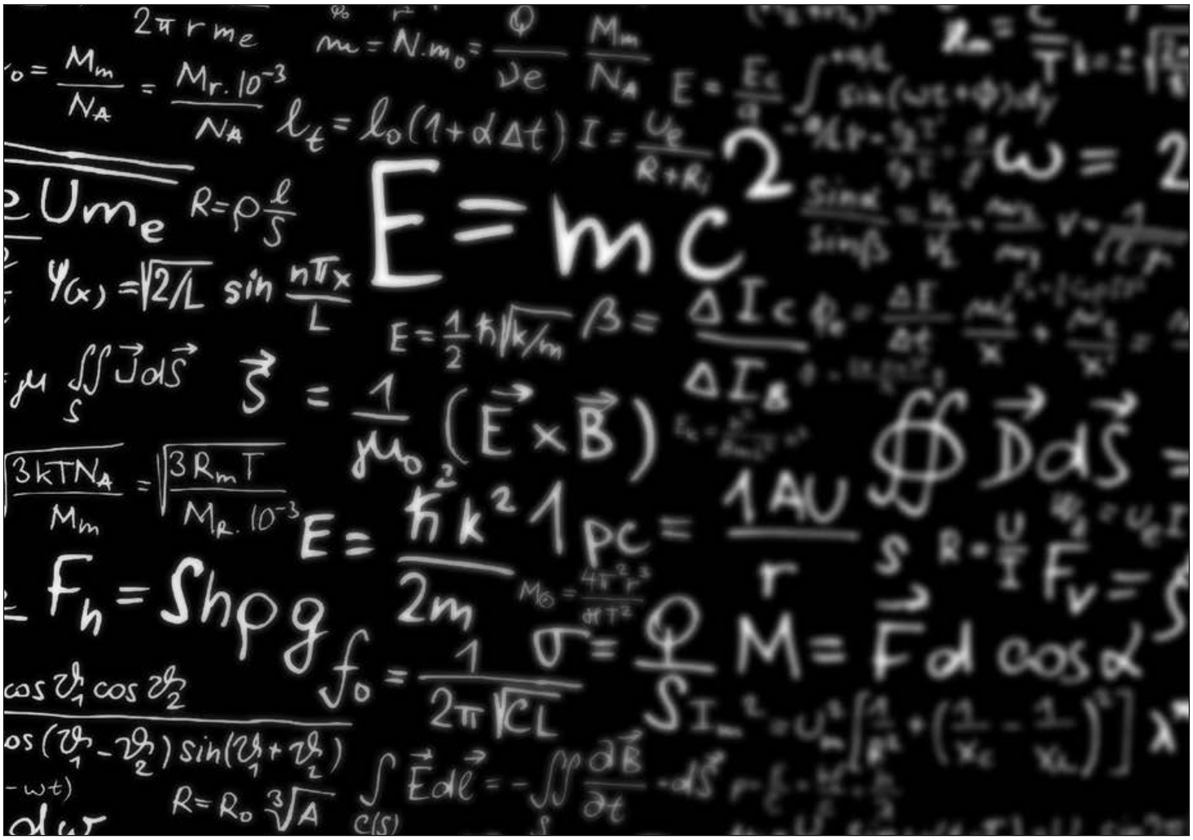

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

OPEN ACCESS Asian Journal of Mathematics and Statistics

ISSN 1994-5418
DOI: 10.3923/ajms.2016.6.10

Research Article
Method of Estimating Missing Values in a Stationary Autoregressive (AR) Process

I.A. Iwok

Department of Mathematics/Statistics, University of Port-Harcourt, P.M.B. 5323, Port-Harcourt, Rivers State, Nigeria

Abstract

Background: This study proposed a method for the estimation of missing values in a stationary autoregressive (AR) process and proved the unbiasedness of the obtained estimate. **Materials and Methods:** Box and Jenkins autoregressive integrated moving average (ARIMA) was employed for order selection in the analysis. Absolute deviation of estimates from actual values was used as basis of comparisons with existing methods. **Results:** The result showed that the proposed method provides better estimates than the existing methods. Minimum mean square error of the estimate was also obtained theoretically and the estimate was found to be unbiased. **Conclusion:** Since the estimate obtained by this method is found to be unbiased, the method has offered another framework of with missing values in a stationary autoregressive (AR) process.

