Thailand Statistician
July 2023; 21(3): 607-615
http://statassoc.or.th
Contributed paper

A Suggested Estimator for AR(1) Model with Missing Observations

Mohamed Abdelsamie Enany [a], Mohamed Khalifa Ahmed Issa [b] and Ahmed Abdelfatah Gad*[a]

[a] Department of Statistics and Insurance, Faculty of Commerce, Zagazig University, Egypt.
[b] The Higher Institute of Cooperative and Managerial Studies, Egypt.
*Corresponding author; e-mail: ahmad.abdalfatah@hotmail.com

Received: 8 November 2020
Revised: 26 June 2021
Accepted: 28 September 2021

Abstract

The paper considers estimation of stationary first-order autoregressive model AR(1) with missing observations. The maximum likelihood method is used to estimate the autoregressive parameter for AR(1) with missing observations. The efficiency of the estimation is affected by treating the initial value required to compute the first value of residuals. The conventional methods treat the initial value as fixed. Therefore, we present new method to estimate AR(1) with missing observations based on treating the initial value as random. The likelihood function is uniquely maximized and a new closed-form estimator for AR(1) in case of missing observations is developed. Monte Carlo simulation studies and a real data analysis showed that the bias and efficiency of the new estimators are more reliable than the conventional estimators. Moreover, the proposed method provides better estimates of missing values than the existing methods.

Keywords: AR(1), missing observations, stationary, unconditional maximum likelihood.

1. Introduction

Autoregressive (AR) models consider one of the most popular time series models. These models predict future values based on past values from the same time series. AR(1) is the first order autoregression meaning that the current value is based on the immediately preceding value. Consider the AR(1) model

$$x_k = \rho x_{k-1} + \varepsilon_k, k = 1, 2, \ldots, n, \tag{1}$$

where ρ represents autoregressive parameter, ε_k is i.i.d. $N(0,\sigma^2)$, the error term ε_k and the independent variable x_{k-1} are independent as $E(x_{k-1},\varepsilon_k)=0,$ and it is assumed that the model is stationary which $|\rho|<1.$ Unconditional autoregressive model (UAR) treats the initial value x_0 (which is required to compute ε_1) as random and as a result, x_1 has a random distribution $N(0,(\sigma^2/(1-\rho^2))).$ While if the initial value x_0 is treated as fixed, the model is called conditional autoregressive model (CAR) (Wei 2006). The problem of missing values is often observed in time series analysis. In this case the standard estimation methods cannot be applied to such AR models.

To deal with this problem, various methods have been proposed for estimation of the missing values. In particular, Pourahmadi (1989) obtained the best linear “interpolator” of a missing value for a stationary time series regard to the mean square error. For estimation of missing values when a time series is regarded as a state-space model, see Shumway and Stoffer (2017). The literatures have several techniques to estimate CAR with missing observations. For example, Dunsmuir and Robinson (1981) presented Yule Walker method to estimate the parameter of AR(1) model with missing observations. Takeuchi (1995) developed ordinary least squares (OLS) method to estimate the parameter of AR(1) model without constant with missing values. Hamaz and Ibazizen (2009) used Pitman-Closeness criterion to compare between two estimation methods of missing observations in AR(1). Ding et al. (2010) proposed extended stochastic gradient algorithm to fit AR with missing observations. El-Sayed et al. (2016) provided OLS estimation for the first order auto-regressive model AR(1) with (without) constant term in case of missing values. Moreover, the properties of the estimates are discussed. Abdelwahab (2016) provided Yule Walker, OLS and weighted symmetric methods to estimate AR(1) and AR (2) models with (without) constant with missing observations. Abdelwahab and Issa (2019) derived forms of the moments of AR(P) model with missing observations. For UAR with missing observations, Enany et al. (2020) introduced a closed form estimator for ρ in case of missing observations using maximum likelihood (ML) estimator. However, they ignored the term $\ln(1-\rho^2)$ in logarithmic function which causes the nonlinearity as in Box and Jenkins (1976). In this paper we devolve closed-form estimation for AR(1) in case of missing observations. We first employ approach of Parzen (1963) which formulated AR(1) with missing observations as a specific case of an amplitude modulated stationary process. The maximum likelihood (ML) method is introduced, and the initial value is treated as random. The likelihood function is uniquely maximized without ignoring any terms using the theorem of Hasza (1980). The rest of the paper is organized as follows; in the next section, the proposed estimator is devolved. While in Section 3, Monte Carlo simulation study is conducted to compare the performance of the proposed estimator with the classical estimators. A real data set is analyzed in Section 4. In Section 5 concluding remarks are presented.

