THE IMPACT OF MISDIAGNOSING A STRUCTURAL BREAK ON STANDARD UNIT ROOT TESTS: MONTE CARLO RESULTS FOR SMALL SAMPLE SIZE AND POWER

E Moolman and S K McCoskey*

Abstract

As discussed by Perron (1989), a common problem when testing for unit roots is the presence of a structural break that has not been accounted for in the testing procedure. In such cases, unit root tests are biased to non-rejection of the null hypothesis of non-stationarity. These results have been discussed using asymptotic theory and large samples in papers by Leybourne and Newbold (2000), Montanes and Reyes (1998) and Lee, Huang, and Shin (1997). In this paper we investigate the impact of ignoring structural breaks on sample sizes of more interest to empirical economists and show the results on power and size for both tests of the null of non-stationarity (ADF and Phillips-Perron) and the null of stationarity (KPSS). We are also able to give some guidelines on break placement which can cause the rapid flipping of rejection probabilities as discussed in Leybourne and Newbold (2000). Finally, we provide examples from time series data in South Africa to show the danger of misdiagnosis and the resulting misspecifications that can occur.

1. Introduction

The most important implication of the unit root revolution is that under this hypothesis random shocks have a permanent effect on the system. However, many such series may also contain structural breaks, and therefore it is important to assess carefully the reliability of unit root tests in the presence of a structural break. This paper compares the performance of unit root and stationarity tests in the presence of a structural break, allowing for both the null of stationarity and the null of non-stationarity. Such a comparison can be of use as the debate continues as to the appropriate null hypothesis for different applications.

*Respectively Senior Lecturer at the Department of Economics, University of Pretoria, Pretoria 0001, Republic of South Africa and Associate Professor at the United States Naval Academy, Maryland, USA and Honorary Professor at the University of Pretoria.
Email: emoolman@hakuna.up.ac.za

The most influential contribution with respect to the effect of structural breaks on unit root tests is found in Perron (1989). He considers the null hypothesis that a time series has a unit root against the alternative that the process is trend-stationary. Under both the null and alternative hypotheses a one-time change in the level and/or in the slope of the trend function is allowed, assuming that the time of the break is known. He shows that asymptotically standard tests of the unit root hypothesis against trend stationarity cannot reject the unit root hypothesis if the true data generating process (DGP) is that of stationary fluctuations around a trend function which contains a one-time break. Therefore these tests are biased towards accepting the null whether it is true or not, which means that it decreases the potential power of the tests.

Montanes and Reyes (1998) extend the results of Perron (1989) for the Dickey Fuller (DF) test and show with Monte Carlo studies that the unit root tests are biased in favour of non-rejection of the unit root hypothesis. In contrast with Perron (1989) where it is supposed that the effects of different types of structural breaks are identical, they showed that the DF tests react differently to different types of structural breaks. They also show that the tests react differently to breaks at the beginning of the sample and breaks at the end of the sample, and that the tests also react differently to different sizes of the breaks. The paper by Leybourne and Newbold (2000) also considers the ADF t-test and, like Montanes and Reyes, considers the importance of break placement. In their paper, the authors find that there is a potential break placement which causes a rapid flip from very many rejection to very few with a stationary DGP containing a structural brea. They further predict that in smaller samples, which they do no investigate, this flipping could occur in breaks which are in the first half of the data set. Identifying the exact placement of this break flipping could be an important tool for applied econometricians in interpreting testing results.

When the unit root is the null hypothesis to be tested, then the way in which classical hypothesis testing is carried out ensures that the null hypothesis is accepted unless there is strong evidence against it. Therefore an alternative explanation for the common failure to reject a unit root is simply that most economic time series are not very informative about whether or not there is a unit root, or equivalently, that standard unit root tests are not very powerful against relevant alternatives. For example, De Jong *et al.* (1989) provide evidence that the Dickey-Fuller tests have low power against stable autoregressive alternatives with roots near unity, and Diebold and Rudebusch (1990) show that they also have low power against fractionally integrated alternatives. By testing both the unit root hypothesis and the stationarity hypothesis, it is possible to distinguish series that appear to be stationary, series that appear to have a unit root and series for which the data (or the tests) are not sufficiently informative to be sure whether they are stationary or integrated. Kwaitkowski *et al.* (1992) (hereafter KPSS), considered Lagrange Multiplier (LM) tests with a stationary or trend stationary null hypothesis rather than a unit root hypothesis. However, the KPSS test of the null of stationarity faces a parallel problem as the null of non-stationarity in the presence of a structural break.

Lee, Huang and Shin (1997) clarify the nature of the effects of a structural break on the KPSS stationarity tests. They show that stationarity tests ignoring the existing break diverge and are biased toward rejecting the null hypothesis of stationarity in favor of the false alternative unit root hypothesis. This result can be compared with the findings of Perron (1989) and Montanes and Reyes (1998) that unit root tests are biased toward accepting the false unit root null hypothesis. The size distortion problem of stationarity tests parallels the power loss problem of unit root tests. However, like Perron (1989) they only looked at the asymptotic consequences and not the small sample case. In addition, Lee *et al*. (1997) only consider breaks in the first halve of the sample, but since the KPSS statistic does not treat the errors symmetrically the effect of a break in the second half of the sample may be different. Therefore, in this study the effect of breaks in both halves of the sample will be tested.

However, all the results described above have been established asymptotically and might differ substantially for small samples that are of practical interest to the applied economist. In this paper we expand on the results of Perron (1989) for unit root tests, and the results of Lee *et al*. (1997) for stationarity tests to show the small sample results on power and size in the presence of a structural break. Our testing will be more general than these tests, allowing for different types of structural breaks, different break magnitudes, different placements of the break and also more general error structures. The paper is outlined as follows: Section 2 summarizes tests and explains the Monte Carlo design for the comparison of the tests. Section 3 summarizes the results of the Monte Carlo experiments. Section 4 presents an application of the tests for South Africa, and Section 5 provides some concluding thoughts.

