International Journal of Mathematical Research
2020 Vol. 9, No. 1, pp. 72-80.
ISSN(e): 2306-2223
ISSN(p): 2311-7427
DOI: 10.18488/journal.24.2020.91.72.80

METHODS FOR ESTIMATING MISSING VALUES IN DESCRIPTIVE TIME SERIES STATISTICS: NOVELTY AND EFFICIENCY UNDER BUYS-BALLOT

Ugochinyere I. Nwosu[1+]
Chukwudi P. Obite[2]

[1,2]Department of Statistics, Federal University of Technology Owerri, Owerri, Imo State, Nigeria.
[1]Email: nwosuugochinyere@yahoo.com Tel: +2347061118381
[2]Email: chukwudi.obite@futo.edu.ng Tel: +2347031143410

(+ Corresponding author)

Article History
Received: 20 August 2020
Revised: 6 October 2020
Accepted: 27 October 2020
Published: 16 November 2020

Keywords
Estimation methods
Statistical information
Single-missing
Model structure
Buys-Ballot table.

ABSTRACT

There is dearth of information in the field of statistics on the innovative estimation methods that can replace missing values in descriptive time series data. Therefore, this review work provides information on the existing and new methods of estimating missing values in descriptive time series data. The work provides new insight on the comparative performance of the recently-developed methods and the existing ones and discussed model structure and trending curves as important parameters in estimation of missing values. It is expected that the present contribution will assist statisticians seeking to solve the problem of missing values in descriptive time series data. The application of this work should be restricted to time series data with trend (linear, quadratic and exponential) and seasonal components combined in the additive and multiplicative forms. The contribution covers data missing at one point at a time in a row or column when data are arranged in a Buys-Ballot table. Use of the Buys-Ballot table arrangement in the estimation of missing values is new, convenient and merits scientific analysis.

Contribution/Originality: This review work is one of the few papers that discussed the novelty and efficiency of Buys-Ballot table in the estimation of missing values in descriptive time series statistics.

1. INTRODUCTION

1.1. Background

A time series is a sequential set of data recorded over time on a particular variable [1]. Missing observations at certain points within the data collected has been identified as a frequent problem facing data analysis [1-3]. When an observation is missing, it is pertinent to estimate the missing value for thorough understanding of data nature [2]. Unquestionably, obtaining a good estimate leads to a more accurate forecast. The inability to account for missing values can culminate in serious misinterpretation of the phenomenon under investigation Owili, et al. [3]. Abraham and Thavaneswaran [4] noted that unaccounted missing values can cause severe problem in the estimation and forecasting of linear and nonlinear time series. Therefore, it is necessary to replace the missing value using appropriate estimation methods.

Data may be missing due to several reasons which include equipment malfunctioning, bad weather and incorrect data entry. These reasons lead to different natures of missing values, namely missing completely at random, missing at random and missing not at random. Missing values can lead to erroneous conclusions about data and unfortunately this does not favour the progress of an economy that depends on the result of a forecast from such data. Many countries are currently experiencing economic crises and there is every need to enhance the accuracy of data needed for solving economic problems. The use of the Buys-Ballot table arrangement in the estimation of missing values is new and merits further analysis Iwueze, et al. [5]. Iwueze and Nwogu [6] recorded that the use of the Buys-Ballot table arrangement is convenient and enhances accuracy of the estimates. Investigating claims by Iwueze and Nwogu [6] should form part of new studies, the review recommends.

1.2. Components and Properties of a Time Series

The components of time series are defined as follows:

i. The Trend Component

Trend is generally thought of as a smooth and slow movement over a long term. The concept of "long" in the connection is relative to what is identified as trend for a given series. Tests for trend are given by Kendal and Ord [7]. Correlation analysis can also be used to assess trend. According to Chatfield [8] the autocorrelation value will not be zero except for very large values of the lag if a time series contains a trend.

ii. Seasonal Component

Seasonal components, denoted by S_t, are short-term fluctuation in a time series which occur periodically in a year. This continues to repeat year after year. The seasonal component, S_t, is associated with the property that:

$$S_{(i-1)s+j} = S_j, \; i = 1, 2, \ldots \qquad (1)$$

where s is the periodicity.

iii. The Cyclical Component

Cyclical component, denoted by C_t, is defined as the recurrent upward or downward movements in a time series, but the period of cycle is greater than a year. For short duration of data, trend and cyclical components are customarily combined into a trend-cycle component denoted as M_t [8].

iv. Irregular Component

Irregular components, denoted by I_t, are erratic movements in a time series that follow no regular pattern.

