Size and power of two recent unit root tests that allow for structural breaks

Author: Marcus Nordström

Supervisor: Joakim Westerlund

Seminar date: 31 May 2017

NEKN01, Master Essay I

Department of economics, Lund University

Size and power of two recent unit root tests that allow for structural breaks

Abstract

This paper examines the properties of the two recent structural break unit root tests developed in Harvey, Leybourne and Taylor (2013) and Narayan and Popp (2010). The properties are investigated by Monte Carlo simulations in an environment where two trend breaks of small to large magnitudes are present. We find that the Harvey, Leybourne and Taylor (2013) test has superior size and power properties compared to the Narayan and Popp (2010) test. In addition, we investigate the accuracy of the break detection of the two procedures. The results show that the former test is more accurate than the later test except for when the breaks are very large and the null is true.

Keywords: Unit root test, Structural breaks, Multiple trend breaks, Endogenous breaks, Monte Carlo simulations

Table of contents

1. Introduction

Ever since Dickey and Fuller (1979) presented the Dickey-Fuller unit root test a significant part of the time series econometrics literature has been concerned with unit roots. Over the years there has been a growing body of literature on the subject where the Dickey-Fuller test has been a cornerstone. The augmented Dickey-Fuller test proposed by Said and Dickey (1984) is still often is used as a benchmark. Although, the test is widely used it is, in small samples, known to have low power, see for example Choi (2015), DeJong *et al.* (1992), and Harris (1992), and size distortions, see for example Schwert (1989). Numerous alternative tests with improved properties have been proposed, for example, two popular procedures are developed by Elliott, Rothenberg and Stock (1996) and Phillips and Perron (1988). However, the unit root tests that tend to outperform the augmented Dickey-Fuller test do not do so by much in small to moderate samples Choi (2015), Vougas (2007). Therefore, the small sample problems still persist.

The practitioner may try to overcome the small sample problems by extending the time series. Although, with longer time spans new problems may arise as structural breaks are more likely to have occurred. Perron (1989) showed that the Dickey-Fuller test almost always fails to reject the null for trend stationary series with breaks in level and/or trend. A further complication, see for example Franses and Haldrup (1994), Leybourne, Mills and Newbold (1998) and Psaradakis (2001), is that unit root tests in some cases even spuriously reject the null when breaks are present. Consequently, when breaks are not accurately accounted for unit root tests suffer from both size and power problems.

As a result, much of the unit root literature has been concerned with various methods to account for breaks. Perron (1989) provides a solution by including break variables in the test regression assuming one known break point. He introduces two different methods that allow for the change to occur in different ways. One models the break assuming it occurs instantaneously, this is the additive outlier model, and the other models the break assuming it occurs gradually, this is the innovational outlier model. These two modelling approaches have served as the groundwork of the structural break unit root testing. However, in most applied cases, the researcher does not know whether any breaks are present in the process

or not. Moreover, even if one would be certain that breaks are present their exact timing may be difficult to determine. Therefore, it is essential to be able to endogenously determine breaks within the testing procedure without any priori information about the putative breaks.

Many different approaches to determine possible break points have been developed over the years. One popular approach is to use minimum unit root statistics. In these types of tests the breaks are located such that the test yields the least favourable support of the null hypothesis. Some widely used tests that employ this procedure are, for example, Zivot and Andrews (1992), Lumsdaine and Papell (1997) and Lee and Strazicich (2003). Another common approach is to estimate the break points such that the significance of the level or trend break variables are maximized. In contrast to the minimum statistics tests in this approach the estimated breaks are allocated where they are, in terms of parameter significance, the most likely to have occurred. Some examples of such tests are Christiano (1992), Perron and Vogelsang (1992), Perron (1997) and Vogelsang and Perron (1998).

Vogelsang and Perron (1998) investigate the power and size properties of endogenous unit root tests under various model specifications. They derive asymptotic distributions for both the innovational outlier and the additive outlier models and show that they are invariant to breaks in level but not to breaks in trend. The authors argue, though, that the asymptotic problems may not translate to small sample cases. This is because in most applications the breaks should be small enough to keep the size within reasonable bounds. However, in their simulations they find that size and power vary significantly with the break magnitudes.

Harvey, Leybourne and Newbold (2001) further investigate power and size in the context of Vogelsang and Perron (1998) and the level break case. Their analysis shows that the procedures suffer from size distortions when breaks are present in samples of realistic sizes. They pay special attention to the innovational outlier model and find that the estimated break dates are correct very rarely. Instead, the most frequently estimated break point lies one time point before the true. Deriving upon the results in Kim, Leybourne and Newbold (2000), that exogenous breaks erroneously positioned just before the true break can cause spurious rejections, they suggest that the problematic break detection is a source of the size

distortions. Likewise, Lee and Strazicich (2001) argue that much of the problems in unit root tests without structural breaks may persist in those which allow for breaks when they are estimated with inadequately precision. Using simulated data they find that both the Zivot and Andrews (1992) and the Perron (1997) tests perform poorly in estimating the break dates in general. Furthermore, neither test is invariant to the break magnitudes. Finally, they suggest that erroneous break estimation may be a source of both size and power distortions. Also in this study they consider exogenously determined flawed break points to abstract the consequences of inaccurate break estimation. Note that, although such illustration is interesting in its own right, it is not the same thing as considering endogenously estimated break points. The reason for this is that the endogenous procedure selects the break points systematically based the characteristics of the data. For a given sample and estimated break point the test statistic is the same whether the break point is estimated exogenously or endogenously. However, because of the systematic nature of the break detection this is not the case for averages such as empirical size and power. That is, with the exception from the trivial cases where the endogenous breaks are estimated with 100% accuracy or allocated randomly. However, because complete accuracy is not feasible in practice it is also interesting to investigate size and power for endogenously estimated break points, whether they are accurate or not.

To solve the problem of break estimation inaccuracy in the innovational outlier test Popp (2008) introduces a new test with one endogenous break. Although the method only differs slightly from the previous tests, it enhances the accuracy of the break detection. In the Perron-type tests typically the regression model accounts for the break by including an impulse, level and trend break variable at the time of the presumed break. Instead, Popp (2008) only places an impulse dummy at the time of the break, whereas the level and trend break variables are lagged one time point. In this way, both the level and trend break variables are nested in the parameter corresponding to the impulse dummy. Furthermore, instead of choosing the breaks to maximize the significance of the level or trend break variables he maximizes the significance of the impulse dummy. By simulation he shows that the procedure finds the correct break date as the break magnitude increases. Narayan and Popp (2010), henceforth NP, extend the test to allow for two structural breaks in level and

trend. In a Monte Carlo study Narayan and Popp (2013) show that the NP test has superior small sample size and power properties and estimates the break dates more accurately compared to the Lee and Strazicich (2003) and Lumsdaine and Papell (1997) test. In recent years the NP test has gained popularity in the applied research literature. It is, perhaps, most frequently used in the field of energy economics, see for example Apergis and Payne (2010), Salisu and Mobolaji (2013), Mishra and Smyth (2014) and Tiwari, Shahbaz and Adnan Hye (2013), but is also used in many other fields of applied econometric time series, see for example Chang and Su (2014), Chen and Shen (2015) and García-Cintado, Romero-Ávila and Usabiaga (2015).

Perron and Rodríguez (2003) extend the GLS-detrended unit root test developed by Elliott, Rothenberg and Stock (1996) to allow for one structural break in trend. To estimate the break dates they apply both a minimum Dickey-Fuller t statistic approach and maximize the absolute value of the t statistic of the changing slope. They conclude that, in finite samples, the minimum Dickey-Fuller approach yields higher power. Harvey, Leybourne and Taylor (2013), henceforth HLT, draw extensively on Perron and Rodríguez (2003) and develop a GLS minimum Dickey-Fuller type test that allows for a flexible number of breaks. Being a fairly new contribution to the literature the HLT test has not been applied as frequently as the NP test. However, some prominent applications of the test are Camarero, Carrion-i-Silvestre and Tamarit (2015), Liddle and Messinis (2015) and De Vita and Trachanas (2016) where, in the two later studies, the test is applied together with the NP test. Up to now there are no simulation studies that examine the properties of the two tests under the same data generating process. Since the tests are applied in similar cases it is important to examine them in the same environment to be able to compare their properties. This gap in the literature calls for further research.

Harvey, Leybourne and Taylor (2012) and HLT address the problem of break estimation in unit root tests when break magnitudes are small to moderate in finite samples. They consider local break magnitudes, that is, normalized versions of the magnitudes that account for the time span and the long-run variance. This is a natural way of thinking about the break magnitudes since their size relative to the errors and time span determines to what degree

they contaminate the stochastic process. With this approach they show that many previous unit root tests have significant power valleys for small to moderate sized breaks. Break magnitudes that fall in this region are, for example, magnitudes smaller than one standard deviation of the long run errors in realistic time spans. Moreover, HLT show that their test is robust to breaks of such magnitudes. To demonstrate the relevance of power valleys in applied research they apply their test to commodity prices. Annual data on copper, hides, lead and silver prices for the period 1900-2003 is considered. All the resulting estimated break magnitudes lie within the region of the power valleys emphasizing the importance of tests robust to small to moderate break magnitudes.

