Estimation of Right-censored SETAR-type Nonlinear Time-series Model

Syed Ejaz Ahmed[1], *Dursun* Aydın[2], and *Ersin* Yılmaz[2,*]

[1]Department of Mathematics and Statistics, Faculty of Mathematics and Science, Brock University, St. Catharines, ON, L2S 3A1, Canada
[2]Department of Statistics, Faculty of Science, Mugla Sitki Kocman University, 48000, Mugla, Turkey

Abstract. This paper focuses on estimating the Self-Exciting Threshold Autoregressive (SETAR) type time-series model under right-censored data. As is known, the SETAR model is used when the underlying function of the relationship between the time-series itself (Y_t), and its p delays $\left(Y_{t-j}\right)_{j=1}^{p}$ violates the linearity assumption and this function is formed by multiple behaviors that called regime. This paper addresses the right-censored dependent time-series problem which has a serious negative effect on the estimation performance. Right-censored time series cause biased coefficient estimates and unqualified predictions. The main contribution of this paper is solving the censorship problem for the SETAR by three different techniques that are kNN imputation which represents the imputation techniques, Kaplan-Meier weights that is applied based on the weighted least squares, synthetic data transformation which adds the effect of censorship to the modeling process by manipulating dataset. Then, these solutions are combined by the SETAR-type model estimation process. To observe the behavior of the nonlinear estimators in practice, a simulation study and a real data example are carried out. The Covid-19 dataset collected in China is used as real data. Results prove that although the three estimators show satisfying performance, the quality of the estimate SETAR model based on the kNN imputation technique dominates the other two estimators. **Keywords:** Censored time-series, regime-switching model, regression analysis, imputation.

1 Introduction

In the real world, datasets examined using time-series analysis often involve issues, such as censorship and nonlinearity, that directly prevent accurate analysis unless appropriately solved. In practice, these problems in the datasets are generally ignored. For econometric or financial time-series data, it is very common that the variance and autocorrelation of the series may change, which makes it necessary to use nonlinear time-series analysis. Data obtained during financial crises or radical changes in the economic situation of a country can be used as examples of a nonlinear time-series. In addition, we know that due to different reasons, including technical issues, individual mistakes, device errors, and so on, it may not be possible to observe all data points. For instance, for a fixed period of a risk assessment study, borrowers may join the study at various stages, and they may or may not default before

[*]e-mail: ersinyilmaz@mu.edu.tr

the study ends. In this case, the observation is censored from the right randomly (see [1]). In this example, we can see two problems, nonlinearity, and right-censorship, that need to be solved to facilitate accurate analysis of the data.

Regarding the nonlinear time-series model, there are several types of models referred to as bilinear models by Granger and Andersen [2]. These models include the threshold autoregressive model (TAR) and its specific case Self-Exciting TAR (SETAR) proposed by Tong [3] and discussed further by Fan and Yao [4], the smooth transition autoregressive models (STAR) of Terasvirta and Anderson [5] and Markov-switching, introduced by Hamilton [6]. Notice that these models are capable of reflecting the true behavior of the time-series by representing the different regimes in it. This paper considers the estimation of the SETAR model under right-censored time-series data. There are several important studies about estimating SETAR models without censorship; these include Petruccelli [7], Gooijer [8], Milheiro-Oliveira [9], Naik and Mohan [10], and Aydın and Mermi [11], among others. However, in the literature, SETAR models are not considered under censored time-series. This paper aims to fill that gap.

There have been several important studies about right-censored time-series models, including Park et al. [12], Khardani et al. [13], Aydın and Yılmaz [14]. These studies do not consider nonlinearity and instead focus on different aspects of modeling. However, useful solution techniques are introduced for the right-censorship problem, such as synthetic data kNN imputation, Kaplan-Meier weights (see [14]), and synthetic data transformation [15], which is adapted to the time series modeling by [13].

The main purpose of this study is to introduce SETAR model estimation based on the aforementioned three censorship solution techniques: kNN imputation, Kaplan-Meier weights (KMW), and synthetic data transformation (SDT). Both the problems of nonlinearity and censorship can be solved appropriately using these techniques, but the advantages and disadvantages of each of the three estimation procedures are discussed to determine the best procedure for nonlinear and right-censored time-series modeling.

The organization of the paper is as follows: Section 2 introduces the methodology of the solution techniques for the right-censorship problem and their integration into the SETAR model. Section 3 presents the statistical properties of the introduced model estimation. Evaluation metrics are given in Section 4. A simulation study and a real data case study are carried out in Section 5. Conclusions are provided in Section 6.

2 Material and Methods

To estimate the SETAR model with right-censored time-series data, the censorship needs to be solved first because censorship solution techniques change the observations of the series.

Consider the time-series variable as $\{Y_t\}_{t=1}^{n}$ with sample size n. In many cases, one may not be able to observe all Y_t values completely. Instead, some Y_t values will be accurately recorded and others incomplete or censored. Y_t is censored from the right by censoring variable C_t, which means that Y_t is observed partly and is recorded as C_t. Censoring of the response measurements occurs in different situations in business, finance, and economics. If censored observations are ignored, the estimated time-series models are usually unreliable regarding parameter estimates.

To express the censoring mechanism specifically, let us consider that Z_t is the incomplete (right-censored) observed time-series instead of Y_t due to C_t. This case can be formulated as follows:

$$Z_t = min\left(Y_t, C_t\right), \delta_t = I\left(Y_t \leq C_t\right), \tag{1}$$

where $I(.)$ is an indicator function that involves the information of censorship existence and δ_t is a bivariate variable that involves the censoring information. From Eq. (1), it is assumed that the censored time-series are denoted with pairs $\{Z_t, \delta_t), t = 1, 2, \ldots, n\}$.