Key words: AR process, stationarity condition, autocorrelation function, partial autocorrelation function

Received: May 11, 2016 Accepted: May 27, 2016 Published: June 15, 2016

Citation: I.A. Iwok, 2016. Method of estimating missing values in a stationary autoregressive (AR) process. Asian J. Math. Stat., 9: 6-10.

Corresponding Author: I.A. Iwok, Department of Mathematics/Statistics, University of Port-Harcourt, P.M.B. 5323, Port-Harcourt, Rivers State, Nigeria

Competing Interest: The author has declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

One of the serious problems faced in data collection and analysis is missing observations. Nature does not always provide a complete data set, hence, the mechanism for observing time series is often imperfect. Equipment failure, human error or the disregarding of inaccurate measurements can introduce missing values. In statistics, analysis is often carried out with complete observations. Where any observation is missing either by natural or human error, the missing value has to be estimated to complete the observations before an analysis is carried out.

According to Abraham and Thavaneswaran[1], data that are known to have been observed erroneously can fairly be categorised as missing. By this classification, erroneous data is believed to wreak havoc with the estimation and forecasting of linear and non linear time series models. In their study, two methods for estimating missing observations; one using prediction and fixed point smoothing algorithms and the other using optimal estimating equation theory were identified. Recursive estimation of missing observations in an autoregressive conditionally heteroscedastic (ARCH) model and the estimation of missing observations in a linear time series were shown as special cases. Using these methods, construction of optimal estimates of missing observations was obtained.

Phong and Singh[2] proposed modelling gene expression profiles as simple linear and Gaussian dynamical systems and applied the Kalman filter to estimate missing values. In their study, they discovered that while other current advanced estimation methods are either sensitive to parameters with no theoretical means of selection. Their approach was advantageous because it makes minimal assumptions that are consistent with the biology. The efficiency of the approach was demonstrated by evaluating its performance in estimating artificially introduced missing values in two different time series data set and was compared with a Bayesian approach dependent on the eigen vectors of the gene expression matrix as well as a gene wise average impartation for missing values.

Tight *et al.*[3] established the applicability of time series and influence function techniques in the estimation of missing value and detection of outliers. They went further and made cooperative assessment of new techniques with those used by traffic engineers in practice for local, regional or national traffic count systems. Their result showed that the replacement values derived from the ARIMA model using residuals were found to be most accurate.

Maravall and Pena[4] proposed the estimation of missing observations in possibly non stationary ARIMA models, where the models was assumed known and the structure of the interpolation filter was analysed using the inverse autocorrelation function, it was shown that the estimation of a missing observation is analogous to the removal of an outlier effect. Both were closely related with the signal plus noise decomposition of the series. The results were extended to the case of missing observation near the two extremes of the series, the case of a sequence of missing observations and finally to the general case of any number of sequences of any length of missing observations.

For a stationary series, the problem of interpolating missing values given an infinite realization of the stochastic process was solved by Grenander and Rosenblatt[5] and Wei[6]. The interpolator was obtained as the expected value of the missing observation given the observed infinite realization of the series. For many years, however, their result was not extended to the more general problem of interpolation in a finite realization of a possibly non stationary time series generated by a model with unknown parameters[4].

Xia *et al.*[7] examined 6 methods for estimating missing climatological data for different time scales at 6 German weather stations and 3 Bavarian forest climate stations. It was discovered that the multiple regression technique predominantly gave the best estimation under different topographical conditions.

Anava *et al.*[8] studied the problem of time series prediction using AR model in the presence of missing data. The signal and missing data were set to be arbitrary. The problem was cast as an online learning problem in which the goal of the learner was to minimize prediction error. An efficient algorithm for the problem was devised which was based on autoregressive model and does not assume any structure on the missing data nor on the mechanism that generates the time series. The result showed that the algorithm's performance asymptotically approaches the performance of the best AR predictor in hind right and corroborate the theoretic results with an empirical study on synthetic and real world data.

Ahmed and Al-Khazaleh[9] proposed new method of estimating missing data using the filtering process. The filtering process involved substituting various correlations in the weights of moving average model. The results obtained were checked with Naïve test and were found to be good.

