econometrics

MDPI

Article

Selecting the Lag Length for the M^{GLS} Unit Root Tests with Structural Change: A Warning Note for Practitioners Based on Simulations[†]

Ricardo Quineche [1,2] and Gabriel Rodríguez [2,*]

1. Central Reserve Bank of Peru, 441–445 Antonio Miró Quesada Street, Lima 1, Lima, Peru; rquinecheu@pucp.pe
2. Department of Economics, Pontificia Universidad Católica del Perú, Av. Universitaria 1801, Lima 32, Lima, Peru

* Correspondence: gabriel.rodriguez@pucp.edu.pe; Tel.: +511-626-2000 (ext. 4998)

† This Note is drawn from the Thesis of Ricardo Quineche at the Department of Economics, Pontificia Universidad Católica del Perú (PUCP). This is a substantial revised version of an earlier version circulated under the title *Data-Dependent Methods for the Lag Length Selection in Unit Root Tests with Structural Change* which appear as Working Paper 404 of the Department of Economics, PUCP. The current version has been improved thanks to the relevant comments received from the Guest Editor of the Review (Professor Pierre Perron) and three anonymous Referees. Further, we thank comments of Luis García, José Tavera and Jorge Rojas (PUCP). Any remaining errors are our responsibility.

Academic Editor: Pierre Perron
Received: 31 August 2016; Accepted: 4 April 2017; Published: 16 April 2017

Abstract: This is a simulation-based warning note for practitioners who use the M^{GLS} unit root tests in the context of structural change using different selection lag length criteria. With $T = 100$, we find severe oversize problems when using some criteria, while other criteria produce an undersizing behavior. In view of this dilemma, we do not recommend using these tests. While such behavior tends to disappear when $T = 250$, it is important to note that most empirical applications use smaller sample sizes such as $T = 100$ or $T = 150$. The ADF^{GLS} test does not present an oversizing or undersizing problem. The only disadvantage of the ADF^{GLS} test arises in the presence of $MA(1)$ negative correlation, in which case the M^{GLS} tests are preferable, but in all other cases they are very undersized. When there is a break in the series, selecting the breakpoint using the Supremum method greatly improves the results relative to the Infimum method.

Keywords: unit root tests; structural change; truncation lag; GLS detrending; information criteria; sequential general to specific t-sig method

JEL Classification: C22; C52

1. Introduction

Testing for the presence of a unit root in a time series (i.e., whether or not a structural change can be identified) is now a common starting point in advanced models frequently used in macroeconomics and finance. Recent efficient unit root tests are the ADF^{GLS} and the P_T^{GLS} tests proposed by Elliott et al. (1996), and the M^{GLS} tests proposed by Ng and Perron (2001).[1] All these (GLS-based) tests have been extended to the unit root with one unknown structural change as suggested by Perron

[1] For excellent surveys, see Stock (1994), Maddala and Kim (1998), Phillips and Xiao (1998), Haldrup and Jansson (2006), Perron (2006), and Choi (2015).

and Rodríguez (2003), who show that these tests enjoy the same efficiency characteristics. M^{GLS} tests have become increasingly popular in the literature. For example, Haldrup and Jansson (2006) argue that practitioners should abandon the use of ADF tests altogether in favor of M^{GLS} tests because of their excellent size properties and nearly optimal power properties. However, this note arrives at the opposite conclusion, suggesting that the choice of the most suitable testing method should be carefully assessed.

Currently, it is widely accepted that the selection of the lag length (denoted by k) has important implications for the (size and power) behavior of the different unit root tests. See, for instance, Schwert (1989), Ng and Perron (1995), Agiakloglou and Newbold (1992), Agiakloglou and Newbold (1996), Elliott et al. (1996), Ng and Perron (2001), Del Barrio Castro et al. (2011), and Fossati (2012). The consensus is to use data-dependent methods. These rules include *AIC* (Akaike Information Criterion), *BIC* (Bayesian Information Criterion), Modified *AIC* (*MAIC*), Modified *BIC* (*MBIC*), and the t-sig method, which are briefly explained below.

Recently, we performed a routine empirical application of the M^{GLS} tests and obtained strange results. For example, applying the $MZ_{\widehat{\alpha}}^{GLS}$ and the *AIC* method to the labor market of the Spanish region of Cantabria,[2] we obtained an unemployment rate of $-3'140{,}463$, a huge (explosive) negative value with $k = 9$. Using the t-sig procedure , we obtained $-50'078{,}041$ with $k = 10$, which is even more impressive. A straightforward interpretation implies an overwhelming rejection of the null hypothesis, given any of the asymptotic or finite critical values tabulated in Perron and Rodríguez (2003). However, it is clear that the magnitude of this value is counter-intuitive and inadmissible, because its magnitude is very far from standard values. In contrast, other rules yield opposite results (very small values in absolute value). When applied to other three time series (unemployment rates in the Spanish regions of Galicia and Murcia, and to Peru's monetary policy rate), similar results are obtained.[3] In consequence, we consider that it is worth analyzing the source of the poor behavior of the M^{GLS} tests in the cases mentioned above. Hence, we perform extensive finite sample simulations for the M^{GLS} tests using different lag-length criteria, where the size performance is our primary interest.

This note (to our best knowledge) represents the first simulation-based attempt to study the size and the eccentric behavior of the M^{GLS} unit root tests in the context of structural change. We do not pretend to perform an exhaustive analysis of each rule. Rather, this document is only a simulation-based note of caution for users of these unit root tests.[4]

This note is structured as follows. In Section 2, the GLS approach with structural break, the test statistics, the rules used to select k, and the two methods to select the break date are briefly reviewed. In Section 3 we present simulation evidence about the size of the $MZ_{\widehat{\alpha}}^{GLS}$ test linking the results with an explosive behavior of the test. Section 4 provides some conclusions.