As indicated in Section 1, the TAR model considers the different regimes at different times, meaning that the series exhibit threshold behavior (see [3, 4]). Accordingly, the regimes in the series are determined by a threshold or transition variable S_t, which depends on threshold value m. Here, S_t is one of the lags of the right-censored time-series Z_t as $S_t = Z_{t-d}$ where d is a lag parameter. Thus, the "self-exciting" definition is realized in the SETAR model, which is defined as follows:

$$Z_t = \left[\phi_{(1,0)} + \sum\nolimits_{j=1}^{q_1} \phi_{1,j} Z_{t-j}\right] (1 - I(Z_{t-d} > m)) + \left[\phi_{(2,0)} + \sum\nolimits_{i=1}^{q_2} \phi_{2,j} Z_{t-i}\right] (1 - I(Z_{t-d} > m)) + \varepsilon_t, \tag{2}$$

where ε_t's are random error terms that are independent and identically distributed with zero mean and constant variance σ_ε^2. $\phi_1 = \left\{\phi_{1,j}, j = 1, \ldots, q_1\right\}$ and $\phi_2 = \{\phi_{2,i}, i = 1, \ldots, q_2\}$ are coefficients of the SETAR models to be estimated with autoregressive degrees, where q_1 and q_2 are degrees of the lower and upper regimes of the autoregressive (AR) model, respectively. Notice that threshold variable S_t depends on the lag (d) and threshold (m), which should therefore be chosen suitably (See [16]). In Eq. (2), the SETAR model is provided with two regimes.

The main interest here is estimating the vectors of autoregressive coefficients $\mathbf{\Phi}(m) = \left(\phi_1^T, \phi_2^T\right)$ of model Eq. (2) for both regimes. In this matter, Tong [3] proposed a maximum likelihood estimator (MLE) under the assumption of normally distributed error terms and uses the Akaike information criterion (AIC) to choose the threshold constant (m) and optimal lag (d). However, MLE works well only when both regimes of Eq. (2) are first-order autoregressive models, which limits the advantages of the SETAR model (see [17]). On the other hand, [7] and [17] show that the conditional least squares (CLS) estimate of $\mathbf{\Phi}(m) = \left(\phi_1^T, \phi_2^T\right)$ has consistency for SETAR models with different autoregressive degrees and the number of regimes. Also, [11] applied this method to the different data examples successfully. This study, therefore, focuses on the CLS approach for right-censored SETAR model estimation.

Let us rewrite model Eq. (2) as follows:

$$Z_t = \phi_1^T \mathbf{X}_{1t} I(Z_{t-d} \leq m) + \phi_2^T \mathbf{X}_{2t} I(Z_{t-d} > m) + \varepsilon_t, \tag{3}$$

where $\mathbf{X}_{1t} = \left(1, Z_{t-1}, \ldots, Z_{t-q_1}\right)$, $\mathbf{X}_{2t} = \left(1, Z_{t-1}, \ldots, Z_{t-q_2}\right)$ are covariate matrices that are formed by the lags of the dependent variable $\phi_1 = \left(\phi_{(1,0)}, \ldots, \phi_{(1,q_1)}\right)$ and $\phi_2 = \left(\phi_{(2,0)}, \ldots, \phi_{(2,q_1)}\right)$ are the vectors of the coefficients for lower and upper regimes of Eq. (3) and ε_t's are the stationary random error terms with a constant variance.

Model Eq. (3) can be simply written and given by

$$Z_t = \mathbf{X}_t^T(m) \mathbf{\Phi}(m) + \varepsilon_t(m), 1 \leq t \leq n, \tag{4}$$

where $\mathbf{X}_t(m) = \left[\mathbf{X}_t^T I(Z_{t-d} \leq m), \mathbf{X}_t^T I(Z_{t-d} > m)\right]$, and as mentioned above, $\mathbf{\Phi}(m) = \left(\phi_1^T, \phi_2^T\right)$. For determined m values, CLS estimates $\hat{\mathbf{\Phi}}(m) = \left(\hat{\phi}_1^T, \hat{\phi}_2^T\right)$ can be obtained as follows:

$$\hat{\mathbf{\Phi}}(m) = \sum\nolimits_{t=1}^{n} \left[\mathbf{X}_t^T(m) \mathbf{X}_t(m)\right]^{-1} \sum\nolimits_{t=1}^{n} \left[\mathbf{X}_t^T(m) Z_t\right], \tag{5}$$

Notice that the estimation of Eq. (4) depends on the optimal values of m and d parameters. Choosing m and d is achieved by calculating ε_t. From equation Eq. (5) , $\hat{\varepsilon}_t(m) =$

$Z_t - \mathbf{X}_t(m)\boldsymbol{\Phi}(m)$ and the variance of the model is obtained as $\hat{\sigma}^2_\varepsilon(m) = n^{-1}\sum_{t=1}^n \hat{\varepsilon}^2_t(m)$. In addition, determining the threshold variable Z_{t-d} and optimal lag d is an important problem to be solved. Therefore, regarding the SETAR model, m and d are selected by minimizing $\hat{\sigma}^2_\varepsilon(m)$ by doing an appropriate grid search in the following minimization problem:

$$\left(\hat{m}, \hat{d}\right) = \arg\min \hat{\sigma}^2_\varepsilon(m, d), m \in Z^+, d \in Z^+, \tag{6}$$

where $\hat{m}$and$\hat{d}$ are the chosen threshold and lag parameters and the lag parameter should ensure the condition $d < (q_1, q_2)$. Instead of optimizing Eq. (6) , as shown by [18], AIC is used to choose optimal m and d as below:

$$AIC\left(\hat{m}, \hat{d}\right) = n_1 \ln\left(\sigma^2_{\varepsilon_1}(m, d)\right) + n_2 \ln\left(\sigma^2_{\varepsilon_2}(m, d)\right) + 2(q_1 + 1) + 2(q_2 + 1), \tag{7}$$

where n_1 and n_2 are the sample sizes in lower and upper regimes, respectively. Similarly, $\sigma^2_{\varepsilon_1}$ and $\sigma^2_{\varepsilon_2}$ are the variances of these regimes. It should also be noted that the selection of $\left(\hat{m}, \hat{d}\right)$ is realized in three steps that are discussed by [19]. Thus, after the selection of $\left(\hat{m}, \hat{d}\right)$ using Eq. (7) , $\hat{\boldsymbol{\Phi}}(m)$ is obtained using Eq. (5) . However, due to right-censored time-series Z_t, equation Eq. (5) cannot be used directly. Therefore, the following three solution techniques are introduced: KMW, SDT, and kNN imputation.

