Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 73 a 94

Testing Linearity against a Univariate TAR Specification in Time Series with Missing Data

Sobre una prueba de linealidad en presencia de datos faltantes contra la alternativa de no linealidad especificada por un modelo TAR

Fabio H. Nieto[1,a], Milena Hoyos[2,b]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

[2]Facultad de Economía, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

Nowadays, procedures for testing the null hypothesis of linearity of a (univariate or multivariate) stochastic process are well known, almost all of them based on the assumption that their paths (i.e. observed time series) are complete. This paper describes an approach for testing this null hypothesis in the presence of missing data, using an extension of one of the test statistics used in the literature. The alternative hypothesis is that the univariate stochastic process of interest follows a threshold autoregressive (TAR) model. It is found that if the missing-data percentage is low, the null distribution of the proposed test statistic is maintained; while if it is high, it is not. A threshold value for the missing-data percentage is detected, which can be utilized in practice.

***Key words*:** Linearity test, Missing data, Nonlinear time series, Threshold autoregressive model.

Resumen

Las pruebas estadísticas que se conocen actualmente para examinar la hipótesis nula de linealidad de un proceso estocástico (univariado o multivariado) están basadas, casi todas, en el supuesto de que las series temporales observadas son completas. En este trabajo, se presenta un nuevo procedimiento para examinar esta hipótesis nula, en presencia de datos faltantes, el cual es una extensión de un método muy citado en la literatura. La hipótesis alternativa especifica que el proceso estocástico de interés obedece a un modelo autoregresivo de umbrales (TAR). Se encuentra que si el porcentaje de observaciones faltantes es bajo, la distribución nula de la estadística de prueba se mantiene; en otro caso no. El estudio arroja un valor umbral para este porcentaje, el cual puede ser usado en la práctica.

***Palabras clave*:** datos faltantes, modelos autoregresivos de umbrales, prueba de linealidad, series de tiempo no linales.

[a]Profesor titular. E-mail: fhnietos@unal.edu.co
[b]Profesora auxiliar. E-mail: nmhoyosg@unal.edu.co

1. Introduction

Nieto (2005) developed a procedure for modeling a univariate threshold-autoregressive processes (TAR) in the presence of missing data. The approach was based on the assumption that one knows a priori that the dynamic relationship between the two stochastic processes is nonlinear. This model can be seen as a particular case of Tsay's (1998) multivariate threshold model, where a test for the null hypothesis of linearity was considered. An important contribution of Nieto's (2005) paper is the development of a smoother for estimating the missing data in the two time series involved.

There are several methods for testing the null hypothesis of linearity in a univariate or multivariate stochastic process. However, almost all of these methods have been developed on the basis that the time series are complete or equally spaced. Sometimes, this is not the case and one is faced with the problem of performing those tests in the presence of partial or missing observations. Tong & Yeung (1991*a*, 1991*b*), Brockwell (1994) and Tsai & Chan (2000) have worked on this topic, but only Tong & Yeung (1991*a*) have considered *discrete* time series, while the other authors have addressed the problem under the continuous-time context. Specifically, Tong & Yeung (1991*b*) have studied the case of partially observed time series, where the main characteristic is that, by nature, the observations are not equally spaced, as happens with financial variables that are not observed in the weekends or holidays. The underlying model in that paper was a univariate self-exciting threshold (SETAR) model. In this paper, we will consider the case where the missing data appear because, for different reasons, the values of a variable were not recorded although they actually occurred. Of course, this situation causes unequally-spaced time series.

Unfortunately, Tong & Yeung's (1991*a*) procedure has a drawback, in the sense that the state space model they used for basing their *adapted* tests is not appropriate, as we will show in Section 3 below. Then, their arranged-autoregression ideas cannot be extended to the case of TAR models via state space forms. Instead, in this paper, we extend Tsay's (1998) test statistic and look for its null distribution under three scenarios: (1) complete data, (2) low missing-data percentage, and (3) medium and high missing-data percentage. Our goal is to find a threshold value for the missing-data rate up to which our extended test statistic maintains its null distribution.

The idea behind our work is the following: under the null hypothesis of linearity, one can estimate the missing data in the time series, using a linear-model based procedure as that of Gómez & Maravall (1994). Now, if the so-called input time series is nonlinear, we use a simplification of Nieto's (2005) smoother; if not, we also use the same linear-model based procedure. Then, one completes the time series with the estimated values and computes the test statistic. At the bottom line, we will find the distribution of the proposed test statistic under the null hypothesis of linearity taking into account the uncertainty of the missing data estimates. This work is done by means of Monte Carlo simulations.

The paper is organized as follows. In Section 2, we present the basic TAR model and its simplification under the null hypothesis. Section 3 describes Tsay's (1998) nonlinearity test, the extended test statistic and its null distribution for complete time series. In Section 4, we include a theoretical example that shows the drawback of Tong & Yeung's (1991*a*) procedure and analyze the effect that the missing-data-estimates uncertainty has on the null distribution of the proposed test statistic. Section 5 presents a real-data application and Section 6 concludes.

2. Specification of the TAR Model

Let $\{X_t\}$ and $\{Z_t\}$ be stochastic processes related by the equation (TAR model)

$$X_t \quad = \; a_0^{(j)} + \sum_{i=1}^{k_j} a_i^{(j)} X_{t-i} + h^{(j)} \varepsilon_t \;,\; r_{j-1} < Z_t \leq r_j \tag{1}$$

where $j = 1, \ldots, l-1$ indicate the presence of l regimes in the process $\{X_t\}$, which are determined by the threshold values r_0,r_1,...,r_{l-1}, and r_l of process $\{Z_t\}$, with $r_0 = -\infty$ and $r_l = \infty$. Here, $a_i^{(j)}$ and $h^{(j)}$; $j = 1, \ldots, l$; $i = 0, 1, \ldots, k_j$; are real numbers and $\{\varepsilon_t\}$ is a Gaussian zero-mean white noise process with variance 1. Additionally, the nonnegative integer numbers $k_1, \ldots, k_l$ denote, respectively, the autoregressive orders of $\{X_t\}$ in each regime. We shall use the symbol TAR$(l; k_1, \ldots, k_l)$ to denote this model and call l,r_1,...,r_{l-1}, k_1,...,k_{l-1} and k_l the model structural parameters.

These models were introduced by Tong (1978) and Tong & Lim (1980), specifically, in the case where the threshold variable is the lagged variable X_{t-d}, where d is some positive integer. In this case, the model is known as the self-exciting TAR (SETAR) model and, at present, there is a lot of literature about the topic of analyzing these models, under the frequent assumption that we know the number l of regimes and the autoregressive orders $k_1, \ldots, k_l$.