NP run simulations to investigate size and power properties of their test. The trend break magnitudes in the case where no level breaks are present are 0, 5 and 10 standard deviations of the long-run errors. Narayan and Popp (2013) only consider the case where trend breaks appear along with level breaks. Since, at least, the break detection should not be invariant to level breaks we cannot assume that these results translate to the case without level break. Furthermore, the trend break magnitudes are studied in steps of one standard deviation of the long-run errors. In both studies, for realistic time horizons and errors, in terms of the local break magnitudes most of these trend breaks are huge. All of which, apart from those of a magnitude of 1 or perhaps 2 standard deviations, would arguably be very rare in most applications. More reasonable break magnitudes between 0 and 1, on the other hand, are left out. Therefore, the tests performance when trend breaks alone are present has not yet been properly investigated. Because one would never fully be aware of the characteristics of the breaks such investigation is relevant to any empirical application, whether one suspects level breaks to be present or not. This calls for an investigation of the NP test under trend breaks alone.

The aim of this paper is to fill the gaps in the literature in that the NP and HLT test have never been examined alongside with each other before and, furthermore, the NP test has not yet been properly examined when small to moderate sized trend breaks are present. This is done by Monte Carlo simulations imposing two breaks in trend. The break magnitudes range from zero to large and have a good coverage of the small and moderate

sized magnitudes in between. Additionally, following Popp (2008), NP and Narayan and Popp (2013) we investigate the tests abilities to estimate the break dates. However, in contrast to these articles, we do not only present the probabilities of estimating the true breaks but the corresponding probabilities for the entire set of possible break dates. In order to make sense of the precision of the break estimation size and power when endogenously estimated breaks are allocated at various time points, correct as well as incorrect, are studied. This approach differs from similar studies such as Lee and Strazicich (2001), Kim, Leybourne and Newbold (2000) and Harvey, Leybourne and Newbold (2001) where exogenously determined breaks are considered. The results are a relevant contribution to both applied and theoretical research. In the applied field it provides guidance for which test to apply under certain conditions. For theoretical research the contribution lies in the implications of the break detection in the two tests.

The paper is organized as follows. Section 2 provides a detailed description of the two testing procedures. In section 3 the Monte Carlo study is setup and the underlying data generating processes are defined. In section 4 the results are presented and discussed. Finally, section 5 consists of concluding remarks and suggestions for further research.

2. Overview of the tests

2.1 The HLT test

As briefly discussed above HLT construct a GLS-based minimum Dickey-Fuller unit root tests with a flexible number of breaks in trend. Ultimately the test is performed on detrended data with possible lags allowed in the stochastic part but not the deterministic part of the process. This assumes that the presumed breaks occur instantaneously, which is in accordance with the additive outlier model. HLT consider a data generating process consisting of one deterministic and one stochastic part as follows:

$$y_t = d_t + u_t \qquad t = 1, \dots, T$$

$$d_t = \alpha + \beta t + \boldsymbol{\gamma}' \boldsymbol{DT_t}(\boldsymbol{\tau_0}) \qquad t = 1, \dots, T$$

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad t = 2, \dots, T$$

where $\boldsymbol{DT_t}(\boldsymbol{\tau_0}) := \left[DT_t(\tau_{0,1}), \dots, DT_t(\tau_{0,m})\right]'$ is a vector of trend breaks with elements defined as $DT_t(\tau) := 1(t > \lfloor \tau T \rfloor)(\tau - \lfloor \tau T \rfloor)$, and the operator $\lfloor \cdot \rfloor$ takes the integer part of its argument. $\boldsymbol{\tau_0} = \left[\tau_{0,1}, \dots, \tau_{0,m}\right]'$ is a vector of unknown supposed break fractions such that $\tau_{0,i} \in [\tau_L, \tau_U]$ for $i = 1, \dots, m$ where τ_L and τ_U represent the lower and upper bounds of the break fractions. Furthermore, we assume that $\left|\tau_{0,i} - \tau_{0,j}\right| > \eta > 0$ for all $i \neq j$. $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_m]'$ is a vector of break magnitudes. Finally, we have the error $\varepsilon_t = \mathrm{C}(L)v_t$ with $\mathrm{C}(L) := \sum_{i=0}^{\infty} \mathrm{C}_i L^i$ and $\mathrm{C}_0 := 1$ where v_t is assumed to be independent identically distributed (IID) with zero mean and finite second and fourth moment. Hence, the short-run and long-run variance of ε_t are $\sigma_\varepsilon^2 = E(\varepsilon_t^2)$ and $\omega_\varepsilon^2 := \lim_{T\to\infty} T^{-1} E(\sum_{t=1}^{T} \varepsilon_t)^2$ respectively.

HLT use a minimum GLS-detrended Dickey Fuller approach to unit root testing and allow for up to m breaks under both the null and the alternative. The procedure performs an augmented Dickey-Fuller test on GLS-detrended data for all possible combinations of break dates and chooses the break dates that yield the minimum resulting test statistics.

In more detail, the procedure starts by running a GLS regression of $\boldsymbol{y}_{\bar{\rho}} = [y_1, y_2 - \bar{\rho}y_1, \dots, y_T - \bar{\rho}y_{T-1}]'$ on $\boldsymbol{Z}_{\bar{\rho},\tau} = [\boldsymbol{z}_1, \boldsymbol{z}_2 - \bar{\rho}\boldsymbol{z}_1, \dots, \boldsymbol{z}_T - \bar{\rho}\boldsymbol{z}_{T-1}]'$, where $\boldsymbol{z}_t = [1, t, \boldsymbol{DT_t}(\boldsymbol{\tau})']'$ and $\bar{\rho} = 1 - \bar{c}/T$ for some user supplied $\bar{c} > 0$, to obtain $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\boldsymbol{\gamma}}$. Using the GLS parameter estimates the GLS residuals are generated as $\tilde{u}_t = y_t - \tilde{\alpha} - \tilde{\beta}t - \tilde{\boldsymbol{\gamma}}'\boldsymbol{DT_t}(\boldsymbol{\tau})$. Having obtained the detrended data the usual augmented Dickey-Fuller regression is set up as follows:

$$\Delta\tilde{u}_t = \hat{\pi}\tilde{u}_{t-1} + \sum_{j=1}^{k} \hat{\psi}_j \Delta\tilde{u}_{t-j} + \hat{e}_{t,k} \qquad t = k+2, \dots, T$$

from which the test statistics is generated from. This procedure is done for every combination of break dates within the allowed bounds. Finally, the statistic of the test is the minimum of the Dickey-Fuller statistics calculated above, that is:

$$MDF_m = \inf_{\tau_1,\dots,\tau_m} DF_{\bar{c}}^{GLS}(\boldsymbol{\tau})$$

where $\tau_1, \ldots, \tau_m \epsilon (\tau_L - \tau_U)$ and $|\tau_i - \tau_j| > \eta$ for all $i \neq j$. The authors suggest using the modified Akaike information criterion proposed by Ng and Perron (2001). Furthermore, for $m = 2$ they suggest using $\bar{c} = 21.5$ and provide corresponding asymptotic critical values.

2.2 The NP test

NP developed a unit root test with two putative structural breaks in either level or in level and trend under the null and the alternative. Because, in this study, we are interested in the case where trend breaks occur we only consider the model with a break in level and trend, this corresponds to model M2 in their article. As pointed out in the introduction the test corresponds to the innovational outlier model. This can be seen in the way it, as opposed to the additive outlier model, allows for lags in the break structure. In this way the change in the deterministic part of the process can be considered to occur gradually. With all variables defined as above, unless stated otherwise, the data generating process is similarly divided into one deterministic and one stochastic part:

$$y_t = d_t + u_t \qquad t = 1, \ldots, T$$

$$d_t = \alpha + \beta t + \mathrm{C}(L)\left(\boldsymbol{\theta}' \boldsymbol{DU_t}(\boldsymbol{\tau_0}) + \boldsymbol{\gamma}' \boldsymbol{DT_t}(\boldsymbol{\tau_0})\right) \qquad t = 1, \ldots, T$$

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad t = 2, \ldots, T$$

where $\boldsymbol{DU_t}(\boldsymbol{\tau_0}) = \left[DU_t(\tau_{0,1}), DU_t(\tau_{0,2})\right]'$ is a vector of level breaks with elements defined as $DU_t(\tau) = 1(t > \lfloor \tau T \rfloor)$, and $\boldsymbol{\theta} = [\theta_1, \theta_2]'$ being its associated break magnitudes, $\boldsymbol{\gamma}$ and $\boldsymbol{DT_t}(\boldsymbol{\tau_0})$ are defined as above with $m = 2$.

NP perform an augmented Dickey-Fuller test controlling for the breaks in level and trend. The break dates are detected using a sequential search method choosing the break dates with the statistically most significant associated parameters. More specifically, the procedure boils down to estimating an OLS regression as:

$$y_t = \rho y_{t-1} + \alpha + \beta t + \vartheta_1 D_t(\tau_1) + \vartheta_2 D_t(\tau_2) + \theta_1^* DU_{t-1}(\tau_1) + \theta_2^* DU_{t-1}(\tau_2)$$

$$+ \gamma_1^* DT_{t-1}(\tau_1) + \gamma_2^* DT_{t-1}(\tau_2) + \textstyle\sum_{j=1}^{k} \phi_j \Delta y_{t-j} + e_t$$

where $D_t(\cdot)$ is an impulse dummy such that $D_t(\tau) = 1(t = \lfloor \tau T \rfloor + 1)$. Note, that the break variables for the level and trend break are lagged one time period but not the impulse dummy. The test is carried out using the usual t-statistic of the estimated parameter $\hat{\rho}$. Note that a possible gradual change in the breaks is captured by the lagged differenced dependent variables in the OLS regression.

