International Journal of Science, Environment and Technology, Vol. 8, No 3, 2019, 430 – 453

ISSN 2278-3687 (O)
2277-663X (P)

EFFICIENCY OF IMPUTATION TECHNIQUES IN UNIVARIATE TIME SERIES

Sampson Twumasi-Ankrah[1*], Benjamin Odoi[2], Wilhemina Adoma Pels[1] and Eric Herrison Gyamfi[1]

[1]Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

[2]Department of Mathematical Sciences, University of Mines and Technology, Tarkwa, Ghana

E-mail: stankrah2017@gmail.com ([*]*Corresponding Author*)

Abstract: In this paper, we are interested in two main issues concerning how missing values should be treated in univariate time series. Firstly, three different error metrics are examined to know which one is appropriate for the different characteristics in univariate time series data in context of imputation techniques. Secondly, the performance of nine different imputation techniques with respect to the two main missing data imputation mechanisms (namely missing at random (MAR) and missing completely at random (MCAR)) are considered. Four original datasets exhibiting different features in time series data are used. We use different missing rate values ranging from 10% to 90% at equal interval of 10, assuming both MAR and MCAR. For the first objective, it is observed that the appropriate error metric for datasets having both trend and seasonality and also dataset with trend but no seasonality, is the MAPE. However, the RMSE is the appropriate error metric measure for data that exhibits very high seasonality but no trend and also dataset with no trend and no seasonality. For the second objective, the "best" imputation technique for dataset which shows both trend and seasonality is the STL (Seasonal and Trend decomposition using Loess) Based Interpolation ("interp") technique in both MAR and MCAR. Again, when a dataset exhibits seasonality but no trend, the "best" imputation technique is "interp". However, when dataset has trend but not seasonal, the "best" imputation technique with respect to MAR is Kalman and "interpolation" in MCAR. However, it is observed in both MAR and MCAR that, the two "best" imputation techniques for dataset that exhibits seasonality, but no trend are the "mean" and "Replace".

Keywords: Missing values, Imputation Techniques, Missing Data Mechanisms, Time Series and Error Metric.

1. INTRODUCTION

Univariate time series data are measured is almost every domain; for example, biology (Bar-Joseph et al., 2003), finance (Taylor, 2007), and social science (Gottman, 1981). Issues with missing values normally occur at any time that data are measured and recorded. In time series data, a most notable feature is the time-dependent correlations between observations, and this may indicate that a current value depends on the values of past or future observations (Shumway and Stoffer, 2011; Box et al., 2015). Therefore, missing values can lead to serious

*Received April 25, 2019 * Published June 2, 2019 * www.ijset.net*

problems in time series analysis, and there is the need to replace the missing values with reasonable values. In statistics, this process is called imputation.

The type and amount of missing values depend on characteristics of the data, nevertheless missing observations can occur for several reasons (Schafer, 1997; Schafer and Graham, 2002). For example, a more common scenario for time series data is missing observations in 'chunks' or data missing in sequence for a period (Schafer, 1997; Donders et al.,2006). Thus, the accuracy of imputation methods can vary considerably depending on characteristics of the dataset (Yozgatligil et al., 2013).

According to *Marcus* et. al (2018), identifying an appropriate imputation method is often the first step towards more formal time series analysis. Different imputation methods will have differing precision in reproducing missing values. Several studies like Saunders (2006), Jörnsten et al. (2007), Nguyen et al., (2013), Li et al., (2015), Ran et al., (2015), Schmitt et al., (2015), Moritz et al. (2015), and Tak et al.(2016) have used a similar workflow to compare the performance of imputation methods.

There are several challenges for adopting a standardized approach to compare imputation methods. According to Moritz (2017) a list of concerns for a complete benchmark assessment of imputation techniques would include: (1) different missing data percentages, (2) different datasets, (3) different Missing data mechanisms.

Good overview articles comparing different imputation techniques using the complete benchmark assessment with respect to univariate time series are yet missing. Also, according to Yozgatligil et al. (2013), interpretations of the different imputation techniques may be influenced by the choice of error metric (e.g., *RMSE*) as different metrics have different objectives. Thus, in this study, the performances of nine (9) imputation techniques are compared by following the complete benchmark assessment and also by taking into consideration different error metric.

2. METHODS AND MATERIALS

2.1 Data Source and Nature

In this study, the performance of nine imputation methods is compared on four reference time series datasets available in the TSA package (Chan and Ripley, 2012). These datasets are frequently used in literature and exhibit known characteristics which are common in all time series data.

The four datasets used in this study are given below:

- Air passengers: The Air Passengers dataset contains monthly total international airline passengers from 01/1960 - 12/1971 with 144 observations and the frequency of time series as 12. The dataset exhibits both trend and seasonality.
- Beer sales: The dataset contains monthly beer sales in millions of barrels, 01/1975 - 12/1990. The number of observations is 192 with a time series of frequency 12. Beer sales dataset possesses very high seasonality, but no trend.
- SP - Quarterly S&P Composite Index, 1936Q1 - 1977Q4 which has trend, no seasonality
- Google - Daily returns of the google stock from 08/20/04 -9/13/06 which exhibits no trend, no seasonality

2.2 Missing Data Imputation Mechanisms

While considering the potential effect of the missing data, it is vital to consider the fundamental explanations behind why the data are missing. The three types of missing data grouped by Gelman and Hill (2007) are Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). However, we did not take into consideration that of MNAR because studies like King et al. (2001) and Yarandi (2002) suggest that presence or absence of MNAR can hardly be demonstrated using only the observed data.

Missing Completely at Random (MCAR)

When data are MCAR, there is no systematic contrasts that exist between members with missing data and those with complete data. This means the data points in a missing data usually happen totally at random. This implies the probability of missingness of the data is equal for all the data points.

$$\boldsymbol{P(r|Yobserved, Ymissing)} = \boldsymbol{P(r)} \quad (1)$$

Missing at Random (MAR)

MAR assumes that the probability a data point is missing depends only on the available data. At the point when data are MAR, the way that the data are missing is deliberately identified with the observed yet not the unobserved data. This is represented in a probability form as:

$$\boldsymbol{P(r|Yobserved, Ymissing)} = \boldsymbol{P(r|Yobserved)} \quad (2)$$

2.3 Imputation Techniques

Imputation is the way of supplanting missing data points with substituted estimates. When substituting for a single data point, it is called "unit imputation"; when substituting for a part

of a data point, it is called "item imputation". The following imputation techniques are used in the study.

i. Mean Imputation

Essentially, this technique calculates the mean of the observed values of the non-missing values for that variable and replaces any missing value of that variable by the mean for all other cases. This method is from the zoo and impute TS packages in R. It fills the missing values with mean value of a time series using “na.aggregate” or “na.mean”. This technique is estimated using the formula:

$$\hat{\mu}_k = \sum_{i=k} {}^{y_i}/_{n_k} \qquad (1)$$

where y_i is the observed values, n_k is the number of observations

ii. Spline Interpolation Imputation

This technique assumes a linear relationship between data points and utilizes non-missing values from adjacent data points to compute a value for a missing data point. The algorithm of the interpolation requires identifying the previous value (x_0) and subsequent value (x_1)of the non-missing cells observations. It uses impute TS package in R and the algorithm “na.interpolation” is used to replace the missing values. The estimated formula is given as:

$$f(x) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x_1 - x_0) \qquad (2)$$

iii. Last Observation Carried Forward (LOCF)Imputation

In LOCF imputation, orderly datasets are sorted out according to the number of variables. The technique uses algorithm to find the first missing value and uses the non-missing value immediately prior to the data that are missing to impute the missing value. In R, the method uses “zoo” and impute TS packages. The algorithm used is “na.locf”.