1.3. Model Structure

A model describes the properties of a system. In time series forecasting, one tries to find a model of a system suitable for the realization of the time series variables. Model is used to simulate the system, analyze it and to derive future values of the time series. As a model is inferred from the time series and not from the system itself, it is a description of time series rather than the system and it can be hoped that a model says something useful about the underlying system [9].

Apart from identifying the pattern (the components) of time series data, for a good analysis of a time series data to be done, the correct model to be used is very important. The specific functional relationship between these components can assume different forms. However, two main possibilities are that they combine in an additive or a multiplicative form;

1. Additive model (when trend, seasonal and cyclical components are additively combined):

$$X_t = M_t + S_t + e_t \qquad (2)$$

where

$$\sum_{j=1}^{s} S_j = 0 \qquad (3)$$

and $e_t \sim N(0, \sigma_1^2)$

2. Multiplicative model (when trend, seasonal and cyclical components are multiplicatively combined):

$$X_t = M_t S_t e_t \qquad (4)$$

where

$$\sum_{j=1}^{s} S_j = s \qquad (5)$$

and $e_t \sim N(0, \sigma_2^2)$

2. BUYS-BALLOT TABLE

The Buys-Ballot table arranges a seasonal time series data of length n conventionally into m rows and s columns as shown in Table 1. The rows represent the periods/year while the columns are the seasons [6, 10]. The Buys-Ballot procedure was developed by Iwueze and Nwogu [6] for short period data in which trend and cyclical component are jointly estimated. In their findings, the row and column means are unbiased estimates of the observations found in the respective rows and columns. Based on this assumption, the row and column mean imputation methods of estimating missing value were developed by Nwosu [11]. The information presented in this work concerns data that have trend (linear, quadratic and exponential) and seasonal components combined in the additive and multiplicative forms. In addition, during estimation, the power of the methods developed through the Buys-Ballot procedure are measured by comparing them with already existing methods using appropriate accuracy measures.

Table-1. The buys-ballot table.

P (period)	**Season**								
	1	**2**	**…**	**J**	**…**	**S**	**Total $T_{i.}$**	**Mean $\bar{X}_{i.}$**	**Std $\hat{\sigma}_{i.}$**
1	X_1	X_2	…	X_j	…	X_s	$T_{1.}$	$\bar{X}_{1.}$	$\hat{\sigma}_{1.}$
2	X_{s+1}	X_{s+2}	…	X_{s+j}	…	X_{2s}	$T_{2.}$	$\bar{X}_{2.}$	$\hat{\sigma}_{2.}$
3	X_{2s+1}	X_{2s+2}	…	X_{2s+j}	…	X_{3s}	$T_{3s.}$	$\bar{X}_{3.}$	$\hat{\sigma}_{3.}$
…	…	…	…	…	…	…	…	…	…
I	$X_{(i-1)s+1}$	$X_{(i-1)s+2}$	…	$X_{(i-1)s+j}$	…	$X_{(i-1)s+s}$	$T_{i.}$	$\bar{X}_{i.}$	$\hat{\sigma}_{i.}$
…	…	…	…	…	…	…	…	…	…
M	$X_{(m-1)s+1}$	$X_{(m-1)s+2}$	…	$X_{(m-1)s+j}$	…	X_{ms}	$T_{m.}$	$\bar{X}_{m\cdot}$	$\hat{\sigma}_{m.}$
$T_{.j}$	$T_{.1}$	$T_{.2}$	…	$T_{.j}$	…	$T_{.s}$	$T_{..}$		
$\bar{X}_{.j}$	$\bar{X}_{.1}$	$\bar{X}_{.2}$	…	$\bar{X}_{.j}$	…	$\bar{X}_{.s}$		$\bar{X}_{..}$	
$\hat{\sigma}_{.j}$	$\hat{\sigma}_{.1}$	$\hat{\sigma}_{.2}$	…	$\hat{\sigma}_{.j}$	…	$\hat{\sigma}_{.s}$			$\hat{\sigma}_{..}$