To identify the break dates the regression is run to test all possible positions for one break. The resulting most significant break date is chosen as one of the two break dates. As a second step the regression is run controlling for the previously identified break date and repeats the procedure to identify a second break. Consequently, we have that

$$\widehat{TB}^{*} = arg \max_{TB} \left| \hat{t}_{\vartheta_{*}}(TB) \right|$$

$$\widehat{TB}^{**} = arg \max_{TB} \left| \hat{t}_{\vartheta_{**}}\left(TB | \widehat{TB}^{*}\right) \right|$$

where TB is the set of all possible break locations, $\widehat{TB}^{*}$ and $\widehat{TB}^{**}$ are estimated break dates and $\hat{t}_{\vartheta_{*}}$ and $\hat{t}_{\vartheta_{**}}$ are the resulting t values of the impulse dummies from the first and second run of the regression respectively. The first and second break date is the smaller and the larger of the two identified breaks respectively. The authors suggest using the lag selection procedure proposed by Hall (1994). Small sample critical values that are simulated assuming no break and standard normal IID errors are provided. Furthermore, when generating the critical values they set the lag length equal to the true, that is, equal to zero.

3. Monte Carlo setup

In this section the Monte Carlo study in which the properties of the two tests are investigated in is setup. Since the innovational and additive outlier models imply different characteristics of the breaks the two tests are not equivalent in general. However, in the special case where the true lag length is equal to zero the gradual change in the innovational outlier model becomes sudden. Hence, it coincides with the additive outlier model. Furthermore, although the NP test allows for changes in level applying it to the case of sole trend breaks does not violate the assumptions of the test. Consequently, in order to operate

in an environment where both tests apply we assume that the true lag length is equal to zero and only trend breaks are allowed to occur. The data generating process is setup in a similar manner as above imposing, without loss of generality, $\alpha = \beta = 0$, and in order to make the tests comparable $m = 2$, $\boldsymbol{\theta} = [0, 0]'$, and $\mathrm{C}(L) = 1$. Thus we have:

$$y_t = d_t + u_t \qquad t = 1, \dots, T$$

$$d_t = \gamma_1 DT_t(\tau_{0,1}) + \gamma_2 DT_t(\tau_{0,2}) \qquad t = 1, \dots, T$$

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad t = 1, \dots, T$$

with ε_t being standard normal IID where we impose varying magnitudes of γ_1, γ_2, and the autoregression coefficient ρ. Furthermore, we only consider the case where the break fractions are equal to $(\tau_{0,1}, \tau_{0,2}) = (0.4, 0.6)$. In both tests the trimming factor is set to 0.2, which implies that the region of possible breaks is $\lfloor 0.2T, 0.8T \rfloor$. Additionally, the minimum distance between the estimated break dates is set equal to 1. In the simulations the true lag length is equal to zero, however, since we never know the true lag length in practice we perform the tests both with and without lag selection. When the lag selection is applied the maximum lag length is set to $k_{max} = \lfloor 12 \times (T/100)^{1/4} \rfloor$ apart from when the sample size equals to 50, in which case we set $k_{max} = 7$. Size and power are calculated using critical values at 5 % level of significance. All simulations are performed using the standard Monte Carlo method.

To begin with, simulations of sample sizes $T = 50$, $T = 100$, and $T = 300$ with $\rho \in (1, 0.9, 0.7)$ for various break magnitudes $\boldsymbol{\gamma}$ are performed. Since we pay special attention to breaks magnitudes and break detection the specifications of the elements in the vector $\boldsymbol{\gamma}$ are central to the analysis. Following the arguments above, that it is important to investigate tests properties when the break magnitudes are relatively small, for $T = 50$ and $T = 100$ we set $\gamma_1 = \gamma_2 = (0, 0.1, 0.2, \dots, 4)$. For computational reasons we only consider $\gamma_1 = \gamma_2 = (0, 0.25, 0.5, \dots, 4)$ for $T = 300$. To put the magnitudes into a context we use the local break magnitudes approach as a vehicle to understand their implications. Following HLT, the vector of local break magnitude, $\boldsymbol{\kappa} = [\kappa_1, \dots, \kappa_m]'$, is defined as $\boldsymbol{\kappa} = \boldsymbol{\gamma} \omega_\varepsilon T^{1/2}$. Hence, with standard normal IID errors the long-run variance is equal to one, and thus we have

$\boldsymbol{\kappa} = \boldsymbol{\gamma} T^{1/2}$. Essentially what this does is that it translates what a fixed magnitude for a given time span under a certain data generating process would imply in another time span or data generating process. In our cases the implied local break magnitudes are $\kappa_1 = \kappa_2 = (0, 1, 2, \dots, 40)$ for $T = 100$, $\kappa_1 = \kappa_2 \approx (0, 0.71, 1.42, \dots, 28.28)$ for $T = 50$, and $\kappa_1 = \kappa_2 \approx (0, 4.33, 8.66, \dots, 69.28)$ for $T = 300$. According to HLT, both the power valleys and the realistic break magnitudes lie within the span such that $0 < \kappa_i < 15$. Therefore, given the magnitudes we study, at least for $T = 50$ and $T = 100$, we have good coverage for the relevant span. For the cases where $T = 100$ and $T = 50$ we perform 5000 repetitions. As a result of the computational burden we only perform 1000 repetitions for the case where $T = 300$. The generated size and power are presented in Figure 1 to 3. Additionally, normalized histograms of the estimated break dates are illustrated in Figure 4 to 7.

On the basis of being a realistic time span in applied research whilst being large enough to make reasonably powerful tests plausible $T = 100$ is treated as a benchmark case. To shade some light on how erroneous endogenously estimated breaks relate to power and size the benchmark case is studied further. For this purpose size and power for specific endogenously estimated break dates are generated. This is done by generating a sample in the same manner as above, where we did consider the average size and power over all estimated break points. The only difference now is that we divide the sample into subsamples by sorting it according to the estimated break dates. In order to get reasonable estimates for each possible break point, however, this requires a much larger sample compared to when the average size and power are examined. Consequently, for this case, we generate a sample of 60000 observations. For example, this implies that for break dates that are estimated with a frequency of 1% we have a sample of 600 observations. The simulations are made under the null and the alternative with $\rho = 0.7$ and breaks of magnitudes $\gamma_1 = \gamma_2 \in (0, 0.5, 1)$. The results are found in Table 1 and 2.

In each sample 50 initial observations are generated for the period before the considered time periods start. This is to overcome initial value problems and the initial observations are discarded before running the tests. The simulations are performed in MATLAB based on codes written by the author of this paper. However, the codes for the HLT and NP test are

largely based on GAUSS codes originally provided by their respective authors.[1] In a preliminary analysis we studied both positive and negative breaks of different magnitudes. We also studied the case where one break was fixed and the other varied from positive to negative. The results are qualitatively similar to those reported in this paper and are available upon request.

4. Results

4.1 Size and power curves

In this section size and power of the tests with respect to the break magnitudes for various sample sizes are considered. The results are presented in Figure 1 to 3. Note that the vertical axes of the graphs take different values. The scaling was done to enable one to visually distinguish the curves form one another in all cases.

The first graph in Figure 1 illustrates the size of the tests in the benchmark case where $T = 100$.The only test that has the appropriate size when no breaks occur is the NP test when the lag length is set equal to zero. Note, however, that the critical values for this test are generated under the very same data generating process, and thus the size in this case is correct by construction. However, when the break magnitudes increase the size of the zero lag NP test declines and reaches its lowest level of just over 1.5% for break magnitudes of about 1. Thereafter it slowly increases to surpass 5% for magnitudes of 2.5 and reaches 10% for magnitudes equal to 4. When including lags in the test it is substantially oversized at 11% when no breaks occur. Thereafter the size follows a similar pattern as it did in the zero lag case. To conclude, we see that neither version of the NP test has stable size as the break magnitudes change. That is, neither test has the appropriate size in general.

With lag length set to zero the HLT test is also oversized at a level of 10% when no breaks occur. However, the pattern when the break magnitudes increase is quite different from that of the NP test. At first, the rejection rate is quite stable for small to moderate breaks of

[1] The original GAUSS codes for the HLT and NP test respectively are available at: https://www.nottingham.ac.uk/research/groups/grangercentre/research/gauss-code.aspx and https://www.researchgate.net/publication/290147745_Narayan_and_Popp_2010_Journal_of_Applied_Statistics_two_endogenous_structural_break_test_GAUSS_CODE

magnitude 0.5. Thereafter the size decreases and lies stable between 6 and 6.5% for break magnitudes larger than 1. When lag selection is introduced, in contrast to the NP test, the HLT tests size curve is shifted slightly downwards. In the no break case its size is about 8% and it is converging to a level between 5.5 and 6% when the break magnitudes grow larger. Consequently, it is a bit less oversized when lag selection is applied but in general it follows a similar pattern. Most striking about the two HLT tests is that, although they are somewhat oversized, they are reasonably invariant to the break magnitudes. Furthermore, since asymptotic critical values are used in this relatively small sample, it is reasonable to expect the tests to be a bit oversized.