2. The Monte Carlo design

The goal of this Monte Carlo study is to compare the size and power of unit root and stationarity tests in the presence of a structural break, with special reference to the small sample properties. Three tests are considered: the Dickey Fuller (DF), Phillips-Perron (PP) and Kwaitkowsky *et al*. (KPSS). The first two tests are constructed under the null of a unit root and the last one under the null of stationarity. The power and size of these tests are compared across both the null of stationarity and the null of non-stationarity.

2.1 The test statistics

The DF and PP tests are tests of the null hypothesis of non-stationarity against the alternative of stationarity. The DF test involves estimating the autoregressive (AR) coefficient (ρ) of the dependent variable and then determining whether to accept or reject the null hypothesis of a unit root (ρ=1) by comparing the following statistic with the appropriate DF critical value:

DF t-statistic: $\frac{(\hat{\rho}-1)}{\hat{\sigma}_{\hat{\rho}_T}}$ … (1)

By definition, the DF test assumes that the data generating process (DGP) is an AR(1) process. To allow for a higher AR order, the augmented Dickey-Fuller (ADF) test is used. Both DF tests assume that the errors are i.i.d. Said and Dickey (1984) showed that the ADF test can be used when the error process is a moving average (MA). In this study the DF test will be used except when the errors contain an AR component, in which case the ADF test will be used.

The PP test is a generalization of the DF-procedure that allows for fairly mild assumptions concerning the error distribution, by allowing the errors to be weakly dependent and heterogeneously distributed. The test is performed by comparing the following test statistic to the relevant critical value:

$$t = \frac{\gamma_0^{1/2} t_0}{\omega} - \frac{(\omega^2 - \gamma_0) T s_0}{2\omega\hat{\sigma}} \quad \ldots (2)$$

where: $\omega^2 = \gamma_0 + 2\sum_{j=1}^{q}\left(1 - \frac{j}{q+1}\right)\gamma_j$ … (3)

$$\gamma_j = \frac{1}{T}\sum_{t=j+1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-j} \quad \ldots (4)$$

Although the ADF and PP tests have very low power especially for a near unity AR term or trend stationarity, Monte Carlo studies have shown that the PP test has greater power than the ADF test (Enders 1995:242). However, Monte Carlo studies have also shown that in the presence of negative MA terms, the PP test tends to reject the null of a unit root whether or not the actual DGP contains a negative unit root (Enders 1995, p. 242). It is therefore preferable to use the ADF test when the true model contains negative MA terms and the PP test when the true model contains positive MA terms. However, in practice the true DGP is never known, and therefore a safe choice is to use both tests.

Kwaitkowsky *et al.* (1992) (hereafter KPSS) developed a test of the null hypothesis that an observable series is stationary around a deterministic trend, against the alternative that the series is difference-stationary. They derived a one-sided LM-test under the assumption that the errors are white noise, however, since this is assumption is not credible in many applications they derived a modified version of the LM statistic that is valid under more general conditions. In the modified version they use a similar autocorrelation correction to the PP corrections.

Test statistic: $\hat{\eta}_\mu = \eta_\mu / s^2(\ell) = T^{-2}\sum S_t^2 / s^2(\ell)$ … (5)

where $s^2(\ell) = T^{-1}\sum_{t=1}^{T}\varepsilon_t^2 + 2T^{-1}\sum_{s=1}^{\ell} w(s,\ell)\sum_{t=s1}^{T}\varepsilon_t\varepsilon_{t-s}$ … (6)

Bartlett window: $w(s,\ell) = \ell - s/(\ell+1)$ … (7)

With i.i.d. errors, the KPSS test has approximately the correct size, except when T is small and the lag truncation parameter (ℓ) is large. The power of the test increases as T increases. There is a trade-off between correct size and power: choosing ℓ large enough to avoid size distortions in the presence of realistic amounts of autocorrelation will make the tests have very little power.

2.2 Construction of the hypotheses

As in Perron (1989), three different models are considered under both the null hypotheses: one that permits an exogenous change in the level of the series, one that permits an exogenous change in the time trend, and a mixed model that allows both changes. These hypotheses are parameterized as follows:

Under the null of a unit root:

Model A: $y_t = \mu + dD(TB)_t + y_{t-1} + \varepsilon_t$ … (8)

Model B: $y_t = \mu + y_{t-1} + \delta_1 DU_t + \varepsilon_t$ … (9)

Model C: $y_t = \mu_1 + y_{t-1} + dD(TB)_t + \delta_1 DU_t + \varepsilon_t$ … (10)

Under the null of trend stationarity:

Model A: $y_t = \mu_1 + \beta t + \delta_1 DU_t + \varepsilon_t$ … (11)

Model B: $y_t = \mu + \beta_1 t + \delta_2 DT_t^* + \varepsilon_t$ … (12)

Model C: $y_t = \mu_1 + \beta_1 t + \delta_1 DU_t + \delta_2 DT_t + \varepsilon_t$ … (13)

where:

T_B is the time of the break

$D(TB)_t$ = 1 if t = T_B+1, 0 otherwise … (14)

DU_t = 1 if >T_B, 0 otherwise … (15)

DT_t* = t – T_B if t > T_B, 0 otherwise … (16)

DT_t = t if $t > T_B$, 0 otherwise … (17)

In model A, the null hypothesis of a unit root is characterized by a dummy variable that takes the value one at the time of the break. Under the alternative hypothesis of trend stationarity, the model allows for a one-time change in the intercept of the trend function. Model B allows a change in the drift parameter at the time of the break under the null, as opposed to a change in the slope of the trend under the alternative. Model C allows for both effects to take place simultaneously, in other words a break consisting of a change in the level as well as a change in the time trend. All the testing will be performed under all three models, to test whether different types of breaks have significantly different effects on the unit root and stationarity testing.