2. The Proposed Estimator

By considering model (1), to account for the missing observations, the approach of Parzen (1963) can be employed. Consider that the observed data $\{y_k\}(k=1,2,3,\ldots,n)$ can be expressed as

$$y_k = a_k x_k, \tag{2}$$

where $\{a_k, k=1,2,3,\ldots,n\}$ is stochastic Markov process. It is assumed that a_k and ε_k are independent, a_k represents the state of observation such that

$$a_k = \begin{cases} 1 & \text{if } x_k \text{ is observed,} \\ 0 & \text{if } x_k \text{ is missing.} \end{cases} \tag{3}$$

Multiplying (1) by a_k, we get

$$a_k x_k = \rho a_k x_{k-1} + a_k \varepsilon_k. \tag{4}$$

Substituting the value of (2) in (4), it gives

$$y_k = \rho a_k x_{k-1} + a_k \varepsilon_k. \tag{5}$$

By using (2), we obtain

$$y_{k-1} = a_{k-1} x_{k-1}. \tag{6}$$

Then multiplying (6) by a_k and using the assumption that $a_k a_{k+1} = 1$ as in Abdelwahab (2016) we get

$$x_{k-1} = a_k y_{k-1}. \tag{7}$$

Substituting (7) in (5) and simplifying, we obtain

$$a_{k+1} y_k = \rho a_k y_{k-1} + \varepsilon_k. \tag{8}$$

Lemma *The maximum likelihood estimator for AR(1) in case of missing observations is obtained by*

$$\hat{\rho}_{UML} = \frac{2}{3}\delta_3^{-1}(\delta_2^2 - 3\delta_1\delta_3)^{1/2}\cos(\theta) - \frac{1}{3}\delta_2\delta_3^{-1}, \tag{9}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left\{(1-\hat{\rho}^2)a_2 y_1^2 + \sum_{k=2}^{n}(a_{k+1}y_k - \hat{\rho}a_k y_{k-1})^2\right\},$$

where

$$\delta_3 = \left(\frac{n-1}{n}\right)\sum_{k=2}^{n-1} a_{k+1}y_k^2, \quad \delta_2 = -\left(\frac{n-2}{n}\right)\sum_{k=2}^{n} y_k y_{k-1}, \quad \delta_1 = -\sum_{k=2}^{n-1} a_{k+1}y_k^2 - \frac{1}{n}\sum_{k=1}^{n} a_{k+1}y_k^2, \quad \delta_0 = \sum_{k=2}^{n} y_k y_{k-1},$$

$$\theta = \frac{1}{3}\arccos\left[-\frac{1}{2}(2\delta_2^3 - 9\delta_1\delta_2\delta_3 + 27\delta_0\delta_3^2)(\delta_2^2 - 3\delta_1\delta_3)^{-3/2}\right] + \frac{4\Pi}{3}.$$

Proof: Since $x_1 \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$, the marginal probability density function for y_1 can be written as

$$f(y_1) = \left[\frac{(1-\rho^2)}{2\pi\sigma^2}\right]^{1/2} \exp\left[-\frac{(1-\rho^2)a_2 y_1^2}{2\sigma^2}\right]. \tag{10}$$

The probability density function for the observations $y_2, y_3, \ldots, y_n$ with assumed $Y_1 = y_1$ is given by

$$f(y_2, y_3, \ldots, y_n | Y_1 = y_1) = \frac{1}{(\sigma\sqrt{2\pi})^{n-1}} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=2}^{n}(a_{k+1}y_k - \rho a_k y_{k-1})^2\right]. \tag{11}$$

The unconditional joint probability density function for $y_1, y_2, \ldots, y_n$ can be found by multiplying the marginal probability density function for y_1 in (10) by the probability density function in (11) this gives

$$f(y_1, y_2, \ldots, y_n) = \frac{(1-\rho^2)^{1/2}}{(2\sigma^2\pi)^{n/2}} \exp\left[-\frac{S(\rho)}{2\sigma^2}\right], \tag{12}$$

where

$$S(\rho) = \left[(1-\rho^2)a_2 y_1^2 + \sum_{k=2}^{n}(a_{k+1}y_k - \rho a_k y_{k-1})^2\right]. \tag{13}$$