iv. Replace Imputation

With this imputation method, missing values are replaced by randomly sampling between two bounds. The default bounds are the minimum and the maximum value in the non-missing values from the time series data. Replace imputation uses “na.replace” to replace missing values in R package called impute TS.

v. Kalman Smoothing Imputation

This imputation method uses Kalman filters to operate on state-space models of the form

$$\begin{aligned} y_t &= Z\alpha_t + \varepsilon_t \varepsilon_t \sim N(0, H) \\ \alpha_{t1} &= T\alpha_t + n_t n_t \sim N(0, Q) \\ \alpha_t &\sim N(a_t, P_t) \end{aligned} \qquad (3)$$

where y_t = observed series (possibly with missing values) andα_t = unobserved

The measurement equation, y_t means the observed data is related to the unobserved states Transition equation, α_{t1}implies the unobserved states evolve over time in a particular way. Kalman filter uses algorithm to find best estimates of α_t.Once the Kalman filter has been applied to the entire time period, you have best estimates of the states a_t, P_tat $t = 1,2, \dots, T$. In simplification, it can be computed as $\hat{y}_t = Za_t$. Kalman Smoothing uses impute TS package in R and "na.kalman" to replace the missing values in R.

vi. Moving Average (MA) Imputation

MA imputation operates using moving average values to replace the missing values. The average in this algorithm is taken from the same number of data points on either side of the central value. This implies for a missing value at position$\boldsymbol{i}$ of a time series data, the observations $\boldsymbol{i-1},\ \boldsymbol{i+1}\ \boldsymbol{and}\ \boldsymbol{i+1},\ \boldsymbol{i+2}$ are used to compute the average. In R, this method uses impute TS package and "na.ma" to replace the missing values.

vii. Least Squares Approximations or Linear Interpolation Imputation

The least square principle minimizes the sum of the squares of the errors for a polynomial of given degree. In R, it uses zoo package or impute TS package. It replaces the missing values with interpolated values using "na.approx". Algorithm of the imputation is given as:

Let $\boldsymbol{k}$ be the number of data points missing in a given time series data such that $\boldsymbol{a_0, a_1, a_2, \dots, a_{k-1}}$be a constant, and $\boldsymbol{y(t)}$ represents the missing observation at time $\boldsymbol{t}$ then consider fitting a polynomial of fixed degree k:

$$\boldsymbol{y(t) = a_0 + a_1 t + a_2 t^2 + \dots + a_{k-1} t^k} \qquad (4)$$

This is computed using the observed values of the time series data. The values of $\boldsymbol{a_0, a_1, a_2, \dots, a_{k-1}}$ are obtained using matrix approach. Missing values are computed at each time, $\boldsymbol{t = 1, 2, 3, \dots, T}$.

viii. Random Imputation

Random imputation selects a simple random sample of size $\boldsymbol{m}$ with replacement from $\boldsymbol{s_r}$and then uses the associated y-values as donors, that is, $\boldsymbol{y_i^* = y_j}$ for some $\boldsymbol{j \in r}$. Random imputation uses imputeTS package in R and supplant the imputed values with the algorithm "na.random". The imputed values for random imputation are estimated as:

$$\boldsymbol{Z_j^* = y_i^* + \hat{S}_j/\hat{T}_j - \hat{S}/\hat{T}} \qquad (5)$$

ix. STL Based Interpolation:

It is a method of imputation which uses linear interpolation for non-seasonal series and a periodic STL decomposition with seasonal series. This method uses impute TS package in R and replaces the missing values with the algorithm "na.interp".

2.4. Error Metric

We compare three error metrics in order to check for their effect on imputation techniques.

i. Root Mean Square Error (RMSE)

RMSE is a measure of spread of the forecast errors about the actual data points. This implies that the RMSE informs how far or near the forecasted values of an estimated model are from the real data points. The formula is given as:

$$RMSE_{forecast} = \sqrt{\sum_{i=1}^{N}\left(\frac{Y_t - \hat{Y}_t}{N}\right)^2} \quad (6)$$

Where $\hat{Y}_t$ is the forecasted values, $Y_t$the actual data points, and $N$is the sample size.

ii. Mean Absolute Percentage Error (MAPE)

MAPE is a measure of the size of error of a forecast in percentage. It is used to measure the accuracy of a forecast using the formula below:

$$\boldsymbol{MAPE_{forecast}} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum\frac{|\boldsymbol{Y_t} - \hat{\boldsymbol{Y}}_{\boldsymbol{t}}|}{|\boldsymbol{Y_t}|}\right) \times \mathbf{100}\% \quad (7)$$

iii. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is the simplest measure of forecast accuracy. That is, the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error is expect from the forecast on average.

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum\left|\boldsymbol{Y_t} - \hat{\boldsymbol{Y}}_{\boldsymbol{t}}\right| \quad (8)$$

2.5 Comparison Test

In this study, the Analysis of Variance (ANOVA) is used to check whether there is significant difference among the three error measures. Mathematically, the model is given as:

$$x_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1, 2, 3; \qquad j = 1, 2, 3, \ldots, 9 \quad (9)$$

where $\mu_i = ith\ treament\ mean, \quad and\ \varepsilon_{ij} = error\ term\ ,\ such\ that\ \varepsilon_{ij} \sim N(0, \sigma^2)$

Hypothesis for the Model

$$H_0\colon \mu_{rmse} = \mu_{mape} = \mu_{mae}$$

$$H_1\colon \mu_i \neq \mu_j, \quad i \neq j, for\ at\ least\ one\ i\ at\ 0.05\ level\ of\ significant$$

where $\mu_{rmse} = population\ mean\ for\ rmse$

$$\mu_{mape} = population\ mean\ for\ mape$$

$$\mu_{mae} = population\ mean\ for\ mae$$

Tukey Multiple Comparison Test

If the ANOVA leads to a conclusion that there is evidence that the error measure means differ, then we will be interested in investigating which of the means are different. The Tukey's multiple comparison test is used for this purpose. Algebraically, the test statistic is given as:

$$HSD = \frac{\mu_i - \mu_j}{\sqrt{\frac{MS_W}{n_h}}},\ \ i \neq j \quad with\ \ i,j = RMSE, MAPE\ and\ MAE \qquad (10)$$

where:

- $\mu_i - \mu_j$is the difference between the pair of meansto calculate this, μ_i should be larger than μ_j.
- MS_w is the Mean Square Within, and n is the number in each error measure.

2.6 Principle of the Analysis

The general principle of the analysis in this study is given as:

i. Four original datasets without missing values exhibiting different features in time series data are used.

ii. Introduce in each data a varying percentage of missing values (ranging from 10% to 90%) in the datasets which are generated under both MAR and MCAR assumptions.

iii. Apply the nine (9) imputation techniques in step (ii)

iv. Estimate the error metric.

