African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 4, Issue 2, 2021 (pp. 1-9)

www.abjournals.or

KALMAN FILTER ALGORITHM VERSUS OTHER METHODS OF ESTIMATING MISSING VALUES: TIME SERIES EVIDENCE

Adejumo Oluwasegun Agbailu*, Albert Seno and Onifade Oluwafemi Clement

Department of Statistics, University of Abuja.

*Email: olushegzy006@gmail.com

Cite this article:

Adejumo O.A., Onifade O.C., Albert S. (2021), Kalman Filter Algorithm versus Other Methods of Estimating Missing Values: Time Series Evidence. African Journal of Mathematics and Statistics Studies 4(2), 1-9. DOI: 10.52589/AJMSS-VFVNMQLX.

Manuscript History

Received: 7 April 2021

Accepted: 21 April 2021

Published: 3 May 2021

ABSTRACT: *Ideally, we think data are carefully collected and have regular patterns with no missing values, but in reality, this does not always happen. This study examines four (4) methods—mean imputation (MI), median imputation (MDI), linear imputation (LI) and Kalman filter algorithm (KAL)—of estimating missing values in time series. The study utilized pairs of nine (9) simulated series; each pair constitutes "actual series" and "12% missingness series". The three (3) sample sizes i.e. small (50), medium (200) and large (1000) were varied over the additive models linear, quadratic and exponential forms of trend. The 12% missingness series were estimated using MI, MDI, LI and KAL. The performances of the method were checked using the root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), while the overall performances of the estimating methods were accessed using the average of the accuracy measures (RMSE, MAE and MAPE). The results of the average-accuracy measures show that KAL outperformed other methods (MI, MDI and LI) at the three sample sizes when the trend was linear; also, MDI outperformed other methods at the three (3) sample sizes when the trend was exponential. Furthermore, MI outperformed others at small and large sample sizes when the trend was quadratic. However, the Kalman filter algorithm proved better when the sample size was medium. Hence, KAL, MI and MDI methods are recommended to estimate missing data in time series when the trend is linear, quadratic and exponential respectively, until further study proves otherwise.*

KEYWORDS: Missingness Series, Time Series, Kalman Filter Algorithm, Mean Imputation, Median Imputation, Linear Interpolation.

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 4, Issue 2, 2021 (pp. 1-9)

www.abjournals.or

INTRODUCTION

Practically, we habitually analyse and make inferences using real data. In an ideal world, we would think that the data are carefully collected and have consistent patterns with no outliers or missing value. In reality, this is not always true, therefore an important part of the initial examination of the data is to assess the eminence of the data and to consider modifications where necessary. The handling of missing data has been an issue in statistics for some time, but it has come to the fore in recent years as it has gained the attention of many studies such as [1], [2], [6], [7], [10] and so on. The missing values occur for the reason that some observations may not be made at a particular time due to faulty equipment, lost records, natural disaster, or a mistake, which cannot be rectified until a time in the future. This problem is frequently encountered in time series data, since the data are records taken over time. Missing observations possibly will make the identification of the nature of time series data problematic. Many missing observations might be virtually impossible to obtain, either because of time or cost limitations [7].

Diverse methods of estimating missing observations for time series data have been employed in literature. In early studies, [9] examined classical filtering and prediction in relation to missing observations in time series. Jones [8] later extended [9] methods to observational error. Harvey and Pierse [5] highlighted the importance of state space modelling and Kalman filter to the problems of missing data in times series. Recently, several methods of determining missing observations have continued to evolve. Chiewchanwattana [3] developed a new algorithm for the imputation of missing samples of time series data. The algorithm was based on the observation that time series data that are manifestations of natural phenomena contain several sets of similar time series subsequence. The algorithm was achieved by finding a complete subsequence that is similar to the missing sample subsequence and imputing the missing samples from the complete subsequence.