There are quite a number of methods for dealing with missing observations. The commonest ones are: (i) Replacing the missing observation with the mean of the series,

(ii) Replacing it with the Naïve forecast which uses the current time periods value for the next time period, (iii) Replacing it with a simple trend forecast and (iv) Replacing it with an average of the last two known observations that bound the missing observation[9]. These methods are simple and most times provides better estimates than some of the complex methods outlined by some authors above.

MATERIALS AND METHODS

Stationarity: A time series is said be stationary if the statistical properties are essentially constant through time. In order words, a series X_t is said to be stationary if for any admissible time points t_1, t_2, ..., t_n and any k, all the joint moments up to order 2 of $\{X_{t_1}, X_{t_2}, ..., X_{t_n}\}$ exist and is equal to the corresponding joint moments up to order 2 of $\{X_{t_{1+k}}, X_{t_{2+k}}, ..., X_{t_{n+k}}\}$.

Given an autoregressive process of order p [denoted AR (p)]:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t$$

The stationarity condition is that the roots of the characteristic equation:

$$\phi\,(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) = 0$$

must lie outside the unit circle.

Autocorrelation function (ACF) and partial autocorrelation (PACF): The order (p) of a model is identified by examining the behaviour of the autocorrelation and partial autocorrelation function for the values of a stationary time series. For an AR process, the pacf cut off after lag p, while its autocorrelation function is infinite in extent and consists of a mixture of damped exponentials and/or damped sine waves[10].

Missing value approach: Suppose the data with the missing value is being generated by an autoregressive (AR) process of order p. That is:

$$\phi\,(B)\,X_t = \varepsilon_t$$

where, $\phi\,(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, the ϕ_i's are the AR parameters and B is the backward shift operator.

Let us consider a time series X_t with n observations. Let the ith value of X_t be missing. Let the missing value x_i be such that the values x_{i-1}, x_{i-2}, ... , x_{i-p} exist; where, p is the order of the series x_i, x_{i+1}, ... , x_n.

This method proposes that an AR (p) model be fitted to the series x_i, x_{i+1}, ... , x_n. Suppose the resulting model is:

$$X_t = \phi_1^{(1)} X_{t-1} + \phi_2^{(1)} X_{t-2} + ... + \phi_p^{(1)} X_{t-p} + \varepsilon_t$$

Then the missing value is estimated at the first step as:

$$\widehat{x_i^{(1)}} = \widehat{\phi_1^{(1)}} x_{i-1} + \widehat{\phi_2^{(1)}} x_{i-2} + ... + \widehat{\phi_p^{(1)}} x_{i-p}$$

The value $\widehat{x_i^{(1)}}$ obtained is then substituted at the missing position in the data and the analysis is repeated using the entire data $x_1, x_2, ..., \widehat{x_i^{(1)}}, ..., x_n$ to obtain a new model with different parameters:

$$X_t = \phi_1^{(2)} X_{t-1} + \phi_2^{(2)} X_{t-2} + ... + \phi_p^{(2)} X_{t-p} + \varepsilon_t$$

The final value of x_i is re-estimated as:

$$\widehat{x_i^{(2)}} = \widehat{\phi_1^{(2)}} x_{i-1} + \widehat{\phi_2^{(2)}} x_{i-2} + ... + \widehat{\phi_p^{(2)}} x_{i-p}$$

RESULTS

The above proposed method is illustrated using Blowfly time series data[6]. There are n = 82 number of observations. Now, suppose the 5th observation $x_5 = 4424$ is missing. A time series model is first fitted to the series $x_6 = 2784$ to $x_{82} = 4066$. Following the nature of the autocorrelation function and partial function (using Minitab software), the series is said to follow a AR (1) process giving the model:

$$X_t = 0.986430 X_{t-1} + \varepsilon_t$$

This gives the first estimate of x_5 to be:

$$\widehat{x_5^{(1)}} = 0.986430 x_4 = 0.986430\ (4639) \approx 4576$$

The second step involves replacing this estimate in its missing position and re-applying the Box-Jenkins method to the entire data to obtain the new model:

$$X_t = 0.982933 X_{t-1} + \varepsilon_t$$

This gives the final estimate of x_5 to be:

$$\widehat{x_5^{(1)}} = 0.982933 x_4 \approx 4560$$

Table 1: Estimates of the missing values and their errors

x_i	Actual value	KF	IF	M	NF	TF	A	I	NM
x_5	4424	3342 (1082)	5123 (699)	4139 (285)	4639 (215)	4573 (149)	3712 (712)	4591 (167)	4560 (136)
x_{11}	3835	4131 (296)	3947 (112)	4139 (304)	3920 (85)	3861 (26)	3769 (66)	3941 (106)	3853 (18)
x_{12}	3618	3794 (176)	3812 (194)	4139 (521)	3835 (217)	3791 (173)	3443 (175)	3795 (177)	3770 (152)
x_{24}	2717	3011 (294)	2234 (483)	4139 (1422)	2432 (285)	2517 (200)	2441 (276)	2613 (104)	2660 (57)
x_{29}	3156	3675 (519)	3742 (586)	4139 (983)	3677 (521)	3621 (467)	3975 (819)	3634 (478)	3614 (458)

Thus, any missing observation x_i can be estimated using:

$$\widehat{x_i^{(2)}} = 0.982933x_{i-1}$$

Stationarity of the AR (1) process: For this process, the root of auxiliary equation is:

$$(1\text{-}0.982933B) = 0$$

$$\Rightarrow B = 1.01736334>1$$

Thus, the process is stationary.

Comparison of some old methods with the new method: A comparison of the various methods of estimating missing values in time series with the new proposed method is made. This is going to be assessed by the Absolute Deviations (AD) of the estimated values of the different methods from the actual values in the raw data. This absolute deviations will be called errors. However, few existing methods that are simple in computation and are within the reach of our software will be considered for comparison. The following methods with abbreviations will be compared: Kalman Filter (KF), Influence Function (IF), replacing with mean of the series (M), Naïve Forecast (NF), simple Trend Forecast (TF), average of the last two known observations that bound the missing observation (A), Interpolation (I) and the New Method (NM).

The missing values created at different positions in the Blowfly data set are presented in Table 1. The actual values removed at the different positions to create the missing observations were randomly selected. The estimates provided by the different estimation methods are also presented and the calculated errors (AD) are placed in brackets under each estimate.

In Table 1, it is clearly shown that the New Method (NM) has the smallest errors in all the estimates. This means that amongst all the estimation methods considered, the new method provides the best estimates. Next, the minimum mean square error of the estimate is obtained theoretically and the estimate is shown to be unbiased.

Minimum Mean Square Error (MMSE) and unbiasedness of the estimate: To obtain the MMSE, the variables are treated as random variables and the series X_t is assumed to be stationary with mean, $\mu = 0$. It is already known that if $\{X_t\}$ is stationary and follows an AR (p) process, then it can be converted to an infinite Moving Average (MA) process. That is an AR (p) series X_i can be written:

$$X_i = \varepsilon_i+\psi_1\varepsilon_{i-1}+\psi_2\varepsilon_{i-2}+.... \tag{1}$$

$$= \psi\ (B)\ \varepsilon_i$$

where, $\psi\,(B) = \sum_{j=0}^{\infty}\psi_j B^j$ and $\psi_0 = 1$.

$$\Rightarrow X_i = \sum_{j=0}^{\infty}\psi_j B^j\varepsilon_i = \sum_{j=0}^{\infty}\psi_j\varepsilon_{i-j} \tag{2}$$

where, ε_i ~NIID (0, σ_ε^2).

Since, X_i is a linear of the current and previous random shocks (the ε_i's), then the estimate $\widehat{X}_i$ is also a linear of the current and previous shocks. Thus, it can be written:

$$\widehat{X}_i = \psi_1^*\varepsilon_i + \psi_2^*\varepsilon_{i-1} + \psi_3^*\varepsilon_{i-2} + \ldots$$

Using (2) and noting that E [ε_i ε_i] = σ_ε^2 and E [ε_i ε_j] = 0, the mean square error of the estimate can be obtained as:

$$E\left[X_i - \widehat{X}_i\right]^2 = \left(1+\psi_1^2+\ldots+\psi_p^2\right)\sigma_\varepsilon^2 + \sum_{J=0}^{\infty}\left[\psi_j - \psi_j^*\right]\sigma_\varepsilon^2$$

$$= \sigma_\varepsilon^2\sum_{j=0}^{p}\psi_j^2 + \sigma_\varepsilon^2\sum_{j=0}^{\infty}[\psi_j - \psi_j^*]$$

which is minimised by setting $\psi_j = \psi_j^*$.