3 RESULTS AND DISCUSSION

In this study, there are two issues that are of interest namely (1) the effect of error metric on the performance of imputation techniques, and (2) the performance of imputation techniques with respect to rate of missing values. Different missing percentages (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%)at random and completely at random are extrapolated from each dataset. The imputation methods are moving average, linear interpolation, spline interpolation, STL based interpolation, Kalman smoothing, replace, mean, last observation carried forward, and random imputation.

3.1 ERROR METRICS AND IMPUTATION TECHNIQUES

In this section we compare the three commonly used error metrics namely RMSE, MAPE and MAE of each imputation method on the four time series data. We are interested to know which of the error metrics is better in terms of recording the lowest value. The four different datasets are Air Passengers dataset, Beer sales dataset, SP dataset and Google dataset that

exhibit specific data patterns are considered. The Analysis of Variance (ANOVA) is used to test whether the three error metrics are different and if at least one error metric is different, the Turkey test is used to indicate which one is different.

Case 1: Air Passengers dataset

The Air Passengers dataset contains monthly total international airline passengers from 01/1960 - 12/1971 with 144 observations and the frequency of time series is 12. The dataset exhibits both trend and seasonality. From Table 1, it is observed that there is no difference between the two missing data imputation mechanisms (MAR and MCAR). In terms of the frequency of minimum error in all the nine imputation methods, the MAPE gives the minimum error measure irrespective of the missing data imputation mechanisms (MAR and MCAR), and the RMSE gives the maximum error measure. This implies that, if a dataset has a trend and seasonality, the appropriate error measure is MAPE.

Table 1: Test of difference among Error Metrics regarding Imputation Techniques for Trend and Seasonal Dataset

Imputation Technique		MAR	MCAR
Approximation	ANOVA test Turkey test	Significant RMSE>MAE>MAPE	Significant MAPE=MAE; RMSE=MAE; RMSE>MAPE
Interp	ANOVA test Turkey test	Not Significant RMSE=MAE=MAPE	Not Significant RMSE=MAE=MAPE
Interpolation	ANOVA test Turkey test	Significant RMSE>MAE>MAPE	Significant MAPE=MAE; RMSE=MAE; RMSE>MAPE
Kalman	ANOVA test Turkey test	Significant MAPE=MAE;RMSE= MAE RMSE>MAPE	Significant MAPE=MAE; RMSE=MAE; RMSE>MAPE
LOCF	ANOVA test Turkey test	Significant MAPE=MAE;RMSE= MAE RMSE>MAPE	Significant MAPE=MAE; RMSE=MAE; RMSE>MAPE
MA	ANOVA test Turkey test	Significant MAPE=MAE;RMSE= MAE RMSE>MAPE	Significant MAPE=MAE; RMSE=MAE; RMSE>MAPE
Mean	ANOVA test Turkey test	Significant MAPE=MAE;RMSE= MAE RMSE>MAPE	Significant RMSE>MAE>MAPE
Random	ANOVA test Turkey test	Significant MAPE=MAE;RMSE= MAE RMSE>MAPE	Significant RMSE>MAE>MAPE
Replace	ANOVA test Turkey test	Significant MAPE=MAE;RMSE= MAE RMSE>MAPE	Significant MAPE=MAE; RMSE=MAE; RMSE>MAPE

Case 2: Beer Sales Dataset

The dataset contains monthly beer sales in millions of barrels, 01/1975 - 12/1990. The number of observations is 192 with a time series of frequency 12. Beer sales dataset possesses very high seasonality, but no trend. In Table 2, the RMSE is appropriate, in terms of the error metric that records the highest number of minimum error for data. Moreover, there is no difference between the appropriate error metric in MAR and MCAR on the nine imputation techniques. In other words, the appropriate error metric in MAR is the same in MCAR.

Table 2: Test of difference among Error Metrics regarding Imputation Techniques forno Trend but Seasonal Dataset

Imputation Technique		MAR	MCAR
Approximation	ANOVA test Turkeytest	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE
Interp	ANOVA test Turkey test	Not Significant RMSE=MAE=MAPE	Not Significant RMSE=MAE=MAPE
Interpolation	ANOVA test Turkey test	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE
Kalman	ANOVA test Turkey test	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE
LOCF	ANOVA test Turkey test	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE
MA	ANOVA test Turkey test	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE;MAPE>MAE; MAPE>RMSE
Mean	ANOVA testTurkey test	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE
Random	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE
Replace	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE

Case 3: SP Dataset

The SP dataset contains 168 quarterly observations. For the SP datasetisa series with just trend and no seasonality with a time series of frequency 4. In terms of the frequency of

minimum error in all the nine imputation methods, the MAPE gives the minimum error measure irrespective of MAR and MCAR as presented in Table 3. In other words, if a data points has no trend but seasonal irrespective of the data MAR and MCAR, the MAPE is the appropriate error metric.

Table 3: Test of difference among Error Metrics regarding Imputation Techniques forTrend but no Seasonal Dataset

ImputationTechnique		MAR	MCAR
Approximation	ANOVA test Turkeytest	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
Interp	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
Interpolation	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
Kalman	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
LOCF	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
MA	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
Mean	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
Random	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE
Replace	ANOVA test Turkey test	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE	Significant MAPE=MAE; RMSE=MAE: MAPE>RMSE

Case 4: Google Dataset

This dataset contains 521 daily returns of the google stock observations. The dataset exhibits no trend and seasonality. In Table 4, the RMSE is appropriate, in terms of the error metric

that records the highest number of minimum error for data that exhibit no trend and seasonal. Moreover, there is no difference between data MAR and MCAR on the nine imputation techniques in terms of the error measures, as both give the same results.

Table 4: Test of difference among Error Metrics regarding Imputation Techniques for no Trend and Seasonal Dataset

Imputation Technique		MAR	MCAR
Approximation	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE; MAPE>RMSE
Interp	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
Interpolation	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
Kalman	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
LOCF	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
MA	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
Mean	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
Random	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE
Replace	ANOVA test Turkey test	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE	Significant RMSE=MAE; MAPE>MAE MAPE>RMSE

Discussion

It is observed that the choice of error metric depends on the characteristics or nature of the dataset, which is consistent with Yozgatligil et al (2013). Again, the choice of a specific error

metric is not affected or influenced by the missing data imputation mechanisms (i.e., MAR and MCAR).

3.2 EFFECT OF MISSING RATE ON IMPUTATION TECHNIQUES

Here, the performance of the imputation techniques across the various rate of missing value in both the MAR and MCAR are discussed by taking into consideration the four time series datasets. The error metric used for each dataset is based on the recommended error metric from section (3.1). The "best" imputation technique is the technique that records the minimum error.