We also assume that $\{Z_t\}$ is exogenous in the sense that there is no feedback of $\{X_t\}$ towards it and that $\{Z_t\}$ is a homogeneous pth order Markov chain with initial distribution $F_0(z, \boldsymbol{\theta}_z)$ and kernel distribution $F_p(z_t \mid z_{t-1}, \ldots, z_{t-p}, \boldsymbol{\theta}_z)$, where $\boldsymbol{\theta}_z$ is a parameter vector in an appropriate numerical space. Furthermore, we assume that these distributions have densities in the Lebesgue-measure sense. Let $f_0(z, \boldsymbol{\theta}_z)$ and $f_p(z_t \mid z_{t-1}, \ldots, z_{t-p}, \boldsymbol{\theta}_z)$ be, respectively, the initial and kernel density functions of the distributions above. In what follows, we assume that the p-dimensional Markov chain $\{Z_t\}$ has an invariant or stationary distribution $f_p(z, \boldsymbol{\theta}_z)$.

Nieto's (2005) algorithms are based strongly in the regime-switching state-space form of the TAR model, given by the following: let $k = \max\{k_1, \ldots, k_l\}$, $\alpha_t = (X_t, X_{t-1}, \ldots, X_{t-k+1})'$, $\omega_t = (\varepsilon_t, 0, \ldots, 0)'$, and $\{J_t\}$ be a sequence of indicator variables such that $J_t = j$ if and only if $Z_t \in B_j$ for some j, $j = 1, \ldots, l$. Now, let $\boldsymbol{H} = (1, 0, \ldots, 0)'$ and for $j = 1, \ldots, l$, let $\boldsymbol{C}_j = (a_0^{(j)}, 0, \ldots, 0)'$,

$$\boldsymbol{A}_j = \left(\begin{array}{cccc|c} a_1^{(j)} & a_2^{(j)} & \cdots & a_{k-1}^{(j)} & a_k^{(j)} \\ & \boldsymbol{I}_{k-1} & & & \boldsymbol{0} \end{array} \right)$$

where $a_i^{(j)} = 0$ for $i > k_j$ and $\boldsymbol{I}_{k-1}$ denotes the identity matrix of order $k-1$, and

$$\boldsymbol{R}_j = \begin{pmatrix} h^{(j)} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Then, the state space form for the TAR$(l; k_1, \ldots, k_l)$ model becomes

$$X_t = \boldsymbol{H}\boldsymbol{\alpha}_t \tag{2}$$

as the observation equation, and

$$\boldsymbol{\alpha}_t = \boldsymbol{C}_{J_t} + \boldsymbol{A}_{J_t}\boldsymbol{\alpha}_{t-1} + \boldsymbol{R}_{J_t}\omega_t \tag{3}$$

as the system or state equation, where it is understood that $\boldsymbol{C}_{J_t} = \boldsymbol{C}_j$ if at time t, $J_t = j$. The same remark holds for the values of the matrices $\boldsymbol{A}_{J_t}$ and $\boldsymbol{R}_{J_t}$. This kind of *nonlinear* state space models, where appart from the observation and system equations there is an underlying *indicator process* that defines the structure of these equations and the probability distributions of the error terms, have been studied in the literature by Shumway & Stoffer (1991), Carter & Kohn (1994, 1996) and Kim & Nelson (1999), among others.

The situation of interest we shall consider is that there are missing observations in the two time series, in such a way that the observed data are located at the unequally-spaced time points $t_1, \ldots, t_N$, with $1 \leq t_1 \leq \cdots \leq t_N \leq T$, for $\{X_t\}$, and at $s_1, \ldots, s_M$, $1 \leq s_1 \leq \cdots \leq s_M \leq T$, for $\{Z_t\}$, where T is the sample size. Nieto (2005) solved the problem of estimating both the model parameters, including the structural parameters, and the missing observations on the basis that $l > 1$. In particular, for estimating missing values in the observed time series of process $\{Z_t\}$, he found that the posterior densities for the variables Z are given by

$$p(\mathbf{z}_T \mid \boldsymbol{\alpha}, \mathbf{x}) \propto \prod_{j=T-p+1}^{T} p(\alpha_j \mid \mathbf{z}_T, \alpha_{j-1}) f_p(\mathbf{z}_T) \tag{4}$$

and

$$p(z_t \mid \mathbf{z}_{t+p}, \boldsymbol{\alpha}_t, \mathbf{x}_t) \propto p(\alpha_t \mid \mathbf{z}_{t+p-1}, \alpha_{t-1}) f_p(z_{t+p} \mid \mathbf{z}_{t+p-1}) f_p(\mathbf{z}_{t+p-1}) \tag{5}$$

for $t = T-p, \ldots, 1$, where, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$, $\boldsymbol{\alpha}_t = (\alpha_1, \ldots, \alpha_t)$, $\mathbf{x}_t = (x_1, \ldots, x_t)$, $\mathbf{x} = (x_1, \ldots, x_T)$, and, in general, $\mathbf{z}_t = (z_{t-p+1}, \ldots, z_t)$. Then, for estimating the missing data at the time points $s_1, \ldots, s_M$, one obtains draws from their corresponding posterior densities given by expressions (4) and (5) via MCMC procedures.

Now, for estimating the missing data in the time series of process $\{X_t\}$, one has to take into account that (see Nieto's (2005) paper)

$$p(\boldsymbol{\alpha} \mid \mathbf{z}, \mathbf{x}) = p(\boldsymbol{\alpha}_T \mid \mathbf{z}, \mathbf{x}) \prod_{t=1}^{T-1} p(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t+1}, \mathbf{z}, \mathbf{x}_t) \tag{6}$$

where $\mathbf{z} = (z_1, \ldots, z_T)$. Since the first component of $\boldsymbol{\alpha}_t$ is X_t, one obtains draws from the posterior density $p(\boldsymbol{\alpha} \mid \mathbf{z}, \mathbf{x})$, then marginalizes it at the time points $t_1, \ldots, t_N$ and picks the first component up. Under the assumption of Gaussianity for the process $\{\varepsilon_t\}$, each factor in (6) is the density of a multivariate normal distribution (see Carter & Kohn's (1994) paper for details)

If $l = 1$, $\{X_t\}$ reduces to a linear AR(k_1) model and there is no influence of $\{Z_t\}$ onto $\{X_t\}$, in the sense of the dynamic causality explained by the TAR model (1). That is to say, for any value of the variable Z, the variable X has the same dynamic autoregressive answer, with parameters $a_i = a_i^{(1)}$ for all i; $i = 0, 1, \ldots, k_1$. Likewise, one has that the white-noise-process weight is $h = h^{(1)}$. Importantly, the process $\{Z_t\}$ can be either linear or nonlinear.