2.1 Kaplan-Meier Weights

This section introduces adapting the SETAR model estimation given in Eq. (5) based on right-censored time-series Z_t. To overcome the right-censored observations, KMW is used, as suggested and discussed by [18] and [20], respectively. In the case of SETAR model estimation, the weight matrix $\mathbf{W}$ is added to equation Eq. (5). Here, the Kaplan-Meier weights are given by:

$$W_{(t)} = \frac{\delta_{(t)}}{n - t + 1} \prod_{j=1}^{t-1} \left(\frac{n - j}{n - j + 1}\right)^{\delta_{(t)}}, \tag{8}$$

where $\mathbf{W} = diag(W_{(1)}, \ldots, W_{(n)})$ is $n \times n$ diagonal matrix computed based on $\{Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}\}$. $\delta_{(t)}$ denotes the values of δ_t in Eq. (1) related to ordered values $Z_{(t)}$'s. Then, using Eq. (8), equation Eq. (5) is rewritten as follows:

$$\hat{\boldsymbol{\Phi}}_W(m) = \sum_{t=1}^n \left[\mathbf{X}_t^T(m)\mathbf{W}\mathbf{X}_t(m)\right]^{-1} \sum_{t=1}^n \left[\mathbf{X}_t^T(m)\mathbf{W}Z_t\right], \tag{9}$$

where $\hat{\boldsymbol{\Phi}}_W(m)$ denotes the estimated coefficients based on KMW. Accordingly, a fitted model is obtained with Eq. (9) and Eq. (5):

$$\hat{Z}_t^W = \mathbf{X}_t^T(m)\hat{\boldsymbol{\Phi}}_W(m), 1 \leq t \leq n, \tag{10}$$

where $\hat{Z}_t^W$ denotes the fitted values obtained by KMW. Hence, the right-censored SETAR model is estimated by the modified estimator $\hat{\boldsymbol{\Phi}}_W(m)$ based on KMW.

2.2 Synthetic Data Transformation

Another solution for a right-censored time-series is synthetic data transformation (SDT), which is an unbiased way to transform right-censored Z_t to synthetic variable $Z_{t\hat{G}}$ to make equivalent their expected values as $E(Z_t) \cong E(Z_{t\hat{G}})$ (see [21]). Similar transformation techniques have been studied by [22, 23], and [24]. The SDT procedure proposed by [15] can be shown as:

$$Z_{tG} = \frac{\delta_t Z_t}{1 - G(Z_t)} = \frac{\delta_t Z_t}{\bar{G}(Z_t)}, \tag{11}$$

where G is the distribution function of censoring variable C_t. However, in real-world applications, because G is generally unknown, instead of G, [15] suggested using its Kaplan-Meier estimator $\hat{G}$, which is given by arbitrary data point r:

$$\hat{G}(r) = 1 - \prod_{t=1}^{n} \left(\frac{n-t}{n-t+1} \right)^{I[Z_{(t)} \le r, \delta_{(t)}=0]}, (r \ge 0), \tag{12}$$

where $\{Z_{(t)}, \delta_{(t)}\}_{t=1}^{n}$ is ordered observation pairs as mentioned after equation (2.8). Note that, due to the main property of the SDT, it is clear that $E(Z_{t\hat{G}}) = E(Y_t) = \mathbf{X}_t^T(m)\,\hat{\mathbf{\Phi}}_{\mathrm{SDT}}(m)$ where $\hat{\mathbf{\Phi}}_{\mathrm{SDT}}(m)$ is the modified estimator of $\mathbf{\Phi}(m)$ based on the SDT approach. Accordingly, the estimator and fitted model can be obtained as follows:

$$\hat{\mathbf{\Phi}}_{SDT}(m) = \sum\nolimits_{t=1}^{n} \left[\mathbf{X}_t^T(m)\,\mathbf{X}_t(m)\right]^{-1} \sum\nolimits_{t=1}^{n} \left[\mathbf{X}_t^T(m)\,\mathbf{Z}_{\hat{G}}\right], \tag{13}$$

and

$$\hat{Z}_t^{SDT} = \mathbf{X}_t^T(m)\,\hat{\mathbf{\Phi}}_{SDT}(m), 1 \le t \le n, \tag{14}$$

where $\hat{Z}_t^{SDT}$ values are fitted values obtained based on the SDT technique.

2.3 kNN Imputation

kNN imputation for a right-censored time-series is introduced by [14] to overcome the right censorship. The main function of kNN imputation is to replace the right-censored observations with their kNN estimates. Notice that for any censored point, imputation is realized by an averaged value of k uncensored nearest neighbors. For a detailed discussion of kNN imputation, see [24] and [14]. Notice that the kNN imputation technique does not need any distributional assumption and does not touch uncensored observations, as in KMW and SDT. These properties are the main difference between kNN imputation from the other two solution methods. In this paper, to measure the distance between the neighbors, a commonly used Euclidean distance is used, which can be given by:

$$D(Z_1, Z_2) = \sqrt{\sum\nolimits_{t=1}^{n} |Z_{1t} - Z_{2t}|^2}, \tag{15}$$

An algorithm provided for *k*NN imputation by [14] is given below in table 1.

From the algorithm given in table 1, the estimator and fitted model based on kNN imputation are provided by:

$$\hat{\mathbf{\Phi}}_k(m) = \sum\nolimits_{t=1}^{n} \left[\mathbf{X}_t^T(m)\,\mathbf{X}_t(m)\right]^{-1} \sum\nolimits_{t=1}^{n} \left[\mathbf{X}_t^T(m)\,\mathbf{Z}^k\right], \tag{16}$$

and

$$\hat{Z}_t^k = \mathbf{X}_t^T(m)\,\hat{\mathbf{\Phi}}_k(m), 1 \le t \le n, \tag{17}$$

where $\hat{Z}_t^k$'s are the fitted values of the SETAR model obtained based on the kNN imputation technique.