2.3 Experiment dimensions

Different dimensions of the structural break will be tested in four experiments. The dimensions that will be considered are the sample size, the placement of the break, the size of the break, different error structures and the distance between the null and the alternative. To test the distance between the null and the alternative, a more general form of the stationary models will be used by adding an autoregressive dependent variable. The coefficient of this variable (ρ in (18) to (20)), will then be varied to change the distance between the null and the alternative.

The focus of this paper is on sample sizes that are of practical interest to the applied economist. In Experiment 1, the power and size of the tests are compared for different sample sizes, as the sample size (T) will take on the following values: $T \in \{15, 25, 50, 100\}$.

Montanes and Reyes (1998) and Leybourne and Newbold (2000) showed that the power and size of the DF tests are influenced by the placement of the break. To test whether small samples have the same results, and to test whether this is also the case with the PP and the KPSS tests, the effect of the placement of the break will also be tested. This is done by assigning the following values to the break fraction (λ), i.e. the ratio of pre-break sample size to total sample size: $\lambda \in \{0{,}1, 0{,}2, 0{,}3, 0{,}4, 0{,}5, 0{,}6, 0{,}7, 0{,}8, 0{,}9\}$.

Again, the results of Montanes and Reyes (1998) and Leybourne and Newbold (2000) indicate that the magnitude of the break has a significant impact on the rejection rate of the DF tests. In Experiment 2 the magnitude of the break will be varied tot test whether it is also the case in small samples and for the other tests, by using the following values for the break magnitudes δ_1 and δ_2: $\delta_1 \in \{1, 2, 3, 4\}$, $\delta_2 \in \{0{,}2, 0{,}4, 2, 4\}$.

The AR(1) parameter ρ is a convenient nuisance parameter to consider, since it naturally measures the distance of the null from the alternative. Since Monte Carlo studies have shown that unit root tests can not distinguish between non-stationary and stationary series when the AR coefficient (ρ) is close to unity (Enders 1995: 251), it will be tested whether ρ has a significant impact on the tests. The coefficient (ρ) will take on the following values: $\rho \in \{0, 0,25, 0,5, 0,6, 0,7, 0,8, 0,9\}$. In Experiment 3, a lagged dependent variable will be added to model A, B and C under stationarity, which give the following null hypotheses:

Model A: $y_t = \mu_1 + \beta t + \delta_1 DU_t + \rho y_{t-1} + \varepsilon_t$ … (18)

Model B: $y_t = \mu + \beta_1 t + \delta_2 DT_t^* + \rho y_{t-1} + \varepsilon_t$ … (19)

Model C: $y_t = \mu_1 + \beta_1 t + \delta_1 DU_t + \delta_2 DT_t + \rho y_{t-1} + \varepsilon_t$ … (20)

Without a structural break the tests perform differently with different error structures, for example, PP tends to always reject the null when the errors are negative MA (Enders 1995, p. 242), and Lee *et al.* (1997) have shown that the power of KPSS increases as σ increases with i.i.d. N(0, σ) errors. Therefore the tests' performances in the presence of a structural break will be tested with different error structures. In Experiment 4 different error structures will be used: an AR(p) structure, a MA(q) structure and an i.i.d. N(0, σ) structure where: $p \in \{0,75, 0,85, 0,95\}$, $q \in \{-0,8, 0, 0,8\}$, $\sigma \in \{0,25, 1, 4\}$.

Unless otherwise specified, the study considers λ=0.5 and T=25, and the errors that are i.i.d. N(0,1). The random number generator used for this study is URN22 from Karian and Dudewicz (1991). All the experiments were done with a 5% level of significance, and each experiment was repeated 10 000 times.

3. Monte Carlo results

All the results of the Monte Carlo study are presented as rejection rates, where a rejection rate is the percentage of times that a test rejects the null hypothesis. Because the tests are not all derived under the same null hypothesis, it is difficult to compare their performances directly. When the DGP is stationary, the rejection rate of the KPSS measures its size, while the rejection rate of the DF and PP measure their power. On the other hand, when the DGP is non-stationary, the rejection rate of the KPSS measures its power while the rejection rates of DF and PP measure their size. Table 1 provides an overall summary of the results.

3.1 Experiment 1

In Experiment 1 the size and power of the tests were compared for different sample sizes (T) to see whether the results of small samples corresponds to the asymptotic results. At the same time the influence of the break fraction (λ) was also tested. The rejection rates of the ADF, PP and KPSS tests for the models, A, B and C explained in Section 3, are given for a stationary DGP in Table 2 and for a non-stationary DGP in Table 3.

With model A, the ADF and PP tests performs very well under both hypotheses, as both has constant power of 0,9999 while the size was also within reasonable ranges.

For example, the size of ADF ∈ (0,048, 0,076) and the size of PP ∈ (0,05, 0,12). For both tests, the size distortions get worse as the T decreases. In other words, a break in the intercept doesn't entirely mislead the tests although it is a little bit more misleading in smaller samples. However, it seems as if the break makes the data look non-stationary to KPSS, since it is biased towards rejecting stationarity whether the DGP is stationary or not. For example, under stationarity the rejection rate of KPSS (i.e. size) ∈ (0,22, 0,99), while the rejection rate under non-stationarity (i.e. power) ∈ (0,45, 0,99). In model A, the break fraction (λ) did not have a significant impact on any of the tests, apart from slight size distortions with KPSS.

