. 3861–3869, 2013.

[17] I. Eekhout, R. M. De Boer, J. W. R. Twisk, H. C. W. De Vet, and M. W. Heymans, "Missing data: A systematic review of how they are reported and handled," *Epidemiology*, vol. 23, no. 5, pp. 729–732, 2012.

[18] S. Van Buuren, *Flexible Imputation of Missing Data*, Chapman & Hall/CRC, New York, NY, USA, 2012.

[19] Y.-S. Su, A. Gelman, J. Hill, and M. Yajima, "Multiple imputation with diagnostics (mi) in R: Opening windows into the black box," *Journal of Statistical Software*, vol. 45, no. 2, pp. 1–31, 2011.

[20] J. Honaker, G. King, and M. Blackwell, "Amelia II: a program for missing data," *Journal of Statistical Software* , vol. 45, no. 7, pp. 1–47, 2011.

[21] F. E. Harrell Jr et al., "Hmisc: Harrell Miscellaneous," R package version 4.0-3, 2017, https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf.

[22] L. R. Landerman, K. C. Land, and C. F. Pieper, "An empirical evaluation of the predictive mean matching method for imputing missing values," *Sociological Methods & Research*, vol. 26, no. 1, pp. 3–33, 1997.

[23] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study," *American Journal of Epidemiology*, vol. 179, no. 6, pp. 764–774, 2014.

[24] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests," *Psychological Methods*, vol. 14, no. 4, pp. 323–348, 2009.

[25] C. C. Turrado, M. D. C. M. López, F. S. Lasheras, B. A. R. Gómez, J. L. C. Rollé, and F. J. D. C. Juez, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors*, vol. 14, no. 11, pp. 20382–20399, 2014.

[26] K. Henrickson, Y. Zou, and Y. Wang, "Flexible and robust method for missing loop detector data imputation," *Transportation Research Record*, vol. 2527, pp. 29–36, 2015.

[27] R. C. Fernandes, P. S. Lucio, and J. H. Fernandez, "Data imputation analysis for Cosmic Rays time series," *Advances in Space Research*, vol. 59, no. 9, pp. 2442–2457, 2017.

[28] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*, Springer-Verlag, Berlin, Germany, 2008.

[29] T. Mahmud, M. Hasan, A. Chakraborty, and A. K. Roy-Chowdhury, "A poisson process model for activity forecasting," in *Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP '16)*, pp. 3339–3343, September 2016.

[30] Z. H. Munim and H.-J. Schramm, "Forecasting container shipping freight rates for the Far East-Northern Europe trade lane," *Maritime Economics & Logistics*, vol. 19, no. 1, pp. 106–125, 2017.

[31] Y. Zou, X. Hua, Y. Zhang, and Y. Wang, "Hybrid short-term freeway speed prediction methods based on periodic analysis," *Canadian Journal of Civil Engineering*, vol. 42, no. 8, pp. 570–582, 2015.

[32] Y. Zou, X. Zhu, Y. Zhang, and X. Zeng, "A space-time diurnal method for short-term freeway travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 33–49, 2014.

[33] S. Hu, Y. Ouyang, Y. Yao, M. H. Fallah, and W. Lu, "A study of LTE network performance based on data analytics and statistical modeling," in *Proceedings of the 23rd Wireless and Optical Communication Conference*, Newark, NJ, USA, May 2014.

[34] P. Torres, P. Marques, H. Marques et al., "Data analytics for forecasting cell congestion on LTE networks," in *Proceedings of the Network Traffic Measurement and Analysis Conference (TMA '17)*, pp. 1–6, Dublin, Ireland, June 2017.

[35] S. Albasheir and M. Kadoch, "Enhanced Control for Adaptive Resource Reservation of Guaranteed Services in LTE Networks," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 179–189, 2016.

[36] C. Yue, R. Jin, K. Suh, Y. Qin, B. Wang, and W. Wei, "LinkForecast: Cellular Link Bandwidth Prediction in LTE Networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1582–1594, 2018.

[37] V. Zarikas et al., "Wireless telemetry: channel characterization and statistical imputation of missing values," *Communications*, vol. 58, no. 1, pp. 58–65, 2015.

[38] Rohde & Schwarz (R&S), "ROMES4 Drive Test Software – Mobile Coverage and QoS Measurements in Mobile Networks," https://scdn.rohde-schwarz.com/ur/pws/dl_downloads/dl_common_library/dl_brochures_and_datasheets/pdf_1/ROMES4_broen_5214-2062-12_v1400.pdf.

[39] S. Moritz and T. Bartz-Beielstein, "imputeTS: Time series missing value imputation in R," *The R Journal*, vol. 9, no. 1, pp. 207–218, 2017.

[40] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?" *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011.

[41] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, NY, USA, 1987.

[42] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric MAPE," *International Journal of Forecasting*, vol. 15, no. 4, pp. 405–408, 1999.

[43] M. Theodosiou, "Forecasting monthly and quarterly time series using STL decomposition," *International Journal of Forecasting*, vol. 27, no. 4, pp. 1178–1195, 2011.

[44] A. M. De Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1513–1527, 2011.

[45] D. Tikunov and T. Nishimura, "Traffic prediction for mobile network using Holt-Winter's exponential smoothing," in *Proceedings of the 15th International Conference on Software, Telecommunications and Computer Networks*, pp. 1–5, Split-Dubrovnik, Croatia, September 2007.