Table 1. k-NN imputation for right censored data

Input. Right-censored time-series Z_t; Censoring indicator δ_t associated with Z_t
Number of nearest neighbors k; Determined lagged $Z_t^{'}$s : $Z_{t-1}, \ldots, Z_{t-p}$ to calculate distances.
Output: Imputed dataset Z_t^k c
1: Begin
2: for ($t = 1$to$n$) do
3: if ($\delta_t = 0$) do (if data point is censored)
4: for ($j = 1$to$n$) do
5: Find the distances between Z_{t-p_1} and Z_{t-p_2} for each censored data point with (2.15)
6: Sort the distances from small to large
7: for ($j = 1$to$k$) do
8: Take the first uncensored k values of Z_t associated to sorted distances
9: Calculate the t^{th} imputed value $\left(Z_t^k\right)$ with the average of nearest k records of Z_t
10: Replace the imputed value $\left(Z_t^k\right)$ with censored data point ($Z_t, \delta_t = 0$)
11: Return Z_t^k
12: End.

3 Evaluation Metrics

This section is prepared to present the evaluation metrics for the introduced three SETAR model estimates based on the given three censorship solution techniques. The fits of these models are notated as $\hat{Z}_t^W$, $\hat{Z}_t^{SDT}$ and $\hat{Z}_t^k$. For our purposes, three commonly used metrics in time-series analysis are preferred. These are root means squared error ($RMSE$), mean squared error (MSE) and mean absolute percentage error ($MAPE$). Calculations of these measurements are given based on joint notation $\left\{\hat{Z}_t\right\}_{t=1}^n$ of the mentioned three fitted values:

$$MSE\left(\hat{Z}_t\right) = n^{-1} \sum\nolimits_{t=1}^{n} \left(Z_t - \hat{Z}_t\right)^2, \tag{18}$$

$$RMSE\left(\hat{Z}_t\right) = \sqrt{n^{-1} \sum\nolimits_{t=1}^{n} \left(Z_t - \hat{Z}_t\right)^2}, \tag{19}$$

$$MAPE\left(\hat{Z}_t\right) = n^{-1} \sum\nolimits_{t=1}^{n} \left| \frac{\left(Z_t - \hat{Z}_t\right)}{Z_t} .100 \right|, \tag{20}$$

By using these criteria, a comparison of the three model estimates is realized and the effect of censorship on the estimated SETAR models is measured.

4 Numerical Studies

4.1 Simulation Study

The purpose of this section is to investigate the performances of the censorship solution methods on SETAR model estimation using simulation evidence. Accordingly, right-censored, stationary times-series and SETAR models as in Eq. (2) are generated as follows based on the work of Chan and Tsay [25]:

$$Z_t = \begin{cases} 0.3 - 0.7Z_{t-1} - 0.4Z_{t-2} + \varepsilon_t, if Z_{t-d} \leq m = 0.8 \\ 1.2 + 0.5Z_{t-1} + 0.1Z_{t-2} + \varepsilon_t, if Z_{t-d} > m = 0.8 \end{cases}, \tag{21}$$

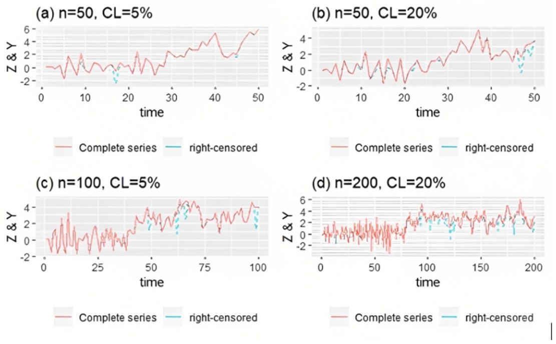


Figure 1. Scatter plot of Z_t to observe the low and high regimes and censoring levels for different configurations

where Z_t is obtained by censoring variable C_t which is generated based on μ_{Y_t} and $\sigma^2_{Y_t}$. Hence, pair of time-series (Z_t, δ_t) is obtained as mentioned in Eq. (1). To clearly express the model parameters and related matrices, model Eq. (21) should be written in harmony with equation Eq. (3) as:

$$Z_t = \phi_1^T \mathbf{X}_{1t} I\left(Z_{t-d} \leq 0.8\right) + \phi_2^T \mathbf{X}_{2t} I\left(Z_{t-d} > 0.8\right) + \varepsilon_t, \tag{22}$$

where $\mathbf{X}_{1t} = (1, Z_{t-1}, Z_{t-2})$ and $\mathbf{X}_{2t} = (1, Z_{t-1}, Z_{t-2})$, $\phi_1 = (0.3, 0.7, 0.4)^T$ and $\phi_2 = (1.5, -0.5, -0.2)^T$. Note that error terms are generated i.i.d. as ε_t
$sim\left(\mu_\varepsilon = 0, \sigma^2_\varepsilon = 0.5\right)$. Accordingly, our aim here is to estimate $\mathbf{\Phi}(m) = (\phi_1, \phi_2)$ under different scenarios by using pairs (Z_t, δ_t) and related censoring solution techniques. Therefore, in our simulation study, the performances of $\hat{Z}_t^W$, $\hat{Z}_t^{SDT}$, and $\hat{Z}_t^k$ are measured based on how they converge to the real vector of parameters $\mathbf{\Phi}(m)$. To achieve this, the performance metrics given in equations (3.1-3.3) are used. In this simulation study, we generate 500 random samples of size $n = 50, 100, \text{and} 200$ and censoring levels $CL = 5\%\text{and}, 20\%$.

The obtained results are presented in the following tables and figures. Figure 1 presents the generated dataset for four configurations. As can be seen in panels (b) and (d), the right-censored observations distort the data structure. Also, the lower and upper regimes can be seen clearly. It can be said that censorship makes it challenging to detect the threshold value (m), which indicates the separation of the regimes.