3. A Nonlinearity Test for Complete Multivariate Stochastic Processes

Let $\{\mathbf{X}_t = (X_{1t}, \ldots, X_{kt})'\}$, $\{\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{vt})'\}$ and $\{Z_t\}$ be stochastic processes, where the first two are multivariate and the last one is univariate. Tsay (1998) proposed the following multivariate threshold model for $\{\mathbf{X}_t\}$ with threshold process $\{Z_t\}$ and delay $d > 0$:

$$\mathbf{X}_t = \mathbf{a}_0^{(j)} + \sum_{i=1}^{p} \mathbf{a}_i^{(j)} \mathbf{X}_{t-i} + \sum_{i=1}^{q} \mathbf{b}_i^{(j)} \mathbf{Y}_{t-i} + \boldsymbol{\varepsilon}_t^{(j)} \tag{7}$$

if Z_{t-d} belongs to the real interval $B_j = (r_{j-1}, r_j]$ for some j; $j = 1, \ldots, l$; where $-\infty = r_0 < r_1 < \cdots < r_{l-1} < r_l = \infty$, $\mathbf{a}_0^{(j)}$ are constant vectors, $\mathbf{a}_i^{(j)}$ and $\mathbf{b}_i^{(j)}$ are constant matrices, and p and q are nonnegative integers. The innovations satisfy $\boldsymbol{\varepsilon}_t^{(j)} = \Sigma_j^{1/2} \mathbf{u}_t$, where $\Sigma_j^{1/2}$ is a symmetric positive definite matrix, $j = 1, \ldots, l$, and $\{\mathbf{u}_t\}$ is a zero-mean vector white noise process with covariance matrix $\mathbf{I}$, the identity matrix. The threshold process $\{Z_t\}$ is assumed to be stationary and have a continuous distribution. Notice the presence of the process $\{\mathbf{Y}_t\}$ in the autoregressive equation for $\{\mathbf{X}_t\}$. This is to explain for exogenous variables.

Model (1) can be seen as a particular case of model (7) if one puts $k = 1$, no exogenous variables, and $d = 0$ (although this value is not strictly covered by Tsay's (1998) model, in a mathematical sense). However, in model (1), one can have different autoregressive orders $k_1, \ldots, k_l$ in each regime and the threshold process is specified to be an invariant Markov chain, a more general concept than that of a stationary process. This point is important under the null hypothesis, as noted in the previous section, because, at present, the only method for estimating missing data in process $\{Z_t\}$, when it is nonlinear, is Nieto's (2005) approach.

Now, we describe Tsay's (1998) test. Consider the null hypothesis that $\{\mathbf{X}_t\}$ is linear, i.e. $l = 1$, versus the alternative hypothesis that it follows the multivariate threshold model given in (7). Using the arranged regression scheme, one has the following: given observations $\mathbf{x}_t$, $\mathbf{y}_t$, and z_t, $t = 1, 2, \ldots, n$, the goal is to detect the threshold nonlinearity of $\{\mathbf{X}_t\}$, assuming that p, q, and d are known.

Let $h = \max\{p, q, d\}$, $\mathbf{W}_t = (1, \mathbf{x}'_{t-1}, \ldots, \mathbf{x}'_{t-p}, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-q})$ (a $(pk + qv + 1)$–dimensional vector) and $\mathbf{\Phi}$ an unknown matrix. If the null hypothesis holds, then the model collapses to

$$\mathbf{X}'_t = \mathbf{W}'_t \mathbf{\Phi} + \boldsymbol{\varepsilon}'_t \tag{8}$$

as explained by Tsay (1998), for $t = h + 1, \ldots, n$, independent of the values of variable Z. Let $S = \{z_{h+1-d}, \ldots, z_{n-d}\}$ be the set of values of Z_{t-d}. Consider the order statistics of S and denote its ith smallest element by $z_{(i)}$. Then, the arranged regression based on the increasing order of the threshold variable Z_{t-d} is

$$\mathbf{X}'_{t(i)+d} = \mathbf{W}'_{t(i)+d} \mathbf{\Phi} + \boldsymbol{\varepsilon}'_{t(i)+d} \tag{9}$$

for $i = 1, \ldots, n - h$.

Now, let $\widehat{\mathbf{\Phi}}_m$ be the least square estimator of $\mathbf{\Phi}$ of equation (9) based on the first m observations, that is, those associated with the m smallest values of S. Let

$$\widehat{\mathbf{e}}_{t(m+1)+d} = \mathbf{X}_{t(m+1)+d} - \widehat{\mathbf{\Phi}}'_m \ \mathbf{W}_{t(m+1)+d} \tag{10}$$

and

$$\widehat{\boldsymbol{\eta}}_{t(m+1)+d} = \widehat{\mathbf{e}}_{t(m+1)+d} / [1 + \mathbf{W}'_{t(m+1)+d} \mathbf{V}_m \mathbf{W}_{t(m+1)+d}]^{1/2} \tag{11}$$

where $\mathbf{V}_m = \left[\sum_{i=1}^{m} \mathbf{W}_{t(i)+d} \mathbf{W}'_{t(i)+d}\right]^{-1}$, be the predictive residual and the standardized predictive residual of regression (9). These quantities can be obtained by the recursive least square algorithm. Next, consider the regression

$$\widehat{\boldsymbol{\eta}}'_{t(l)+d} = \mathbf{W}'_{t(l)+d} \mathbf{\Psi} + \boldsymbol{\epsilon}'_{t(l)+d} \tag{12}$$

for $l = m_0 + 1, \ldots, n - h$, where m_0 denotes the starting point of the recursive least squares estimation. The problem of interest is then to test the hypothesis $H_0 : \boldsymbol{\psi} = \mathbf{0}$ versus the alternative $H_a : \boldsymbol{\psi} \neq \mathbf{0}$. Tsay (1998) proposed the test statistic

$$C(d) = [n - h - m_0 - (kp + vq + 1)][\ln|\mathbf{S}_0| - \ln|\mathbf{S}_1|] \tag{13}$$

where the argument d signifies that the test depends strongly on the delayed threshold variable Z_{t-d}, $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$,

$$\mathbf{S}_0 = \frac{1}{n - h - m_0} \sum_{l=m_0+1}^{n-h} \widehat{\boldsymbol{\eta}}_{t(l)+d} \widehat{\boldsymbol{\eta}}'_{t(l)+d}$$

and

$$\mathbf{S}_1 = \frac{1}{n - h - m_0} \sum_{l=m_0+1}^{n-h} \widehat{\boldsymbol{\epsilon}}_{t(l)+d} \widehat{\boldsymbol{\epsilon}}'_{t(l)+d}$$

where $\widehat{\boldsymbol{\epsilon}}_{t(l)+d}$ is the least squares residual of regression (12). Under the null hypothesis that $\mathbf{X}_t$ is linear, Tsay (1998) showed that $C(d)$ is asymptotically a chi-squared

random variable with $k(pk + qv + 1)$ degrees of freedom. This paper shows that this test statistic has good optimal properties, which are reflected in having a greater power function than other statistical tests for the same null hypothesis. As a by-product, the statistic $C(d)$ can be used for choosing adequate threshold variables, when one has several candidate variables. The idea is to select that variable for which its corresponding value of $C(d)$ is the largest. Furthermore, if $k = 1$, i.e. $\{X_t\}$ is univariate, and there are not exogenous variable, i.e. $q = 0$, then the corresponding chi-squared distribution has $p + 1$ degrees of freedom.