2. DGP, GLS Detrending, M^{GLS} Tests with Structural Change, Rules for Selecting the Lag Length, and Methods for Selecting the Breakpoint

2.1. The DGP

Following Perron and Rodríguez (2003), the data generating process (DGP) is:

$$\begin{aligned} y_t &= d_t + u_t, \\ u_t &= \alpha u_{t-1} + v_t, \end{aligned} \tag{1}$$

[2] Quartely data covering the period Q3 1976–Q2 2012 ($T = 144$ observations).
[3] The sample size for Galicia and Murcia are the same as for Cantabria. For Peru's monetary policy rate, the data are monthly for February 2002–August 2010 ($T = 92$ observations).
[4] We recognize the limitations of this note, which is only based on simulations. We agree with a Referee that formal proofs are needed in the spirit of Del Barrio Castro et al. (2013). Hence, further work in the direction of a formal treatment will be addressed in a future research project.

for $t = 0, 1, 2, ..., T$, where $v_t = \sum_{j=0}^{\infty} \gamma_j e_{t-j}$, $\gamma(L) = \sum_{j=0}^{\infty} \gamma_j L^j$, that is, v_t is an unobserved stationary zero-mean process, where $\sum_{j=0}^{\infty} j|\gamma_j| < \infty$ and e_t is a martingale difference sequence. We assume that $u_0 = 0$ throughout, although the results generally hold for the weaker requirement that $E(u_0^2) < \infty$ (even as $T \to \infty$). The process e_t has a non-normalized spectral density at frequency zero given by $\sigma^2 = \sigma_\epsilon^2 \gamma(1)^2$, where $\sigma^2 = \lim_{T\to\infty} T^{-1} \sum_{i=0}^{\infty} E(e_t^2)$.

In the first equation of (1), $d_t = \psi' z_t$, where z_t is a set of deterministic components. Perron and Rodríguez (2003) consider two models in the context of an unknown structural break: (i) Model I, where there is a single structural change in the slope, that is, $z_t = \{1, t, \mathbf{1}(t > T_B)(t - T_B)\}$ where $\mathbf{1}(.)$ is the indicator function and T_B is the time of change and can be expressed as a fraction of the whole sample as $T_B = \delta T$ for some $\delta \in (0, 1)$; and (ii) Model II, which includes a single structural change in intercept and slope, that is, $z_t = \{1, \mathbf{1}(t > T_B), t, \mathbf{1}(t > T_B)(t - T_B)\}$.[5]

2.2. GLS Detrending and M^{GLS} Statistics

The class of M^{OLS} tests are due to Stock (1999) and further analyzed by Perron and Ng (1996). These tests are shown to have far less size distortions in the presence of important negative serial correlation. The M^{GLS} tests are constructed using $\widetilde{y}_t = y_t - \widehat{\psi}^{GLS\prime} z_t$, where $\widehat{\psi}_{GLS} = (z_t^{\bar{\alpha}\prime} z_t^{\bar{\alpha}})^{-1}(z_t^{\bar{\alpha}\prime} y_t^{\bar{\alpha}})$, with $y_t^{\bar{\alpha}} \equiv [y_1, (1 - \bar{\alpha}L)y_t]$, and $z_t^{\bar{\alpha}} \equiv [z_1, (1 - \bar{\alpha}L)z_t]$, for $t = 2, 3, 4....., T$, and for a chosen $\bar{\alpha} = 1 + \bar{c}/T$ and where z_t has been defined in Section 2.1. We also use the P_T^{GLS} test, as defined in Perron and Rodríguez (2003). Hence, defining $S(\rho, \delta) = \sum_{t=1}^{T} (y_t^{\rho} - \widehat{\psi}^{GLS\prime} z_t^{\rho})^2$ for $\rho = \bar{\alpha}, 1$, the M^{GLS} and the P_T^{GLS} are:

$$
\begin{aligned}
MZ_{\hat{\alpha}}^{GLS}(\delta) &= \frac{T^{-1}\widetilde{y}_T^2 - s^2}{2T^{-2}\sum_{t=1}^{T} \widetilde{y}_{t-1}^2},\ MSB^{GLS}(\delta) = \left[\frac{T^{-2}\sum_{t=1}^{T} \widetilde{y}_t^2}{s^2}\right]^{1/2}, \\
MZ_{t_{\hat{\alpha}}}^{GLS}(\delta) &= \frac{T^{-1}\widetilde{y}_T^2 - s^2}{[4s^2 T^{-2}\sum_{t=1}^{T} \widetilde{y}_{t-1}^2]^{1/2}},\ MP_{T,\mu}^{GLS}(\delta) = \frac{\bar{c}^2 T^{-2}\sum_{t=1}^{T} \widetilde{y}_{t-1}^2 - \bar{c}T^{-1}\widetilde{y}_T^2}{s^2}, \\
MP_{T,\tau}^{GLS}(\delta) &= \frac{\bar{c}^2 T^{-2}\sum_{t=1}^{T} \widetilde{y}_{t-1}^2 + (1 - \bar{c})T^{-1}\widetilde{y}_T^2}{s^2},\ P_T^{GLS}(\delta) = \frac{S(\bar{\alpha}, \delta) - \bar{\alpha}S(1, \delta)}{s^2}.
\end{aligned}
$$