Typically, $S(\rho)$ is called the unconditional sum of squares and the function in (12) represents the unconditional likelihood $L(\rho, \sigma^2 | y_1, y_2, \ldots, y_n)$. To get the estimate of ρ and σ^2; logarithm of likelihood function should be differentiated with respect to ρ and σ^2 and setting the result equal to zero it gives

$$\frac{\partial \ln L(\rho,\sigma^2)}{\partial \rho} = -\frac{\hat{\rho}}{(1-\hat{\rho}^2)} - \frac{1}{2\sigma^2}\left[-2\hat{\rho}a_2y_1^2 - 2\sum_{k=2}^{n} a_k y_{k-1}(a_{k+1}y_k - \rho a_k y_{k-1})\right] = 0, \tag{14}$$

$$\frac{\partial \ln L(\rho,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(1-\hat{\rho}^2)a_2y_1^2 + \sum_{k=2}^{n}(a_{k+1}y_k - \rho a_k y_{k-1})^2 = 0. \tag{15}$$

Equation (14) can be rewritten as

$$\left\{(1-\hat{\rho}^2)a_2y_1^2 - \sigma^2\right\}\hat{\rho} + \left[(1-\hat{\rho}^2)\sum_{k=2}^{n} a_k y_{k-1}(a_{k+1}y_k - \rho a_k y_{k-1})\right] = 0. \tag{16}$$

And from (15) it gives

$$\sigma^2 = \frac{1}{n}\left[(1-\hat{\rho}^2)a_2y_1^2 + \sum_{k=2}^{n}(a_{k+1}y_k - \rho a_k y_{k-1})^2\right]. \tag{17}$$

Substituting for σ^2 in (16) and simplifying we get

$$g(\rho) = \delta_3\rho^3 + \delta_2\rho^2 + \delta_1\rho + \delta_0 = 0. \tag{18}$$

As in Hasza (1980), Equation (18) has exactly one zero in each of the intervals $(-\infty,-1),(-1,1)$ and $(1,\infty)$ and it has three real and unequal roots they may be expressed as

$$r_j = 2\left(\frac{\alpha}{3}\right)^{1/2}\cos\left[\frac{1}{3}\arccos\left(\frac{b}{2}\left(\frac{27}{\alpha^2}\right)^{1/2}\right) + \frac{2(j-1)\pi}{3}\right] - \frac{p}{3},\ j = 1,2,3, \tag{19}$$

where $\alpha = \frac{1}{3}(p^2 - 3q),\ b = -\frac{1}{27}(2p^3 - 9pq + 27u),\ p = \delta_2\delta_3^{-1},\ q = \delta_1\delta_3^{-1},\ u = \delta_0\delta_3^{-1}.$

It is noted that $r_2 \le r_3 \le r_1$ because $x \in (-1,1)$ and $\arccos(x) \in (0,\pi),$ we find that

$$\cos\left[\frac{1}{3}\arccos(x) + \frac{2\pi}{3}\right] \le \cos\left[\frac{1}{3}\arccos(x) + \frac{4\pi}{3}\right] \le \cos\left[\frac{1}{3}\arccos(x)\right].$$

It follows that $r_1 \in (1,\infty),\ r_2 \in (-\infty,-1),\ r_3 \in (-1,1)$ and as a result the maximum likelihood estimator of ρ in (9) is obtained by (19) for $j = 3.$

Remark 1: It is noted that (14) is not linear in $\rho.$ The nonlinearity is due to presence of quantity in likelihood function (13). According to Box and Jenkins (1976) the effect of $\ln(1-\rho^2)$ in likelihood function can be neglected as it tends to zero when $n \to \infty.$ By ignoring this term and solving (14), a closed form estimator for ρ can be obtained as

$$\hat{\rho}_{CML} = \frac{\sum_{k=2}^{n} y_k y_{k-1}}{\sum_{k=3}^{n} a_k y_{k-1}^2}. \tag{20}$$

This estimator is obtained by Enany et al. (2020).

3. The Simulation Study

This section aims to investigate the properties of the proposed estimation through the simulation study in small, moderate, and large samples. R software is used to perform the Monte Carlo simulation study. In our simulation study, Monte Carlo experiments were carried out based on (1) and under the assumptions given in Section 1 above. The simulated model is generated as follows:

1. The AR(1) model without constant is generated using the function arima.sim in R software. The errors are generated from the normal distribution with mean equals to μ and standard deviation $\sigma,$ and the autoregressive parameter ρ is chosen to be 0.3 and 0.6.