Hence, from Wold's decomposition[11]:

$$X_i = \left[\varepsilon_i + \psi_1\varepsilon_{i-1} + \ldots + \psi_p\varepsilon_{i-p}\right] + \psi_1^*\varepsilon_i + \psi_2^*\varepsilon_{i-1} + \psi_3^*\varepsilon_{i-2} + \ldots$$

Which can be written as:

$$X_i = e_i + \widehat{X}_i$$

where, $e_i = \varepsilon_i + \psi_1\varepsilon_{i-1} + \ldots + \psi_p\varepsilon_{i-p}$ is the error of the estimate. Since E $[e_i] = 0$, the estimate $\hat{x}_i$ is unbiased.

DISCUSSION

The study of Anava *et al.*[8] clearly shows that their algorithm must have been applied to a wrong data. This is because they did not consider the underlying mechanism of the data before they cast an online learning problem in which the goal of the learner was to minimize prediction error. This study, however has determined the underlying structure of the data to be of ARIMA type before the application was made. Besides, the aforementioned researches of the subject matter in the review were based only on empirical methods. This study has substantiated its result with theoretical backings.

CONCLUSION

In essence, this study has proposed a method of estimating missing values in a stationary autoregressive (AR) process. Comparative study was carried out with other existing methods and the new method was found to provide the best estimates. To support this, a theoretical proof showed that the estimate obtained is unbiased. Hence there is no gain saying that this method offers another possibility of estimating missing values in a stationary autoregressive (AR) process.

REFERENCES

1. Abraham, B. and A. Thavaneswaran, 1991. A nonlinear time series model and estimation of missing observations. Ann. Instit. Statist. Mathem., 43: 493-504.
2. Phong, C. and R. Singh, 2012. Missing value estimation for time series. Microarray data using linear Dynamical system modelling. Proceedings of the 22nd International Conference on Advanced Information Networking and Applications, March 26-29, 2012, Fukuoka, Japan.
3. Tight, M.R., E.J. Redfern, S.M. Watson and S.D. Clark, 1993. Outlier detection and missing value estimation in time series traffic count data: Final report of SERC project GR/G23180. Working Paper No. 401, Institute of Transport Studies, University of Leeds, Leeds, UK.
4. Maravall, A. and D. Pena, 1996. Missing observations and additive outliers in time series models. Banco de Espana-Servicio de Estudios Documento de Trabajo No. 9612. Deposito Legal: M-16340-1996, Imprenta del Banco de Espana, Spain.
5. Grenander, U. and M. Rosenblatt, 1957. Statistical Analysis of Stationary Time Series. John Wiley, New York.
6. Wei, W.W.S., 2006. Time Series Analysis: Univariate and Multivariate Methods. 2nd Edn., Pearson Addison Wesley, USA., ISBN-13: 9780321322166, Pages: 614.
7. Xia, Y., P. Fabian, A. Stohl and M. Winterhalter, 1999. Forest climatology: Estimation of missing values for Bavaria, Germany. Agric. Forest Meteorol., 96: 131-144.
8. Anava, O., E. Hazan and A. Zeevi, 2015. Online time series prediction with missing data. Proceedings of the 32nd International Conference on Machine Learning, Volume 37, July 6-11, 2015, France, pp: 2191-2199.
9. Ahmad, M.R. and A.M.H. Al-Khazaleh, 2008. Estimation of missing data by using the filtering process in a time series modeling. Electron. J. Statist.
10. Box, G.E.P. and G.M. Jenkins, 1976. Time Series Analysis: Forecasting and Control. Holden-Day Publisher, Oakland, CA., USA., ISBN-13: 9780816211043, Pages: 575.
11. Wold, H.O., 1954. The Analysis of Stationary Time Series. 2nd Edn., Almqvist and Wiksell, USA., Pages: 236.