Note: Std: Standard deveation.
Source: Iwueze and Nwogu [6].

$$T_{i.} = \sum_{j=1}^{s} X_{(i-1)s+j}, i = 1, 2.., m$$

where (6)

$$\mathrm{X}_{i.} = \frac{T_{i.}}{s} = \frac{1}{s}\sum_{j=1}^{s} \mathrm{X}_{(i-1)s+j}, i = 1, 2.., m \tag{7}$$

$$T_{.j} = \sum_{i=1}^{m} \mathrm{X}_{(i-1)s+j}, \mathrm{j} = 1, 2, \ldots \mathrm{s} \tag{8}$$

$$\bar{\mathrm{X}}_{.j} = \frac{\mathrm{T}_{.j}}{\mathrm{m}} = \frac{1}{\mathrm{m}}\sum_{i=1}^{m} \mathrm{X}_{(i-1)s+j}, \mathrm{j} = 1, 2,, \ldots \mathrm{s} \tag{9}$$

$$T_{..} = \sum_{i=1}^{m} \mathrm{T}_{i.} = \sum_{j=1}^{s} \mathrm{T}_{.j} = \sum_{i=1}^{m}\sum_{j=1}^{s} \mathrm{X}_{(i-1)s+j} \tag{10}$$

$$\bar{X}_{..} = \frac{T_{..}}{ms} = \frac{T_{..}}{n}, n = ms \tag{11}$$

$$\hat{\sigma}_{i.} = \sqrt{\frac{1}{s-1}\sum_{j=1}^{s} (\mathrm{X}_{(i-1)s+j} - \bar{\mathrm{X}}_{i.})^2} \text{ ,i = 1,2,..,m} \tag{12}$$

$$\hat{\sigma}_{.j} = \sqrt{\frac{1}{\mathrm{m}-1}\sum_{i=1}^{m} (\mathrm{X}_{(i-1)s+j} - \bar{\mathrm{X}}_{.j})^2}, \mathrm{j} = 1, 2,.., \mathrm{s} \tag{13}$$

$$\hat{\sigma}_{..} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{m}\sum_{j=1}^{s} (\mathrm{X}_{(i-1)s+j} - \bar{\mathrm{X}}_{..})^2} \tag{14}$$

2.1. Important Equations

In using the methods, simulated or real life data that fit additive (2) and multiplicative (4) models with the linear (15), quadratic (16) and exponential (17) trend curves should be considered as follow:

i. For the linear trend curve,

$M_t = a+bt$, t = 1, 2, …, m (15)

ii. For the quadratic trend curve,

$M_t = a+bt+ct^2$, t = 1, 2, …, m (16)

iii. For the exponential trend curve,

$M_t = be^{ct}$, t = 1, 2, …, m (17)

3. ESTIMATION METHODS FOR REPLACING MISSING VALUES

The existing methods for estimating missing values in time series data and the recently developed estimation methods are as follow. These methods are adaptable to single missing condition and suitable in buys-ballot arrangement. However, their adaptability to double missing condition or more, has not been tested in a buys-ballot arrangement.

1. Mean imputation
2. Linear interpolation
3. Linear trend at point
4. Series mean

Existing methods

5. Row mean imputation
6. Column mean imputation
7. Decomposition without the missing value

Recently-developed when value is missing at a point
but never tested under double missing condition.

Notably, methods 1 to 4 are some of the existing methods in literature [3, 12, 13]. Methods 5 to 7 are the recently-developed methods by Iwueze, et al. [5] used to estimate missing value in time series when data is missing at a point. The recent estimation methods by Iwueze, et al. [5] have not been tested under multiple missing conditions. This is one of the gaps which Statisticians should bridge through further scientific studies. Assuming an observation is missing in the Buys - Ballot table at one point (say $X_{(i-1)s+j}$), it is estimated using the different methods as follows:

3.1. Mean Imputations (MI)

Mean imputation replaces the missing value with the mean of the values below the missing position. This is achieved by taking the sum of the non-missing values below the missing value and dividing by the corresponding number of observations.

$$MI = \hat{X}_{(i-1)s+j} = \frac{1}{(i-1)s+j-1}\left[X_1+X_2+X_3+\ldots+X_{(i-1)s+j-1}\right] \tag{18}$$

Equation 18 gives the missing observation.