Now, we consider the test statistic $C(d)$ above and set $d = 0$, then the previous regressions can still be conducted and thus the statistic $C(0)$ is computed. To find its null distribution for complete time series, we proceed via simulation and obtained tha,t even for small sample sizes, it is practically a chi-squared distribution with $p + 1$ degrees of freedom. Table 1 presents the results for a Monte Carlo simulation experiment, where the autoregressive linear model under the null is $X_t = 2 + 0.5X_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}$ is a Gaussian zero-mean white noise process with variance 1. The sample size was $n = 150$ and we run 5000 replicates. In the body of the table appear the quantiles of the $\chi^2(2)$ distribution and of the empirical distribution of $C(0)$. The p-value for the Kolmogorov-Smirnov test statistic was 0.36, approximately, which signals a no rejection of the null hypothesis of equal distributions. Additionally, we used a sample size $n = 10000$ and another AR(1) models with coefficients -0.5 and 1 (nonstationary process), and found analogous results. These are presented in Tables 4 and 5 of the Appendix. Furthermore, we considered AR(2) and AR(3) models with coefficients that make the processes stationary and nonstationary and, in the first case, we considered scenarios where the roots are either real numbers or some of them complex numbers. In this way, we take into account different characteristics in the time and frequency domain. We can provide these results upon request. The overall conclusion was the same, i.e. the distribution of $C(0)$ is practically a $\chi^2(p + 1)$ distribution, $p = 1, 2, 3$. We feel that the maximum value 3 for the AR order is enough in this simulation study because of model parsimony and that the exercise can be extended to seasonal AR processes obtaining the same global result.

Table 1: Comparison of empirical quantiles of the null distribution of $C(0)$ with those of the χ^2, in the case of an AR model with parameter 0.5.

	Quantiles								
Distribution	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
$\chi^2(2)$	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21
$C(0)$	0.02	0.05	0.10	0.22	1.42	4.59	5.97	7.31	9.19

4. The Null Distribution of the Proposed Test Statistic in the Case of Missing Data

4.1. A Tray to Use State-Space-Model Based Approaches

In the SETAR-model univariate context, Tong & Yeung (1991*a*) presented a state-space-model based procedure for implementing known tests of the null hypothesis $H : l = 1$, in the presence of partial data. Following the arranged-regression philosophy, they argued that under the null hypothesis the arranged regression can be cast in state space form. Setting $t(i)$ in place of t everywhere in equations 2 and 3, one would obtain

$$\boldsymbol{X}_{t(i)} = \boldsymbol{H}\boldsymbol{\alpha}_{t(i)} \tag{14}$$

as the observation equation, and

$$\boldsymbol{\alpha}_{t(i)} = \boldsymbol{C}_{J_{t(i)}} + \boldsymbol{A}_{J_{t(i)}}\boldsymbol{\alpha}_{t(i)-1} + \boldsymbol{R}_{J_{t(i)}}\omega_{t(i)} \tag{15}$$

as the system or state equation, where it is remarkably noted that $J_{t(i)} = 1$ for all $i = 1, \ldots, n-h$. Hence, the system equation becomes

$$\boldsymbol{\alpha}_{t(i)} = \boldsymbol{C}_1 + \boldsymbol{A_1}\boldsymbol{\alpha}_{t(i)-1} + \boldsymbol{R}_1\omega_{t(i)} \tag{16}$$

Equations (14) and (16) define Tong & Yeung's (1991*b*) state space model and they will be referred as an arranged state space model.

Apparently, the usual statistical assumptions and properties of state space models continue to be valid (see Harvey's (1989) book, for example), and thus the Kalman filter, its associated smoothing algorithms and the Nieto's (2005) approach might still be used. However, this is not possible. Indeed, (i) an important argument in deducting the Kalman filter and then the well-known smoothing algorithms (see, among others, Harvey (1989), Catlin (1989) and Brockwell & Davis (1991)) is that the time points at which observations are made need to be in a monotone order although not necessarily equally spaced. In the present scheme, it can happen that $i < j$ and even $t(i) > t(j)$. (ii) The so-called predictive residuals are not orthogonal among them and orthogonal to lagged variables of the output process $\{X_t\}$ neither. Hence, the probabilistic behaviour of the so-called *adapted* test statistics of Tong & Yeung (1991*a*), which is necessary for implemeting their tests, is not necessarily guaranteed. The following example illustrates these facts.

An AR(1) model will be considered for process $\{X_t\}$ given by $X_t = a_1 X_{t-1} + h\varepsilon_t$ (as happens under the null hypothesis), where a_1 and h are real numbers with $h > 0$, and $\{\varepsilon_t\}$ is a zero-mean white noise process for which $\mathrm{E}(X_s\varepsilon_t) = 0$ for $s < t$. Then, trivially, the state-space-model elements are $\boldsymbol{\alpha}_t = (X_t)$, $\boldsymbol{\omega}_t = (\varepsilon_t)$, $\boldsymbol{H} = 1$, $\boldsymbol{C}_1 = 0$, $\boldsymbol{A}_1 = a_1$, and $\boldsymbol{R}_1 = h$. We assume that the sample size is $n = 1000$, and that $t(1) = 200$, where there is a missing data, $t(2) = 27$, and $t(3) = 379$. After some simple calculations, using the algorithms presented by Tong & Yeung (1991*a*) for computing the state-space-model based predictive errors $\boldsymbol{\eta}$'s, one founds that $\eta_{27} = X_{27} - a_1^2 X_{199}$ and $\eta_{379} = X_{379} - a_1^2 X_{27}$, which are clearly

correlated. Consequently, we leave Tong & Yeung's (1991*a*) state-space-model based approach and consider the alternative of using directly the test described in Subsection 3.1.

4.2. The Null Distribution of $C(0)$

Now, the idea is to assess the influence that the uncertainty in the missing data estimates has on the distribution of $C(0)$. Let $\widehat{C}(0)$ be the statistic that is obtained when we use missing data estimates to compute $C(0)$. We proceed via Monte Carlo simulation as in the case of complete data, maintaining the same AR(p) models for process $\{X_t\}$, that is with $p = 1, 2, 3$ and different values for the autoregressive parameters, and considering different sample sizes. The new element is to consider several rates of missing observations, going from low to high percentages, and to detect a threshold rate up to which the χ^2 null distribution is preserved. In this paper, we consider values in the set $\{0\%, 10\%, 20\%, \ldots, 80\%\}$.

The design of the simulation experiment was the following: we fix a stationary AR(1) model for process $\{Z_t\}$; in this way, $\{Z_t\}$ is a Markov chain of order 1 with invariant distribution. The chosen model is $Z_t = 0.25Z_{t-1} + \alpha_t$, where $\{\alpha_t\}$ is a Gaussian zero-mean white noise process with variance 1.5^2. Then,

(1) we draw time series for each stochastic process $\{X_t\}$ and $\{Z_t\}$, say $\{x_t\}$ and $\{z_t\}$, in an independent way.

(2) We select randomly two sets of time points in the set $\{1, \ldots, T\}$. The first one of size $T - N$ for $\{x_t\}$ and the second one of size $T - M$ for $\{z_t\}$. These time points are fixed. Then, we discard the observations in the time series $\{x_t\}$, located at the first set of time points, and those for $\{z_t\}$ that correspond to the second set.