In table 2, the nonlinearity test results for right-censored time-series Z_t are provided. Here, Tsay's F-test for nonlinearity is used under type-I error $\alpha = 0.05$. The calculation of the F-statistic and further details are provided by Tsay (1986) [26]. The null hypothesis of the test is as follows:

H_0 : Z_t follows some AR process

H_1 : Z_t has a nonlinear structure

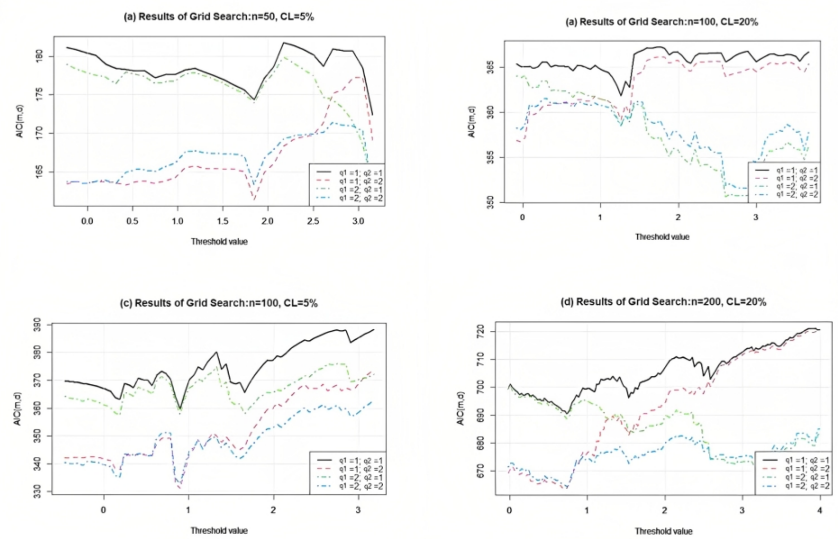
According to table 2, due to the p-value $(p = 0.020)$ being smaller than $\alpha = 0.05$, H_0 is rejected and it is accepted that Z_t has a nonlinear structure. After deciding the nonlinearity of the series, the optimal pair of $\left(\hat{m}, \hat{d}\right)$ should be selected. To achieve this, a grid search is carried out based on AIC. The results are provided in table 3 and Figure 2.

Figure 2 presents a grid search for a single generated SETAR series from a determined configuration. The autoregressive degrees of low and high regimes $(q_1,\ q_2)$ are generally chosen as $q_1 = 1, q_2 = 2$ or $q_1 = 2, q_2 = 1$. The optimal threshold value takes a value between $m = 0.4$ and $m = 2.70$. Notice that if panels (b) and (c) are carefully inspected, the

Table 2. Nonlinearity test for Z_t

n	CL	F-statistic	p-value
50	5%	2.575	0.020*
	20%	4.349	p<0.001*
100	5%	2.120	0.008*
	20%	2.019	0.0416*
200	5%	8.259	p<0.001*
	20%	2.252	p<0.001*

Note: H_0 is rejected with 95% confidence.


(a) Results of Grid Search:n=50, CL=5%

(a) Results of Grid Search:n=100, CL=20%

(c) Results of Grid Search:n=100, CL=5%

(d) Results of Grid Search:n=200, CL=20%

Figure 2. Grid search with *AIC* for *d* and *m*

censoring level affects the choice of the optimal threshold value (m) significantly. Therefore, results prove that a high level of censorship makes it difficult to decide the regimes. Therefore, the range of m is wide. We expect to solve this problem using the three censorship solution techniques, *KMW*, *SDT*, and *kNN*.

From table 3, the optimally chosen orders of low and high regimes ($\hat{q}_1, \hat{q}_2$) of the SETAR models to be estimated, the threshold value ($\hat{m}$), and the delay of the threshold variable ($\hat{d}$) are presented for all simulation configurations. Notice that, based on *AIC*, the behavior of $\hat{m}$ and ($\hat{q}_1, \hat{q}_2$) can be observed according to sample size and censoring level. Therefore, it is obvious that under heavy censorship, m is selected far from the real value of $m = 0.8$. On the other hand, when $n = 200$, *AIC* selects both ($\hat{q}_1, \hat{q}_2$) and $\widehat{(m)}$ more accurately than other configurations. After obtaining the optimal parameters of SETAR models, the estimation can be realized.

Tables 4-5 involve the estimated autoregressive coefficients for both low and high regimes. In both tables, it can be observed that *SDT* and *kNN* obtained similar estimates, whereas *KMW* estimates relatively smaller coefficients. The difference between *KMW* and the others is that the weight matrix in KMW makes estimates more stable than kNN and SDT. These inferences are ensured by the scores of tables 4-5. The changes in estimated coefficients from lower to higher censoring levels are smaller for *KMW* compared to the *SDT*

Table 3. Selection of *d* and *m* based on *AIC* criterion

n	CL	$\hat{q}_1$	$\hat{q}_2$	$\hat{m}$	$\hat{d}$	$AIC\left(\hat{m},\hat{d}\right)$
50	5%	2	1	2.12	0	151.38
	20%	2	1	1.39	0	149.09
100	5%	2	1	1.47	0	290.01
	20%	2	1	2.68	0	340.42
200	5%	2	2	0.43	0	658.89
	20%	2	2	1.83	0	707.20

Table 4. Estimated low-regime coefficients and associated statistics of right-censored $SETAR(2,1)$ and $SETAR(2,2)$ models

		Low regime		
		KMW	*SDT*	*KNN*
n	*CL*	$\phi_{01};\phi_{11};\phi_{21}$	$\phi_{01};\phi_{11};\phi_{21}$	$\phi_{01};\phi_{11};\phi_{21}$
50	5%	-0.07;-0.05;0.01	0.46;-0.26;-0.01	0.41;-0.40;-0.13
	20%	-0.07; -0.03;0.01	0.56;-0.11;0.22	0.46;-0.20;-0.09
100	5%	-0.07;-0.06;-0.01	0.46;-0.38;-0.07	0.42;-0.49;-0.20
	20%	-0.10;-0.04;0.01	0.48;-0.28;0.08	0.34;-0.47;-0.15
200	5%	-0.10;-0.07;-0.01	0.32;-0.52;-0.20	0.34;-0.59;-0.30
	20%	-0.07; -0.04;-0.01	0.49;-0.20;0.07	0.38;-0.41;-0.14

Table 5. Estimated high-regime coefficients and associated statistics of censored $SETAR(2,1)$ and $SETAR(2,2)$ models