Turning our attention to the second plot in Figure 1 we see that neither test looks promising in detecting near unit roots. We can see that the NP test collapses for moderate sized breaks. The power of the NP test with zero lags falls as low as to a level of 2% for break magnitudes around 1 only to slowly regain some power as the break magnitudes increase. Similarly, the NP test with lag selection falls and hits its bottom at 5% for break magnitudes between 1 and 1.5. Following a similar pattern as its zero lag version it slowly regain some power as the break magnitudes increase. However, it never reaches a power higher than 8% for large breaks. The power curves of the two HLT tests in this case are shifted upwards between 5 to 7 percentage points compared to the size.

The third plot in Figure 1 shows the power curves with respect to break magnitudes when $\rho = 0.7$. Now, the NP test shows a quite large improvement in power for the for the no break case, in particular when the lag length is set to zero. However, again the test collapses when the break magnitudes are increased. For both versions of the test the power reaches a bottom at 10% for break of magnitudes between 1 and 2. Similar to the previous cases, after reaching its bottom the power increases slowly as the break magnitudes increase. The HLT test shows a substantial improvement in power, both with and without lag selection. However, introducing lags now punishes the test a bit harder than it did in the previous cases. The curve still has its characteristic appearance being larger for small breaks but already for moderate break magnitudes of about 0.5 the power has settled at a stable level.

Figure 1: Size and power with respect to the break magnitudes, $T = 100$

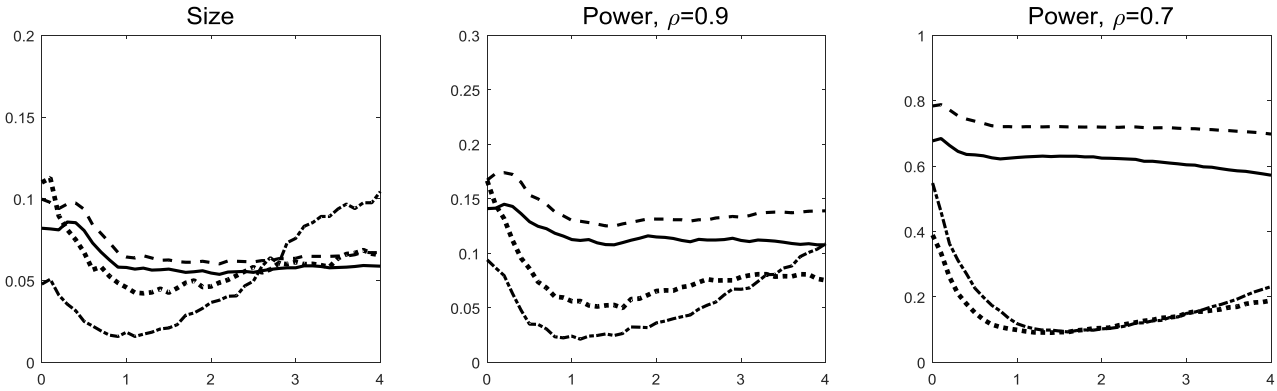


Figure 1 shows size and power plots of the HLT and NP test. The data is based on Monte Carlo simulations with 5000 repetitions under the null and the alternative with $\rho \in (0.7, 0.9, 1)$, changing break magnitudes and sample size $T = 100$. The horizontal axes indicate the break magnitudes and the vertical axes correspond to the rejection frequency. The vertical axes are scaled for visual clarity. HLT lags: $-$, HLT zero lags: $- -$, NP lags: $\cdots$, NP zero lags: $- \cdot -$.

For small sample power and size we turn to Figure 2 where the results for $T = 50$ are pictured. As before, we see that the only test that has the appropriate size is the NP test when the lag length is set to zero and no breaks are present. Again, when lags are introduced into the procedure it is oversized, now somewhat more than it was in the $T = 100$ case. As the break magnitudes increase the power curves show a similar behavior as in the benchmark case, however, a bit less pronounced. Relying on asymptotic critical values both of the HLT tests are now quite significantly oversized. For these tests the size curves are more or less shifted 9 percentage points upwards compared to the benchmark case.

The second plot of Figure 2 shows that both procedures only marginally increase the rejection frequencies when we move from the null to the alternative with $\rho = 0.9$. For the case where $\rho = 0.7$ in the third plot the HLT test is quite substantially improved. At the same time, the NP test does not change by more than a few percentage points when no breaks occur. When breaks are present the change is almost absent. Although, the HLT test does show some improvement in the $\rho = 0.7$ case none of the tests can be considered performing well when the sample size is as small as $T = 50$.

Figure 2: Size and power with respect to the break magnitudes, $T = 50$

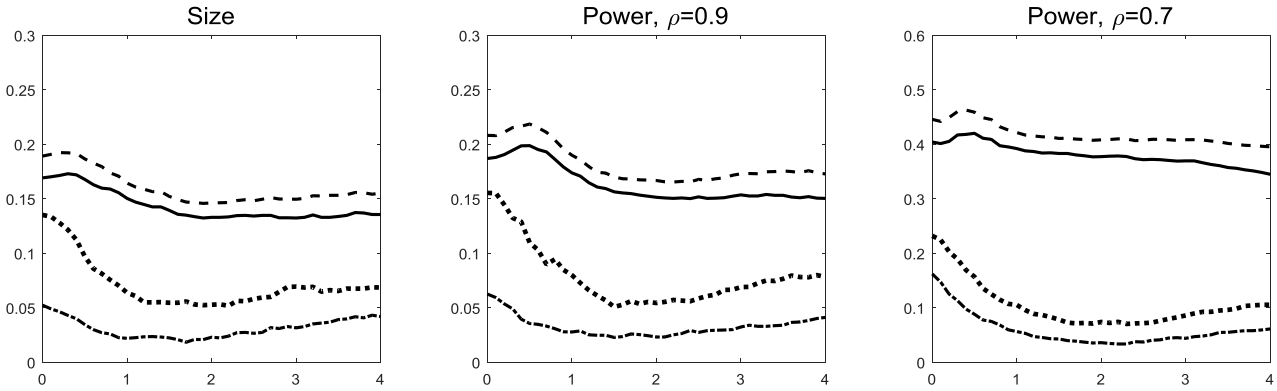


Figure 1 shows size and power plots of the HLT and NP test. The data is based on Monte Carlo simulations with 5000 repetitions under the null and the alternative with $\rho \in (0.7, 0.9, 1)$, changing break magnitudes and sample size $T = 50$. The horizontal axes indicate the break magnitudes and the vertical axes correspond to the rejection frequency. The vertical axes are scaled for visual clarity. HLT lags: $-$, HLT zero lags: $- -$, NP lags: $\cdots$, NP zero lags: $- \cdot -$.

The large sample case where $T = 300$ is displayed in Figure 3 and the first plot pictures the size. The HLT test is now less oversized when no break occurs compared to the smaller sample cases. As expected, since we are using asymptotic critical values the test is closer to its nominal level for larger sample sizes. However, because the decline in size when the break magnitudes increase persists the test is now somewhat undersized for large break magnitudes. As before, including lags to the procedure results in a downward shift of the size curve. The NP test, on the other hand, is subject to some major fluctuations. The shape of the curve is still similar to the smaller sample cases. However, the extent to which it is oversized is now worsened for large break magnitudes. The test when assuming zero lags is significantly oversized as it reaches its peak at 30%. The version of the test with lags also suffers from size problems, although, while the size reaches its peak at about 10% the problems are not as severe. Worth noting, however, is that the test is less oversized when no break occurs.

The second picture shows the power when $\rho = 0.9$. Regarding the HLT test the shape of the curve is more or less the same as in the smaller sample cases. Striking, though, is that the test now has reasonably high power. The NP tests also achieve a substantial improvement in power in the no break case. However, the large fall in power as a consequence of the breaks still persists and it reaches a bottom at around 7%. This is the case whether lags are included or not. The power is increased after reaching its bottom and lies at about 10 percentage

points above the size. The third picture illustrates the case when $\rho = 0.7$. In this instance the curves are shifted upward about 40 percentage points in the no break case for the NP test and in general for the HLT test. For the HLT test without lags this implies 100% power. When breaks are present the power of the NP test falls as it did in the previous cases.

Figure 3: Size and power with respect to the break magnitudes, $T = 300$

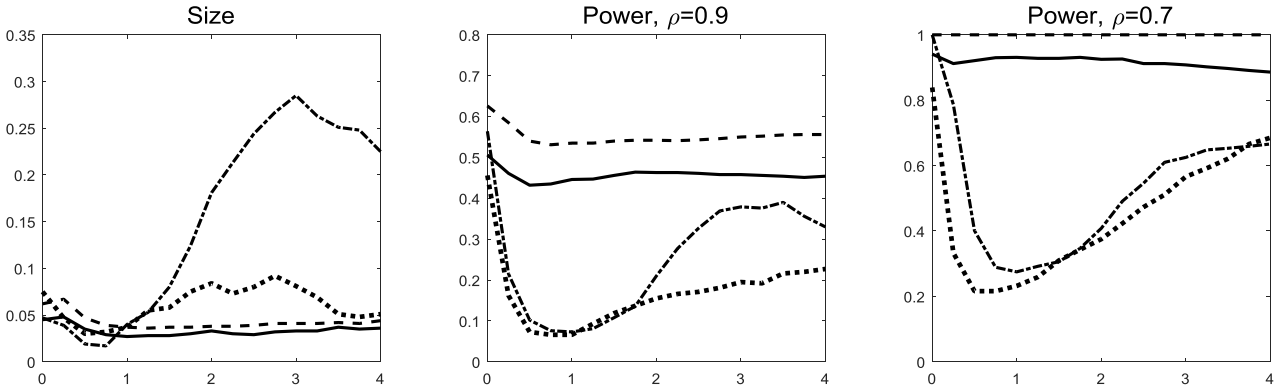


Figure 1 shows size and power plots of the HLT and NP test. The data is based on Monte Carlo simulations with 5000 repetitions under the null and the alternative with $\rho \in (0.7, 0.9, 1)$, changing break magnitudes and sample size $T = 300$. The horizontal axes indicate the break magnitudes and the vertical axes correspond to the rejection frequency. Note that the vertical axes are scaled for visual clarity. HLT lags: —, HLT zero lags: — —, NP lags: ⋯, NP zero lags: — · —.