Following Perron and Rodríguez (2003), we use $\bar{c} = -22.5$.[6] The statistics are modified versions of the $Z_{\hat{\alpha}}$ test of Phillips and Perron (1988), Bhargava (1986)'s R_1 statistic, and the $Z_{t_{\hat{\alpha}}}$ test proposed by Phillips and Perron (1988), respectively. The term s^2 is an autoregressive estimate of (2π times) the spectral density at frequency zero of u_t, suggested by Perron and Ng (1998), and defined by $s^2 = s_{ek}^2/[1 - \hat{b}(1)]^2$, where $s_{ek}^2 = (T - k_{\max})^{-1} \sum_{t=k+1}^{T} \widehat{e}_{tk}^2$, $\hat{b}(1) = \sum_{j=1}^{k} \hat{b}_j$, with $\hat{b}_j$ and $\{\widehat{e}_{tk}\}$ obtained from the autoregression:

$$\Delta\widetilde{y}_t = \alpha_0 \widetilde{y}_{t-1} + \sum_{j=1}^{k} b_j \Delta\widetilde{y}_{t-j} + e_{tk}. \tag{2}$$

Another test is the so-called $ADF^{GLS}(\delta)$ test, which is the t-statistic for testing the null hypothesis that $\alpha_0 = 0$ in (2).

2.3. Rules for Selecting the Lag Length (k)

In the derivation of the asymptotic distributions of the different unit root tests, the theoretical conditions provide little practical guidance for choosing k. The literature suggests to use data-dependent rules like the AIC and the BIC where k is chosen by minimizing: $IC_k = \ln \hat{\sigma}_k^2 + \frac{kC_T}{T - k_{\max}}$, where $\hat{\sigma}_k^2 = (T - k_{\max})^{-1} \sum_{t=k+1}^{T} \widehat{e}_{tk}^2$, $\frac{C_T}{T - k_{\max}}$ is the penalty attached to an additional regressor, and

---

5 See Rodríguez (2007) for the *crash* model proposed by Perron (1989).

6 Following Elliott et al. (1996) and Ng and Perron (2001), the parameter $\bar{c}$ is selected in such a way that 50% of the Gaussian power envelope is attained.

$T - k_{\max}$ is the number of observations effectively available.[7] The AIC and the BIC are obtained when $C_T = 2$ and $C_T = \ln(T - k_{\max})$, respectively. Another procedure is the sequential t-sig procedure described in Campbell and Perron (1991).[8] Selecting a value for $k_{\max}$, the lag k is selected in a general to specific recursive procedure based on a two-tailed t-statistic on the coefficient associated with the last lag in (2). This approach is denoted by t-sig(10). In a more recent contribution, Ng and Perron (2001) proposed a class of Modified Information Criteria (MIC) that selects k satisfying: arg min $MIC_k = \ln \hat{\sigma}_k^2 + \frac{C_T[\hat{\tau}_T(k)+k]}{T-k_{\max}}$, with $\hat{\tau}_T(k) = (\hat{\sigma}_k^2)^{-1}\hat{\alpha}_0^2 \sum_{t=k_{max+1}}^{T} \tilde{y}_{t-1}^2$. The modified Akaike ($MAIC$) is obtained when $C_T = 2$, and the modified BIC ($MBIC$) is obtained when $C_T = \ln(T - k_{\max})$. Recently, in order to improve finite (size and power) sample performance, Perron and Qu (2007) have proposed a hybrid approach consisting of two steps: (i) OLS detrended data are used to select k using AIC, BIC, $MAIC$ or $MBIC$; and (ii) estimating (2) using GLS detrended data to construct s^2. In the simulations, we consider this hybrid approach and the methods used are classified as AIC^{OLS}, BIC^{OLS}, $MAIC^{OLS}$ and $MBIC^{OLS}$, respectively.

2.4. Selecting the Breakpoint

Given that the break date (δ) is considered to be unknown, we follow Perron and Rodríguez (2003) using two methods for selecting the break date. The first is to define the break date as the point that minimizes the statistic $t_{\hat{\alpha}_0}$ in (2). This procedure is known as the Infimum method; see Zivot and Andrews (1992) and Perron and Rodríguez (2003) for further details. The second method is based on the maximum absolute value of the t-statistic associated with the dummy variable of the break in the slope. This procedure is known as the Supremum method, which is equivalent to minimizing the SSR; see Perron (1997) and Perron and Rodríguez (2003) for further details.

3. Finite Sample Simulations

3.1. Setup

The DGP is $y_t = \alpha y_{t-1} + u_t$ with three scenarios for the autocorrelation of u_t: (i) the *i.i.d.* case: $u_t = e_t$; (ii) the AR(1) case: $u_t = \phi u_{t-1} + e_t$; and (iii) the MA(1) case: $u_t = e_t + \theta e_{t-1}$. For all cases, $e_t \sim$ *i.i.d.* $N(0,1)$, 1000 replications, $T = 100$ and 250, $\phi = -0.8, -0.4, 0.4, 0.8$ and $\theta = -0.8, -0.5, 0.3, 0.8$ and $\alpha = 1$ (null hypothesis). We performed extensive simulations for all M^{GLS} tests, using both models and both ways to select the break point. We present a selected set of results. We have selected the $MZ_{\hat{\alpha}}^{GLS}$ test as the representative test for the entire family of the M^{GLS} tests. Furthermore, the Infimum method is used to select the break date and results are only reported for Model I. All other results or Tables are available upon request.[9]

3.2. The Problem of Size

Table 1 shows the size of the $MZ_{\hat{\alpha}}^{GLS}$ test for $T = 100$ and for the different criteria for selecting k. The $k_{max} = int[10 \times (\frac{T}{100})^{1/4}]$, that is, $k_{max} = 10$. For the *i.i.d.* case, the results indicate that the test constructed using BIC and BIC^{OLS} have a size around 3.0%, suggesting an undersized test. Testing based on all $MAIC$ (OLS and GLS versions) seems to be extremely conservative (with an exact size of 0.0%). On the other side, testing constructed with AIC, AIC^{OLS} and the t-sig(10) present values implying an extremely oversized test (22%, 27% and 63%, respectively). This same result

7 Note that in all experiments we use $T - k_{\max}$ as the available number of observations, which is fixed, as suggested by Ng and Perron (2005).