Case 1: Air Passengers Dataset

For the Air Passengers dataset, the performance of imputation techniques is assessed based on both MAR and MCAR respectively.

i. Performance of Imputation Techniques in context of MAR

From Table 5, Kalman smoothing technique has the minimum error when the missing rate value is 10%. When the missing rate values are 20%, 30%, 40%, 50%, 60% and 70% , the STL based interpolation ("interp") turns out to be the "best" imputation technique followed by both Spline interpolation ("interpolation") and Linear interpolation ("Approximation"). And when the missing rate values are80% and 90%, the "interpolation" and "approximation" techniques are both the "best". Thus, when using a dataset which has both trend and seasonality, the appropriate imputation techniques when missing rate value is more than 10% but less than 80%is the "interp". As expected, the performances decreased with increasing rate of missing values. The MAPE error measure is used since it turned out to be the appropriate error metric for the Air passengers' dataset in section (3.1).

Table 5: Comparison of various Imputation techniques of Air Passengers dataset for Different Missing Percentage Values of MAR. Smaller values are better. Best values are shown in boldface

	Imputation Techniques								
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	1.0963	0.6981	1.0963	1.4984	4.5091	9.7222	1.4022	8.7569	**0.6311**
20	2.6629	**2.1699**	2.6629	2.8739	12.0711	20.1388	3.1960	16.9395	3.2325
30	4.0747	**3.3312**	4.0747	4.5189	15.7437	29.1388	4.4097	23.4111	4.9955
40	5.2610	**4.0715**	5.2610	8.8984	17.3980	40.2778	6.6583	35.9330	14.0812
50	6.9293	**4.5412**	6.9293	11.2950	17.8848	50.0000	7.9976	31.4600	15.7243
60	9.3211	**7.4596**	9.3211	13.8254	29.3640	59.7222	12.0113	41.5586	19.7670
70	10.1820	**7.2878**	10.1820	20.3285	27.6030	70.1389	16.1084	48.4679	22.9813
80	**10.7155**	11.3473	**10.7155**	24.4388	47.2396	79.8611	16.2255	58.0892	22.3409
90	**14.7817**	514733.3	**14.7817**	36.6730	54.0400	90.2778	26.2666	57.4916	24.1137

Figure 1 shows the average performances of each technique as a function of the percentage of missing values for the Air Passenger dataset, the "interp" technique is more effective and reliable.

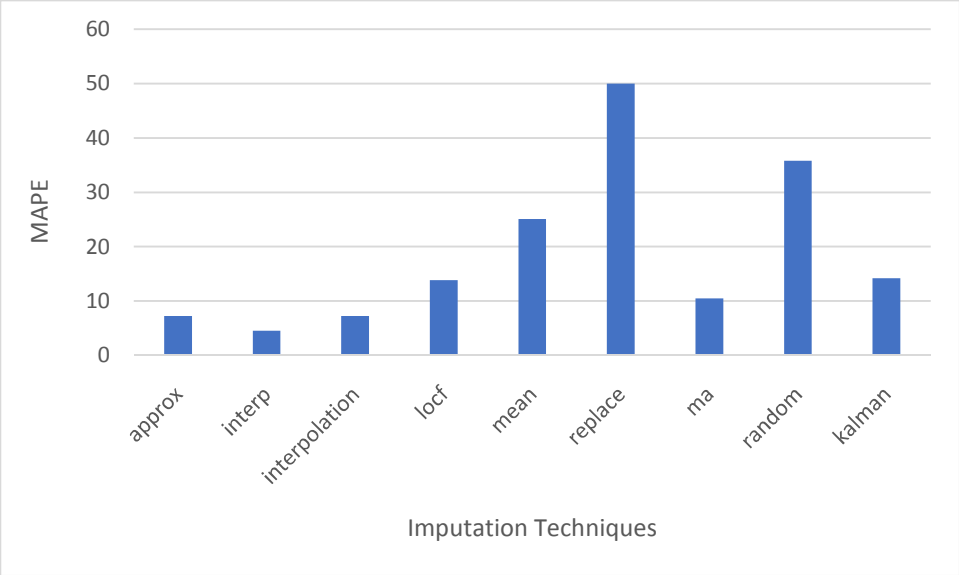

Figure 1: Average performances of each technique as a function of the percentage of missing values for MAR

ii. Performance of Imputation Techniques in context of MCAR

In Table 6, the Kalman smoothing technique is the "best" technique when the missing rate values are 10%, 30%, 40%, 60%, 70% and 90%; however, when the missing rates are 20%, 50%and 80%, the STL based interpolation ("interp") turns to be the "best" technique.

Table 6: Comparison of various Imputation techniques of Air Passengers dataset for Different Missing Percentage Values of MCAR. Smaller values are better. Best values are shown in boldface

Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	0.4926	0.4500	0.4926	0.8914	4.4046	9.7222	0.7049	8.0754	**0.3694**
20	1.3194	**0.8544**	1.3194	1.9647	8.2054	20.1388	1.6021	14.4600	0.8608
30	2.0412	1.4846	2.0412	3.0438	12.5089	29.8611	2.4685	26.2194	**1.2556**
40	3.1630	2.3670	3.1630	4.8259	17.8852	40.2778	3.6040	27.9494	**1.8148**
50	4.1159	**3.0560**	4.1159	5.9580	21.9658	50.0000	4.7345	41.7852	2.8230
60	5.5529	4.4212	5.5529	7.7605	26.6148	59.7222	6.0549	33.1277	**3.6374**
70	7.7845	6.2467	7.7845	10.3239	31.8065	70.1389	8.2222	62.6914	**5.3945**
80	9.8623	**9.0150**	9.8623	12.5332	34.5866	79.8611	10.5260	49.3437	7.6280
90	12.2395	111.7466	12.2395	16.1712	42.9713	90.2778	14.1882	46.5005	**11.5774**

Generally, the Kalman technique gives more effective and reliable imputation as compared to the rest of the imputation techniques followed by "interp" as given in Figure 2.

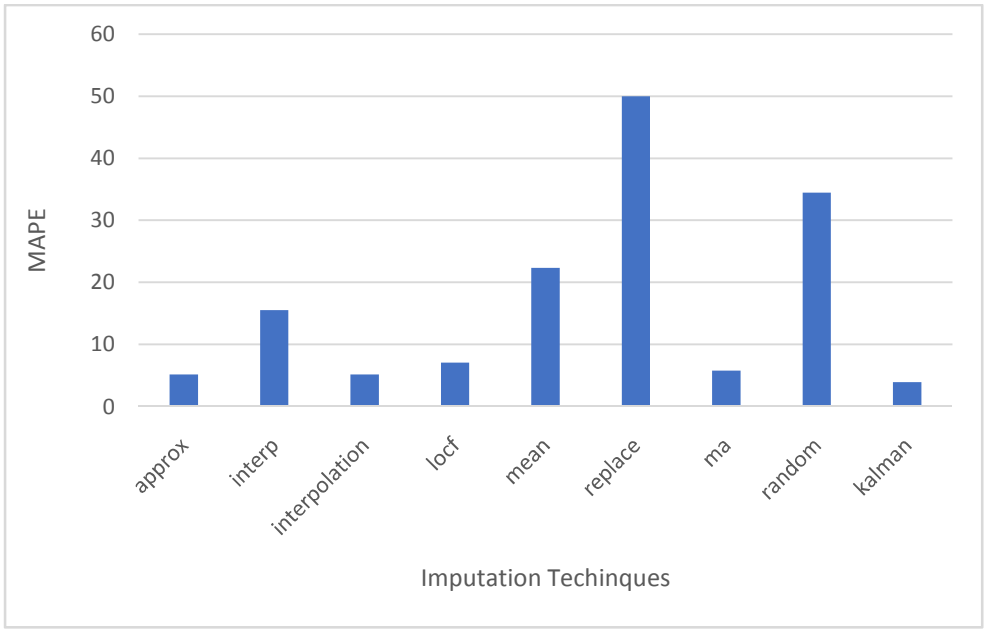

Figure 2: Average performances of each technique as a function of the percentage of missing values for MCAR

Discussion

The type of missing data imputation mechanisms (i.e. MAR and MCAR) has effect on the performance of imputation techniques if the dataset shows both trend and seasonality. However, we recommend the usage of "interp" technique since it's the "best" in MAR and second "best" in MCAR.