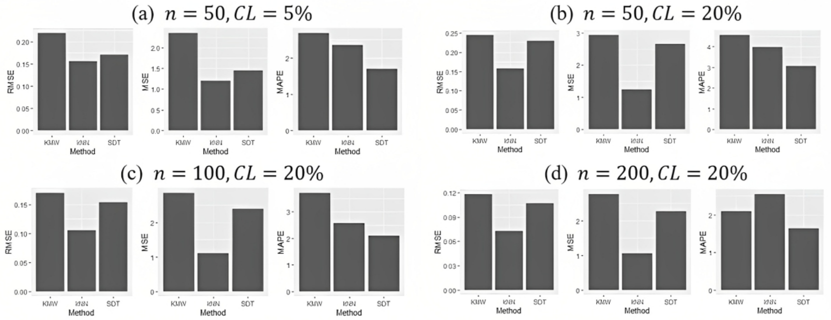
High regime		
KMW	*SDT*	*kNN*
$\phi_{02};\phi_{12};\phi_{22}$	$\phi_{02};\phi_{12};\phi_{22}$	$\phi_{02};\phi_{12};\phi_{22}$
0.24;0.36; –	1.83;0.23; –	1.15;0.31; –
0.21;0.34; –	1.87; 0.31; –	1.58; 0.19; –
0.21;0.35; –	1.04;0.36; –	0.58;0.36; –
0.20;0.29; –	1.10; 0.33; –	0.73; 0.31; –
0.17;0.29;0.30	0.36;0.42;0.36	0.10;0.43;0.45
0.16;0.27;0.27	1.18;0.33;0.16	0.39;0.36;0.41

and *kNN* methods. Although *KMW* is more robust against censorship, the performance scores given in table 6 show that its performance is worse than *kNN* and *SDT*. Hence, one may select the *KMW*-based *SETAR* model estimation for more stable estimates but for low performance. On the other hand, *kNN* and *SDT* give qualified results but they are affected by censorship more than *KMW*.

Table 6 presents the scores of the evaluation metrics *MAPE*, *RMSE*, and *MSE*, respectively, for all simulation configurations. The best scores are indicated in bold. At first glance, the two expected results can be observed. These are the incremental change in performance when *n* increases, and the negative effect of high censoring levels on the methods. The results show that kNN dominates the other two methods in general, but there are some points to be emphasized. *SDT* shows good performance in terms of *MAPE* criterion when $n = 50$ and $n = 200$. Also, when $CL = 5\%$, even though *kNN* gives the best results, *SDT* is right behind. However, for $CL = 20\%$, *SDT* fails most of the time, and it is affected by the censoring level more than *kNN* and *KMW*. Note that, as mentioned above, the values of *KMW* increase less as the censoring level increases in terms of *MAPE*. Additionally, because *kNN* has no restrictions and is a fully nonparametric technique, it seems kNN has a good strength against censorship that can be understood from the *MSE* and *RMSE* values. Figure 3 introduces barplots of the scores provided in table 6. The mentioned inferences can be easily observed from the figure.

Table 6. *MSE*, *RMSE* and *MAPE* scores for estimated models

		MAPE			*RMSE*			*MSE*		
n	*CL*	*KMW*	*SDT*	*kNN*	*KMW*	*SDT*	*kNN*	*KMW*	*SDT*	*kNN*
50	5%	2.686	1.680	2.361	0.219	0.171	0.156	2.355	1.449	1.205
	20%	4.568	3.228	3.978	0.244	0.229	0.158	2.941	2.655	1.232
100	5%	1.784	1.713	1.210	0.151	0.113	0.104	2.262	1.297	1.092
	20%	3.708	2.840	2.579	0.169	0.153	0.105	2.856	2.386	1.116
200	5%	1.035	0.838	1.109	0.103	0.078	0.073	2.134	1.232	1.080
	20%	2.099	1.657	2.549	0.118	0.106	0.072	2.771	2.279	1.163


(a) $n = 50, CL = 5\%$

(b) $n = 50, CL = 20\%$

(c) $n = 100, CL = 20\%$

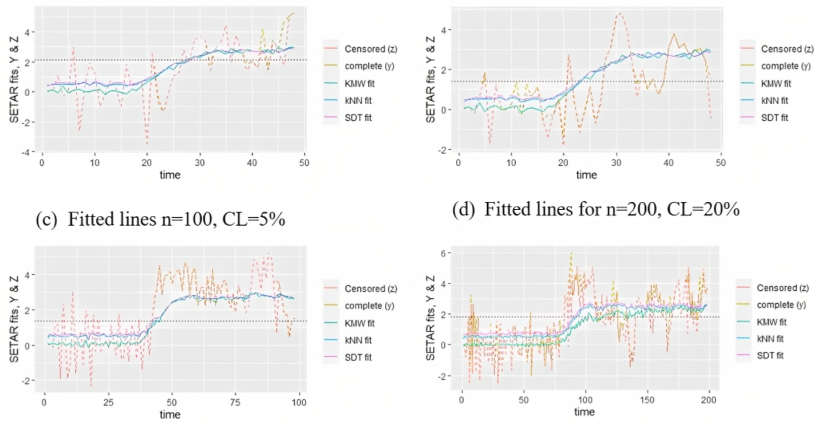
(d) $n = 200, CL = 20\%$

Figure 3. Barplots of performance scores given in table 6

In figure 4, fitted $SETAR(2,1)$ and $SETAR(2,2)$ models are given with generated censored time-series (Z_t) and the completely observed series (Y_t). In these panels, horizontal dashed lines denote the optimally chosen threshold value (m) for each configuration. It is obvious that the regimes are determined correctly in the SETAR model, and the three fits represent the data well. In detail, the poor performance of the *KMW* fit can be distinguished from *kNN* and *SDT*, especially in panels (b) and (d), due to heavy censorship. The closeness of *kNN* and *SDT* is also observed. Note that to save space, only certain configurations are illustrated in the figure. However, the performances of the methods can be monitored from tables 4-5 and table 6. The simulation results show that all three censorship solution methods work well with SETAR, and they can be easily integrated into each other. Regarding the comparison, *kNN* produces the best results *SDT* and *KMW* shows satisfying results under specific conditions. *KMW* also has a different advantage, which is it diminishes the effect of the censoring level on the estimation performance.