8 See also Hall (1994) and Ng and Perron (1995).

9 We are agree with the Editor that our scenario is the worst possible scenario because we are using the Infimum method jointly (in some cases) with the t-sig(10) rule. However, this worst scenario is widely used in typical empirical applications. Furthermore, it is a regular or natural option in many statistical packages used by practitioners. Minimizing SSR (or Supremum) is better, as we mention later.

appears when we use some fixed values of k ($k = 5, 6, ..., 10$), where sizes go from 43% to 82%. Indeed, the size is greater when the selected k is higher. For the $AR(1)$ case, very similar results are found. In the $MA(1)$ case, we observe the standard result that the test is oversized. In fact, when $\theta = -0.80$, all selection criteria yield an oversized test. Even when using $MAIC$ and $MBIC$, the sizes are 23% and 24%, respectively.

Table 1. Size of the $MZ_{\hat{\alpha}}^{GLS}$ Test, Model I, $T = 100$.

	i.i.d.	AR(1) Case				MA(1) Case			
		$\phi = -0.8$	$\phi = -0.4$	$\phi = 0.4$	$\phi = 0.8$	$\theta = -0.8$	$\theta = -0.5$	$\theta = 0.3$	$\theta = 0.8$
AIC	0.22	0.17	0.23	0.30	0.49	0.73	0.30	0.30	0.66
BIC	0.03	0.01	0.07	0.09	0.19	0.90	0.41	0.09	0.40
$MAIC$	0.00	0.00	0.00	0.01	0.10	0.23	0.04	0.01	0.06
$MBIC$	0.00	0.00	0.00	0.00	0.10	0.24	0.04	0.00	0.00
AIC^{OLS}	0.27	0.21	0.27	0.33	0.53	0.80	0.36	0.33	0.70
BIC^{OLS}	0.03	0.01	0.09	0.09	0.20	0.93	0.45	0.10	0.42
$MAIC^{OLS}$	0.00	0.00	0.00	0.01	0.11	0.32	0.04	0.00	0.05
$MBIC^{OLS}$	0.00	0.00	0.00	0.00	0.09	0.33	0.04	0.00	0.00
$t - sig(10)$	0.63	0.54	0.60	0.67	0.82	0.64	0.57	0.66	0.76
$k = 5$	0.43	0.23	0.36	0.46	0.64	0.42	0.30	0.45	0.58
$k = 6$	0.53	0.36	0.48	0.57	0.71	0.44	0.40	0.55	0.46
$k = 7$	0.64	0.48	0.60	0.67	0.79	0.49	0.51	0.65	0.71
$k = 8$	0.70	0.57	0.65	0.73	0.84	0.51	0.58	0.71	0.68
$k = 9$	0.77	0.66	0.75	0.80	0.89	0.55	0.67	0.79	0.81
$k = 10$	0.82	0.73	0.81	0.82	0.91	0.62	0.73	0.83	0.80

In Table 2, the results are presented for $T = 250$, where $k_{max} = 13$. The values of the distortions decrease, meaning that the explosiveness (oversizing) problem decreases. For the *i.i.d.* case, the tests constructed with BIC and BIC^{OLS} yield 2.6% and 2.7%, respectively which are very similar when $T = 100$. With MIC and MIC^{OLS}, the test has sizes of 1.7% and 1.6%, respectively which are better than for $T = 100$, but are still very undersized. Tests using the AIC, AIC^{OLS} and t-sig(10) have sizes of 9%, 11.2%, and 37.9%, respectively, which are smaller than the values for $T = 100$, but they still indicate an oversized test, in particular the t-sig(10) criterion. With a fixed k ($k = 5, 6, ..., 13$), sizes are greater when k is higher, although smaller compared with $T = 100$.

Table 2. Size of the $MZ_{\hat{\alpha}}^{GLS}$ Test, Model I, $T = 250$.

	i.i.d.	AR(1) Case				MA(1) Case			
		$\phi = -0.8$	$\phi = -0.4$	$\phi = 0.4$	$\phi = 0.8$	$\theta = -0.8$	$\theta = -0.5$	$\theta = 0.3$	$\theta = 0.8$
AIC	0.091	0.054	0.102	0.129	0.197	0.345	0.154	0.155	0.373
BIC	0.026	0.002	0.024	0.053	0.077	0.697	0.246	0.087	0.203
$MAIC$	0.017	0.000	0.008	0.033	0.056	0.038	0.025	0.040	0.089
$MBIC$	0.015	0.000	0.012	0.008	0.058	0.046	0.029	0.003	0.007
AIC^{OLS}	0.112	0.074	0.124	0.155	0.217	0.445	0.173	0.173	0.390
BIC^{OLS}	0.027	0.001	0.026	0.053	0.078	0.802	0.259	0.088	0.211
$MAIC^{OLS}$	0.016	0.000	0.010	0.035	0.057	0.057	0.028	0.041	0.077
$MBIC^{OLS}$	0.015	0.000	0.013	0.006	0.060	0.059	0.029	0.001	0.005
$t - sig(10)$	0.379	0.266	0.345	0.395	0.514	0.233	0.312	0.392	0.467
$k = 5$	0.161	0.051	0.122	0.188	0.229	0.229	0.075	0.173	0.259
$k = 6$	0.204	0.081	0.173	0.223	0.261	0.143	0.110	0.202	0.136
$k = 7$	0.244	0.126	0.216	0.251	0.320	0.152	0.159	0.244	0.315
$k = 8$	0.283	0.168	0.265	0.293	0.359	0.152	0.202	0.278	0.221
$k = 9$	0.304	0.204	0.284	0.333	0.421	0.140	0.219	0.317	0.371
$k = 10$	0.357	0.243	0.329	0.381	0.461	0.162	0.258	0.363	0.335
$k = 11$	0.407	0.287	0.378	0.424	0.507	0.157	0.302	0.419	0.459
$k = 12$	0.459	0.343	0.431	0.479	0.540	0.173	0.354	0.452	0.447
$k = 13$	0.496	0.406	0.462	0.533	0.602	0.191	0.416	0.517	0.537