(3) Using Gómez & Maravall (1994) procedures, specifically their fixed-point smoother algorithm, we estimated the missing observations in the time series $\{x_t\}$. Since $\{Z_t\}$ is linear, the same procedure is used to estimate the missing data in its simulated time series.

(4) Compute $\widehat{C}(0)$ with these "completed" time series.

***Note* 1.** Thinking in practice, if process $\{Z_t\}$ is not linear, we can use the smoother given by equations (4) and (5) for estimating its missing data, with the following modification: since there is no influence of $\{Z_t\}$ onto $\{X_t\}$, the posterior densities presented in equations (4) and (5) are reduced, respectively, to

$$p(\mathbf{z}_T \mid \boldsymbol{\alpha}, \mathbf{x}) \propto f_p(\mathbf{z}_T) \tag{17}$$

for $t = T - p + 1, \ldots, T$, and

$$p(z_t \mid \mathbf{z}_{t+p}, \boldsymbol{\alpha}_t, \mathbf{x}_t) \propto f_p(z_{t+p} \mid \mathbf{z}_{t+p-1}) f_p(\mathbf{z}_{t+p-1}) \tag{18}$$

for $t = T-p, \ldots, 1$. In this way, drawings for $\mathbf{Z}_T = (Z_{T-p+1}, \ldots, Z_T)$ are obtained directly from the invariant distribution of the Markov chain $\{Z_t\}$, and for $t =$

$T-p,\ldots,1$, drawings for Z_t are obtained from the distribution given by the product of the kernel density with the invariant density.

The above procedure is repeated I times, $I \geq 1$, to obtain a sample of size I for the statistic $\widehat{C}(0)$, maintaining fixed the missing-data percentages $(T-N)/T$ and $(T-M)/T$ through all the iterations. With this sample for $\widehat{C}(0)$, we obtained its empirical cumulative distribution function and then compare it with that of the χ^2_{p+1} distribution.

The results of the simulation experiment, with the AR(1) model with parameter 0.5 for $\{X_t\}$, are presented in Table 2, using samples of size $n = 150$ and $I = 5000$, a fixed value that we will use in all the remaining simulations. We can see the following important facts: (i) when there are not missing observations in the time series $\{x_t\}$, for any percentage of missing data in $\{z_t\}$, the empirical quantiles are practically equal to the $\chi^2(2)$ distribution. This reflects that under H_0, the process $\{Z_t\}$ does not influence $\{X_t\}$. (ii) Fixing the missing-data percentage of $\{x_t\}$ and varying that of $\{z_t\}$, the corresponding empirical quantiles are very similar, reflecting once more again the observation made in (i). (iii) When the rate of missing data in the time series $\{x_t\}$ gets larger, the discrepancy between empirical and theoretical quantiles gets larger, too, independent of the missing-data percentage in the time series $\{z_t\}$. This fact suggests that the null distribution of $\widehat{C}(0)$ departs from the $\chi^2(2)$ distribution.

Table 2: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ with those of the $\chi^2(2)$ for $n = 150$ and $\phi = 0.5$.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.04	0.10	0.21	1.37	4.58	5.92	7.43	9.31
20	0	0.02	0.05	0.09	0.20	1.34	4.48	5.80	7.06	8.94
30	0	0.02	0.05	0.10	0.22	1.51	5.12	6.67	8.04	10.36
0	10	0.02	0.05	0.10	0.21	1.40	4.62	5.98	7.33	8.80
10	10	0.02	0.05	0.09	0.21	1.35	4.50	5.88	7.28	8.72
20	10	0.02	0.04	0.09	0.19	1.33	4.37	5.65	6.87	8.59
30	10	0.02	0.05	0.12	0.22	1.51	5.10	6.48	8.22	10.23
0	20	0.02	0.05	0.10	0.21	1.41	4.63	6.27	7.73	9.57
10	20	0.02	0.05	0.10	0.20	1.39	4.74	6.17	7.56	9.40
20	20	0.02	0.05	0.11	0.22	1.29	4.25	5.61	6.91	8.84
30	20	0.02	0.05	0.11	0.22	1.32	4.48	5.75	7.09	9.30
0	30	0.02	0.05	0.10	0.22	1.37	4.64	6.06	7.51	9.22
10	30	0.02	0.05	0.10	0.21	1.40	4.70	6.15	7.57	9.88
20	30	0.02	0.04	0.09	0.20	1.30	4.24	5.61	6.94	8.80
30	30	0.02	0.04	0.09	0.18	1.29	4.21	5.43	6.71	8.65
$\chi^2(2)$		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

In Table 3 we present the Kolmogorov-Smirnov statistics for the same missing-data rates considered in Table 2 and, as we can see, when the percentage of missing data in $\{x_t\}$ is greater or equal than 20%, approximately, the null hypothesis of

equal distributions is rejected in almost every case. Consequently, we always halt the simulations at the rate 30%.

TABLE 3: p-values of the Kolmogorov-Smirnov statistic for the distributions in Table 2.

% of missing data		
x-data	z-data	p-value*
10	0	0.14
20	0	0.28
30	0	0.00
0	10	0.80
10	10	0.26
20	10	0.01
30	10	0.00
0	20	0.76
10	20	0.85
20	20	0.00
30	20	0.02
0	30	0.86
10	30	0.90
20	30	0.00
30	30	0.00

* Rounded to two decimal digits.

We repeated the simulation exercise for this AR model with sample sizes $n = 1000, 10000$ and obtained very similar results to the last ones. Also, we used the other AR(1) models proposed in Section 3 with the same sample sizes $n = 150$ and $n = 10000$, and we find analogous results, which are in Tables 6-17 in the Appendix. In this last case, we omitted the sample size $n = 1000$ because we do not observe important differences with respect to the sample size $n = 150$. Furthermore, we consider the same AR(2) and AR(3) models of Section 3 with the same sample sizes and, once more again, we obtained similar results, which can be provided upon request.

As a global conclusion, when the missing data percentage in the time series $\{x_t\}$ is less than 20%, approximately, we can say that the null distribution of $\widehat{C}(0)$ is still a $\chi^2(p+1)$ distribution. For rates of missing data greater than this approximate threshold, the influence of the missing-data estimates uncertainty on the distribution of $\widehat{C}(0)$ is so relevant that its empirical distribution departs from the χ^2 distribution. It is important to remark here that one could extend this simulation study via the selection of percentage values between 10% and 20%, to find a more precise bound for the percentage of missing values in $\{x_t\}$ up to which the χ^2 distribution continues to be valid as the null distribution of $\widehat{C}(0)$.

5. An Empirical Example

Nieto (2005) considers an application with actual data. The time series considered were daily rainfall (in mm), as the threshold variable, and a daily river flow (in m^3/s), as the response variable, in a certain Colombian geographical region. The data set corresponds to the sample period from January 1, 1992, up to November 30, 2000 (3256 data), and it was assembled by IDEAM, the official Colombian agency for hydrological and meteorological studies. In Figure 1, one can see the two time series, where is clear the dynamical relationship between the two variables. Additionally, one can see certain stable path in both variables and bursts of large values, specifically in the river flow.