Case 2: Beer sales Dataset

The performance of imputation techniques is assessed based on MAR and MCAR respectively**.**

i. Performance of Imputation Techniques in context of MAR

It is observed in Table 7 that, the STL based interpolation ("interp") technique turns to be the "best" imputation technique for missing rate values of 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80%. However, when the missing rate value is 90%, the "mean" imputation technique is the "best". The RMSE error measure is used since it turned out to be the appropriate error metric for the Beer sales dataset.

Table 7: Comparison of Imputation techniques of Beersales dataset for Different Missing Percentage Values of MAR. Smaller values are better. Best values are shown in boldface

Imputation Techniques									
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	0.7050	**0.2110**	0.7050	0.7191	0.5662	4.4869	0.7313	0.8986	0.5183
20	0.9403	**0.2996**	0.9403	1.0331	0.8509	6.4573	1.1557	1.2550	1.0174
30	1.1077	**0.3575**	1.1077	1.3191	1.0403	8.0430	1.2316	1.6238	2.1836
40	1.3620	**0.4241**	1.3620	1.5755	1.1858	9.2531	1.5771	1.7687	3.5069
50	1.6657	**0.4945**	1.6657	1.8728	1.3730	10.3134	1.6819	2.1319	5.5153
60	1.7434	**0.5596**	1.7434	1.9648	1.5294	11.2459	1.8661	2.5349	2.3075
70	1.8892	**0.6564**	1.8892	2.4253	1.6739	12.0757	2.1094	2.1411	4.4204
80	1.9518	**0.8963**	1.9518	2.5934	1.9163	13.0430	2.2518	2.7878	4.6630
90	2.0106	4.2622	2.0106	2.2459	**1.9724**	13.7073	2.2292	2.8548	1.9983

Generally, the "interp" technique gives more effective and reliable imputation as compared to the rest of the imputation techniques followed by "mean" technique as given in Figure 3.

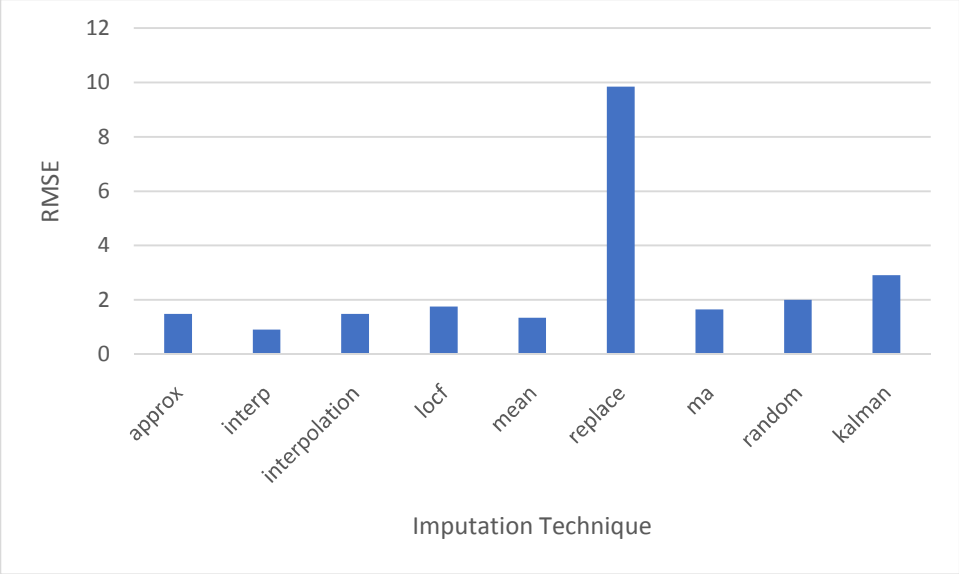

Figure 3: Average performances of each technique as a function of the percentage of missing values for MAR

ii. Performance of Imputation Techniques in context of MCAR

The Kalman Smoothing technique performs better when the missing rate value is 10% as given in Table 8. However, the STL based interpolation ("interp") technique is the "best" when the missing rate values are 20%, 30%, 40%, 50%, 60%, 70% and 80%. But the "mean" imputation technique turns out to be better than the other imputation techniques when the missing rate value is 90%.

Table 8: Comparison of Imputation techniques of Beer sales dataset for Different Missing Percentage Values of MCAR. Smaller values are better. Best values are shown in boldface

	Imputation Techniques								
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	0.2995	0.1883	0.2995	0.4447	0.6050	4.4578	0.3617	0.8140	**0.2182**
20	0.4163	**0.2908**	0.4163	0.6232	0.8362	6.4491	0.4697	1.3493	0.3048
30	0.5040	**0.3651**	0.5040	0.8333	1.0563	7.9039	0.6076	1.6617	0.4830
40	0.6877	**0.4272**	0.6877	1.1138	1.1745	9.1287	0.7605	1.6031	0.5649
50	0.8139	**0.4728**	0.8139	1.3420	1.3316	10.1462	0.8944	1.9776	0.6413
60	1.0609	**0.5264**	1.0609	1.6377	1.4651	11.1476	1.1221	2.3677	0.8189
70	1.3391	**0.6010**	1.3391	1.8210	1.6030	12.1023	1.4004	2.1902	1.0169
80	1.7591	**0.7015**	1.7591	2.0561	1.7246	12.9011	1.8481	2.6830	1.3370
90	2.0573	2.2030	2.0573	2.2693	**1.9367**	13.7324	2.1659	2.3967	1.9570

In general, when using a dataset with no trend but seasonality, the best imputation technique to use is STL based interpolation ("interp") technique followed by "Kalman" technique as presented in Figure 4.