If we increase k_{max}, the size of the test for higher k values increases considerably. We may emphasize this issue comparing with the same class of test, but without a structural change, that is, with some of the results obtained by Ng and Perron (2001). If we observe Table II.B of Ng and Perron (2001), the $MZ_{\widehat{\alpha}}^{GLS}$ for $\theta = -0.80$ using $k = 10$ yields a size of 18% with $T = 100$. In our case, for the same values, we have a size of 62%. With $T = 250$, Ng and Perron (2001) obtain 3.6%, a size close to the nominal size (5%). However, in our case, for this sample we have a size of 19% (Table 2, $k = 13$). In fact, our simulations suggest that we need $T = 350$ in order to obtain a size close to 5% when $\theta = -0.80$. The results are surely due to the higher number of deterministic components in our models compared with Ng and Perron (2001). However, our conclusion is that practitioners interested in applying the $MZ_{\widehat{\alpha}}^{GLS}$ need a non-trivial number of observations.

A further comparison with Ng and Perron (2001) is possible if we select k using different criteria. Again, in the $MA(1)$ case, where $\theta = -0.80$ and $T = 100$, the test constructed with $MAIC$ and $MBIC$ yields sizes of 23% and 24%, respectively. The OLS versions of these criteria yield 32% and 33%, respectively (see Table 1). However, in the case shown in Table VI.A of Ng and Perron (2001), sizes of 5.9% are obtained using MIC and 12.3% using MIC^{OLS} ($T = 100$). In Table 2, for $T = 250$, the tests constructed with $MAIC$ and $MBIC$ yield sizes of 3.8% and 4.6%, respectively. In the case of Ng and Perron (2001), MIC and MIC^{OLS} yield 1.2% and 1.6%, respectively.

3.3. Some Additional Results[10]

Two values are used in the construction of s^2: s_{ek}^2 and $\widehat{b}(1)$. Available simulations show that the reason why $s^2 \to \infty$ is $\widehat{b}(1) \to 1$. That is, when a higher k is selected, it is possible to incur in overparameterization in (2) and $\widehat{b}(1) \to 1$. If s^2 tends to $+\infty$, then the $MZ_{\widehat{\alpha}}^{GLS}$ and $MZ_{t_{\widehat{\alpha}}}^{GLS}$ statistics tend to $-\infty$ and MSB^{GLS} and P_T^{GLS} converge in probability to zero.

Additional simulations show a link between the excessive size of the test and a high probability of selecting higher values of k. Following Ng and Perron (1995), we examine the number of times that $k = i$ is selected by each rule for $i = 0, 1, 2, ..., 10$ and $T = 100$. In the *i.i.d.* case, the results show that AIC, BIC, $MAIC$ and $MBIC$ have probabilities to select $k = 1$ of 56.2%, 93.2%, 74.4%, and 81.6%, respectively. The t-sig(10) criterion has probabilities of selecting lag lengths that are equally distributed for all values of k. For instance, the recursive t-sig(10) has a probability of around 53% of selecting $k \geq 7$. Until now, a basic conclusion is that the AIC, AIC^{OLS}, and t-sig(10) methods are not recommended, as they have high probabilities of selecting higher values of k, which are associated with the size distortions observed in Tables 1 and 2.

When we calculate the mean value for $MZ_{\widehat{\alpha}}^{GLS}$ (in the *i.i.d.* case), explosive negative values are obtained for $k \geq 5$ in AIC, AIC^{OLS}, BIC^{OLS} and t-sig(10). In contrast, reduced values of the test (in absolute value) are given by $MAIC$, $MBIC$, $MAIC^{OLS}$ and $MBIC^{OLS}$. We also examine the number of times that the $MZ_{\widehat{\alpha}}^{GLS}$ test is smaller than a threshold. We consider six possible values: $-500, -1000, -5000, -10,000, -50,000, -100,000$, and the *i.i.d.* case. For all thresholds considered, we find that the number of explosive values of $MZ_{\widehat{\alpha}}^{GLS}$ increases as the value of k is larger. For example, for $k = 7$, the probability of getting a value of $MZ_{\widehat{\alpha}}^{GLS} \prec -1000$ is 13.4%; and the probabilities for $k = 9$ and $k = 10$ are 31% and 40.2%, respectively. Furthermore, the probabilities of finding values of $MZ_{\widehat{\alpha}}^{GLS} \prec -100,000$ are 18% and 22.7% for $k = 9$ and $k = 10$, respectively.

All previous results are less severe when $T = 250$. Among other things, the probabilities of finding elevated k values are lower. In this regard, the oversizing problem is attenuated (see Table 2). Moreover, when a break is included in the simulations, the improvement is greater when $T = 250$. However, explosive negative values are still observed when the lag is selected with AIC, AIC^{OLS}, and t-sig (10).

---

[10] We present a summary of the Tables from the Working Paper version of this Note (see Quineche and Rodríguez (2015)). All other tables are available upon request.

3.4. The ADF^{GLS} Statistic

While the $MZ_{\widehat{\alpha}}^{GLS}$ test (and the entire family of the M^{GLS} tests) shows either oversizing or undersizing problems, depending on the criteria used to choose k, the ADF^{GLS} statistic works well. In the available Tables, we find that the mean value for ADF^{GLS} is not explosive irrespective of the selection criterion used. There are some slightly large negative values when $\theta = -0.8$, but it is a standard result in the literature.