4.2 Case Study: Covid-19 Data from China

This section is prepared to provide inferences for behaviors of the introduced estimators for the SETAR model for the real dataset which involves right-censored observations. In this



(c) Fitted lines n=100, CL=5%

(d) Fitted lines for n=200, CL=20%

Figure 4. Fitted models versus completely observed time-series (Y_t). The black horizontal line denotes the chosen optimal "m" which separates the regimes (see table 3)

context, their performances are compared to the simulation results. Covid-19 data from China was selected for this purpose. The dataset consists of 104 data points and two variables provided by Afshin and Jorge (2020). In this study, the modeling procedure is realized for the variables of the number of recovered patients from Covid-19 (*recover*) as a right-censored time-series and the censoring level is 6.73%. Accordingly, SETAR model statistics based on the modeling procedures are provided in the following tables and figures.

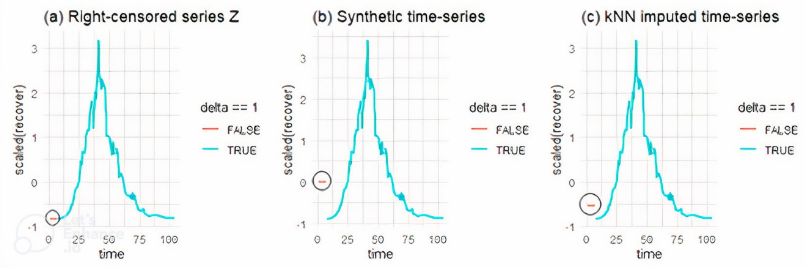


Figure 5. Scaled recover series against time with three panels: (a) Right-censored Z_t, (b) $Z_{t\hat{G}}$ obtained by SDT, (c) Z_t^k obtained by *kNN* imputation

To test the stationarity of the Covid-19 series, the ADF test is made, and the results are shown in table 7. Accordingly, it is seen that the data is stationary when its 4^{th} lag. Hence, the remaining analysis is realized accordingly to ensure the stationarity assumption.

Table 7. Augmented Dickey-Fuller test results for the Covid-19 series

Variable	*t-statistics*	Lag-order	*p-value*
Z_t	-1.701	4	0.701
$Z_{t\hat{G}}$	-1.519	4	0.776
Z_t^k	-1.600	4	0.742

Table 8 involves Tsay's nonlinearity test results, which are needed to prove that the Covid-19 time-series are adequate for the SETAR type model and also that the optimal SETAR model parameters are determined by AIC. As can be seen, nonlinearity is validated for raw (incomplete) series Z_t, transformed series by *SDT*, $Z_{t\hat{G}}$ and imputed by *kNN* Z_t^k. Tsay's *F*-statistic and the *p*-values show that the null hypothesis, which claims that the mentioned series follows some AR process, is strongly rejected. Thus, associated SETAR parameters (q_1, q_2), *m*, and *d* are determined in table 8. From that, the suitable model for the Covid-19 dataset is selected as $SETAR(1, 1)$.

Table 8. Tsay's nonlinearity test for Z_t and determination of *m* and *d* with *AIC*

	n	*CL*	*Tsay's F-statistic*	*p-value*	*m*	d	q1	q2	*AIC*
Z_t	104	6.73%	6.565	p<0.001	-0.74	0	1	1	-183.65
$Z_{t\hat{G}}$			4.358	p<0.001	-0.72	0	1	1	-122.129
Z_t^k			2.376	0.0024	-0.67	0	1	1	-144.216

The estimated coefficients for $SETAR(1, 1)$ are given in table 8 for low and high regimes for all three methods. The best scores are indicated in bold. As in the simulation study, the coefficients of *KMW* are different from the other two due to Kaplan-Meier weights. This difference is also seen in Table 9, which includes the values of performance criteria. Notice that *SDT* and *kNN* provide closer results and $SETAR(1, 1)$ gives the smallest *MSE* and *RMSE* values, whereas *SDT* gives smaller *MAPE* values. These results show that the introduced estimators produce coherent behaviors in both the real data example and the simulation study. Figure 6 includes the barplots of the values given in table 10 to illustrate the difference between the performances of the methods more easily.

Table 9. Results obtained from $SETAR(1, 1)$ models based on three censorship solution methods

		KMW	*SDT*	*kNN*
Low	$\hat{\phi}_{10}$	-0.196	-0.045	-0.047
	$\hat{\phi}_{11}$	0.155	0.967	1.025
High	$\hat{\phi}_{20}$	0.039	0.002	0.008
	$\hat{\phi}_{21}$	0.172	0.873	0.873

Table 10. Performance scores for the estimates

	KMW	*SDT*	*kNN*
MAPE	1.154	0.729	0.996
RMSE	0.039	0.022	0.019
MSE	0.159	0.053	0.040

In figure 7, fitted $SETAR(1.1)$ models obtained by the three methods are given. As in the simulation study, threshold values (*m*) for the three methods are illustrated by horizontal lines. One can here see the low and high regimes easily. In this example, the low regime represents the beginning of the Covid-19 pandemic with right-censored observations, and the high regime indicates the time when the disease peaked and then descended. In the figure, imputed values of *kNN* can be seen in the censored region of the series. Also, the fits of *kNN* and *SDT* are closer to each other, as previously mentioned, and *KMW* shows its difference by representing the real series worse than the other two methods. However, it should be noted

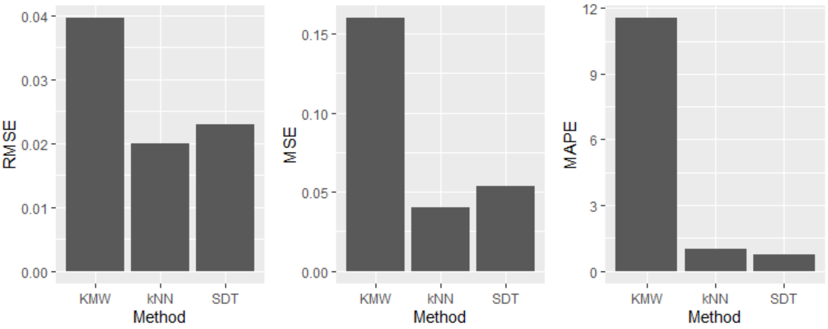


Figure 6. Barplots of performance scores

that although *KMW* performs worse than the other two methods in this example, it still can handle the right-censored data well, which is explained in the simulation study.

