Mathematics and Statistics 10(3): 670-683, 2022
DOI: 10.13189/ms.2022.100324 http://www.hrpub.org

Simulation-Based Assessment of the Effectiveness of Tests for Stationarity

Vasile-Alexandru Suchar[1,*], **Luis Gustavo Nardin**[2]

[1]Washington State Department of Ecology, Lacey, 98503, Washington, United States

[2]Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F-42023 Saint-Etienne France

Received January 29, 2022; Revised April 13, 2022; Accepted May 9, 2022

Cite This Paper in the following Citation Styles

***(a)**: [1] Vasile-Alexandru Suchar, Luis Gustavo Nardin, "Simulation-Based Assessment of the Effectiveness of Tests for Stationarity," Mathematics and Statistics, Vol.10, No.3, pp. 670-683, 2022. DOI: 10.13189/ms.2022.100324*

***(b)**: Vasile-Alexandru Suchar, Luis Gustavo Nardin, (2022). Simulation-Based Assessment of the Effectiveness of Tests for Stationarity. Mathematics and Statistics, 10(3), 670-683. DOI: 10.13189/ms.2022.100324*

Abstract Non-stationarity potentially comes from many sources and they impact the analysis of a wide range of systems in various fields. There is a large set of statistical tests for checking specific departures from stationarity. This study uses Monte Carlo simulations over artificially generated time series data to assess the effectiveness of 16 statistical tests to detect the real state of a wide variety of time series (i.e., stationary or non-stationary) and to identify their source of non-stationarity, if applicable. Our results show that these tests have a low statistical power outside their scope of operation. Our results also corroborate with previous studies showing that there are effective individual statistical tests to detect stationary time series, but no effective individual tests for detecting non-stationary time series. For example, Dickey-Fuller (DF) family tests are effective in detecting stationary time series or non-stationarity time series with positive unit root, but fail to detect negative unit root as well as trend and break in the mean, variance, and autocorrelation. Stationarity and change point detection tests usually misclassify stationary time series as non-stationary. The Breusch-Pagan BG serial correlation test, the ARCH homoscedasticity test, and the structural change SC tests can help to identify the source of non-stationarity to some extent. This outcome reinforces the current practice of running several tests to determine the real state of a time series, thus highlighting the importance of the selection of complementary statistical tests to correctly identifying the source of non-stationarity.

Keywords Stationarity Tests, Non-Stationarity, Time Series Analysis, Simulation

1 Introduction

Time series analysis contributes to the explanation, prediction and control over a wide range of systems in various fields. Many time series models assume that the data have been generated by a stationary stochastic process. But, if the data are not stationary, the uncertainty of the estimates of the model's parameters increases leading to inaccurate predictions.

A common source of non-stationarity is the unit root [1]. Unit root, however, is not the only source of non-stationarity and it is possible for time series to be non-stationary, yet not have a unit root. Figure 1 illustrates several examples of time series that violate the stationarity properties in terms of trend and break in mean, variance and autocorrelation, yet they have no unit root.

Several statistical tests were specifically developed to check for stationarity [2, 3, 4, 5, 6, 7]. Additional specialized statistical tests were proposed aiming to detect particular types of non-stationarity, such as unit roots, trends, structural breaks, or other time dependent patterns [8, 9, 10].

Non-stationarity comes from various sources, yet statistical tests are designed to detect specific sources of non-stationarity. To overcome this limitation, time series practitioners usually check a time series against multiple statistical tests and deem a time series stationary only if it passes all or a subset of these tests. This practice renders the selection of the tests relevant for the correct detection of stationarity and the source of non-stationarity for the proper selection and application of time series models.