2. The different samples size has been used as: n= 50, 100, 150 and 200.

3. To investigate the resistance of the proposed method, we randomly generate different percentages of missing observations $\tau\%$ equals to 5, 15 and 25.

4. All Monte Carlo experiments involved 1000 replications and all the results of all separate experiments are obtained by precisely the same series of random numbers.

We compare performance of the proposed method (UML) defined in (9) with the method (CML) which defined in (20), and the method of Yule Walker (YW) which defined by Dunsmuir and Robinson (1981) which treats the initial value as fixed. a. We compute the bias and mean squared error (MSE) for each estimator. The bias and MSE are calculated by

$$\text{Bias}=\frac{1}{L}\sum_{l=1}^{L}(\hat{\rho}_l-\rho) \quad \text{and} \quad \text{MSE}=\frac{1}{L}\sum_{l=1}^{L}(\hat{\rho}_l-\rho)^2,$$

where ρ is the true autoregressive parameter in (1), L is the number of iterations $l=1,2,...,L,$ and $\hat{\rho}_l$ is the estimated value for ρ in the trial number $l.$

The results of simulation are recorded in Tables 1 and 2. The tables present the bias and MSE values for different estimators and for different sample sizes and for the various percentages of missing observations. We can conclude that in all simulation cases, it is noticeable that the values of bias and MSE for UML estimator is smaller than those for CML and YW estimators. In other word, we can conclude that the proposed method is more efficient than the other methods. Generally, the bias and MSE values for the different methods are decreasing as the sample size is increasing. Also, as the percentage of missing observations increases, the bias and MSE values of all estimators increase.

4. Real Data Application

The feasibility of the proposed estimator is illustrated using Blowfly time series data (Wei 2006). The data is taken from Nicholson (1950). It consists of the numbers of adult flies with balanced sex ratios kept inside a cage and given a fixed amount of food daily. The sample size is 82. We consider $x_k=\ddot{x}_k/1000,$ where $\ddot{x}_k$ is numbers of adult flies at time $k.$ Table 3 presents Phillips-Perron and Augmented Dickey-Fuller tests for stationary assumption. The null hypothesis of the unit root test is rejected for different lags. As a conclusion it is seen that the blowfly data is stationary. The next step is to determine the best model to fit the data. The sample ACF and PACF are shown in Figure 1. It is noted that the sample ACF decays exponentially and the sample PACF cuts of after lag 1. Thus the AR(1) model is more appropriate to fit the data. Moreover, Table 4 compares the AR(1) model with some competitive models based on AIC. It is obvious that AR(1) has the minimum AIC. So that the best model to fit the blowfly data is AR(1).

Table 1 Bias and MSE values of the estimators for different sample sizes and for different percentages of missing observations, ρ =0.3

Estimators	Sample size			
	50	100	150	200
$\tau = 5\%$				
CML	0.11720 (0.02147)	0.08099 (0.01031)	0.06595 (0.00685)	0.05661 (0.00498)
YW	0.11542 (0.02095)	0.08023 (0.01014)	0.06579 (0.00684)	0.05656 (0.00496)
UML	0.11500 (0.02075)	0.08017 (0.01012)	0.06546 (0.00676)	0.05635 (0.00494)
$\tau = 15\%$				
CML	0.12904 (0.02603)	0.08891 (0.01245)	0.07262 (0.00828)	0.06460 (0.00671)
YW	0.12755 (0.02551)	0.08912 (0.01256)	0.07311 (0.00838)	0.06531 (0.00682)
UML	0.12665 (0.02515)	0.08794 (0.01220)	0.07213 (0.00818)	0.06428 (0.00664)
$\tau = 25\%$				
CML	0.14583 (0.01662)	0.10255 (0.01662)	0.08415 (0.01112)	0.07345 (0.00842)
YW	0.14449 (0.03258)	0.10317 (0.01682)	0.08521 (0.01137)	0.07471 (0.00867)
UML	0.14302 (0.03195)	0.10152 (0.01631)	0.08362 (0.01098)	0.07312 (0.00835)

Note: MSE values are written in parentheses beside the values of bias.