Table 1: Summary of results of experiments

Experiment	Parameter(s) tested	Results
1	T, λ	Model A: Tests perform consistent with asymptotic results. Model B & C: ADF and PP reject when break at beginning, ADF and PP don't reject when break at end. KPSS almost always rejects.
2	δ_2 δ_1	Insignificant Only significant under stationarity: ↑δ → ↑bias of tests toward finding series non-stationary.
3	ρ	↑ρ→↑bias toward finding series non-stationary

4	AR(p)
	MA(q)
	
	N(0,σ)
		PP better than DF.
		DF higher power than PP.
		DF better for negative q, PP better for positive q.
		DF better than PP;
		size and power of KPSS increase as σ increases

For data constructed according to model B, the ADF and PP tests seem to always reject the null, whether it is true or not, for breaks at the beginning of the sample. This result is consistent with the findings of Montanes and Reyes (1998) that asymptotically the test statistics converge to values that only permit the rejection of the unit root null hypothesis if the break is very close to the beginning of the sample. For breaks at the beginning of the sample ADF and PP tend to reject less as T decreases, because it converges slower to the theoretical values in smaller samples. For example, when λ=0,2 the rejection rate of DF decreases from 0,99 to 0,24 under stationarity while it decreases from 0,82 to 0,30 under non-stationarity, when T decreases from 100 to 15. However, when the break occurs in the first or second year in a very small sample (T=15 or T=25), the break doesn't mislead the tests and they have the normal characteristics, which means approximately the correct size and the usual low power. For example, when T=15 and λ=0,1 (i.e. a break in the first year) the size of DF is 0,06 while the power is 0,13. For breaks that appear later on, the ADF and PP tests converge to nearly zero rejection in larger samples (as predicted by Montanes and Reyes (1998)), but again they converge slower to the theoretical values in smaller samples.

This discussion allows us to identify crucial break placement values for which the rejection probabilities quickly flip from high to low rates of rejection. The presence of such flip values is an important result of Leybourne and Newbold. It is also interesting to see that these flip values, according to our results, depend on T and even the test used. For example, consider the DF test, if T=100, the rejection rate dramatically decreases from 0,99 to 0,01 as λ decreases from 0,2 to 0,4. For T=50, the rejection rate does not reach 0,01 until λ=0,5. Overall, the PP test seems to flip at higher λ than the DF test. For example, for T=100, the rejection rate is still as high as 0,57 when λ=0,5 and decreases to 0,07 for λ=0,6. In addition, in our simulations for small T (either 25 or 15) there is a return to high rejection rates for λ=0,9. This does not happen in the samples where T=50 or 100. Understanding the importance of the break placements may give applied researchers more incentive to use the PP test (at least in conjunction with the DF tests) as the PP tests have better power and are more useful under a wider variety of break placements. Note that in Model C the flip values occur for earlier break values, especially in the case of the DF test.

In model B, KPSS has very high power, which seems to decrease as the T decreases. For example, its power decreases from 0,99 to 0,54 when T decreases from 100 to 15, when λ=0,1. However, it shows some serious size distortions, consistent with the results of Lee *et al*. (1997). The size distortions tend to be smaller in very small samples, for example it decreases from 0,99 to 0,19 when T decreases from 100 to 15, when λ=0,4.

In model C, the change in the intercept (model A) is added to the change in the time trend (model B), and the results are consistent with that of only a change in the time trend (model B). This is consistent with the initial result for model A, namely that the change in the level doesn't mislead the tests. For all three tests, the power and size only change slightly but still have the same properties than with model B.

Table 2: Rejection rates with stationary DGP, for different T and λ (Experiment 1)

	Model A				Model B				Model C			
T=	**100**	**50**	**25**	**15**	**100**	**50**	**25**	**15**	**100**	**50**	**25**	**15**
Test						λ=0,1						
DF	0,99	0,99	0,99	0,99	0,99	0,98	0,17	0,13	0,99	0,80	0,08	0,06
PP	0,99	0,99	0,99	0,99	0,99	0,99	0,54	0,45	0,99	0,99	0,15	0,25
KPSS	0,96	0,74	0,62	0,54	0,45	0,13	0,86	0,54	0,99	0,93	0,99	0,98
						λ=0,2						
DF	0,99	0,99	0,99	0,99	0,99	0,91	0,78	0,24	0,43	0,20	0,08	0,09
PP	0,99	0,99	0,99	0,99	0,99	0,99	0,95	0,75	0,99	0,96	0,64	0,41
KPSS	0,99	0,92	0,70	0,52	0,99	0,43	0,14	0,12	0,99	0,99	0,89	0,66
						λ=0,3						
DF	0,99	0,99	0,99	0,99	0,54	0,65	0,08	0,10	0,03	0,07	0,00	0,00
PP	0,99	0,99	0,99	0,99	0,99	0,99	0,42	0,38	0,93	0,86	0,03	0,04
KPSS	0,99	0,89	0,86	0,80	0,99	0,89	0,95	0,65	0,99	0,99	0,99	0,99
						λ=0,4						
DF	0,99	0,99	0,99	0,99	0,01	0,22	0,27	0,21	0,00	0,02	0,04	0,06
PP	0,99	0,99	0,99	0,99	0,76	0,95	0,86	0,70	0,26	0,61	0,49	0,39
KPSS	0,97	0,77	0,51	0,39	0,99	0,99	0,47	0,19	0,99	0,99	0,96	0,68
						λ=0,5						
DF	0,99	0,99	0,99	0,99	0,00	0,01	0,03	0,08	0,00	0,00	0,00	0,00
PP	0,99	0,99	0,99	0,99	0,14	0,57	0,28	0,35	0,00	0,17	0,02	0,05
KPSS	0,93	0,66	0,69	0,66	0,99	0,99	0,99	0,69	0,99	0,99	0,99	0,99
						λ=0,6						
DF	0,99	0,99	0,99	0,99	0,00	0,00	0,03	0,10	0,00	0,00	0,00	0,02
PP	0,99	0,99	0,99	0,99	0,00	0,07	0,42	0,49	0,00	0,00	0,11	0,21
KPSS	0,97	0,77	0,51	0,39	0,99	0,99	0,99	0,54	0,99	0,99	0,99	0,95
						λ=0,7						
DF	0,99	0,99	0,99	0,99	0,00	0,00	0,01	0,05	0,00	0,00	0,00	0,00
PP	0,99	0,99	0,99	0,99	0,00	0,00	0,19	0,35	0,00	0,00	0,01	0,07
KPSS	0,99	0,90	0,62	0,50	0,99	0,99	0,99	0,72	0,99	0,99	0,99	0,99
						λ=0,8						
DF	0,99	0,99	0,99	0,99	0,00	0,00	0,00	0,03	0,00	0,00	0,00	0,00
PP	0,99	0,99	0,99	0,99	0,00	0,00	0,08	0,27	0,00	0,00	0,00	0,04
KPSS	0,99	0,92	0,69	0,52	0,99	0,99	0,99	0,80	0,00	0,00	0,67	0,15
						λ=0,9						
DF	0,99	0,99	0,99	0,99	0,00	0,00	0,53	0,21	0,00	0,00	0,67	0,15
PP	0,99	0,99	0,99	0,99	0,00	0,00	0,86	0,84	0,00	0,00	0,76	0,90
KPSS	0,96	0,74	0,33	0,22	0,00	0,99	0,97	0,38	0,99	0,99	0,81	0,99