The difference between the three figures can best be understood in the light of the local break magnitudes framework. For example, under the null when $T = 100$ the size of the NP test without lags reached its bottom when the break magnitudes were roughly equal to 1. This implies local break magnitudes equal to $\kappa_i = \gamma_i T^{1/2} \Leftrightarrow \kappa_i = 1 \cdot 100^{1/2} = 10$. Thus, keeping the local magnitudes fixed when $T = 300$ we have $10 = \gamma_i 300^{1/2} \Leftrightarrow \gamma_i \approx 0.6$. As illustrated in Figure 3 this is, likewise, the point where we find the minimum size in this case. In this way, Figure 3 indicates what we may had seen if we had considered larger magnitudes in the $T = 100$ case in Figure 1. Considering the power the intuition is the same, however, in this case also the autoregression coefficient has to be rescaled.

The overall finding from Figure 1 to 3 is that the HLT test is roughly invariant to the break magnitudes, whereas the NP test exhibits large variations. Relying on asymptotic critical values the HLT test is oversized in small samples. Nevertheless, the only case where the NP test is unambiguously closer to the nominal size is in the smallest sample. However, the size of the NP test with lags in this case still varies more than that of the HLT test, in addition, the

power is extremely low. Therefore, it cannot be considered to possess better size properties. Furthermore, the HLT test has equal or higher power compared to the NP test in all case, and hence it is superior to the NP test in general.

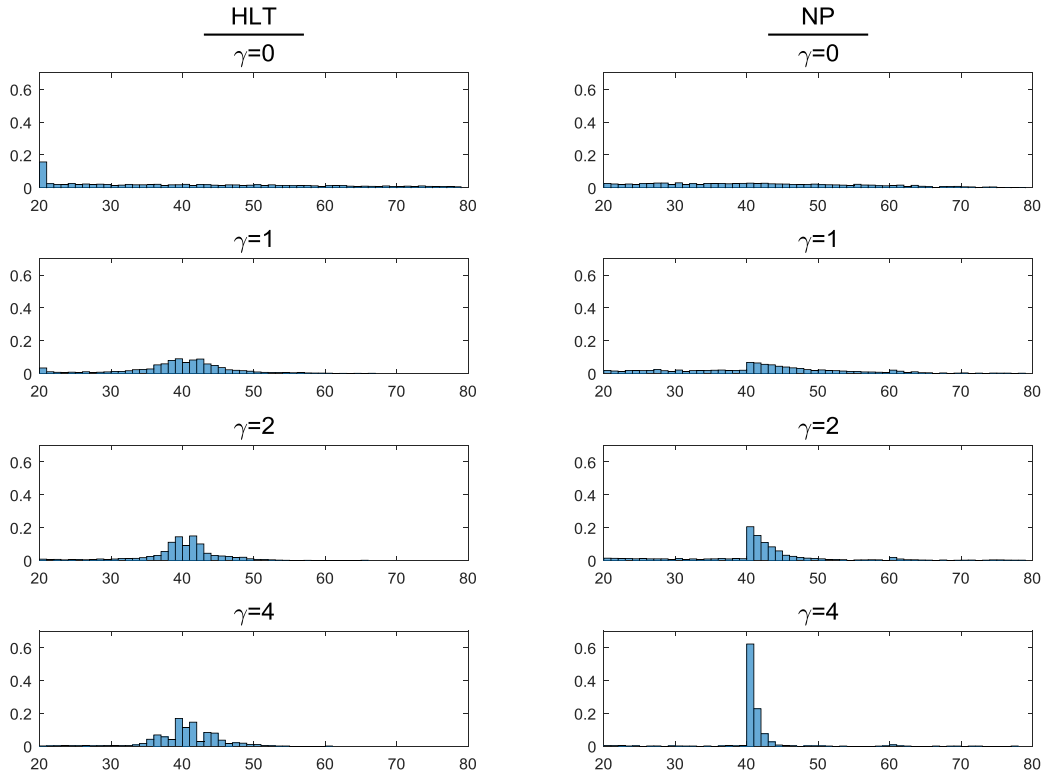
4.2 Break detection

In the following section the break date detection in the benchmark case where $T = 100$ is treated. Note that, in this case the true break dates are positioned at the time points 40 and 60. Due to symmetry in the detection procedures, for both tests, the distributions of the estimated first and second break are merely mirrored versions of each other. Furthermore, the results are roughly the same whether lag selection is applied or not. Therefore, for simplicity we only present the results for break detection of the first break when lag selection is applied.

Figure 4 shows normalized histograms of the estimated break dates under the null for different break magnitudes. In the no break case, as expected, the distribution of the estimated breaks in both tests look quite close to uniform. That is, apart from the peak in the HLT test at the lower bound of the set of possible break points. As the break magnitude increases the break detection in the HLT test attains a bell shape around the true break date. Although it looks symmetric, and hence unbiased in terms of its mean value, the true break date is not the most frequently observed. Instead, the two neighbouring break points are more frequent creating a small dump where the true break date is located.

The distribution of the NP test, on the other hand, is skewed with a sharp increase at the true break date followed by a gradual decline to the right. For breaks of magnitude 1 the two tests are quite equal in their ability to detect the true break date exactly. Although, if instead, one considers the mass of an area around the true break date the HLT test outperforms the NP test. As the break magnitude increases the break detection of the NP test is improved more rapidly than that of the HLT test. For the largest magnitude the NP test is the more accurate of the two tests. Note, however, that even if the test manages to estimate the true break point accurately in some cases, the break estimation is on average biased.

Figure 4: Distribution of the estimated first break date under the null



The figure consists of normalized histograms of the estimated first break resulting from the HLT and NP test with lag selection. The data is based on Monte Carlo simulations with 5000 repetitions under the null with break magnitudes $\gamma_1 = \gamma_2 \in (0, 1, 2, 4)$. The horizontal axes indicate the estimated break dates and the vertical axes resemble their corresponding frequencies.

Figure 5 pictures estimated break dates under the alternative where $\rho = 0.7$. The most striking difference when compared to Figure 4 is the improved break detection in the HLT test. The distribution now has a larger portion of its mass situated close to the true break date. Furthermore, the true break date is now the most frequently estimated when the breaks are large. The NP test, on the other hand, is only affected marginally by the change of the parameter ρ. The pictures indicates that, as ρ is getting smaller there is a slight tendency towards the mass being concentrated closer to its biased estimated mean.

Figure 5: Distribution of the estimated first break date under the alternative, $\rho = 0.7$

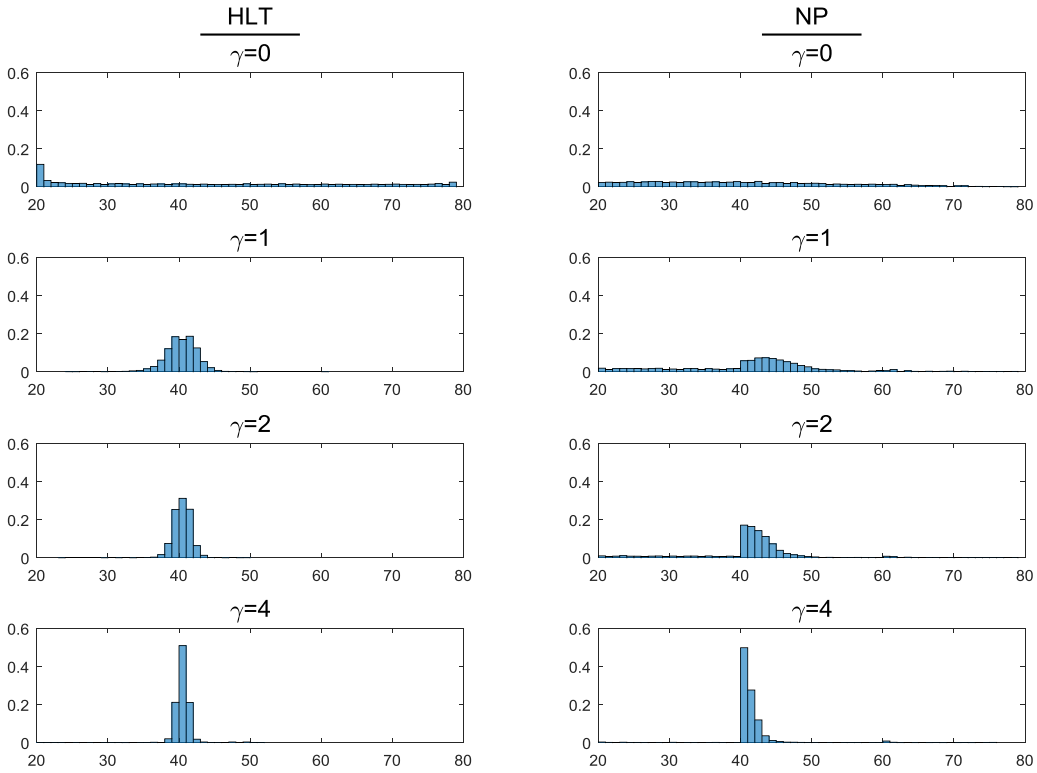


Figure 5 consists of normalized histograms of the estimated first break resulting from the HLT and NP test with lag selection. The data is based on Monte Carlo simulations with 5000 repetitions under the alternative with $\rho = 0.7$ and break magnitudes $\gamma_1 = \gamma_2 \in (0, 1, 2, 4)$. The horizontal axes indicate the estimated break dates and the vertical axes resemble their corresponding frequencies.