Table 2 Bias and MSE values of the estimators for different sample sizes and for different percentages of missing observations, ρ =0.6

Estimators	Sample size			
	50	100	150	200
$\tau = 5\%$				
CML	0.10007 (0.01600)	0.068968 (0.00759)	0.05572 (0.00497)	0.04787 (0.00362)
YW	0.10035 (0.01634)	0.07041 (0.00801)	0.05714 (0.00523)	0.04969 (0.00387)
UML	0.09838 (0.01574)	0.06847 (0.00755)	0.05532 (0.00495)	0.04772 (0.00362)
$\tau = 15\%$				
CML	0.11324 (0.02048)	0.07705 (0.00960)	0.06287 (0.00626)	0.05566 (0.00496)
YW	0.11661 (0.02158)	0.08210 (0.01081)	0.06741 (0.00716)	0.05942 (0.00562)
UML	0.11118 (0.01100)	0.07637 (0.00953)	0.06253 (0.00623)	0.05543 (0.00495)
$\tau = 25\%$				
CML	0.12368 (0.02490)	0.08937 (0.01245)	0.06971 (0.00786)	0.06188 (0.00614)
YW	0.13283 (0.02799)	0.09892 (0.01515)	0.07761 (0.00971)	0.06984 (0.00774)
UML	0.12134 (0.02420)	0.08860 (0.01230)	0.06928 (0.00781)	0.06167 (0.00613)

Note: MSE values are written in parentheses beside the values of bias.

Table 3 Phillips-Perron and augmented Dickey-Fuller unit root tests

lags	Phillips and Perron test		Augmented Dickey-Fuller test	
	Tau	Pr < Tau	t	Pr < t
0	–3.64	0.006	–3.64	0.006
1	–3.74	0.005	–3.54	0.009
2	–3.81	0.004	–3.47	0.011
3	–3.78	0.004	–2.84	0.056

Table 4 AIC results

Model	AIC
AR(1)	1,353.862
AR(2)	1,354.733
AR(3)	1,356.100
MA(1)	1,377.748
MA(2)	1,358.789
MA(3)	1,357.891

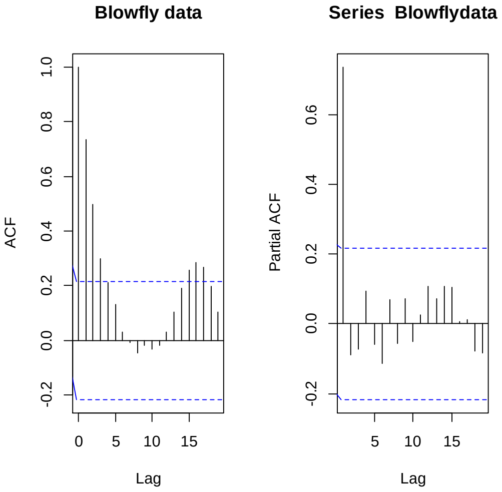

Figure 1 Sample ACF and PACF for Blowfly time series data

In the following, we randomly generate different percentages of missing values (5%, 15%, and 25%) in the Blowfly data set using delete_MCAR function in R software. we compare performance of the proposed method (UML) with the method (CML) which defined in (20), and the method of Yule Walker with missing observations (YW) Dunsmuir and Robinson (1981). To compare between the different methods; we firstly use the three methods to estimate the autoregressive parameter ρ in the presence of the missing observations. Then we use the estimated parameter to provide estimates of missing observations as

$$\hat{x}_k = \hat{\rho} x_{k-1}, k = 1, 2, \ldots, n. \qquad (21)$$

To measure the accuracy of estimating missing data for different methods; we compute mean absolute error (MAE) and mean absolute percentage error (MAPE), which are calculated as

$$\text{MAE}=\frac{1}{n}\sum_{k=1}^{n}|\hat{x}-x| \quad \text{and} \quad \text{MAPE}=\frac{1}{n}\sum_{k=1}^{n}\left|\frac{\hat{x}-x}{x}\right|,$$

where $x's$ are the true values and $\hat{x}'s$ are the predictions. Table 5 shows the values of MAE and MAPE for different methods and for different percentages of missing observations. It is obvious that the new method (UML) has the smallest values of MAE and MAPE for all different percentages of missing observations. This ensures that, the new method provides the best estimates compared with the different estimation methods. Figure 2 illustrates the exact and estimated values for Blowfly time series data using the new method when the percentage of missing observations equals to 25%.