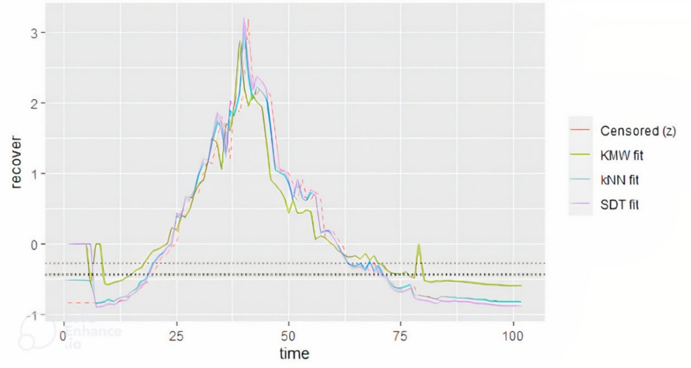


Figure 7. Right-censored $SETAR(1,1)$ model fits estimated based on the three censorship solution techniques

5 Conclusions

This paper aims to show SETAR-type model estimation when the time-series are right-censored. In particular, the censorship solutions KMW, SDT, and *kNN* imputation methods are combined with the SETAR estimation procedure, and their performances are inspected practically. To achieve this purpose, a simulation study and a real-data case study are provided. Based on the results of both studies, the following conclusions are obtained:

- Regarding SETAR models and right-censored nonlinear time-series, it is proven that censorship highly affects the accuracy of the estimations by manipulating data. In Table 3, one can see how the optimal parameter selection of SETAR is affected by the censoring level.
- Especially for the threshold value (m) selection, if the high regime involves more censored observations, then the detection of the regimes may be challenging. It is found that kNN and SDT-based SETAR model estimations solve the problem better than the method.

- Although *KMW* gives worse SETAR fits than *kNN* and *SDT*, it has a stability advantage on estimated models that are demonstrated in the simulation study. In addition, under low censoring levels, *kNN* and *SDT* give very close results, which are also ensured by the Covid-19 data example. However, under heavily censored data, *kNN* dominates the other methods and gives the best performance scores.

As a result, this paper suggests that under right-censored data, the SETAR model is successfully modeled based on the censorship solution techniques and it is seen that *kNN* imputation works well for all simulation configurations and the Covid-19 example.

In the future, as a continuation of this study, it is planned to study non-parametric and semi-parametric estimation methods to make predictions with less risk in the estimation of right-censored SETAR models in the future. In addition, it is planned to use different criteria for the selection of threshold value (m) and lag parameter (d) shown in Eq. (6) and analyze the results, which are selected to determine the upper and lower regimes.

The research of Professor S. Ejaz Ahmed was supported by the Natural Sciences and the Engineering Research Council (NSERC) of Canada.

References

[1] J.J. JABER, Ph.D. thesis, Universiti Kebangsaan Malaysia, Malaysia (2017)
[2] C.W.J. Granger, A. Ap (1978)
[3] H. Tong, Pattern recognition and signal processing pp. 575–586 (1978)
[4] J. Fan, Q. Yao, *Nonlinear time series: nonparametric and parametric methods*, Vol. 20 (Springer, 2003)
[5] T. Terasvirta, H.M. Anderson, Journal of applied econometrics **7**, S119 (1992)
[6] J.D. Hamilton, Econometrica: Journal of the econometric society pp. 357–384 (1989)
[7] J.D. Petruccelli, Journal of Forecasting **9**, 25 (1990)
[8] J.G. De Gooijer, Journal of Time Series Analysis **22**, 267 (2001)
[9] P. Milheiro-Oliveira, Statistics & Probability Letters **184**, 109385 (2022)
[10] N. Naik, B.R. Mohan, Mathematics **9**, 1595 (2021)
[11] D. AYDIN, S. MERMİ, Eskişehir Technical University Journal of Science and Technology A-Applied Sciences and Engineering **23**, 48 (2022)
[12] J.W. Park, M.G. Genton, S.K. Ghosh, Canadian Journal of Statistics **35**, 151 (2007)
[13] S. Khardani, M. Lemdani, E. Ould Saïd, Metrika **75**, 229 (2012)
[14] D. Aydın, E. Yılmaz, Empirical Economics **61**, 2143 (2021)
[15] H. Koul, V. Susarla, J. Van Ryzin, The Annals of statistics pp. 1276–1288 (1981)
[16] M.Y. Khan, Ph.D. thesis, lmu (2015)
[17] K.S. Chan, H. Tong, Journal of time series analysis **7**, 179 (1986)
[18] R.G. Miller, Biometrika **63**, 449 (1976)
[19] E.H. Firat, Mathematics and Statistics **5**, 33 (2017)
[20] J. Orbe, J. Virto, Biometrical Journal **60**, 947 (2018)
[21] M. Talamakrouni, A.E. Ghouch, I. Van Keilegom, Scandinavian Journal of Statistics **42**, 214 (2015)
[22] T. Choi, A.K. Kim, S. Choi, Computational Statistics & Data Analysis **164**, 107306 (2021)
[23] Z. Sun, Y. Liu, K. Chen, G. Li, Annals of the Institute of Statistical Mathematics **74**, 69 (2022)

[24] S.E. Ahmed, D. Aydin, E. Yılmaz, *Nonparametric regression estimates based on imputation techniques for right-censored data*, in *International Conference on Management Science and Engineering Management* (Springer, 2019), pp. 109–120

[25] K.S. Chan, R.S. Tsay, Biometrika **85**, 413 (1998)

[26] R.S. Tsay, Biometrika **73**, 461 (1986)