Figures 6 and 7 show the distribution of the estimated break dates for realistic, and possibly problematic, break magnitudes. Again, we can see that there is a clear improvement in break detection of the HLT test as ρ is getting smaller.

Figure 6 shows the case where the break magnitude is equal to 0.5. Under this magnitude the break dates are difficult to estimate while they are presumably large enough to cause problems to the tests if not accounted for. Under the null the break detection is very inaccurate for both tests, which have distributions close the no break case. As noted above, when ρ changes the break detection is improved in the HLT test. For $\rho = 0.9$ there is some tendency towards enhanced break detection and for $\rho = 0.7$ the improvement is substantial. The NP test, on the other hand, is barely affected at all by ρ.

Figure 6: Distribution of the estimated first break date with fixed break magnitudes, $\gamma = 0.5$

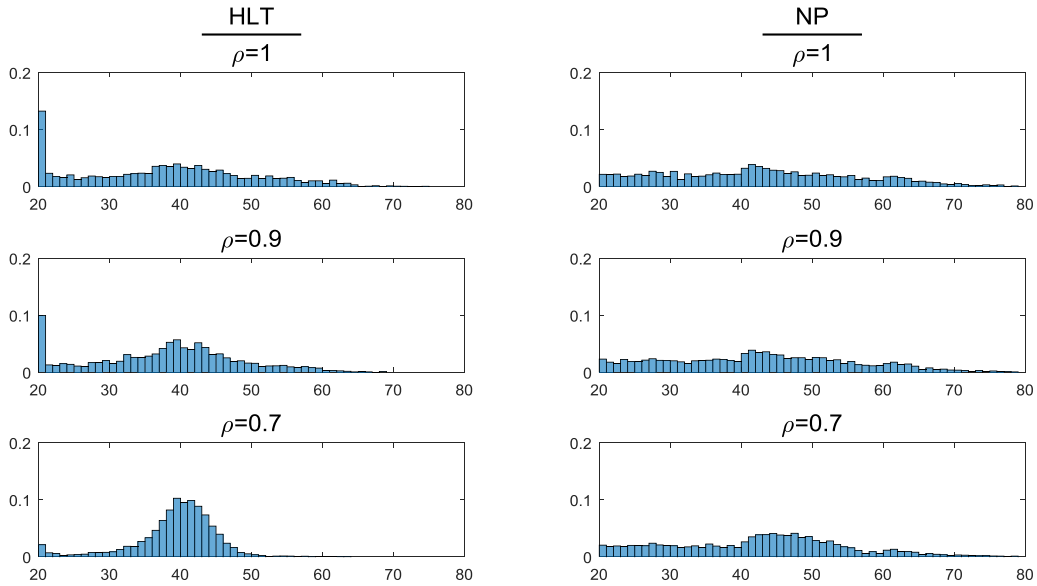


Figure 6 consists of normalized histograms of the estimated first break resulting from the HLT and NP test with lag selection. The data is based on Monte Carlo simulations with 5000 repetitions under the null and alternative with $\rho \in (0.7, 0.9, 1)$ and break magnitudes $\gamma_1 = \gamma_2 = 0.5$. The horizontal axes indicate the estimated break dates and the vertical axes resemble their corresponding frequencies.

Figure 7: Distribution of the estimated first break date with fixed break magnitudes, $\gamma = 1$

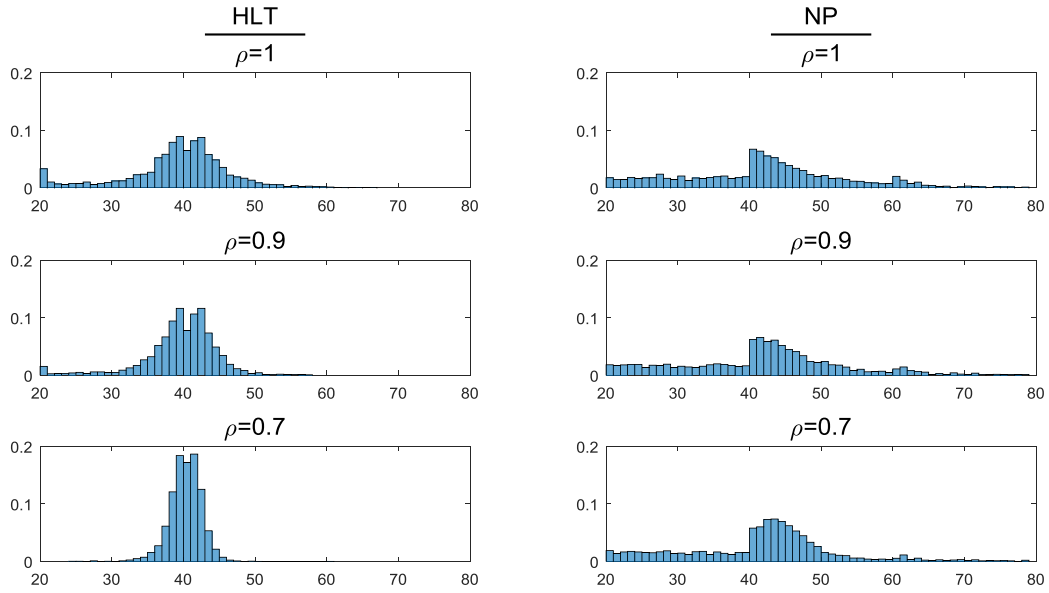


Figure 7 consists of normalized histograms of the estimated first break resulting from the HLT and NP test with lag selection. The data is based on Monte Carlo simulations with 5000 repetitions under the null and alternative with $\rho \in (0.7, 0.9, 1)$ and break magnitude $\gamma_1 = \gamma_2 = 1$. The horizontal axes indicate the estimated break dates and the vertical axes resemble their corresponding frequencies.

For breaks of magnitude 1 in Figure 7 the results are qualitatively similar to the previous cases. In this setting the HLT test is superior to the NP test both under the null and the alternative. The HLT test achieves quite high precision when $\rho = 0.7$ with roughly 90% of the

sample being within three time points from the true break date. The NP test shows, as previously, a tendency towards being cumulated around its biased mean.

In this section we have considered the accuracy of the break detection of the two tests. The main finding is that neither test is very accurate when breaks of magnitudes commonly observed in practice are present. However, the accuracy is improved in the HLT test when we go from a true null to a true alternative. In general, the break detection of the HLT test is more accurate than the NP test except for when the breaks are very large and the null is true.

4.3 Size and power with respect to the estimated break points

To see what the precision in the break estimation implies for the two tests we investigate how they perform depending on the location of the estimated breaks. As we saw in the histograms above, when the break magnitudes increase the precision of the break point estimation increases relatively slow. Consequently, for realistic break magnitudes neither test manages to estimate the break dates very accurately. Therefore, one may suspect that the size and power problems in the procedures may be related to inaccurate break estimation. In this case, we should presumably observe divergence in the size and power for incorrect break dates. Then the severity of inaccurate break estimation depends on both the divergence in size and power for each particular break point and the frequencies for which they are estimated. To investigate the matter size and power for specific estimated break points are tabled along with their corresponding estimation frequency. Again, we only consider the estimated first break point.

Table 1 presents size and power as well as the frequency of the estimated break points of the HLT test. The results are based on simulations with break magnitudes $\gamma_1 = \gamma_2 \in (0, 0.5, 1)$ under the null and the alternative with $\rho = 0.7$. When the estimated break point is equal to the true the test with lags is largely invariant to break magnitudes. For the test without lags we can see a slight decrease in the size when the magnitude is increased from 0.5 to 1, although, the divergence is not very large.

Considering incorrect break estimates when breaks are present we can see that as we move further away from the true break point there is a gradual decrease in size. Moreover, the extent to which this decline takes place depends on the break magnitude. For breaks of magnitude 0.5 the test is still very resilient to incorrect estimates. The size of the test with lags only falls slightly below 5% for the uncommon breaks estimated at time point 60 and later. The test without lags never falls below 5% for any of the considered break points. When the break magnitude is equal to 1 the decline is more distinct. In this case the size reaches problematic levels for estimated break points far away from the true. That is, in particular, the case for estimates that lies at time point 50 and later, in which case the size is equal to 2% or less. However, because of the precision in the break detection the troublesome cases do not amount to a very large portion of the sample. In the present case they only make up 5.3% and 3.9% of the sample for the test with and without lags respectively. Hence, this does not cause any major problems to the testing procedure. The rejection frequency of the break points equal to 60 and later is close to zero, however, these events are rare enough to be negligible.