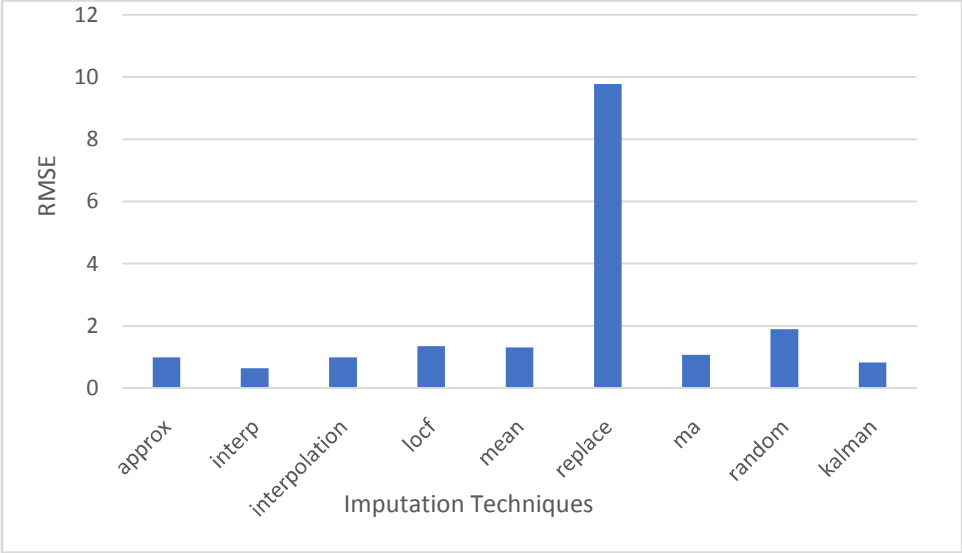

Figure 4: Average performances of each technique as a function of the percentage of missing values for MCAR.

Discussion

It is observed that, when a dataset exhibit seasonality but no trend, the "best" imputation technique is "interp", which is consistent in the two missing imputation mechanisms used in this study (MAR and MCAR).

Case 3: SP Dataset

Here, **t**he performance of imputation techniques is assessed based on MAR and MCAR respectively.

i. Performance of Imputation Techniques in context of MAR

In Table 9, the STL based interpolation ("interp") technique performs better when the missing rate value is 10%. When the missing rate values are 20% and 30%, the Linear interpolation and Spline interpolation techniques are boththe "best". However, the Kalman Smoothing technique is the "best" when the missing rate values are 40%, 50%, 60% and 70%, but the MA and LOCF imputation techniques out-perform the other imputation techniques when the missing rates are 80% and 90%. The RMSE error measure is used since it turned out to be the appropriate error metric for the SP dataset.

Table 9: Comparison of Imputation techniques of SP dataset for Different Missing Percentage Values of MAR. Smaller values are better. Best values are shown in boldface

	Imputation Techniques								
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	1.1799	**1.1649**	1.1799	1.4515	10.7564	10.1190	1.3839	16.2074	1.2356
20	**2.7089**	2.7420	**2.7089**	5.2371	27.0965	20.2380	3.6439	22.8191	2.7316
30	**4.5183**	4.5918	**4.5183**	7.1473	46.8892	29.7619	5.9210	67.1472	4.6187
40	6.5520	6.6846	6.5520	10.9559	52.8414	39.8810	8.6183	47.2231	**5.8161**
50	10.3784	10.8661	10.3784	20.1552	70.8732	50.0000	15.3378	82.1144	**10.0897**
60	17.4580	18.0137	17.4580	27.1571	94.2860	60.1190	22.3276	109.357	**14.8443**
70	22.6442	23.0059	22.6442	30.4834	89.8170	70.2381	25.6093	101.682	**19.0399**
80	31.8485	31.9486	31.8485	43.2201	109.876	79.7619	**30.2077**	128.942	39.3749
90	40.8670	41.2950	40.8670	**38.8007**	111.191	89.8810	34.7910	125.769	38.5512

In general, when using a dataset with trend but no seasonality, the best imputation technique to use is Kalman technique followed by both "interpolation" and "approximation" techniques as presented in Figure 5.

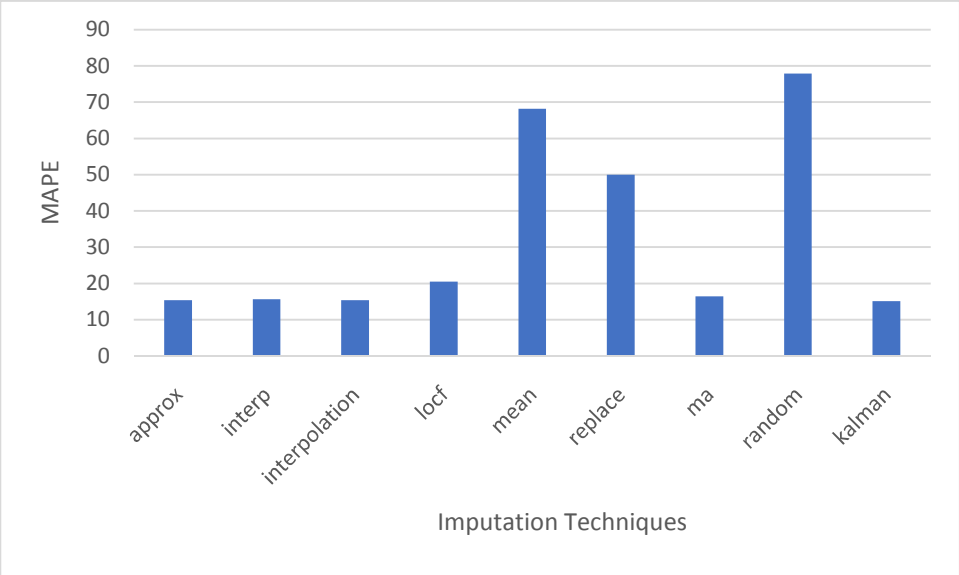

Figure 5: Average performances of each technique as a function of the percentage of missing values for MAR

ii. Performance of Imputation Techniques in context of MCAR

It is observed in Table 10 and Figure 6that; both the Linear interpolation and Spline interpolation techniques are both the "best" imputation techniques for all missing rate values when the dataset exhibits trend but no seasonality.

Table 10: Comparison of Imputation techniques of SP dataset for Different Missing Percentage Values of MCAR. Smaller values are better. Best values are shown in boldface

Imputation Techniques									
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	**0.4538**	0.5073	**0.4538**	0.7945	14.0157	10.1190	0.5579	20.9030	0.5176
20	**0.9323**	1.0220	**0.9323**	1.5111	24.8411	20.2381	1.0949	28.3356	1.0293
30	**1.4372**	1.6295	**1.4372**	2.3409	41.7323	29.7619	1.7303	66.3067	1.7429
40	**2.1126**	2.5336	**2.1126**	3.3176	52.4387	39.8810	2.4123	72.9251	2.4273
50	**2.5896**	3.6392	**2.5896**	4.5501	61.6904	50.0000	2.9093	79.4410	2.9528
60	**3.8125**	6.5237	**3.8125**	6.1447	84.9893	60.1190	4.2654	81.0018	4.4095
70	**5.4460**	7.9865	**5.4460**	9.0318	84.2485	70.2381	6.3110	99.6542	6.6948
80	**7.2243**	10.3867	**7.2243**	12.0574	107.363	79.7619	8.8596	147.625	8.1981
90	**11.0131**	19.0426	**11.0131**	17.4646	123.184	89.8810	13.1469	173.387	12.6083