Table 3: Rejection rates with non-stationary DGP, different T and λ (Experiment 1)

	Model A				Model B				Model C			
T	**100**	**50**	**25**	**15**	**100**	**50**	**25**	**15**	**100**	**50**	**25**	**15**
Test						λ=0,1						
DF	0,05	0,05	0,06	0,08	0,75	0,41	0,05	0,06	0,75	0,40	0,04	0,06
PP	0,06	0,07	0,08	0,13	0,83	0,53	0,19	0,10	0,82	0,52	0,18	0,10
KPSS	0,99	0,99	0,79	0,45	0,99	0,99	0,85	0,54	0,99	0,99	0,84	0,54
						λ=0,2						
DF	0,05	0,05	0,05	0,08	0,82	0,46	0,31	0,30	0,82	0,46	0,32	0,30
PP	0,06	0,07	0,08	0,12	0,90	0,56	0,37	0,41	0,90	0,56	0,37	0,42
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,77	0,99	0,99	0,99	0,75
						λ=0,3						
DF	0,05	0,05	0,05	0,08	0,27	0,11	0,12	0,20	0,27	0,11	0,12	0,21
PP	0,06	0,07	0,08	0,12	0,42	0,17	0,17	0,26	0,42	0,18	0,18	0,27
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,92	0,99	0,99	0,99	0,92
						λ=0,4						
DF	0,05	0,05	0,05	0,08	0,01	0,01	0,01	0,05	0,01	0,01	0,01	0,05
PP	0,06	0,07	0,08	0,12	0,01	0,01	0,02	0,06	0,01	0,01	0,02	0,07
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
						λ=0,5						
DF	0,05	0,05	0,05	0,08	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,02
PP	0,06	0,07	0,08	0,12	0,00	0,00	0,00	0,03	0,00	0,00	0,00	0,03
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
						λ=0,6						
DF	0,05	0,05	0,05	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PP	0,06	0,07	0,08	0,12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
						λ=0,7						
DF	0,05	0,05	0,06	0,08	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01
PP	0,06	0,07	0,08	0,12	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,98	0,99	0,99	0,99	0,99
						λ=0,8						
DF	0,05	0,05	0,05	0,07	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01
PP	0,06	0,07	0,08	0,12	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,99	0,93	0,99	0,99	0,99	0,94
						λ=0,9						
DF	0,05	0,05	0,05	0,08	0,00	0,00	0,04	0,06	0,00	0,00	0,04	0,06
PP	0,06	0,07	0,08	0,12	0,00	0,00	0,03	0,11	0,00	0,00	0,02	0,10
KPSS	0,99	0,98	0,79	0,45	0,99	0,99	0,92	0,59	0,99	0,99	0,93	0,62

3.2 Experiment 2

In Experiment 2 the effect of the magnitude of the break was tested. In Table 4 and 5, the results of the tests are summarized under the null of stationarity and the null of non-stationarity respectively. As δ_1, the size of the once-off change in the intercept, was increased, the power of ADF and PP decreased while the size distortions of KPSS got worse. When the data is already non-stationary, an increase in the magnitude of the break doesn't make the data appear more non-stationary. The results for the ADF and PP tests for model B and C are consistent with the results of Lee *et al*. (1997), namely that the power of the tests are not affected when different values of the magnitude of structural breaks (δ_1 and δ_2) are used.

3.3 Experiment 3

In Experiment 3 the effect of the distance between the null and the alternative was tested by varying the AR(1) coefficient (ρ) of the dependent variable when it was added to each of the three stationary models. The results are summarized in Table 6. The results of the power of the ADF and PP tests in model A are counter-intuitive as it decreases for ρ<0,5, and increases for ρ>0,5. With the changing time trend model, the power of ADF and PP decreases as ρ increases, since the data gives less evidence of not being stationary and therefore the tests reject less often. Overall the PP test seems to perform better than the ADF test, for example with model A power of PP $\in$ (0,195, 0,682) while the power of DF $\in$ (0,021, 0,159).