In addition, [10] developed a novel approach that interpolates within data and imputes across data streams. They named the approach Multi-directional Recurrent Neural Net-work (M-RNN). [7] also proposed new methods of estimating missing values in time series data. The new methods proposed were based on the rows, columns and overall averages of time series data arranged in a Buys-Ballot table with m-rows and s-columns. Their new methods assume that only one value is missing at a time, the trending curve may be linear, quadratic or exponential and the decomposition method is either additive or multiplicative. Reference [4] compared different interpolation algorithms of estimating missing values in time series. They examined three (3) interpolation algorithms, namely: Radial Basis Function (RBF), Moving Least Squares (MLS) and Adaptive Inverse Distance Weighted (AIDW). Their study confirmed Lancaster's MLS as the best and also found that the number of nearest observed values for reference and the distribution of missing values could strongly affect the accuracy of imputation. In view of the aforementioned, treatment of missing data has been an issue in statistics in recent times. Therefore, this paper was set to add to existing literature by comparing the Kalman filter algorithm method of estimating missing values to other methods of estimation such as Mean Imputation, Median Imputation and Linear Interpolation.

The rest of this paper was organized as follows: In section 2, the methodologies employed in this work were discussed. Data and empirical results were presented in section 3 and conclusion was made in the final section.

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 4, Issue 2, 2021 (pp. 1-9)

www.abjournals.or

METHODOLOGY

Methods of Estimating Missing Values

This paper utilized four methods of estimating missing values namely mean imputation (MI), median imputation (MDI), linear interpolation (LI) and Kalman filter (KAL).

A. Mean Imputation

Mean imputation (MI) is a simple method of missing values estimation. It involves the replacement of the missing value with the mean of the values before the missing position(s). This is realized by taking the total i.e. summation, of the values and dividing by the number of observations before the missing position(s). This method maintains the sample size and is easy to use.

$$MI = \hat{X}_{(i-1)s+j} = \frac{1}{(i-1)s+j-1}[X_1 + X_2 + X_3 + \cdots + X_{(i-1)s+j-1}] \quad (1)$$

$$\boldsymbol{MI = \frac{1}{n^*}\sum_{t=1}^{n^*} X_t} \quad (2)$$

where $\boldsymbol{n^* = (i-1)s + j - 1}$ is the number of observations preceding the missing observation(s).

B. Median Imputation

Series median (MDI) estimates the missing value with the median of the remaining series. Symbolically, the series median is given by:

$$\boldsymbol{MDI = \hat{X}_{(i-1)s+j} = X_{\frac{N^*}{2}th}} \quad (3)$$

where N* = total number of observations excluding the missing values.

C. Linear Interpolation

This method replaces missing values using a linear interpolation. It utilizes the last valid value before the missing value and the first value after the missing value for the interpolation. The linear interpolation (LI) for estimating missing values is given by:

$$\boldsymbol{LI = \hat{X}_{(i-1)s+j} = \frac{1}{2}(X_{(i-1)s+j-1} + X_{(i-1)s+j+1})} \quad (4)$$

D. Kalman Filter Algorithm

The Kalman filter (KAL) is a statistical algorithm that enables certain computations to be carried out for a model cast in state space form. However, to obtain a more accurate estimate of missing values, the smoothing algorithm is performed. Let $\boldsymbol{y_{t-1}}$ denote the set of past observations and assume the conditional distribution of μ_t given $\boldsymbol{y_{t-1}}$ is $\boldsymbol{N(\mu_t, p_t)}$ where μ_t and $\boldsymbol{p_t}$ are assumed to have been determined. Hence, the Kalman filter equations for updating the missing values from time t to t+1 are given by:

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 4, Issue 2, 2021 (pp. 1-9)

www.abjournals.or

$$\mu_t = \mu_{t-1} + k_{t-1}v_{t-1};\ v_{t-1} = y_{t-1} - \mu_{t-1};\ p_t = p_{t-1}(1 - k_{t-1}) + \sigma_\eta^2$$

$$k_{t-1} = p_{t-1}/f_{t-1};\ f_{t-1} = p_{t-1} + \sigma_\varepsilon^2 \quad (4)$$

For t = 1,2,…,n, where $\boldsymbol{v_{t-1}}$ is the Kalman filter residual or prediction (signal) error, $\boldsymbol{f_{t-1}}$ is its variance and $\boldsymbol{k_{t-1}}$is the Kalman gain.