3.2. Linear Interpolation (LI)

The linear interpolation method replaces missing value employing a linear interpolation. In this method, the last value before the missing value and the first value after the missing value are used for the interpolation.

$$LI = \hat{X}_{(i-1)s+j} = \frac{1}{2}\left(X_{(i-1)s+j-1}+X_{(i-1)s+j+1}\right) \tag{19}$$

3.3. Linear Trend at Point (LTP)

Linear trend at point replaces the missing value with the linear trend for that point. The remaining series is regressed on an index variable scaled 1, 2, 3,…, (i-1)s+j-1, (i-1)s+j+1,…, (i)s+j-2, (i)s+j,…, ms. Missing value is then replaced with its predicted value. Then, the missing values will be estimated from the regression equations.

$$LTP = \hat{X}_{(i-1)s+j} = \hat{a}+\hat{b}\left[(i-1)s+j\right] \tag{20}$$

3.4. Series Mean (SM)

Series mean achieves replacement of the missing value with the mean of the remaining series, by using the average of the available data values to replace the missing observation.

$$SM = \hat{X}_{(i-1)s+j} = \frac{T^{*}_{..}}{n-1}, n = ms \tag{21}$$

Where,

$$T^{*}_{..} = \left[X_1+X_2+\ldots+X_{(i-1)s+j-1}+X_{(i-1)s+j+1}+X_{ms}\right] \tag{22}$$

3.5. Row Mean Imputation (RMI)

According to Iwueze and Nwogu [6] the row mean of the Buys-Ballot Table 1 is an unbiased estimate of the row observations. Therefore, the row mean imputation replaces the missing values with the row mean and this is achievable by finding the average of the available row values and replacing the missing value with the estimate [11].

$$\text{RMI} = \hat{X}_{(i-1)s+j} = \frac{1}{s-1}\left[X_{(i-1)s+1} + X_{(i-1)s+2} + \ldots + X_{(i-1)s+j-1} + X_{(i-1)s+j+1} + \ldots + X_{(i-1)s+s}\right] \quad (23)$$

$$\text{RMI} = \hat{X}_{(i-1)s+j} = \frac{1}{s-1}\left[\sum_{t=1}^{j-1} X_{(i-1)s+j} + \sum_{t=j+1}^{s} X_{(i-1)s+j}\right] \quad (24)$$

The missing value is replaced using Equation 24.

3.6. Column Mean Imputation (CMI)

The column mean of the Buys-Ballot table was recognized as an unbiased estimate of the column observations. Therefore, the column mean imputation by Iwueze, et al. [5] entails replacing the missing value with the column mean of the Buys-Ballot table. This is achieved by finding the mean of the respective column values and replacing the missing value with this estimate.

From Equation 9

$$\bar{X}_{.j} = \frac{1}{m}\sum_{i=1}^{m} X_{(i-1)s+j} \qquad j = 1,2 \ldots, s$$

Which implies that:

$$\bar{X}_{.j} = \frac{1}{m}\left[X_j + X_{s+j} + X_{2s+}j + \ldots.. + X_{(i-2)s+j} + X_{(i-1)s+j} + X_{is+j} + X_{(i+1)s+j} + \ldots X_{(m-1)s+j}\right] \quad (25)$$

Then

$$\text{CM I} = \hat{X}_{(i-1)s+j} = \frac{1}{m-1}\left[X_j + X_{s+j} + X_{2s+}j + \ldots + X_{(i-2)s+j} + X_{is+j} + X_{(i+1)s+j} + \ldots X_{(m-1)s+j}\right] \quad (26)$$

Summing Equation 26 gives:

$$\text{CMI} = \hat{X}_{(i-1)s+j} = \frac{1}{m-1}\left[\sum_{t=1}^{i-1} X_{(i-1)s+j} + \sum_{t=i+1}^{m} X_{(i-1)s+j}\right] \quad (27)$$