Table 3 shows the exact size of the ADF^{GLS} statistic when $T = 100$. For the *i.i.d.* case, the tests constructed with $MAIC$ and $MAIC^{OLS}$ yield sizes of 3.1% and 3.4%, respectively; that is, they are slightly undersized, but closer to 5%. A similar observation is valid for $MBIC$ and $MBIC^{OLS}$. Other information criteria, like AIC, AIC^{OLS} and t-sig(10), generate oversized tests; but the values are much smaller compared with Table 1 for the $MZ_{\widehat{\alpha}}^{GLS}$ test. For example, for the t-sig (10) procedure, Table 1 (*i.i.d.* case) shows that the statistic $MZ_{\widehat{\alpha}}^{GLS}$ has a size of 63%, which is poor. However, this value is reduced to 14.6% in the case of the ADF^{GLS} test (Table 3). In general, the values in all scenarios are smaller compared with Table 1 for $MZ_{\widehat{\alpha}}^{GLS}$. The only difference (as expected) arises when $\theta = -0.80$. In this case, the $MZ_{\widehat{\alpha}}^{GLS}$ test has sizes of 23% and 24% for the $MAIC$ and $MBIC$, respectively, while for the ADF^{GLS} test the values are 31.5% and 32.6%, respectively.

Table 3. Size (5%) of ADF Test, Model I, $T = 100$.

	i.i.d.	**AR(1) Case**				**MA(1) Case**			
		$\phi = -0.8$	$\phi = -0.4$	$\phi = 0.4$	$\phi = 0.8$	$\theta = -0.8$	$\theta = -0.5$	$\theta = 0.3$	$\theta = 0.8$
AIC	0.136	0.117	0.149	0.128	0.167	0.826	0.362	0.145	0.147
BIC	0.072	0.069	0.173	0.069	0.089	0.976	0.568	0.095	0.151
$MAIC$	0.031	0.024	0.039	0.008	0.042	0.315	0.106	0.005	0.004
$MBIC$	0.034	0.025	0.040	0.000	0.034	0.326	0.109	0.004	0.000
AIC^{OLS}	0.145	0.130	0.163	0.132	0.177	0.881	0.402	0.152	0.152
BIC^{OLS}	0.076	0.070	0.196	0.070	0.092	0.985	0.633	0.097	0.155
$MAIC^{OLS}$	0.033	0.030	0.042	0.008	0.042	0.435	0.123	0.006	0.003
$MBIC^{OLS}$	0.038	0.030	0.043	0.000	0.031	0.444	0.127	0.004	0.000
$t - sig(10)$	0.146	0.129	0.154	0.141	0.193	0.516	0.256	0.148	0.136
$k = 5$	0.058	0.056	0.052	0.061	0.058	0.243	0.063	0.056	0.094
$k = 6$	0.054	0.052	0.053	0.054	0.068	0.148	0.054	0.055	0.029
$k = 7$	0.053	0.055	0.061	0.047	0.059	0.122	0.057	0.052	0.060
$k = 8$	0.041	0.037	0.036	0.039	0.063	0.097	0.039	0.037	0.035
$k = 9$	0.049	0.031	0.037	0.049	0.071	0.075	0.034	0.045	0.060
$k = 10$	0.047	0.038	0.047	0.054	0.069	0.059	0.043	0.050	0.036

Table 4 shows the exact size of the ADF^{GLS} test when $T = 250$. Again, the size distortions are clearly smaller compared to those of the $MZ_{\widehat{\alpha}}^{GLS}$ test (Table 2). As in Table 3, the results using the $MZ_{\widehat{\alpha}}^{GLS}$ test are better when $\theta = -0.80$. In Table 4, the ADF^{GLS} test yields 11.5% and 12.8% when $MAIC$ and $MBIC$ are used, respectively. In the case of the $MZ_{\widehat{\alpha}}^{GLS}$ test, the values are 3.8% and 4.6%, respectively. Furthermore, our calculations show that the ADF^{GLS} test will have a size closer to 5% for $\theta = -0.80$ when $T = 350$. This sample size is even more prohibitive for most empirical applications.

A comparison of Tables 1 and 2 against Tables 3 and 4 suggests that it is recommendable to use the ADF^{GLS} test, except when practitioners are sure that they face a strong $MA(1)$ negative correlation. In this case, practitioners should use $T = 350$ or $T = 250$ for ADF^{GLS} or $MZ_{\widehat{\alpha}}^{GLS}$, respectively.

Table 4. Size (5%) of ADF Test, Model I, $T = 250$.

	i.i.d.	**AR(1) Case**				**MA(1) Case**			
		$\phi = -0.8$	$\phi = -0.4$	$\phi = 0.4$	$\phi = 0.8$	$\theta = -0.8$	$\theta = -0.5$	$\theta = 0.3$	$\theta = 0.8$
AIC	0.088	0.078	0.075	0.089	0.098	0.527	0.190	0.102	0.102
BIC	0.054	0.049	0.052	0.061	0.059	0.831	0.351	0.092	0.095
$MAIC$	0.032	0.028	0.027	0.033	0.041	0.115	0.054	0.036	0.011
$MBIC$	0.036	0.030	0.031	0.007	0.044	0.128	0.062	0.003	0.004
AIC^{OLS}	0.097	0.085	0.084	0.096	0.107	0.612	0.205	0.106	0.111
BIC^{OLS}	0.054	0.050	0.053	0.061	0.059	0.900	0.373	0.093	0.099
$MAIC^{OLS}$	0.034	0.029	0.029	0.035	0.041	0.156	0.064	0.037	0.011
$MBIC^{OLS}$	0.039	0.033	0.033	0.006	0.047	0.162	0.069	0.003	0.004
$t - sig(10)$	0.092	0.091	0.095	0.100	0.107	0.309	0.153	0.104	0.104
$k = 5$	0.054	0.054	0.052	0.059	0.068	0.501	0.059	0.061	0.099
$k = 6$	0.052	0.054	0.055	0.052	0.058	0.361	0.050	0.057	0.036
$k = 7$	0.058	0.052	0.055	0.056	0.056	0.275	0.056	0.055	0.083
$k = 8$	0.059	0.056	0.055	0.058	0.054	0.214	0.063	0.057	0.042
$k = 9$	0.056	0.055	0.054	0.053	0.058	0.171	0.051	0.049	0.064
$k = 10$	0.046	0.056	0.052	0.046	0.066	0.136	0.053	0.048	0.036
$k = 11$	0.044	0.050	0.050	0.050	0.061	0.117	0.049	0.043	0.061
$k = 12$	0.044	0.047	0.052	0.058	0.064	0.117	0.049	0.043	0.061
$k = 13$	0.053	0.053	0.050	0.056	0.054	0.089	0.055	0.054	0.052