Considering the power we can see that it falls when breaks are introduced irrespective of whether they are estimated accurately or not. Thereafter, when the break magnitudes increase, the power is more or less invariant when the breaks are correctly estimated. Note that this reflects the initial decline in the power curves illustrated in Figure 1 to 3. This is interesting because it indicates that the reduction in power when the break magnitudes go from small to moderate may not be caused by inaccurate break estimation. When the estimated break points are far from the true the power is below acceptable levels. However, since these events are increasingly rare as the magnitudes increase, this does not cause any problems to the test. In this way, it appears as the effects of increased precision and declining power due to inaccurate break point estimates cancel out as the break magnitudes increase. This serves as a possible explanation for, as we saw in Figure 1 to 3, why the power and size are approximately invariant to the break magnitudes when they grow from moderate to large.

Table 1

HLT with lag selection

	$\widehat{TB_1}$	~30		35		37		40		43		45		50~		60~	
ρ	γ	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq
1	0	37.1	8.3	1.8	9.7	1.8	10.2	1.8	9.5	1.7	9.4	1.6	7.8	29.4	6.7	15.9	5.5
	0.5	31.8	7.5	2.7	8.6	3.3	10.4	3.3	9.4	3.4	7.6	2.5	7.5	15.7	5.2	3.7	4.4
	1	12.6	2.9	3.4	7.4	5.8	6.8	6.1	9.2	6.6	6.6	3.4	3.3	5.3	1.3	0.5	0
0.7	0	30.7	67.6	1.5	73.6	1.4	73.2	1.5	74	1.4	70.8	1.4	67.3	41.1	67.2	27.1	66.9
	0.5	8.1	58.2	3.5	62.6	6.3	65.1	9.6	67.1	7.1	61.6	3.6	61.8	1.4	39.2	0.1	0
	1	0.3	20.3	1.4	49.8	6.1	59	17.6	68.6	5.3	57	0.8	39.4	0	-	0	-

HLT with zero lags

	$\widehat{TB_1}$	~30		35		37		40		43		45		50~		60~	
ρ	γ	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq
1	0	38.2	10.1	1.8	12.2	1.8	13.5	1.8	11.9	1.7	11.7	1.6	11	28.4	8.3	15.1	7
	0.5	31.9	9.2	2.7	10.8	3.3	11.6	3.4	11.3	3.5	8.5	2.4	8.9	15	6.6	3.2	6.5
	1	10.7	3.9	3.2	8.6	6	7.6	6.6	9.6	7	6.8	3.3	3.9	3.9	2	0.1	0
0.7	0	31.9	79.3	1.4	83.3	1.5	83.4	1.5	84.6	1.4	79.1	1.4	80.2	40.4	78.3	26.7	78.1
	0.5	7.9	72.1	3.3	75.4	6.3	75	10	76.2	7	73	3.3	75.4	0.9	62.1	0	-
	1	0.2	45.5	1	72.5	5.3	72.1	18.8	75.6	4.4	71.6	0.4	72.6	0	-	0	-

Table containing simulated rejection frequencies for estimated trend break one and their corresponding frequencies in the sample denoted in percent. The simulation is based on 60,000 repetitions. Break points with a corresponding part of the sample of less than 0.1 percent are excluded on the basis of being too small to draw meaningful conclusions on. Note that ~30 corresponds to break dates estimated at time point 30 and earlier and that 50~ and 60~ correspond to break dates estimated at time point 50 and later and 60 and later respectively.

Table 2

NP with lag selection

	$\widehat{TB}_1$	~30		35		37		40		43		45		50~		60~	
ρ	γ	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq
1	0	26.3	7.2	2.6	12.8	2.5	13.4	2.5	15.4	2.3	14.7	2.3	14.6	28.4	9.9	11.7	7.8
	0.5	22.7	3.6	2.3	9.1	2.3	9.6	3.4	9.9	3.1	11	2.9	9.9	28	5	12.4	2.6
	1	19.3	2	2	7.7	2	9.8	6.8	6.3	5.1	7.7	4.1	6.3	19.9	1.9	8.3	0.3
0.7	0	28.8	34.6	2.8	40.1	2.4	40.1	2.5	39.7	2.3	41.3	2.2	42.3	26.1	38	10.9	34.6
	0.5	20.9	6	2	22.3	1.8	23.3	2.8	24.1	4	26.6	4.2	25.2	25.1	7.1	8.1	0.6
	1	17.7	1.7	1.5	15.1	1.4	18.5	5.5	18.2	7.4	18.2	6.4	11.4	13.6	0.5	5.9	0

NP with zero lags

	$\widehat{TB}_1$	~30		35		37		40		43		45		50~		60~	
ρ	γ	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq	Part of sample	Reject freq
1	0	30.6	5	2.7	5	2.6	4.3	2.5	5.4	2.2	5.8	2.1	4.9	25.1	3.5	10.9	2.8
	0.5	24.8	2.7	2.3	5.2	2.3	3.6	3.4	4.5	3.1	3.4	3	2.9	26.2	1.2	12.4	0.7
	1	19.1	1.6	1.9	4.4	1.8	3.1	6.8	1.8	5.8	2.8	4.5	2	17.1	0.3	7.7	0
0.7	0	33.1	56.4	2.6	55.7	2.3	54.2	2.4	56.5	2	58	2	54.6	24.5	52.3	10.7	49.6
	0.5	18.2	17.7	1.7	34.4	1.6	35.2	2.7	36.1	4.1	37.6	4.6	35.4	26.7	7.6	8.3	0.3
	1	13.7	3.4	1.3	20.8	1.3	21.9	5.5	20.9	8.5	21.6	7.9	12.6	10.9	0.1	4.4	0

Table containing simulated rejection frequencies for estimated trend break one and their corresponding frequencies in the sample denoted in percent. The simulation is based on 60,000 repetitions. Note that ~30 corresponds to break dates estimated at time point 30 and earlier and that 50~ and 60~ correspond to break dates estimated at time point 50 and later and 60 and later respectively.

In Table 2 size and power with respect to the estimated break points in the NP test are presented. A first observation is that for correctly estimated break points the size is negatively related to the break magnitude. This is the case whether lags are included in the model or not. The test without lags is more or less invariant to the estimated break locations when no break occurs. That is apart from break points at the very right end on of the table, which are still not very far from the nominal size of 5%. The test with lags shows a more problematic behaviour. It varies substantially with respect to the estimates location and is more oversized than it is on average when the estimated break lies at time point 40. As the break points move towards the ends of the set of time points the size gradually declines.

When breaks are present, as we saw in the Figure 1, the size falls substantially when the break magnitudes increase from zero to one. Table 2 gives some idea of what might have gone wrong. For the test with lags we see a substantial overall decline in the size as the break magnitude increases. Although in a region around the true break point the size is still at a reasonable level, whereas it is well below 5% at the ends. When the break magnitude is equal to 1 the estimated break points that are smaller than 30 or larger than 50 exhibit a size of 2% or less. As recently discussed, the degree to which this causes a problem depends on the frequency of these estimates. In the current case the problematic estimates amount to about 40% of the sample. Consequently, they have a substantial impact on the average size. The results for the test without lags are largely qualitatively similar to the test with lags.

When no breaks are present, just as in the case of the size, the power of the test without lags is almost invariant to the estimated break points and lies between 50% and 55%. The test with lags exhibits a similar declining tendency at the ends of the time set as it did for the size, however, to a lesser extent. Both tests experience a substantial decline in power when breaks are present. This is even the case when the breaks are accurately estimated. An interesting observation is that both tests seem to be more or less invariant to the estimated breaks locations when they lie within the interval 35 to 45. However, the rejection frequencies when the break points lie outside of the interval are very low. In some cases, the power is low as zero. Since these instances are also quite frequent they account for a substantial part of the overall power of the tests. The cases with very low power amount to

between 25% and 45% of the sample. The cases with power close to zero amount to about 14% and 11% when the break magnitude is equal to 1 for the test with lags and without lags respectively.

To further put the instances of low power into some perspective we compare it to the size. When the break magnitude is equal to 1, in the test with lags, the size is virtually equal to or larger than the power for breaks estimated at 10 or more time points away from the true. Under the null these instances amount to 40% of the sample and under the alternative they amount for 30%. This implies that, when the breaks are equal to one the procedure does not effectively test for anything at all in 30% to 40% of the cases. The test performs somewhat better when no lags are included, however, the problem is still present to non-negligible degree.

The results in this section have somewhat different implications for the two procedures. From Table 1 we could see that the HLT test is largely invariant to whether the break points are correct or not. Furthermore, since the few instances where the size and power reach problematic levels are relatively rare the precision in the break estimation is arguably sufficient. Instead, the sensitivity of size and power with respect to the break magnitudes seem to be present irrespective of the estimated break points locations. Therefore, the source of the variation appears to be inherent to the model, that is, apart from the precision. For the NP test, on the other hand, the implications are different. Because in this test the break point estimates associated with problematic size and power are in fact very frequent the mere precision in the procedure seems to be a problem. Finally, because even when the estimated break date is correct neither size nor power is invariant to the break magnitudes it appears to be other problems inherent to this method as well.

5. Conclusions

In this paper the properties of the two structural break unit root tests developed in HLT and NP have been compared by Monte Carlo simulations. The simulations show that the HLT test dominates the NP test in general. Furthermore, whereas the HLT test is relatively invariant

to break magnitudes, the NP test exhibits considerable variations in size and power when breaks are introduced.