3.7. Decomposing Without the Missing Value (DWMV)

This method decomposes the remaining data series without the missing value to obtain the trend at point (i-1)s+j [5].

$$\hat{M}_{(i-1)s+j} = \hat{a} + \hat{b}(i-1)s+j \quad (28)$$

Note that:

$$\hat{S}_{(i-1)s+j} = \hat{S}_j \quad (29)$$

Then, the estimate at the aforementioned point for additive model becomes:

$$\text{DWMV} = \hat{X}_{(i-1)s+j} = \hat{M}_{(i-1)s+j} + \hat{S}_j \quad (30)$$

Similarly, the estimate at the aforementioned point for multiplicative model becomes

$$\text{DWMV} = \hat{X}_{(i-1)s+j} = \hat{M}_{(i-1)s+j} \text{ x } \hat{S}_j \quad (31)$$

4. COMPARATIVE NOVELTY AND EFFICIENCY OF ESTIMATION METHODS OF MISSING VALUE IN LITERATURE

Findings of different workers on estimation methods of missing values revealed that an estimate is never completely accurate and usually deviates from the actual value [14, 15]. From the literature reviewed [5, 11, 15] some estimation methods have shown better novelty and efficiency than others. According to Nwosu [11] and Iwueze, et al. [5] the Decomposing without the Missing Value estimation method was most effective using mean absolute error, mean absolute percentage error and root mean square error as criteria. According to Iwueze, et al. [5] the novelty shown by the Decomposing without the Missing Value estimation method is probably because it considers seasonality of the missing value. In the recent ranking by these authors, the Linear Interpolation estimation method was second, Linear Trend at Point third, Row Mean Imputation fourth, Series Mean fifth, Column Mean Imputation sixth and Mean Imputation estimation method seventh in position. Each of these methods in comparison with the others maintained its position 100%, the literature concluded. This indicates that the methods of estimation of missing values highlighted in this review were consistent in their performance without being prone to minimal variations. These are important statistical information.

5. ACCURACY MEASURES FOR DETERMINING EFFICIENCY OF ESTIMATORS

To evaluate the accuracy of estimators, the estimated values obtained from the various estimation methods are compared with the actual values in the data. However, in Buys-Ballot arrangement when data is missing at a point, the estimated values ($\hat{X}_{(i-1)s+j}$) should be subtracted from the actual value ($X_{(i-1)s+j}$) to obtain the estimated error. The estimated error is denoted by

$$\hat{e}_{(i-1)s+j} = X_{(i-1)s+j} - \hat{X}_{(i-1)s+j} \qquad (32)$$

Given a data set of size n = ms, and data missing at one point at a time in a row or column of the Buys-Ballot table for different m_0 positions, $m_0 < n > 1$.

the m_0 estimated errors is denoted by U_k, k = 1,2,…, m_0 .

where

$$U_k = X_k - \hat{X}_k, k = 1,2,\ldots, m_0 \qquad (33)$$

m_0 = the number of estimated missing value, X_k is the actual value of the series at position k and $\hat{X}_k$ is the estimated missing value.

The estimated errors U_k, k = 1,2,…, m_0 are used to define accuracy measures [16] for comparing the different methods of estimation of missing values. The performance criteria used in ascertaining the efficiency of these estimates discussed in this review are: Mean Error (ME), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute percentage Error (MAPE).

5.1. Mean Error (ME)

It indicates the deviation between the actual values and estimates. Mean Error is given as:

$$\text{ME} = \frac{1}{\text{m}_0}\sum_{i=1}^{\text{m}_0} U_k = \frac{1}{\text{m}_0}\sum_{k=1}^{\text{m}_0} U_k \qquad (34)$$

5.2. Mean Absolute Error (MAE)

MAE is the positive derivation between the actual values and estimates, it is denoted as:

$$\text{MAE} = \left[\frac{1}{\text{m}_0}\sum_{k-1}^{\text{m}_0} |U_k|\right] \qquad (35)$$

5.3 Mean Absolute Percentage Error (MAPE)

This accounts for the percentage of deviation between the actual values and estimates. This can be obtained as:

$$\mathrm{MAPE} = \left[\frac{1}{\mathrm{m}_0} \sum_{k=0}^{\mathrm{m}_0} \left| \frac{U_k}{X_k} \right| \right] \times 100 \tag{36}$$