Despite having a high statistical power to check for stationarity or detect specific sources of non-stationarity, the existing statistical tests show low statistical power for detecting all other sources of non-stationarity (see [11, 12, 13]). Besides, dif-

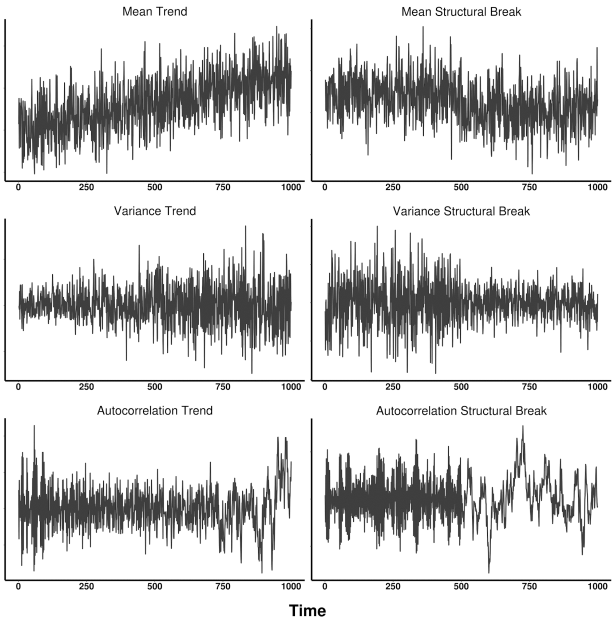

Figure 1. Types of non-stationarity time series.

ferent tests check different sources of non-stationarity, which renders problematic the comparison of their results. But even tests that check the same sources of non-stationarity (e.g., unit root) often produce different results [14], which confuses practitioners. These possible contradictory outcomes increase the importance of correctly choosing the statistical tests to check for stationarity. Such selection should be based both on the full understanding of the tests' assumptions and on the tests abilities to correctly detect the real state of the time series (i.e., stationary or non-stationary).

This study (1) summarizes the most common statistical tests applied to check for time series stationarity and non-stationarity and (2) assesses the ability of these tests to detect the real state of time series and to identify the source of non-stationarity, if applicable.

While the power of statistical tests has been previously assessed [14, 15, 16, 17, 18, 19, 13, 20], to the best of our knowledge, this is the first time that such large set of statistical tests was assessed for various sources of non-stationarity, such as unit roots, and trend and break in mean, variance and autocorrelation.

The paper unfolds as follows. Section 2 briefly describes the various individual statistical tests used in detecting the real state of a time series. The simulation study and the time series data generating model are described in Section 3. Section 4 reports on the results of the individual statistical tests in terms of their detection effectiveness. Section 5 discusses the overall set of results and concluding remarks are provided in Section 6.

2 Tests for Stationarity and Non-Stationarity

A stochastic process $X_t, t \in Z$ is said to be stationary, or second-order stationary, if the first two moments (i.e., the mean and the covariance) are time-invariant [21]. This does not mean that the observed properties of the time series (i.e., mean, variance and autocorrelation structure) do not change over time, but the way they change do not change over time.

Next, we describe a range of individual statistical tests specifically developed or used to check for stationarity or different sources of non-stationarity.

2.1 Unit Root Tests

The most known unit root tests are those from the Dickey-Fuller (DF) family. These tests assess the null hypothesis that the time series contains a unit root (i.e., non-stationary time series) against the alternative hypothesis that it does not contain a unit root (i.e., stationary time series) [22]. The original Dickey-Fuller (DF) test [23, 24] assesses a time series variable characterized by an AR(1) process for

- a unit root,
- a unit root with drift, and
- a unit root with drift and deterministic trend.

In the same family, the Augmented Dickey-Fuller (ADF) test was designed for time series characterized by ARIMA(p, 1, q) processes with unknown orders p and q [25].

Both the DF and ADF tests were designed for autoregressive processes with finite-variance disturbances [26]. For an unspecified autocorrelation and heteroscedasticity in the disturbance process, the Phillips-Perron (PP) test was proposed [27]. The PP test improves on the previous tests because it does not require the specification of the lag length since it addresses the higher order of autocorrelation by making a non-parametric correction to the t-test statistic.