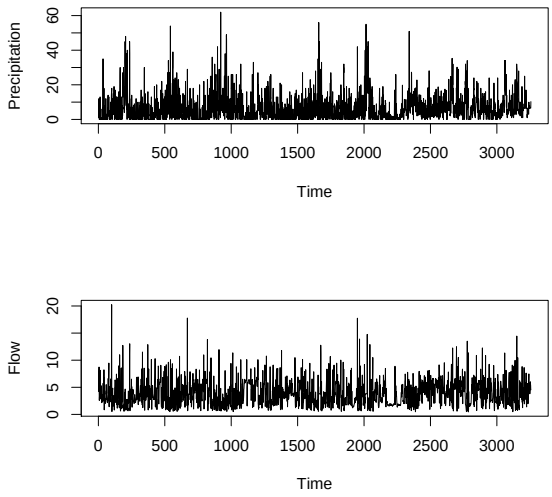

FIGURE 1: Time series for the real data example: (a) Precipitation (b) Flow.

Let P_t and X_t be respectively the rainfall and the river flow at day t. Because of the universal convention for measuring these two variables, he set up $Z_t = P_{t-1}$. That is, the precipitation was lagged one period back for relating it to the river flow. The flow time series was adjusted with two transformations: (1) square root of the data and (2) an adjustment for conditional heteroscedasticity via an ARCH(1) model. From now on, the flow data to be analyzed will be the transformed ones and we denote them as $\{\widetilde{x}_t\}$. The two time series have missing data, 52 in $\{z_t\}$ and 32 in $\{x_t\}$. In percentage terms, these are 1.6% for time series $\{z_t\}$ and 1% for $\{\widetilde{x}_t\}$, which are lesser than 20%.

Under the null hypothesis, $\{X_t\}$ is a linear AR process. Using Akaike's information criterion, we found that $p = 2$ is a reasonable autoregressive order for this process. In this way, we are in the conditions studied in Section 4 and, consequently, we can use the $\chi^2(3)$ distribution for running the statistical test for

the null hypothesis. We note firstly that the process $\{Z_t\}$ has two regimes with $r_1 = 6.0$ mm and that it was modeled as a 1st order Markov chain with approximate intial and transition kernel distributions given by the mixtures

$$f_0(z) = 0.26h_n(z) + 0.74g(z)$$

and

$$f_1(z_t \mid z_{t-1}) = 0.87h_n(z_t) + 0.13g(z_t \mid z_{t-1})$$

respectively, where

$$h_n(z) = \begin{cases} 0, & \text{if } -\infty < z < -1/n, \\ (n\pi/2)\cos(n\pi z + \pi/2), & \text{if } -1/n \leq z \leq 0 \\ 0, & \text{if } z > 0 \end{cases}$$

$g(z)$ denotes the truncated density of a $N(3.24, 7.76^2)$ at the point $z = 0$, $g(z_t \mid z_{t-1})$ is the truncated density of a $N(z_{t-1}, 7.76^2)$ at the same point $z = 0$, $P(Z_t = 0) = 0.26$, and $P(Z_t = 0 \mid Z_{t-1} \in B_1) = 0.87$. For more details on the modeling of process $\{Z_t\}$, the reader can see Gómez & Maravall (1994) paper. Then, we used the procedure described in Subsection 3.3 for estimating the missing data in the two time series, specifically, we used the TSW software (Caporello & Maravall 2003) for estimating the missing data in $\{\widetilde{x}_t\}$ and the smoother given in equations (17) and (18) for estimating the missing observations in $\{z_t\}$. Next, I complete the time series with the estimated data and obtained that $\widehat{C}(0) = 52.57$ with p-value equal to 0. These results signal the strong threshold nonlinearity of $\{X_t\}$, which is explained by $\{Z_t\}$.

6. Conclusions

In this paper, we have shown the feasibility of a well-known statistical procedure for testing the null hypothesis of linearity against the alternative of TAR nonlinearity, when there are missing data in a bivariate time series. The statistical test is an extension of Tsay's (1998) statistic. The extension consists in allowing the number zero to be the delay parameter of the threshold variable. Then, to compute values of this statistic we use estimates of the missing data as observed values. Strictly speaking, this statistic is other than the extended one. Via Monte Carlo simulations, we found that the extended statistic also follows a χ^2 distribution with complete data and that the modified statistic also has this distribution if the proportion of missing data in the output time series is less or equal than 20%, approximately. We feel that if the missing-data percentage is larger than this value, we should use additional variables that help to explain the dynamical behavior of the output one, via for example regression models, to get more observed data and to do a frequentist statistical test. Another alternative might be to use a Bayesian testing approach to compensate the large uncertainty produced by the high percentage of missing data. This route would need to consider appropriate prior distributions. This is a challenging problem for future research.

In the lines of the above recommendation, another interesting problem for future research would be to consider test statistics other than Tsay's (1998) one, as for example Hansen's (1996) test. And then to do a comparison about the size and power of the different tests, under scenarios of missing-data proportions where the known null distributions are preserved.

As a by-product of this study, we have also shown that state-space-model based approaches, which aim to take into account the arranged autoregression philosophy, are not adequate. This means that the appealing idea of detecting change points, via arranged regressions, should be used directly for designing inferential procedures in the context of TAR models.

7. Acknowledgments

We gratefully acknowledge Professor Ruey S. Tsay at the University of Chicago for his disposal to discuss with us the topic and for his valuable advising on the development of the research. We also thank an anonymous referee for useful comments and suggestions on a previous version of the paper, which help to improve substantially the initial manuscript.

[Recibido: enero de 2008 — Aceptado: febrero de 2011]

References

Brockwell, P. J. (1994), 'On continuous-time threshold ARMA processes', *Journal of Statistical Planning and Inference* **39**, 291–303.

Brockwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods*, Springer-Verlag, New York.

Caporello, G. & Maravall, A. (2003), *Software TSW*, Banco de España, Madrid.

Carter, C. K. & Kohn, R. (1994), 'On Gibbs sampling for state space models', *Biometrika* **81**, 541–553.

Carter, C. K. & Kohn, R. (1996), 'Markov chain Monte Carlo in conditionally gaussian state space models', *Biometrika* **83**, 589–601.

Catlin, D. (1989), *Estimation, Control, and the Discrete Kalman Filter*, Springer-Verlag, New York.

Gómez, V. & Maravall, A. (1994), 'Estimation, prediction, and interpolation for nonstationary series with the Kalman filter', *Journal of the American Statistical Association* **89**, 611–624.

Hansen, B. E. (1996), 'Inference when a nuisance parameter is not identified under the null hypothesis', *Econometrica* **64**, 413–460.

Harvey, A. C. (1989), *Forecasting, Structural Time Series, and the Kalman filter*, Cambridge University Press, Cambridge.

Kim, C. & Nelson, C. R. (1999), *State Space Models with Regime Switching*, The MIT Press, Cambridge.

Nieto, F. H. (2005), 'Modeling bivariate threshold autoregressive processes in the presence of missing data', *Communications in Statistics - Theory and Methods* **34**(4), 905–930.