Comparison of Methods of Estimating Missing Values

Numerous measures are available for accessing the performances of the four methods (MI, MDI, LI and KAL). We evaluate the deviation of $\widehat{\boldsymbol{X}}_{(i-1)s+j}$ from the actual $\boldsymbol{X}_{(i-1)s+j}$, which can be calculated as $\hat{\boldsymbol{e}}_{(i-1)s+j} = \boldsymbol{X}_{(i-1)s+j} - \widehat{\boldsymbol{X}}_{(i-1)s+j}$, in order to access the performances of the afore-mentioned methods. We compare the "Accuracy Measures" of the four methods. These "Accuracy Measures" are root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and average accuracy measures (AAM), which are defined as follows:

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{MSE}} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\boldsymbol{A_t} - \boldsymbol{F_t})^2} \quad (4)$$

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{t=1}^{n} |(\boldsymbol{A_t} - \boldsymbol{F_t})| \quad (5)$$

$$\boldsymbol{MAPE} = \left[\frac{1}{m_0}\sum_{k=1}^{m_0} \left|\frac{e_k}{X_k}\right|\right] \times \mathbf{100} \quad (6)$$

where $\boldsymbol{A_t}$is the actual value in time *t*, and $\boldsymbol{F_t}$ is the forecast value in time *t*. We considered one missing value at a time for different $\boldsymbol{m_0 < n}$ position, $\boldsymbol{n > 1}$. The overall performances of the four estimating methods were accessed using the average of the accuracy measures with the minimum average accuracy measures is the best.

DATA AND EMPIRICAL RESULTS

Data Source

The nature of this study necessitated the use of simulated data. The simulated data used are simulated from the Additive Model; $\boldsymbol{X_t = M_t + S_t + e_t}$. The generation of the datasets was done by R statistical software package. Data are simulated from the Additive Model. The trend-cycle component $\boldsymbol{M_t}$ used are: $\boldsymbol{M_t = a + bt,\ at\ a = 1\ \&\ b = 2}$ (Linear); $\boldsymbol{M_t = a + bt + ct^2, at\ a = 1,\ b = 2\ \&c = 3}$ (Quadratic); and $\boldsymbol{M_t = be^{ct}, b = 10\ \&\ c = 0.02}$ (Quadratic). It is assumed that $\boldsymbol{e_t \sim N(0,1)}$for the Additive Model. $\boldsymbol{S_t}$, where $\boldsymbol{t = 1, 2, \ldots, 12,}$ denotes the seasonal indices. Table 1 presents the seasonal indices used for simulation.

Table 1. Seasonal Indices Used for Simulation

t	1	2	3	4	5	6	7	8	9	10	11	12
S_t	1.1	1.2	1.1	1	1	1	0.9	0.9	0.9	0.9	1.1	1.1

Note: S_t denotes Seasonal indices for Additive model

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

First of all, some packages in R library—such as *mcar* (for missing data simulation), *TestDataImputation* (for mean imputation and median imputation) and *interp* (for linear interpolation and Kalman filter algorithm)—were loaded. Then the aforementioned parameters were set out. Thereafter, sample sizes were varied our as follows:

i. The small sample size, which consists of 12% missingness sets of 50

ii. The medium sample size, which consists of 12% missingness sets of 200

iii. The large sample size, which consists of 12% missingness sets of 1000.

The above sample sizes were varied over the additive models linear, quadratic and exponential forms of trend (M_t). This implies that nine (9) series were stimulated with 12% missingness. The missing values in X_{mcar_t} were estimated using mean imputation (MI), median imputation (MDI), linear interpolation (LI) and Kalman filter (KAL) whose estimates were subsequently compared to nine-actual series (X_t). The codes were available in the R environment for mean imputation (MI), median imputation (MDI), linear interpolation (LI) and Kalman filter (KAL) set.

Empirical Results

Table 2 to Table 4 present the summary of the accuracy measures for mean imputation (MI), median imputation (MDI), linear interpolation (LI) and Kalman filter (KAL) methods of estimating missing values. Table 2 depicts results of the accuracy measures for the four methods when sample size is small, for the selected trend (linear, quadratic and exponential). In Table 2, the results show that KAL has the lowest (1.2129) average-accuracy measures (RMSE, MAE and MAPE) for linear trend components. Also, it shows that MI and MDI have the lowest (MI = 923.7945, MDI = 1.5263) average-accuracy measures for the quadratic and exponential trend components respectively. Hence, for small sample size, KAL returns as the best for linear trend structure while MI and MDI return as the best for quadratic and exponential trend structures respectively.