Using different parametrization requirements than the DF family tests, Schmidt and Philips [28] developed the Schmidt-Philips (SP) test with higher power to detect unity roots in the presence of a deterministic trend.

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test was designed to assess whether a time series is stationary around a deterministic trend [29]. But contrary to the other unit root tests, the KPSS test has trend stationarity as the null hypothesis and unit root as the alternative hypothesis. Some studies propose using the KPSS test in addition to the unit root tests of the DF family since the former is less affected by the sample size. However, the results of the KPSS test with a unit root test of the DF family in tandem are difficult to interpret.

ADF, PP and KPSS tests have low power against highly persistent stationary processes (i.e., autoregressive processes where ϕ is close to 1) [30, 17, 31]. The Elliot-Rothenberg-Stock (ERS) test overcomes this drawback by de-trending (de-meaning) the time series using a generalized least-squares regression to estimate the deterministic component over which it performs a usual ADF test [12].

2.2 Stationarity Tests

Stationarity tests determine whether or not a time series is stationary by testing if its spectral density is constant in time.

One of the first tests for stationarity was the Priestley-Subba Rao (PSR) test [6]. The PSR test analyzes the logarithm variance of the spectral density function to determine if the spectrum varies in time. The null hypothesis defines that the spectrum varies in time (i.e., non-stationary) and the alternative hypothesis defines the contrary (i.e., stationary). Using a different approach, von Sachs and Neumann [7] developed a test that examines the Haar wavelet coefficients in time over a set of smoothed frequencies to assess the constancy of the time varying spectrum. Nason [5] also used Haar wavelet to develop a method using the wavelets over a finite set of scales allowing to test for stationarity hypothesis of both Gaussian and non-Gaussian time series. Nason's test has the advantage of identifying the location and time scale where non-stationarity is present.

2.3 Serial Correlation Tests

Serial correlation tests check the presence of autocorrelation in a time series [32]. A common serial correlation test used for this purpose is the Durbin-Watson (DW) test [33]. Although commonly used for checking non-stationarity, this test is not suitable for lagged dependent variables and cannot account for higher orders of serial correlation. The Breusch-Godfrey (BG) test account for higher orders of serial correlation by testing the error terms of a fitted regression model [34, 35]. Similarly, the Ljung-Box (LB) test checks whether any of a group of autocorrelations of the residual time series are different than zero [36].

2.4 Homoscedasticity Tests

Homoscedasticity tests check the homogeneity of the variance of a time series process variable. Two homoscedasticity tests used to check for non-stationarity are Breusch-Pagan (BP) test [37] and ARCH test [38]. The BP test checks if the variance of errors from the model fit are dependent on the values of the independent variables. While the ARCH test checks if a time series exhibits conditional heteroscedasticity or autocorrelation in the squared residuals.

2.5 Structural Change Tests

Structural change tests identify unexpected shifts in the mean, variance or autocorrelation of a time series, which are a clear indication of non-stationarity. Ross [39] implemented several two-sample mean and variance change point tests (i.e., CPM and CPV tests) for a wide variety of unknown distributions (e.g., CPM-STUDENT and CPV-BARTLETT tests). Such tests require the modeling of any underlying autocorrelation in the data followed by the change detection over the residuals of the model. Killick and Eckley [40] implemented single point detection tests using the maximum log likelihood ratio (AMOC), and multiple change point detection tests using the pruned exact linear time (PELT) method and cumulative sum of residuals (CUSUM). These methods can test for both structural changes in mean (i.e., CPTM-AMOC, CPTM-PELT and CPTM-CUMSUM tests) and in variance (i.e., CPTV-AMOC, CPTV-PELT and CPTV-CUMSUM tests).