3.5. The Supremum Method and a Single Breakpoint

The results change favorably when the Supremum method is used to select the breakpoint. Several simulations have been performed under the setup of Section 3.1 for Model I: $z_t = \beta_1 + \beta_2 t + \beta_3 \mathbf{1}(t > T_B)(t - T_B)$ with two scenarios: (i) $\beta_3 = 0$, that is, no break; and (ii) $\beta_3 = 0.5, 1.0, 1.5$ with $\delta = 0.50 \times T$. Similar experiments have been performed for Model II. In the first case, the $MZ_{\widehat{\alpha}}^{GLS}$ test still has explosive values, although less frequently; and the values are negative but of a smaller magnitude (in absolute value) than when using the Infimum method. In the second case, the results show considerable improvement, especially when $T = 250$. The explosive values of the $MZ_{\widehat{\alpha}}^{GLS}$ test practically disappear for the MIC and MIC^{OLS} rules, although the cost is to have small values (in absolute value), which produce a conservative test. On the other hand, the rules AIC, AIC^{OLS}, and t-sig(10) continue to present an $MZ_{\widehat{\alpha}}^{GLS}$ test with explosive values which, however, are very small compared to the previous cases, and occur only when a higher k is selected.

The best results with the Supremum method are important, since this method is recommended in the literature to select the break date. For instance, Vogelsang and Perron (1998) argue that this method is to be preferred, since it allows a consistent estimate of the breaking point, a matter that the Infimum method cannot do.

The evidence suggests that, in the empirical applications, the Supremum method should be used to select the breakpoint along with the MIC and MIC^{OLS} rules, although the potential cost is to have a conservative test. The evidence suggests avoiding the use of rules such as AIC, AIC^{OLS}, and t-sig(10) to select k, as well as the use of the Infimum method to select the breakpoint.

4. Conclusions

This note aims to examine the performance of the size of the M^{GLS} statistics to test for the presence of a unit root using different lag length selection criteria in the context of an unknown structural change. In particular, we have focused on the size performance of the $MZ_{\widehat{\alpha}}^{GLS}$ test. Overall, the results show that there is a strong relationship between the explosive negative values of the $MZ_{\widehat{\alpha}}^{GLS}$ test and the values of the selected k. Using the Infimum method to select the break point jointly with some rule, such as AIC, AIC^{OLS} or t-sig(10), produces the worst scenario, in the sense that the test yields explosive negative values, which generates severe oversizing problems. On the opposite side, using other criteria for k implies conservative tests. These issues seem to improve when $T = 250$ (relative to $T = 100$) or more, which creates sample size difficulties for most macroeconomic applications, especially in Latin American countries.

The results indicate that ADF^{GLS} should be used, because it does does not result in explosiveness. Although for other reasons, this recommendation is in the same vein as Harvey et al. (2013). The advantage of the $MZ_{\widehat{\alpha}}^{GLS}$ test is that it is intrinsically conservative. So, if we obtain a good size when $\theta = -0.80$, this is achieved at the cost of having an undersized test in the other cases, including the *i.i.d.* case. Our results are in line with those obtained in Del Barrio Castro et al. (2011), Del Barrio Castro et al. (2013), and Del Barrio Castro et al. (2015)[11].

The results change for the better when using the Supremum method (minimizing the SSR) to select the breakpoint. However, this result only occurs when there is a break in the series. With this method, the test values are reduced (in absolute value) and no explosiveness is observed. Furthermore, the advantage is that the method offers a consistent breakpoint estimator which is currently suggested in the literature. Although a possible undersizing problem is addressed, then a possible best scenario is to use the Supremum method together with rules for selecting k such as *MIC*. This potential need to perform a pre-testing to see the existence of a break is similar to what is proposed by Kim and Perron (2009) when there is only one break and the proposal of Carrión-i-Silvestre et al. (2009) when there are multiple breaks.

Author Contributions: Both authors contributed eaully to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

Agiakloglou, Christos, and Paul Newbold. 1992. Empirical Evidence on Dickey-Fuller-Type Tests. *Journal of Time Series Analysis* 13: 471–83.

Agiakloglou, Christos, and Paul Newbold. 1996. The balance between size and power in Dickey-Fuller tests with data-dependent rules for the choice of truncation lag. *Economics Letters* 52: 229–34.

Bhargava, Alok. 1986. On the theory of testing for unit root in observed time series. *Review of Economic Studies* 53: 369–84.

Campbell, John Y., and Pierre Perron. 1991. Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots. In *NBER Macroeconomics Annual*. Edited by Olivier J. Blanchard and Stanley Fischer. New York: MIT Press, Volume 6, pp. 141–201.