In the benchmark case with sample size $T = 100$ nether test holds the correct size over the entire set of break magnitudes. The HLT test experiences a small fall in size when breaks are introduced and is largely stable as the break magnitudes grow larger. The size of the NP test, on the other hand, undergoes large fluctuations over the set of break magnitudes. The power is quite low in both tests when a near unit root is considered. Conversely, when the autoregressive coefficient is closer to zero the power of the HLT test is substantially improved relative to the NP test. This is the case for all break magnitudes, although, because the power of the NP test falls significantly when breaks are introduced the difference is markedly larger when breaks are present. For large sample sizes both tests have good sample and power properties when no breaks occur. However, when breaks are present the results are qualitatively similar to the benchmark case, and hence only the HLT test remains reliable. Neither of the tests exhibits good size and power properties for the smallest sample size where $T = 50$.

For sample sizes and break magnitudes that are normally observed in practice both procedures estimate the break dates inaccurately. However, the HLT test outperforms the NP test in break date detection in all cases except for when the break magnitudes are very large and the null is true. Furthermore, the robustness of the HLT test as well as the sensitivity to break magnitudes of the NP test persist whether the breaks are accurately estimated or not. However, the problems in the NP test are worsened when the estimated break points are far from the true.

Finally, because we did only consider one specific location of the trend breaks we cannot abstract what the results would have been if the breaks were positioned differently. Moreover, the simulations were undertaken to cover the special case where both tests apply without violating any assumptions. For a practitioner it would indeed have been interesting to investigate the tests under other data generating processes as well. Consequently, areas of further research are to examine the tests under other error processes, inclusion of level breaks and different locations of the break points.

References

Apergis, N. and Payne, J. E. (2010) 'Structural breaks and petroleum consumption in US states: Are shocks transitory or permanent?', *Energy Policy*, 38(10), pp. 6375–6378.

Camarero, M., Carrion-i-Silvestre, J. L. and Tamarit, C. (2015) 'The relationship between debt level and fiscal sustainability in organization for economic cooperation and development countries', *Economic Inquiry*, 53(1), pp. 129–149.

Chang, M. J. and Su, C. Y. (2014) 'The dynamic relationship between exchange rates and macroeconomic fundamentals: Evidence from Pacific Rim countries', *Journal of International Financial Markets, Institutions and Money*, 30(1), pp. 220–246.

Chen, S. W. and Shen, C. H. (2015) 'Revisiting the Feldstein-Horioka puzzle with regime switching: New evidence from European countries', *Economic Modelling*, 49, pp. 260–269.

Choi, I. (2015) 'Inference on Unit Roots: Basic Methods', in *Almost all about unit roots: foundations, developments, and applications*. 1st edn. Cambridge: Cambridge University Press, pp. 16–57.

Christiano, L. J. (1992) 'Searching for a Break in GNP', *Journal of Business & Economic Statistics*, 10(3), pp. 237–250.

DeJong, D. N., Nankervis, J. C., Savin, N. E. and Whiteman, C. H. (1992) 'The power problems of unit root test in time series with autoregressive errors', *Journal of Econometrics*, 53(1–3), pp. 323–343.

Dickey, D. A. and Fuller, W. A. (1979) 'Distribution of Estimators for Autoregressive Time Series With a Unit Root', *Journal of the American Statistical Association*, 74, pp. 427–431.

Elliott, G., Rothenberg, T. J. and Stock, J. H. (1996) 'Efficient Tests for an Autoregressive Unit Root', *Econometrica*, 64(4), pp. 813–836.

Franses, P. H. and Haldrup, N. (1994) 'The effects of additive outliers on tests for unit roots and cointegration', *Journal of Business & Economic Statistics*, 12(4), pp. 471–478.

García-Cintado, A., Romero-Ávila, D. and Usabiaga, C. (2015) 'Can the hysteresis hypothesis in Spanish regional unemployment be beaten? New evidence from unit root tests with breaks', *Economic Modelling*, 47, pp. 244–252.

Hall, A. (1994) 'Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection', *Journal of Business & Economic Statistics*, 12(4), pp. 461–470.

Harris, R. I. D. (1992) 'Testing for unit roots using the augmented Dickey-Fuller test. Some issues relating to the size, power and the lag structure of the test', *Economics Letters*, 38(4), pp. 381–386.

Harvey, D. I., Leybourne, S. J. and Newbold, P. (2001) 'Innovational outlier unit root tests with an endogenously determined break in level', *Oxford Bulletin of Economics and Statistics*, 63(5), pp. 559–575.

Harvey, D. I., Leybourne, S. J. and Taylor, A. M. R. (2012) 'Unit root testing under a local break in trend', *Journal of Econometrics*, 167(1), pp. 140–167.

Harvey, D. I., Leybourne, S. J. and Taylor, A. M. R. (2013) 'Testing for unit roots in the possible presence of multiple trend breaks using minimum Dickey-Fuller statistics', *Journal of Econometrics*, 177(2), pp. 265–284.

Kim, T.-H., Leybourne, S. J. and Newbold, P. (2000) 'Spurious Rejections by Perron Tests in the Presence of a Break', *Oxford Bulletin of Economics and Statistics*, 62(3), pp. 433–444.

Lee, J. and Strazicich, M. C. (2001) 'Break point estimation and spurious rejections with endogenous unit root tests', *Oxford Bulletin of Economics and Statistics*, 63(5), pp. 535–558.

Lee, J. and Strazicich, M. C. (2003) 'Minimum Lagrange Multiplier Unit Root Test with Two Structural Breaks', *The Review of Economics and Statistics*, 85(4), pp. 1082–1089.

Leybourne, S. J., Mills, T. C. and Newbold, P. (1998) 'Spurious rejections by Dickey–Fuller tests in the presence of a break under the null', *Journal of Econometrics*, 87(1), pp. 191–203.

Liddle, B. and Messinis, G. (2015) 'Revisiting sulfur Kuznets curves with endogenous breaks modeling: Substantial evidence of inverted-Us/Vs for individual OECD countries', *Economic Modelling*, 49, pp. 278–285.

Lumsdaine, R. L. and Papell, D. H. (1997) 'Multiple trend break and the unit-root hypothesis', *The Review of Economics and Statistics*, 79(2), pp. 212–218.

Mishra, V. and Smyth, R. (2014) 'Is monthly US natural gas consumption stationary? New evidence from a GARCH unit root test with structural breaks', *Energy Policy*, 69, pp. 258–262.

Narayan, P. K. and Popp, S. (2010) 'A new unit root test with two structural breaks in level and slope at unknown time', *Journal of Applied Statistics*, 37(9), pp. 1425–1438.

Narayan, P. K. and Popp, S. (2013) 'Size and power properties of structural break unit root tests', *Applied Economics*, 45(6), pp. 721–728.

Ng, S. and Perron, P. (2001) 'Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power', *Econometrica*, 69(6), pp. 1519–1554.

Perron, P. (1989) 'The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis', *Econometrica*, 57(6), pp. 1361–1401.

Perron, P. (1997) 'Further evidence on breaking trend functions in macroeconomic variables', *Journal of Econometrics*, 80(2), pp. 355–385.

Perron, P. and Rodríguez, G. (2003) 'GLS detrending, efficient unit root tests and structural change', *Journal of Econometrics*, 115(1), pp. 1–27.

Perron, P. and Vogelsang, T. J. (1992) 'Nonstationarity and Level Shifts With an Application to Purchasing Power Parity', *Journal of Business & Economic Statistics*, 10(3), pp. 301–320.

Phillips, P. and Perron, P. (1988) 'Testing for a Unit Root in Time Series Regressions', *Biometrika*, 75(2), pp. 335–346.

Popp, S. (2008) 'New innovational outlier unit root test with a break at an unknown time', *Journal of Statistical Computation and Simulation*, 78(12), pp. 1145–1161.

Psaradakis, Z. (2001) 'Markov level shifts and the unit-root hypothesis', *Econometrics Journal*, 4(2), pp. 225–241.

Said, S. E. and Dickey, D. A. (1984) 'Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order', *Biometrika*, 71(3), pp. 599–607.

Salisu, A. A. and Mobolaji, H. (2013) 'Modeling returns and volatility transmission between oil price and US-Nigeria exchange rate', *Energy Economics*, 39, pp. 169–176.

Schwert, W. G. (1989) 'Tests for Unit Roots: A Monte Carlo Investigation', *Journal of Business & Economic Statistics*, 7(2), pp. 147–159.

Tiwari, A. K., Shahbaz, M. and Adnan Hye, Q. M. (2013) 'The environmental Kuznets curve and the role of coal consumption in India: Cointegration and causality analysis in an open economy', *Renewable and Sustainable Energy Reviews*, 18, pp. 519–527.

De Vita, G. and Trachanas, E. (2016) '"Nonlinear causality between crude oil price and exchange rate: A comparative study of China and India" - A failed replication (negative Type 1 and Type 2)', *Energy Economics*, 56, pp. 150–160.

Vogelsang, T. J. and Perron, P. (1998) 'Additional Tests for a Unit Root Allowing for a Break in the Trend Function at an Unknown Time', *International Economic Review*, 39(4), pp. 1073–1100.

Vougas, D. V. (2007) 'GLS detrending and unit root testing', *Economics Letters*, 97(3), pp. 222–229.

Zivot, E. and Andrews, D. W. K. (1992) 'Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis', *Journal of Business & Economic Statistics*, 10(3), pp. 251–270.