Table 4: Rejection rates with trend stationary DGP, for different δ_1 (Experiment 2)

Test	DF	PP	KPSS	DF	PP	KPSS	DF	PP	KPSS	DF	PP	KPSS
	δ_1=1			δ_1=2			δ_1=3			δ_1=4		
Model												
A	0,37	0,91	0,23	0,13	0,67	0,69	0,02	0,32	0,97	0,00	0,10	0,99
B	0,00	0,00	0,99	0,00	0,00	0,99	0,00	0,00	0,99	0,00	0,00	0,99
C	0,00	0,00	0,99	0,00	0,00	0,99	0,00	0,00	0,99	0,00	0,00	0,99

Table 5:Rejection rates with non-stationary DGP, for different δ_1 (Experiment 2)

Test	DF	PP	KPSS	DF	PP	KPSS	DF	PP	KPSS	DF	PP	KPSS
	δ_1=1			δ_1=2			δ_1=3			δ_1=4		
Model												
A	0,05	0,08	0,79	0,05	0,08	0,79	0,05	0,08	0,79	0,05	0,08	0,79
B	0,02	0,03	0,98	0,00	0,00	0,99	0,00	0,00	0,99	0,00	0,00	0,99
C	0,02	0,03	0,98	0,00	0,00	0,99	0,00	0,00	0,99	0,00	0,00	0,99

Table 6: Rejection rates with trend stationary DGP, for different ρ (Experiment 3)

	Model A			Model B			Model C		
	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
ρ									
0,00	0,131	0,671	0,694	0,027	0,284	0,989	0,001	0,022	0,999
0,25	0,070	0,287	0,815	0,013	0,071	0,994	0,000	0,001	0,999
0,50	0,034	0,080	0,918	0,006	0,012	0,997	0,000	0,000	0,999
0,60	0,026	0,048	0,960	0,004	0,007	0,998	0,000	0,000	0,999
0,70	0,021	0,048	0,994	0,003	0,004	0,998	0,000	0,000	0,999
0,75	0,024	0,081	0,999	0,002	0,003	0,998	0,000	0,000	0,999
0,80	0,044	0,195	0,999	0,001	0,002	0,999	0,000	0,000	0,999
0,90	0,159	0,682	0,999	0,000	0,000	0,999	0,000	0,000	0,999

As expected, the size distortions of KPSS get worse as ρ comes closer to 1, which means that the growing AR coefficient made the data appear even more non-stationary to the test. With model A, for example, the size of KPSS increased from 0,694 to 0,999 when ρ increased from 0 to 0,9.

With model C all three tests perform really badly, since the complexity of break structure, the inclusion of both types of breaks and the AR component, is completely misleading. In this case the power of DF are always less than 0,001 and the power of PP are always less than 0,022, while the size of KPSS was constant at 0,999.

3.4 Experiment 4

Most of the studies on unit root testing in the presence of a structural break assume that the errors are i.i.d. N(0,1). However, when different error structures were created in Experiment 4, the results (see Tables 7 and 8) indicate that the performances of the various tests are significantly influenced by different error structures.

When the errors are i.i.d. N(0,σ), DF has larger power and smaller size than PP. Consistent with the results of Lee *et al.* (1997), the power of KPSS increases as σ increases. The size of KPSS also increases as σ increases. For example, with model A the power of KPSS increased from 0,173 to 0,787 when σ increased from 0,25 to 1, and then to 0,997 when σ subsequently increased to 4. With model A, the size of KPSS increased from 0,119 to 0,694 when σ increased from 0,25 to 1, and then to 0,995 when σ increased to 4.

With AR error terms, the size of DF and PP increases and the power decreases as the AR coefficient (p) increases. However, p does not influence the size or power of the KPSS test. When the errors contain an AR component, PP has larger power and smaller size than DF.

Table 7: Rejection rates with trend stationary DGP, different error structures (Experiment 4)

	Model A			**Model B**			**Model C**		
	ADF	PP	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
AR(0.75)	0,039	0,088	0,999	0,019	0,040	0,999	0,003	0,007	0,999
AR(0.85)	0,036	0,067	0,999	0,016	0,03	0,999	0,003	0,008	0,999
AR(0.95)	0,030	0,052	0,999	0,016	0,024	0,999	0,004	0,008	0,999
MA(-0.8)	0,004	0,993	0,999	0,000	0,910	0,999	0,000	0,487	0,999
MA(0.00)	0,023	0,710	0,999	0,003	0,328	0,999	0,000	0,035	0,999
MA(0.80)	0,073	0,152	0,999	0,025	0,058	0,999	0,002	0,008	0,999
N(0,0.25)	0,003	0,099	0,119	0,000	0,003	0,997	0,000	0,000	0,999
N(0,1.00)	0,131	0,671	0,694	0,027	0,284	0,989	0,000	0,022	0,999
N(0,4.00)	0,372	0,910	0,995	0,242	0,786	0,999	0,067	0,473	0,999

Table 8: Rejection rates with non-stationary DGP, different error structures (Experiment 4)