Table 5 Values of MAE and MAPE for different methods and for different percentages of missing observations

Estimators	τ%		
	5	15	25
CML	0.71535 (0.17899)	0.71715 (0.17985)	0.71820 (0.18031)
YW	0.71481 (0.17873)	0.73110 (0.18524)	0.74939 (0.19161)
UML	0.71158 (0.17709)	0.71298 (0.17787)	0.71388 (0.17830)

Note: MAPE values are written in parentheses beside the values of MAE.

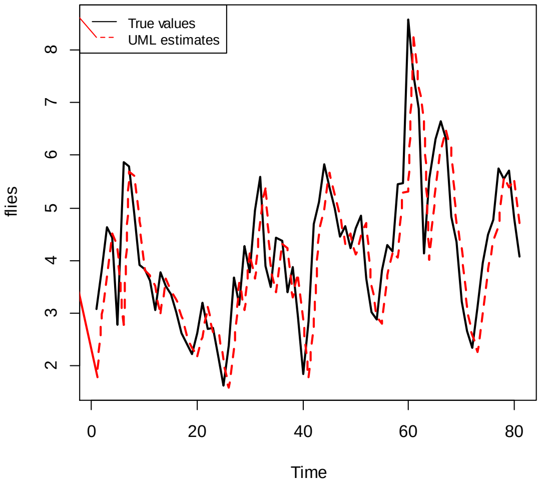

Figure 2 Exact and estimated values for Blowfly time series data using UML method when the percentage of missing observations equals to 25%

5. Concluding Remarks

In this paper, we developed the maximum likelihood method to estimate the autoregressive parameter for stationary first-order autoregressive model AR(1) with missing observations. Firstly AR(1) with missing observations is formulated as a specific case of an amplitude modulated stationary process. The likelihood function is formulated based on treating the initial value as random. Then the likelihood function is uniquely maximized using the Theorem of Hasza (1980). A new closed-form estimator in case of missing observations is developed. Monte Carlo Simulation studies and real dataset have been provided to evaluate the performance of proposed estimator and compared it with the existing methods. Simulation studies and real dataset illustrate that the proposed estimator performs better than the existing estimators.

References

Abdelwahab MM. On Parameter Estimation of Time Series Models with Missing Observations. PhD [thesis]. Egypt: Cairo University; 2016.

Abdelwahab MM, Issa MK. Forms of the moments of AR(P) model with missing observations. Egyptian Stat J. 2019; 63(1): 39-44.

Box GEP, Jenkins GM. Time series analysis: Forecasting and control. San Francisco: Holden-Day; 1976.

Ding J, Han L, Chen X. Time series AR modeling with missing observations based on the polynomial transformation. Math Comput Model. 2010; 51(5-6): 527-536.

Dunsmuir W, Robinson PM. Estimation of time series models in the presence of missing data. J Am Stat Assoc. 1981; 76(375): 560-568.

El-Sayed SM, El-Sheikh AA, Abdelwahab MM. Estmation of AR(1) models with missing values. Adv Appl Stat. 2016; 49(6): 485-492.

Enany, MA, Issa MK, Gad AA. Maximum likelihood estimation of unconditional AR(1) model with missing observations. Working paper. Egypt: Zagazig University; 2020.

Hamaz A, Ibazizen M. Comparison of two estimation methods of missing values using Pitman-closeness criterion. Commun Stat Theory Methods. 2009; 38(13): 2210-2213.

Hasza DP. A note on maximum likelihood estimation for the first-order autoregressive process. Commun Stat Theory Methods. 1980; 9(13): 1411-1415.

Nicholson AJ. Population oscillations caused by competition for food. Nature. 1950; 165(4195): 476-477.

Parzen E. On spectral analysis with missing observations and amplitude modulation. Sankhyā, Series A. 1963; 25: 383-392.

Pourahmadi M. Estimation and interpolation of missing values of a stationary time series. J Time Anal. 1989; 10(2): 149-169.

Shumway RH, Stoffer DS. Time series analysis and its applications: with R examples. New York: Springer; 2017.

Singh, S. Advanced Sampling Theory with Applications. Netherlands: Kluwer Academics Press; 2003.

Takeuchi K. A comment on “Recent Development of Economic Data Analysis” at the 63rd Annual Meeting of Japan Statistical Society; 1995.

Wei WWS. Time series analysis univariate and multivariate methods. New York: Addison Wesley; 2006.