Shumway, R. H. & Stoffer, D. S. (1991), 'Dynamic linear models with switching', *Journal of the American Statistical Association* **86**, 411–430.

Tong, H. (1978), On a threshold model in pattern recognition and signal processing, *in* C. H. Chen, ed., 'Pattern recognition and signal processing', Sijhoff & Noordhoff, Amsterdam.

Tong, H. & Lim, K. S. (1980), 'Threshold autoregression, limit cycles, and cyclical data', *Journal of the Royal Statistical Society, Series B* **42**, 245–292.

Tong, H. & Yeung, I. (1991*a*), 'On tests for self-exciting threshold autoregressive non-linearity in partially observed time series', *Applied Statistics* **40**, 43–62.

Tong, H. & Yeung, I. (1991*b*), 'Threshold autoregressive modeling in continuous time', *Statistica Sinica* **1**, 411–430.

Tsai, H. & Chan, K. S. (2000), 'Testing for nonlinearity with partially observed time series', *Biometrika* **87**, 805–821.

Tsay, R. S. (1998), 'Testing and modeling multivariate threshold models', *Journal of the American Statistical Association* **93**, 1188–1202.

Appendix

TABLE 4: Comparison of empirical quantiles of the null distribution of $C(0)$ with those of a $\chi^2(2)$, in the case of an AR(1) process with parameter -0.5 and sample size $n = 150$.

	Quantiles								
Distribution	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
$\chi^2(2)$	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21
$C(0)$	0.02	0.05	0.11	0.21	1.40	4.62	6.06	7.36	9.15

The Kolmogorov-Smirnov statistical test yields a p-value of 0.67, which implies a no rejection of the null hypothesis of equal distributions. With a sample size $n = 10000$ the p-value of this statistical test is 0.38, which leads to the same decision.

Table 5: Empirical quantiles of the null distribution of $C(0)$ and those of the $\chi^2(2)$, in the case of an AR(1) process with parameter 1.0 and sample size $n = 150$.

	Quantiles								
Distribution	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
$\chi^2(2)$	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21
$C(0)$	0.03	0.06	0.12	0.23	1.40	4.62	6.05	7.42	9.35

Table 6: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 0.5 and sample size 1000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.11	0.22	1.37	4.42	5.83	7.16	9.08
20	0	0.02	0.06	0.10	0.22	1.42	4.63	6.19	7.64	9.54
30	0	0.02	0.05	0.11	0.22	1.46	4.76	6.17	7.59	9.95
0	10	0.02	0.06	0.10	0.21	1.37	4.56	6.06	7.54	9.17
10	10	0.02	0.05	0.11	0.21	1.38	4.39	5.73	7.36	9.23
20	10	0.01	0.05	0.10	0.22	1.43	4.86	6.21	7.67	9.88
30	10	0.02	0.04	0.09	0.20	1.44	4.76	6.20	7.77	9.63
0	20	0.02	0.05	0.10	0.22	1.39	4.63	6.01	7.31	8.93
10	20	0.02	0.05	0.11	0.23	1.40	4.48	5.77	7.26	9.01
20	20	0.02	0.05	0.10	0.21	1.42	4.60	6.00	7.28	9.29
30	20	0.02	0.05	0.11	0.21	1.45	4.77	6.16	7.34	9.78
0	30	0.03	0.06	0.12	0.22	1.43	4.69	6.18	7.53	9.44
10	30	0.02	0.06	0.11	0.23	1.43	4.77	6.15	7.41	9.47
20	30	0.02	0.05	0.10	0.23	1.44	4.82	6.16	7.46	9.67
30	30	0.02	0.05	0.11	0.24	1.52	4.91	6.34	7.96	9.92
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

The Kolmogorov-Smirnov statistical test yields a p-value of 0.74, which signals a no rejection of the null hypothesis of equal distributions. With a sample size $n = 10000$ the p-value of this statistical test is 0.12, which signals the same decision.

TABLE 7: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 0.5 and sample size 1000.

% of missing data		p-value*
x-data	z-data	
10	0	0.56
20	0	0.31
30	0	0.00
0	10	0.78
10	10	0.27
20	10	0.11
30	10	0.11
0	20	0.98
10	20	0.23
20	20	0.41
30	20	0.09
0	30	0.07
10	30	0.01
20	30	0.11
30	30	0.00

* Rounded to two decimal digits.

TABLE 8: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 0.5 and sample size 10000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.10	0.20	1.37	4.68	6.05	7.50	9.15
20	0	0.02	0.05	0.10	0.21	1.40	4.76	6.34	7.67	9.29
30	0	0.02	0.05	0.11	0.22	1.48	4.81	6.24	7.82	9.93
0	10	0.02	0.05	0.11	0.20	1.35	4.54	5.87	7.09	8.65
10	10	0.02	0.05	0.10	0.21	1.35	4.59	5.98	7.28	8.68
20	10	0.02	0.05	0.10	0.20	1.42	4.82	6.23	7.63	9.33
30	10	0.02	0.05	0.11	0.22	1.46	4.77	6.12	7.59	9.91
0	20	0.02	0.05	0.10	0.20	1.35	4.56	5.87	7.25	8.89
10	20	0.02	0.05	0.11	0.21	1.40	4.62	5.85	7.33	9.33
20	20	0.02	0.05	0.10	0.20	1.39	4.82	6.28	7.56	9.57
30	20	0.02	0.06	0.12	0.22	1.43	4.73	6.13	7.52	9.75
0	30	0.02	0.05	0.10	0.20	1.36	4.45	5.94	7.13	8.75
10	30	0.02	0.05	0.11	0.22	1.38	4.57	6.05	7.24	8.75
20	30	0.02	0.05	0.10	0.22	1.38	4.77	6.07	7.46	9.17
30	30	0.02	0.05	0.10	0.22	1.44	4.71	6.12	7.51	9.37
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 9: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 0.5 and sample size 10000.

% of missing data		p-value*
x-data	z-data	
10	0	0.27
20	0	0.23
30	0	0.00
0	10	0.35
10	10	0.43
20	10	0.12
30	10	0.02
0	20	0.15
10	20	0.74
20	20	0.22
30	20	0.30
0	30	0.66
10	30	0.80
20	30	0.44
30	30	0.02

* Rounded to two decimal digits.