Table 3 presents the results of the accuracy measures for the four methods when sample size is medium, for the selected trend (linear, quadratic and exponential). In Table 3, the results show that KAL has the lowest (linear = 0.6248, quadratic = 1570.8530) average-accuracy measures (RMSE, MAE and MAPE) for linear and quadratic trend components. Similarly, the results show that MDI has the lowest (929131.7571) average-accuracy measures for exponential trend components. Hence, for medium sample size, KAL returns as the best for linear and quadratic trend structures while MDI returns as the best for exponential trend structure.

Table 2. Summary result of missing value estimation when sample size is small

Trend Component	Accuracy Measures	Estimation Method			
		MI	MDI	LI	KAL
Linear	RMSE	2.6495	2.6504	2.772	2.6725
	MAE	1.0007	1.0007	0.9873	0.9593
	MAPE	0.0075	0.0074	0.0083	0.0068
	Average	1.2193	1.2195	1.2558	**1.2129**

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

Quadratic	RMSE	2319.233	2562.877	3787.697	2577.596
	MAE	452.1502	403.5013	890.2353	571.0032
	MAPE	0.0004	0.0003	0.0014	0.0007
	Average	**923.7945**	988.7929	1559.3112	1049.5333
Exponential	RMSE	127650.9	3.7539	1895660.8	141424.21
	MAE	31267.95	0.8245	268319.3	34198.6
	MAPE	36.7076	0.0006	140.8656	45.0265
	Average	52985.186	**1.5263**	721373.66	58555.946

Source: Researchers' compilations

Table 3. Summary result of missing values estimation when sample size is medium

Trend Component	Accuracy Measures	Estimation Method			
		MI	**MDI**	**LI**	**KAL**
Linear	RMSE	1.5531	1.759	2.2403	1.5453
	MAE	0.3316	0.3663	0.4371	0.3282
	MAPE	0.0008	0.0006	0.0008	0.0008
	Average	0.6285	0.7086	0.8927	**0.6248**
Quadratic	RMSE	3778.192	4161.102	3982.469	3661.596
	MAE	1101.555	1160.667	987.9734	1050.958
	MAPE	0.0048	0.0036	0.004	0.0051
	Average	1626.5839	1773.9242	1656.8155	**1570.853**
Exponential	RMSE	2534654	2512328.8	2720963.5	2546924.3
	MAE	306517.6	275066.47	426979.97	314754.62
	MAPE	360.5943	0.0014	1097.168	281.9716
	Average	947177.4	**929131.8**	1049680.2	953986.96

Source: Researchers' compilations

Table 4. Summary result of estimation of missing value when sample size is large

Trend Component	Accuracy Measures	Estimation Method			
		MI	**MDI**	**LI**	**KAL**
Linear	RMSE	5717924	10.95917	2.983935	2.35911
	MAE	1667748	9.043338	0.8872693	0.7748485
	MAPE	4227.582	0.009937	0.017085	0.011757
	Average	2463299.9	6.6708	1.2961	**1.0486**

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

Quadratic	RMSE	3055.605	3136.429	3695.043	3099.308
	MAE	976.3247	973.6958	1088.255	984.0337
	MAPE	0.317189	0.26505	0.473352	0.331552
	Average	**1344.082**	1370.13	1594.5905	1361.2244
Exponential	RMSE	1524381	1524916.9	8762248.5	1570153.9
	MAE	200130.3	173719	2279966.1	234634.2
	MAPE	281.448	0.001336	102.4718	802.0694
	Average	574930.92	**566212**	3680772.4	601863.39

Source: Researchers' compilations

Lastly, Table 4 presents the results of the accuracy measures for the four methods when sample size is large, for the selected trend (linear, quadratic and exponential). In Table 3, the results show that KAL has the lowest (1.0486) average-accuracy measures (RMSE, MAE and MAPE) for linear trend components. Similarly, the results show that MI and MDI have the lowest (MI = 1344.0823, MDI = 566211.9671) average-accuracy measures for the quadratic and exponential trend components respectively. Hence, for large sample size, KAL returns as the best for linear trend structure while MI and MDI return as the best for quadratic and exponential trend structures respectively.