Carrión-i-Silvestre, Josep Lluís, Dukpa Kim, and Pierre Perron. 2009. GLS-based Unit Root Tests with Multiple Structural Breaks both Under the Null and Alternative Hypotheses. *Econometric Theory* 25: 1754–92.

Choi, In. 2015. *Almost All about Unit Roots*. Series: Themes in Modern Econometrics. New York: Cambridge University Press.

Del Barrio Castro, Tomás, Paulo M. M. Rodrigues, and A. M. Robert Taylor. 2011. The Impact of Persistent Cycles on Zero Frequency Unit Root Tests. Working Paper 24. Banco de Portugal, Lisbon, Portugal.

Del Barrio Castro, Tomás, Paulo M. M. Rodrigues, and A. M. Robert Taylor. 2013. The Impact of Persistent Cycles on Zero Frequency Unit Root Tests. *Econometric Theory* 29: 1289–313.

Del Barrio Castro, Tomás, Paulo M. M. Rodrigues, and A. M. Robert Taylor. 2015. On the Behaviour of Phillips-Perron Tests in the Presence of Persistent Cycles. *Oxford Bulletin of Economics and Statistics* 77: 495–511.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock. 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64: 813–39.

Fossati, Sebastian. 2012. Covariate unit root tests with good size and power. *Computational Statistics and Data Analysis* 56: 3070–79.

[11] They explain both issues (in particular the undersizing feature) in the context of time series admitting for (near-) unit roots at cyclical frequencies. They suggest that the degree of undersizing is worst when $MAIC$ is used. The problem is aggravated if GLS detrended data are used. See these references for further details.

Haldrup, Niels, and Michael Jansson. 2006. Improving Size and Power in Unit Root Testing. In *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*. Edited by Terence C. Mills, and Kerry Patterson. Basingstoke: Palgrave Macmillan, pp. 252–77.

Hall, Alastair R. 1994. Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection. *Journal of Business & Economic Statistics* 12: 461–70.

Harvey, David I., Stephen J. Leybourne, and A. M. Robert Taylor. 2013. Testing for unit roots in the possible presence of multiple trend breaks using minimum Dickey-Fuller statistics. *Journal of Econometrics* 177: 265–84.

Kim, Dukpa, and Pierre Perron. 2009. Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses. *Journal of Econometrics* 148: 1–13.

Maddala, Gangadharrao S., and In-Moo Kim. 1998. *Unit Roots Cointegration, and Structural Change*. Cambridge: Cambridge University Press.

Ng, Serena, and Pierre Perron. 1995. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90: 268–81.

Ng, Serena, and Pierre Perron. 2001. Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69: 1519–54.

Ng, Serena, and Pierre Perron. 2005. A note on the selection of time series models. *Oxford Bulletin of Economics and Statistics* 67: 115–34.

Perron, Pierre. 1989. The great crash, the oil price shock and the unit root hypothesis. *Econometrica* 57: 1361–401.

Perron, Pierre. 1997. Further evidence of breaking trend functions in macroeconomic variables. *Journal of Econometrics* 80: 355–85.

Perron, P. 2006. Dealing with Structural Breaks. In *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*. Edited by Terence C. Mills, and Kerry Patterson. Basingstoke: Palgrave Macmillan, pp. 278–352.

Perron, Pierre, and Serena Ng. 1996. Useful Modifications to Unit Root Tests with Dependent Errors and their Local Asymptotic Properties. *Review of Economic Studies* 63: 435–65.

Perron, Pierre, and Serena Ng. 1998. An Autoregressive Spectral Density Estimator at Frequency Zero for Nonstationarity Tests. *Econometric Theory* 14: 560–603.

Perron, Pierre, and Zhongjun Qu. 2007. A simple modification to improve the infinite sample properties of Ng and Perron's unit root tests. *Economics Letters* 94: 12–19.

Perron, Pierre, and Gabriel Rodríguez. 2003. GLS Detrending, Efficient Unit Root Tests and Structural Change. *Journal of Econometrics* 115: 1–27.

Phillips, Peter C. B., and Pierre Perron. 1988. Testing for a unit root in time series regression. *Biometrika* **75**, 335–46.

Phillips, Peter C. B., and Zhijie Xiao. 1998. A primer on unit root testing. *Journal of Economic Surveys* 12: 423–70.

Quineche, Ricardo, and Gabriel Rodríguez. 2015. Data-Dependent Methods for the Lag Length Selection in Unit Root Tests with Structural Change. Working Paper 404, Department of Economics, Pontificia Universidad Católica del Perú, Lima, Peru.

Rodríguez, Gabriel. 2007. Finite Sample Behaviour of the Level Shift Model using Quasi-Differenced Data. *Journal of Statistical Computation and Simulation* 77: 889–905.

Schwert, G. William. 1989. Tests for unit roots: A Monte Carlo Investigation. *Journal of Business and Economics Statistics* 7: 147–59.

Stock, James H. 1994. Unit Roots, Structural Breaks and Trends. In *Handbook of Econometrics*. Edited by Daniel L. McFaden and Robert F. Engle. Amsterdam: Elsevier, vol. IV, pp. 2740–841.

Stock, James H. 1999. A Class of Tests for Integration and Cointegration. In *Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive W. J. Granger*. Edited by Engle, Robert F. and Halbert White. Oxford: Oxford University Press, pp. 137–67.

Vogelsang, Timothy J., and Pierre Perron. 1998. Additional Tests for a Unit Root Allowing the Possibility of Breaks in the Trend Function at an Unknown Time. *International Economic Review* 39: 1073–100.

Zivot, Eric, and Donald W. K. Andrews. 1992. Further evidence on the great crash, the oil-price shock and the unit root hypothesis. *Journal of Business and Economics Statistics* 10: 251–70.