5.4. Mean Square Error (MSE)

Mean square error indicates the fluctuations of the deviations and it can be calculated as:

$$\mathrm{MSE} = \frac{1}{\mathrm{m}_0} \sum_{k=1}^{\mathrm{m}_0} U_k^2 \tag{37}$$

5.5. Root Mean Square Error (RMSE)

This is calculated as the square root of the square error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{m}_0} \sum_{k=1}^{\mathrm{m}_0} U_k^2} \tag{38}$$

6. CONCLUSION

This work reviewed seven methods of estimating missing values in descriptive time series data arranged in a Buys-Ballot table. The values in question shall not be missing from the same row or column at the same time in Buys-Ballot table. It was revealed that model structure and trending curves are important parameters which should be considered in the estimation of missing values. Findings [5, 11] revealed that the Decomposing without the Missing Value estimation method is effective, novel and gave the best result when compared with the other six methods. This contribution will be meaningful to statisticians seeking to tackle the problem of missing values in descriptive time series data.

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
Acknowledgement: Both authors contributed equally to the conception and design of the study.

REFERENCES

[1] D. S. Fung, "Methods for the estimation of missing values in time series," M.Sc. Thesis in Faculty of Communications, Health and Sciences, Edith Cowan University, Perth Western Australia, 2006.

[2] D. C. Howell, *The analysis of missing data. In: Handbook of Social Science Methodology (Outhwaite, W. and Turner, S. Eds.)*. London: Sage, 2007.

[3] A. P. Owili, D. Nassiuma, and L. Orawo, "Imputation of missing values for pure bilinear time series models with normally distributed innovations," *American Journal of Applied Mathematics and Statistics*, vol. 3, pp. 199-205, 2015.

[4] B. Abraham and A. Thavaneswaran, "A nonlinear time series model and estimation of missing observations," *Annals of the Institute of Statistical Mathematics*, vol. 43, pp. 493-504, 1991.Available at: https://doi.org/10.1007/bf00053368.

[5] I. Iwueze, E. Nwogu, V. Nlebedim, U. Nwosu, and U. Chinyem, "Comparison of methods of estimating missing values in time series," *Open Journal of Statistics*, vol. 8, pp. 390-399, 2018.

[6] I. Iwueze and E. Nwogu, "bBallot estimates for time series decomposition," *Global Journal of mathematical Sciences*, vol. 3, pp. 83-98, 2004.

[7] M. Kendal and J. K. Ord, *Time series*, 3rd ed. London: Griftin, 1990.

[8] C. Chatfield, *The analysis of time series: An introduction*, 6th ed. London: Chapman and Hall, 2004.

[9] J. Theiler, P. S. Linsaly, and D. M. Rubin, *Exploring the continuum between deterministic and stochastic modeling. In: Time series Prediction: Forecasting the Future and Understanding the Past (Weigend, A. and Gershen, N. eds.)*. New York: Addison-Wesley, 1993.

[10] H. Wold, *A study in the analysis of stationary time series.* Sweden: Almqvist and Wiksell, 1938.

[11] U. I. Nwosu, "Estimation of missing values for time series data arranged in a buys-ballot table," M.Sc. Thesis in the Department of Statistics, Federal University of Technology Owerri, 2016.

[12] E. Damsleth, "Interpolating missing values in a time series," *Scand Journal of Statistics*, vol. 7, pp. 33 – 39, 1979.

[13] M. Pourahmadi, "Estimation and interpolation of missing values of a stationary time series," *Journal of Time Series Analysis*, vol. 10, pp. 149-169, 1989.

[14] M. S. Mario, *Multimedia communications and networking.* U.S.A: CRC Press, Taylor and Francis Group, 2012.

[15] U. E. Chinyem, "Methods of estimating missing values in descriptive time series," M.Sc. Thesis in the Department of Statistics, University of Port Harcourt, 2014.

[16] S. Makridakis and M. Hibon, *Evaluating accuracy (or Error) measures.* France: INSEAD, Fontainebleau, 1995.

Views and opinions expressed in this article are the views and opinions of the author(s), International Journal of Mathematical Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.