CONCLUSION AND RECOMMENDATION

This paper accesses the performances of Kalman filter algorithm method and other methods of missing value's estimation such as mean imputation, median imputation and linear interpolation. The analyses show the comparative results of the missing values estimation methods under three different trend structures (linear, quadratic and exponential) for small, medium and large sample sizes.

Table 5 presents the summary of the analyses results. The analyses result for linear trend structure show that Kalman filter algorithm method of handling missing data, when sample size was either small, medium or large, performed better than mean imputation, median imputation and linear interpolation methods. Also, the results for quadratic trend structure depict that mean imputation method of handling missing data, when sample size was either small or large, performed better than median imputation, linear interpolation and Kalman filter algorithm methods, whereas Kalman filter algorithm proved better when the sample size was medium. Finally, the results for exponential trend structure show that the median imputation method of handling missing data when sample size was either small, medium or large, performed better than mean imputation, linear interpolation and Kalman filter algorithm methods.

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

Table 3. Summary result of estimation

Trend Component	Sample Size		
	Small	**Medium**	**Large**
Linear	KAL	KAL	KAL
Quadratic	MI	KAL	MI
Exponential	MDI	MDI	MDI

In view of the aforementioned, it is recommended that Kalman filter algorithm should be used in estimating 12% missingness sets of observations in time series with linear trend structure at any size (small, medium or large) of sample until further studies prove otherwise. It is also recommended that median imputation should be used in estimating 12% missingness sets of observations in time series with exponential trend structure at any size (small, medium or large) of sample until further studies prove otherwise. Finally, it is recommended that mean imputation should be employed in estimating 12% missingness sets of observations in time series with quadratic trend structure at small and large sample sizes, while Kalman filter algorithm should be adopted at medium sample size until further studies prove otherwise.

REFERENCES

[1] Almed, M.R. and Al-Khazaleh, A.M.H. (2008) Estimation of Missing Data by Using the Filtering Process in a Time Series Modeling.

[2] Cheema, J.R. (2014) Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research. *Journal of Modern Applied Statistical Methods*, 13, Article 3. https://doi.org/10.22237/jmasm/1414814520

[3] Chiewchanwattana S., Lursinsap C. and Chu C.H. (2007). Imputing Incomplete Time Series Data Based on Varied-Window Similarities Measure of Data Sequences. *Pattern Recognition Letters ScienceDirect*, 28, 1091-1103.

[4] Ding Z., Mei G., Cuomo S., Li Y. and Xu N. (2018). Comparison of Estimating Missing Values in IoT Time Series Data Using Different Interpolation Algorithms. *International Journal of Parallel Programming*, 1-15.

[5] Harvey, A.C. and Pierse, R.G. (1984) Estimating Missing Observations in Economic Time Series. *Journal of the American Statistical Association*, 79, 125-131. https://doi.org/10.1080/01621459.1984.10477074

[6] Howell, D.C. (2007) The Analysis of Missing Data. In: Outhwaite, W. and Turner, S., Eds., *Handbook of Social Science Methodology*, Sage, London. https://doi.org/10.4135/9781848607958.n11

[7] Iwueze, I.S., Nwogu, E.C., Nlebedim, V.U., Nwosu, U.I. and Chinyem, U.E. (2018) Comparison of Methods of Estimating Missing Values in Time Series. Open Journal of Statistics, 8, 390-399. https://doi.org/10.4236/ojs.2018.82025

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX

[8] Jones, R.H. (1980) Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations. Technimetrics, 22, 389-395. https://doi.org/10.1080/00401706.1980.10486171

[9] Kalman, R.E. (1960) A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, 81, 35-45. https://doi.org/10.1115/1.3662552

[10] Yoon J., Zume W.R. and Schaar M.V. (2017). Multi-directional Recurrent Neural Networks: A Novel Method for Estimating Missing Data. *Time Series Workshop*, Sydney, Australlia.

Article DOI: 10.52589/AJMSS-VFVNMQLX
DOI URL: https://doi.org/10.52589/AJMSS-VFVNMQLX