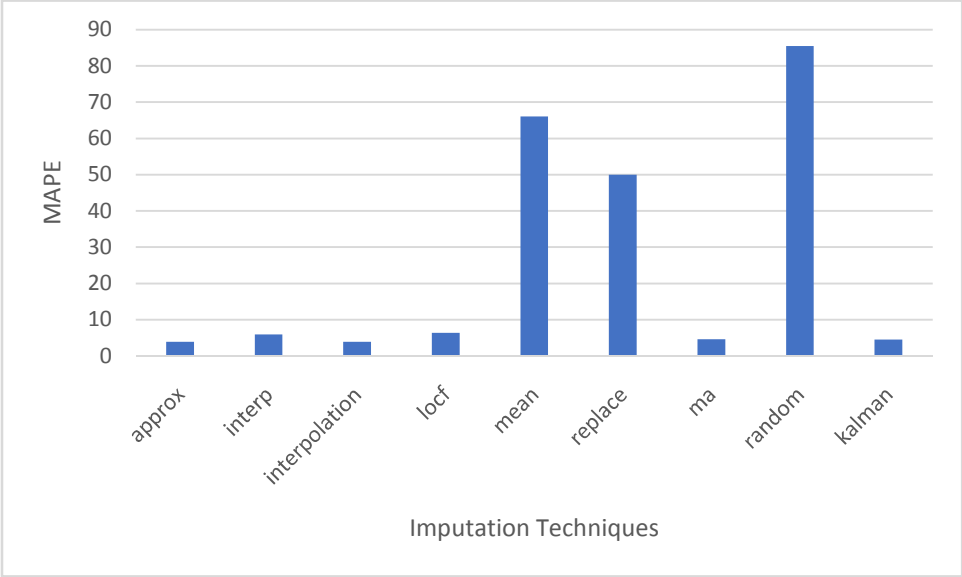

Figure 6: Average performances of each technique as a function of the percentage of missing values for MCAR.

Discussion

It is observed that, the "best" imputation technique with respect to MAR is Kalman followed by "interpolation". However, with respect to MCAR, the "best" technique is "interpolation". Thus, we recommend the usage of "interpolation imputation technique when the dataset has trend but not seasonal regardless of the missing imputation mechanisms.

Case 4: Google Dataset

Here, the performance of imputation techniques is assessed based on MAR and MCAR respectively.

i. Performance of Imputation Techniques in context of MAR

From Table 11, the Mean and Replace imputation techniques both perform the "best" when the missing rate values are 10%, 30%, 40% and 50%. However, the Replace technique is the "best" when the missing rate values are 60%, 70%, 80% and 90%. The RMSE error measure is used since it turned out to be the appropriate error metric for the Google dataset.

Table 11: Comparison of Imputation techniques of google dataset for Different Missing Percentage Values of MAR. Smaller values are better. Best values are shown in boldface

Imputation Techniques									
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	0.0095	0.0095	0.0095	0.0107	**0.0074**	**0.0074**	0.0095	0.0196	0.0077
20	0.0157	0.0157	0.0157	0.0164	**0.0108**	0.0109	0.0155	0.0300	0.0124
30	0.0163	0.0163	0.0163	0.0185	**0.0128**	**0.0128**	0.0155	0.0518	0.0132
40	0.0180	0.0180	0.0180	0.0188	**0.0145**	**0.0145**	0.0169	0.0253	0.0153
50	0.0220	0.0220	0.0220	0.0210	**0.0166**	**0.0166**	0.0212	0.0434	0.0176
60	0.0219	0.0219	0.0219	0.0243	0.0181	**0.0180**	0.0223	0.0434	0.0194
70	0.0243	0.0243	0.0243	0.0276	0.0191	**0.0190**	0.0246	0.0510	0.0207
80	0.0243	0.0308	0.0243	0.0351	0.2140	**0.0213**	0.0305	0.0450	0.0274
90	0.0316	0.0316	0.0316	0.0383	0.0224	**0.0221**	0.0330	0.0562	0.0714

Generally, when using a dataset with a no trend and seasonality, the appropriate imputation technique is the Replace technique followed by the "mean" technique as presented in Figure 7.

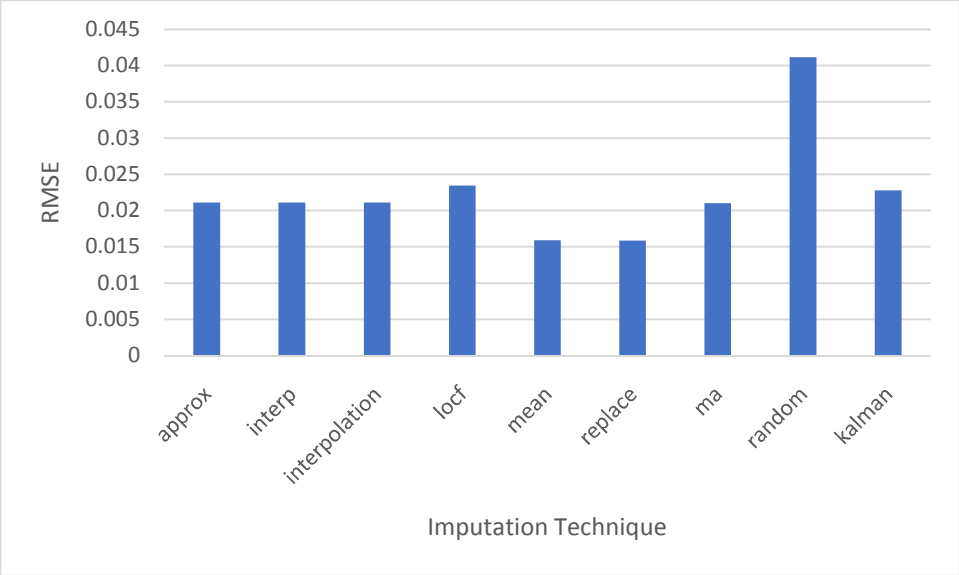

Figure 7: Average performances of each technique as a function of the percentage of missing values for MAR

ii. Performance of Imputation Techniques in context of MCAR

From Table 12, the Mean imputation technique is the "best" when the missing rate values are20%, 30%, 40%, 50% and 60%. However, the Replace technique is the "best" when the missing rate values are 10%, 20%, 80% and 90%.

Table 12: Comparison of Imputation techniques of google dataset for Different Missing Percentage Values of MCAR. Smaller values are better. Best values are shown in boldface

Imputation Techniques									
Missing Percent (%)	Approximation	Interp	Interpolation	LOCF	Mean	Replace	MA	Random	Kalman
10	0.0086	0.0086	0.0086	0.0095	0.0071	0.0070	0.0078	0.0163	0.0072
20	0.0122	0.0122	0.0122	0.0146	0.0101	0.0101	0.0111	0.0277	0.0106
30	0.0156	0.0156	0.0156	0.0183	0.0130	0.0131	0.0144	0.0371	0.0137
40	0.0174	0.0174	0.0174	0.0204	0.0139	0.0140	0.0160	0.0422	0.0147
50	0.0200	0.0200	0.0200	0.0237	0.0162	0.0163	0.0191	0.0451	0.0171
60	0.0222	0.0222	0.0222	0.0256	0.0184	0.0185	0.0214	0.0487	0.0190
70	0.0253	0.0253	0.0253	0.0289	0.0201	0.0201	0.0252	0.0438	0.0218
80	0.0278	0.0278	0.0278	0.0313	0.0211	0.0209	0.0282	0.0419	0.0242
90	0.0295	0.0295	0.0295	0.0334	0.0225	0.0222	0.0315	0.0593	0.0261

Generally, when using a dataset with a no trend and seasonality, the "best" imputation technique is the Mean technique followed by the "Replace" technique as presented in Figure 8.