TABLE 10: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter -0.5 and sample size 150.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.11	0.21	1.39	4.65	6.02	7.15	9.11
20	0	0.02	0.06	0.12	0.22	1.35	4.48	5.96	7.26	8.70
30	0	0.02	0.05	0.11	0.22	1.51	5.14	6.72	8.16	10.36
0	10	0.02	0.04	0.09	0.20	1.40	4.57	5.92	7.17	9.50
10	10	0.02	0.05	0.10	0.21	1.35	4.56	5.92	7.15	8.95
20	10	0.02	0.05	0.11	0.21	1.29	4.23	5.74	7.12	8.40
30	10	0.02	0.05	0.11	0.24	1.49	5.06	6.48	8.03	9.65
0	20	0.02	0.05	0.10	0.21	1.41	4.67	6.13	7.62	9.76
10	20	0.02	0.04	0.10	0.21	1.40	4.65	6.07	7.79	9.65
20	20	0.02	0.04	0.08	0.19	1.25	4.22	5.44	6.64	8.62
30	20	0.02	0.05	0.10	0.22	1.42	4.68	6.33	7.70	9.74
0	30	0.02	0.05	0.11	0.21	1.38	4.52	6.02	7.45	9.25
10	30	0.02	0.05	0.10	0.21	1.43	4.85	6.33	7.73	9.50
20	30	0.02	0.05	0.09	0.19	1.21	4.16	5.36	6.58	8.59
30	30	0.02	0.05	0.10	0.19	1.28	4.44	5.88	7.14	8.71
	χ^2	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 11: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient -0.5 and sample size 150.

% of missing data		p-value*
x-data	z-data	
10	0	0.91
20	0	0.22
30	0	0.00
0	10	0.77
10	10	0.56
20	10	0.00
30	10	0.00
0	20	0.53
10	20	0.97
20	20	0.00
30	20	0.08
0	30	0.79
10	30	0.09
20	30	0.00
30	30	0.00

* Rounded to two decimal digits.

TABLE 12: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter -0.5 and sample size 10000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.04	0.09	0.20	1.38	4.66	5.95	7.20	8.98
20	0	0.02	0.05	0.10	0.22	1.45	4.77	6.20	7.52	9.58
30	0	0.02	0.06	0.12	0.23	1.46	4.82	6.20	7.88	9.93
0	10	0.02	0.05	0.10	0.20	1.37	4.52	5.93	7.11	8.92
10	10	0.02	0.04	0.10	0.21	1.36	4.62	5.97	7.30	9.26
20	10	0.02	0.05	0.12	0.23	1.47	4.82	6.20	7.60	9.75
30	10	0.02	0.05	0.10	0.21	1.44	4.74	6.28	7.68	9.47
0	20	0.02	0.04	0.10	0.20	1.33	4.47	5.76	7.28	8.77
10	20	0.01	0.04	0.10	0.20	1.37	4.45	5.91	7.57	9.40
20	20	0.02	0.06	0.10	0.21	1.43	4.64	6.06	7.44	9.34
30	20	0.02	0.05	0.11	0.23	1.40	4.76	6.16	7.63	9.48
0	30	0.02	0.04	0.10	0.20	1.35	4.47	5.86	7.10	8.61
10	30	0.02	0.04	0.09	0.20	1.39	4.49	5.98	7.43	8.90
20	30	0.03	0.06	0.11	0.21	1.42	4.81	6.25	7.66	9.40
30	30	0.02	0.05	0.11	0.23	1.44	4.72	6.05	7.46	9.43
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 13: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient -0.5 and sample size 10000.

% of missing data		p-value*
x-data	z-data	
10	0	0.96
20	0	0.02
30	0	0.00
0	10	0.60
10	10	0.78
20	10	0.01
30	10	0.01
0	20	0.14
10	20	0.79
20	20	0.06
30	20	0.15
0	30	0.48
10	30	0.50
20	30	0.04
30	30	0.08

* Rounded to two decimal digits.

TABLE 14: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 1.0 and sample size 150.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.10	0.20	1.38	4.55	5.93	7.22	9.09
20	0	0.02	0.06	0.12	0.24	1.40	4.81	6.18	7.58	9.24
30	0	0.02	0.06	0.11	0.23	1.61	5.34	6.94	8.42	10.30
0	10	0.03	0.06	0.11	0.22	1.40	4.60	6.08	7.52	9.12
10	10	0.02	0.05	0.10	0.21	1.39	4.52	5.90	7.33	8.99
20	10	0.02	0.04	0.10	0.22	1.40	4.62	6.09	7.43	9.45
30	10	0.02	0.07	0.13	0.25	1.59	5.24	6.74	8.30	10.01
0	20	0.02	0.05	0.11	0.22	1.41	4.68	6.03	7.40	9.01
10	20	0.02	0.05	0.11	0.23	1.42	4.67	6.13	7.48	9.28
20	20	0.02	0.05	0.11	0.23	1.49	4.77	6.04	7.41	9.47
30	20	0.02	0.05	0.11	0.22	1.45	4.69	6.11	7.68	9.34
0	30	0.02	0.05	0.11	0.22	1.41	4.45	5.77	7.34	9.27
10	30	0.02	0.05	0.10	0.21	1.42	4.46	5.75	7.16	8.96
20	30	0.02	0.05	0.10	0.23	1.49	4.76	6.17	7.64	9.87
30	30	0.02	0.05	0.10	0.20	1.39	4.49	5.67	7.02	8.80
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 15: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 1.0 and sample size 150.

% of missing data		p-value*
x-data	z-data	
10	0	0.83
20	0	0.31
30	0	0.00
0	10	0.25
10	10	0.92
20	10	0.48
30	10	0.00
0	20	0.28
10	20	0.09
20	20	0.00
30	20	0.01
0	30	0.48
10	30	0.34
20	30	0.00
30	30	0.05

* Rounded to two decimal digits.

TABLE 16: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 1.0 and sample size 10000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.01	0.04	0.09	0.16	1.14	3.79	5.03	6.20	7.78
20	0	0.02	0.05	0.11	0.22	1.48	4.88	6.39	7.90	10.21
30	0	0.02	0.06	0.12	0.23	1.42	4.79	6.26	8.14	10.38
0	10	0.02	0.06	0.11	0.22	1.32	4.69	6.22	7.58	9.36
10	10	0.01	0.04	0.10	0.19	1.21	3.96	5.14	6.42	7.84
20	10	0.03	0.06	0.12	0.23	1.46	4.91	6.49	7.88	10.04
30	10	0.02	0.05	0.10	0.22	1.44	4.87	6.37	8.07	10.08
0	20	0.03	0.05	0.10	0.21	1.37	4.72	6.37	7.78	9.51
10	20	0.03	0.05	0.10	0.22	1.30	4.03	5.29	6.78	8.54
20	20	0.02	0.06	0.12	0.23	1.51	5.03	6.56	7.96	9.77
30	20	0.02	0.05	0.10	0.22	1.48	4.96	6.53	8.04	10.23
0	30	0.02	0.04	0.10	0.20	1.37	4.60	5.90	7.32	9.31
10	30	0.02	0.03	0.08	0.16	1.03	3.63	4.79	5.94	7.53
20	30	0.02	0.05	0.11	0.23	1.48	4.86	6.32	7.64	9.82
30	30	0.02	0.05	0.10	0.22	1.47	5.09	6.54	8.31	11.03
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 17: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 1.0 and sample size 10000.

% of missing data		p-value*
x-data	z-data	
10	0	0.12
20	0	0.01
30	0	0.08
0	10	0.09
10	10	0.00
20	10	0.03
30	10	0.00
0	20	0.72
10	20	0.00
20	20	0.00
30	20	0.00
0	30	0.99
10	30	0.00
20	30	0.00
30	30	0.01

* Rounded to two decimal digits.