	Model A			Model B			Model C		
	ADF[2]	PP[2]	KPSS	ADF	PP	KPSS	ADF	PP	KPSS
AR(0.75)	0,067	0,015	0,999	0,056	0,023	0,999	0,055	0,023	0,999
AR(0.85)	0,070	0,022	0,999	0,061	0,025	0,999	0,060	0,025	0,999
AR(0.95)	0,070	0,043	0,999	0,062	0,039	0,999	0,061	0,039	0,999
MA(-0.8)	0,102	0,927	0,999	0,001	0,000	0,999	0,102	0,927	0,999
MA(0.00)	0,049	0,078	0,999	0,016	0,007	0,999	0,049	0,078	0,999
MA(0.80)	0,091	0,016	0,999	0,039	0,018	0,999	0,091	0,016	0,999
N(0,0.25)	0,055	0,081	0,173	0,000	0,000	0,999	0,000	0,000	0,999
N(0,1.00)	0,054	0,081	0,787	0,002	0,004	0,999	0,002	0,004	0,999
N(0,4.00)	0,054	0,081	0,997	0,016	0,028	0,999	0,016	0,028	0,999

Normally the PP test rejects H_0 whether or not it is true in the case of a negative MA error structure. Our results in the previous experiments that the break of model A is not very misleading are confirmed, since PP has the normal property of over-rejecting with negative MA errors. But in model B and C, PP rejects less if the true DGP has a unit root, which means that the effect of the break is dominating the effect of the negative MA errors. When the errors contain a negative MA component, DF has smaller size than PP, while PP has smaller size then DF for a positive MA component. With model A, for example, the size of DF and PP are 0,102 and 0,927 for MA(-0,8) errors, while their sizes are 0,091 and 0,016

[2] For the ADF and PP tests 3 lags were used.

respectively for MA(0,8). The power of DF is always higher than that of PP when the errors contain a MA component. With model A for instance, the power of DF ∈ (0,004, 0,073) while the power of PP ∈ (0,152, 0,993).

The results of this experiment indicate that the relative performance of the tests is significantly influenced by the error structure. This is additional evidence that more than one test should be used in applied work when the true structure of the errors is usually unknown.

4. Does the South African economy really walk randomly?

The graph of the South African long-term interest rate (see Figure 1) confirms that a possible structural break occurred in 1985 when monetary policy in South Africa changed from direct control to market orientated monetary instruments (Botha 1997). The results of the Perron test for testing unit roots in the presence of a structural break, which are summarized in Table 10, confirms that the long-term interest rate is trend stationary with a structural break in 1985. The naïve researcher that misspecifies the break will do the DF, PP and KPSS tests, of which the results are given in Table 9. The DF and PP tests does not reject the null of a unit root when the variable is tested in levels, but rejects when the variable is first-differenced. In other words the researcher will regard the variable to be integrated of order one. The KPSS test, on the other hand, rejects the null of stationarity for the interest rate in levels. This will confirm the researcher's conclusion from the DF and PP tests that the interest rate is integrated of order one. However, if the break is correctly specified, and the Perron test for a unit root in the presence of a structural break is done accordingly, the researcher will reject the null of a unit root and therefore regard the variable trend stationary with a structural break.

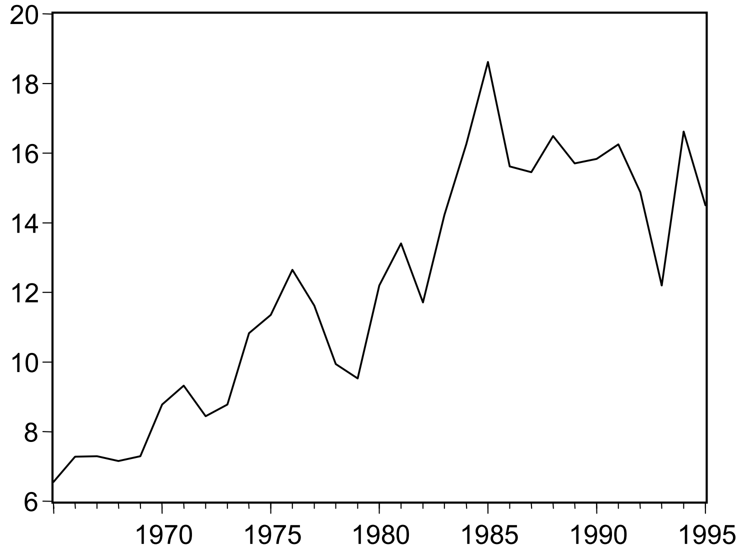

Figure 1: The South African long-term interest rate

The results of the DF, PP and KPSS tests for consumer price inflation are given in Table 9. Both the ADF and PP tests do not reject the null of a unit root in levels, but it rejects the null when inflation is first-differenced. KPSS rejects the null of stationarity in levels, therefore inflation is rendered non-stationary, which does not conform to economic theory. However, if there is a structural break in inflation, it invalidates the results of the ADF and PP tests. The results of the Perron test for unit roots in the presence of a structural break renders inflation trend stationary with a structural break in 1985. The inflation rate is therefore regarded as stationary, consistent with expectations that it cannot contain a unit root.

Table 9: Augmented Dickey-Fuller and Phillips-Perron unit root tests and KPSS test, levels

Series	Model	ADF Lags[3]	τ_τ,τ_μ,τ	ϕ_3,ϕ_1	PP[4]	KPSS
Long-term interest rate	Trend	3	-1,28	2,40	-3,24*	143,66***
	Constant	3	-1,36	2,81	-1,05	
	None	3	1,59	2,22	1,32	
Inflation rate	Trend	3	0,61	2,70	-0,35	1352***
	Constant	3	-1,69	2,01	-1,61	
	None	3	-0,16	1,50	-0,31	
Δlong-term interest rate	Trend	3	-3,39*	14,13***	-8,75***	0,18**
	Constant	3	-3,25**	17,43***	-8,84***	
	None	3	-2,70***	21,03***	-8,17***	
Δinflation rate	Trend	3	-3,61**	7,72***	-5,21***	1,28
	Constant	3	-2,53	7,07***	-4,79***	
	None	3	-2,55**	9,66***	-4,85***	

*/**/*** Significant at a 10%/5%/1% level

According to the Monte Carlo results in the previous section, DF and PP will be biased toward non-rejection and KPSS will be biased toward rejection in the presence of a structural break. The examples above are consistent with this Monte Carlo result, since DF and PP did not reject while KPSS rejected when the break was misspecified. The results of the Perron test for a unit root in the presence of a structural break were in both examples consistent with the *a priori* expectations of stationarity in the presence of a structural break.