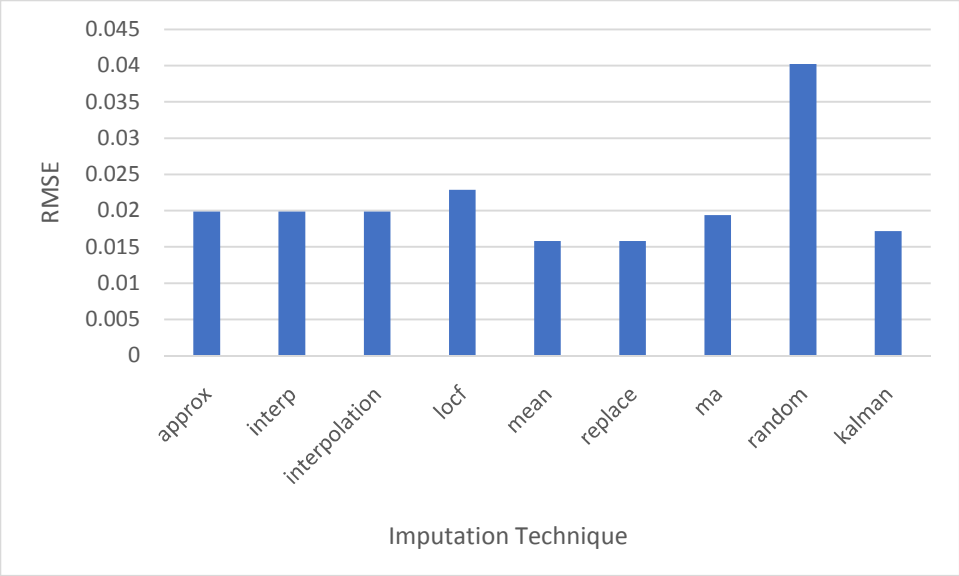

Figure 8: Average performances of each technique as a function of the percentage of missing values for MCAR.

Discussion

It is observed in both MAR and MCAR that, the two "best" imputation techniques for dataset that exhibits seasonality, but no trend are the "mean" and "Replace".

CONCLUSION

The purpose of this study is to provide statistical knowledge on how to treat missing values in univariate time series data. Thus, focusing on the effect of three different error metric on the

performance of imputation techniques, and the performance of nine different imputation techniques with respect to rate of missing values. Four original datasets exhibiting different features in time series data was used. Different missing rate values ranging from 10% to 90% at equal interval of 10, assuming both MAR and MCAR. Conclusions are drawn according to the two focus of this study as indicated earlier.

Firstly, it is observed that the appropriate error metric for datasets having both trend and seasonality and also dataset with trend but no seasonality, is the MAPE. However, the RMSE is the appropriate error metric measure for data that exhibits very high seasonality but no trend and also dataset with no trend and no seasonality. Thus, it is concluded that the choice of error metric depends on the characteristics or nature of the dataset. Again, the choice of a specific error metric is not affected or influenced by the missing data imputation mechanisms (i.e., MAR and MCAR).

Secondly, it is observed that the type of missing data imputation mechanisms (i.e. MAR and MCAR) has effect on the performance of imputation techniques if the dataset shows both trend and seasonality. We conclude that for such dataset, it is appropriate to use the "interp" technique since it's the "best" in MAR and second "best" in MCAR. Again, when a dataset exhibits seasonality but no trend, the "best" imputation technique is "interp", which is consistent in the two missing imputation mechanisms used in this study (MAR and MCAR).However, when dataset has trend but not seasonal, the "best" imputation technique with respect to MAR is Kalman followed by "interpolation". With respect to MCAR, the "best" technique is "interpolation". Thus, we recommend the usage of "interpolation" imputation technique regardless of the missing imputation mechanisms. However, it is observed in both MAR and MCAR that, the two "best" imputation techniques for dataset that exhibits seasonality, but no trend are the "mean" and "Replace".

This paper suggests the usage of a particular error metric measure and imputation technique on missing values of univariate time series data depending on the specific characteristics that the data is exhibiting.

REFERENCES

[1] Bar-Joseph, G.K. Gerber, D.K. Gifford, T.S. Jaakkola, and I. Simon. Continuous representations of time-series gene expression data. Journal of Computational Biology, 10(3-4):341–356, 2003.

[2] Box, G.E., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons. ISBN 978-1-118-67502-1 [p218]

[3] Chan, K. S., & Ripley, B. (2012). TSA: time series analysis. R package version 1.01. *URL: http://CRAN. R-project. org/package= TSA*.

[4] Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., & Moons, K.G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087-1091.

[5] Gelman, A., & Hill, J. (2007). Data analysis using regression and multi-level/hierarchical models. Columbia University, NY: Cambridge University Press.

[6] Gottman, J.M. (1981). Time-series Analysis: A comprehensive introduction for social scientists (No. 519.55 G6).

[7] Jörnsten, R., Ouyang, M., & Wang, H.Y. (2007). A meta-data based method for DNA microarray imputation. *BMC bioinformatics*, *8*(1), 109.

[8] King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review,* 95, 1, pp.49-69.

[9] Li, H., Zhao, C., Shao, F., Li, G. Z., & Wang, X. (2015). A hybrid imputation approach for microarray missing value estimation. *BMC genomics*, *16*(9), S1. [p219]

[10] Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924*.

[11] Moritz, S., & Bartz-Beielstein, T. (2017). Impute TS: time series missing value imputation in R. *The R Journal*, *9*(1), 207-218.

[12] Nguyen, C.D., Carlin, J.B., & Lee, K.J. (2013). Diagnosing problems with imputation models using the Kolmogorov-Smirnov test: a simulation study. *BMC medical research methodology*, *13*(1), 144.

[13] Ran, B., Tan, H., Feng, J., Liu, Y., & Wang, W. (2015). Traffic speed data imputation method based on tensor completion. *Computational intelligence and neuroscience*, *2015*, 22

[14] Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, *6*(1), 1.

[15] Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC. ISBN 978-0412040610 [p218].

[16] Schafer, J.L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147. URL https://doi.org/10.1037//1082-989x.7.2.147. [p218, 224].

[17] Stoffer, D.S., & Shumway, R.H. (2006). Time series analysis and its applications: With R examples.

[18] Tak, S., Woo, S., & Yeo, H. (2016). Data-Driven Imputation Method for Traffic Data in Sectional Units of Road Links. *IEEE Trans. Intelligent Transportation Systems*, *17*(6), 1762-1771.

URL https://doi.org/10.1109/tits.2016.2530312. [p219,227].

[19] Taylor, S. J. (2007). *Modelling financial time series*. world scientific. (second edition). World Scientific Publishing [p1].

[20] Yarandi, H.N. (2002). "Handling Missing Data with Multiple Imputation Using PROC MI in SAS." *Proceedings of the Southeast SAS User Group, Savannah, GA*

[21] Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and applied climatology*, *112*(1-2), 143-167. URL https://doi.org/10.1007/s00704-012-0723-x. [p218, 219].