[3] The number of lags used in the estimated equations was determined according to the method suggested by Said and Dickey (1984), which means a maximum of $T^{1/3}$ lags.

[4] The number of truncation lags used in the Bartlett kernel was determined as suggested by Newey-West. For this sample size Newey-West suggested 3.

Table 10: Perron Test[5] for non-stationarity in the presence of a structural break[6], levels

Series	T_B	λ	K	$\tilde{\mu}$ $(t_{\tilde{\mu}})$	$\tilde{\beta}$ $(t_{\tilde{\beta}})$	$\tilde{\gamma}$ $(t_{\tilde{\gamma}})$	$\tilde{\alpha}$[7] $(t_{\tilde{\alpha}})$	$T(\tilde{\alpha}-1)$
Long-term interest rate	1985	0,64	0	-851,71 (-14,04)	0,44 (14,21)	-0,42 (-4,75)	0,23 (-4,81***)	-29,26*
Inflation rate	1985	0,64	1	-1330,27 (-16,13)	0,68 (16,23)	-1,39 (-11,65)	0,3 (-4,68***)	-26,6*

*/**/*** Significant at a 10%/5%/1% level

5. Conclusion

In this paper we have shown with Monte Carlo simulations and practical examples from the South African economy that misdiagnosed structural breaks in small samples have serious consequences for unit root and stationarity testing. In the presence of a structural break the DF and PP tests are biased toward non-rejection of non-stationarity, while the KPSS test is biased toward rejection of stationarity. The DF and PP have low power in the presence of a structural break, while the KPSS test has serious size distortions. However, the different types of structural breaks are not misleading the tests to the same extent. A structural break consisting only of a change in the intercept is generally less disruptive as a structural break that includes a change in the slope.

The impact of certain dimensions of possible structural breaks has also been tested. When the break consists only of a change in the intercept, the effect of the small sample is dominating the effect of the break since DF and PP are still low power tests while KPSS has serious size distortions. However, when the break includes a change in the time trend, the effect of the break placement is dominating, since DF and PP are biased toward rejection when the break is at the beginning of the

[5] The version that tests H$_0$: $y_t = \mu_t + y_{t-1} + \delta_1 DU_t + \varepsilon_t$ against H$_a$: $y_t = \mu + \beta t + \delta_2 DT_t * + \varepsilon_t$, where DU_t=1 if t>T_B and 0 otherwise and DT_t*=t if t>T_B and 0 otherwise, was used.

[6] The parameters given are from the model:

$$y_t = \mu + \beta t + \gamma DT*_t + y_t ; y_t = \alpha y_{t-1} + \sum_{i=1}^{K} c_i \Delta y_{t-i} + \varepsilon_t$$

[7] Phillips and Ouliaris (1990) showed that t-ratio procedures diverge under that alternative at a slower rate than direct coefficient tests, which means that direct coefficient tests should have superior power properties over t-ratio tests. Therefore the T(α-1) test might have higher power than the α test so that both are reported. However, in this study they gave the same results.

sample, while they are biased toward non-rejection when the break occurs at the end of the sample. This is true regardless the sample size, although it is less biased in smaller samples. KPSS is not influenced by the placement of the break, and has serious size distortions for any type of break.

The magnitude of the break is only significant when the true DGP is stationary, since the data appears less stationary as the break increases. All the tests are increasingly biased towards finding the data non-stationary as the magnitude of the break increases. The performances of the tests are also significantly influenced by the error structure. This is additional motivation to use a combination of tests rather than only one to test the stationarity of a series.

References

De Jong, D N, Nankervis, J C, Savin, N E and Whiteman, C H (1989): "Integration versus Trend Stationarity in Macroeconomic Time Series", *Working paper no. 89-99, Department of Economics*, University of Iowa, Iowa City, IA.

Diebold, F X and Rudebusch, G D (1990): "On the Power of Dickey-Fuller Tests Against Fractional Alternatives", *Economics Letters*, 35**,** 155-160.

Enders, W (1995): *Applied Econometric Time Series*, John Wiley & Sons, Inc., New York.

Karian, Z A, and Dudewicz, E J (1991): *Modern Statistical, Systems, and GPSS Simulation: The First Course*, W.H. Freeman and Company, New York.

Kinderman, A J and Ramage, J G (1976): "Computer Generation of Normal Random Numbers," *Journal of the American Statistical Association*, 71, 893-896.

Lee, J, Huang, C J and Shin, Y (1997): "On Stationarity Tests in the Presence of Structural Breaks," *Economics Letters*, 55: 165-172.

Leybourne, S J and Newbold, P (2000): "Behaviour of Dickey-Fuller t-tests When There is a Break Under the Alternative Hypothesis," *Econometric Theory*, 16, 779-789.

Montanes, A and Reyes, M (1998): "Effect of a Shift in the Trend Function on Dickey-Fuller Unit Root Tests," *Econometric Theory*, 14, 355-363.

Perron, P (1989): "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, 57, 1361-1401.

Said, S and Dickey, D (1984): "Testing for Unit Roots in Autoregressive-Moving Average Models with Unknown Order", *Biometrics*, 71, 599-607.