Using a different approach, Killick et al. [41] also implemented tests for structural changes (i.e., SC) in regression. These tests first generate a linear model of the data and then tests if the coefficient vector varies in time. Three methods are proposed to conduct this hypothesis testing: cumulative sum of residuals (CUSUM), moving sum of residuals (MOSUM) and moving estimates (ME) tests, with both ordinary least squares (SC-OLS-CUSUM, SC-OLS-MOSUM, SC-OLS-ME) and recursive residuals (SC-RR-CUSUM, SC-RR-MOSUM, SC-RR-ME) [42, 43]. The SC-OLS-MOSUM method evaluates a fixed sum of residuals and the SC-OLS-ME method estimates the regression coefficients based on a moving bandwidth or data window of size proportional to the entire time series sample size [43, 44]. Despite the effectiveness of the SC tests in detecting structural changes, these tests cannot identify the actual source of change present in the time series.

3 Simulation Study

In this section we describe the simulation experiment performed to assess the effectiveness of various individual statistical tests in correctly detecting the real state of time series (i.e., stationary or non-stationary).

The study used artificial time series data to perform the assessment. The artificial time series data were generated using the following data generating model

$$y_t = \delta + \frac{\tau t}{T} + \sum_{t>1,i=1}^{p} \phi_{t,i} y_{t-i} + \epsilon_t + \sum_{t>1,j=1}^{q} \theta_{t,j} \epsilon_{t-j}, \quad (1)$$

$$\epsilon_t = \mathcal{N}\left(\mu, \sigma + \frac{\omega t}{T}\right), \quad (2)$$

where y_t is the dependent variable, $t = 1, 2, \ldots, T$ is the time, δ is the mean constant, τ is the trend in mean constant over time, ϕ is the autoregressive parameter, p is the order of autoregressive, θ is the moving average parameter, q is the order of the moving average and ϵ_t is the normal error with mean μ, variance σ and constant trend in variance over time ω.

Table 1 lists the individual statistical tests with their respective R functions and packages used in this study.

We assessed the effectiveness of these individual statistical tests for stationary AR(1) time series without drift and seven other types of AR(1) time series with different sources of non-stationarity:

1. unit root (Unit Root),
2. trend in mean (Trend Mean),
3. trend in variance (Trend Variance),
4. trend in autocorrelation (Trend AC),
5. structural break in mean (Break Mean),
6. structural break in variance (Break Variance) and
7. structural break in autocorrelation (Break AC).

Table 1. Individual statistical tests assessed and their function in R programming language [45]. `x` denotes the simulated data, `xlag` is lag 1, `res` are the residuals of AR(1) model and `reslag` is lag 1 of residuals of AR(1) model.

Test name	R function	R package
Dickey-Fuller (DF)	`adf.test(x, k=0)`	`tseries`
Augmented Dickey-Fuller (ADF)	`adf.test(x)`	`tseries`
Phillips-Peron (PP)	`pp.test(x)`	`tseries`
Schmidt-Phillips (SP)	`ur.sp(x)`	`urca`
Elliot-Rothenberg-Stock (ERS)	`ur.ers(x)`	`urca`
Kwiatkowski-Phillips-Schmidt-Shin (KPSS)	`kpss.test(x)`	`tseries`
Priestley-Subba Rao (PSR)	`stationarity(x)`	`fractal`
Wavelet (WAVELET)	`hwtos2(x)`	`locits`
Bootstrap (BOOTSTRAP)	`BootTOS(x)`	`costat`
Breusch-Godfrey (BG)	`bgtest(x ~ 1 + xlag)`	`lmtest`
Ljung-Box (LB)	`Box.test(res, lag=1)`	`stats`
ARCH Lagrange Multiplier (ARCH)	`ArchTest(res, lags=1)`	`FinTS`
Breusch-Pagan (BP)	`bptest(res ~ reslag)`	`lmtest`
Detect Change Point (CPM-STUDENT, CPV-BARTLETT)	`detectChangePoint(res)`	`cpm`
Change Point Mean (CPTM-[AMOC, PELT, CUMSUM])	`cpt.mean(x)`	`changepoint`
Change Point Variance (CPTV-[AMOC, PELT, CUMSUM])	`cpt.var(x)`	`changepoint`
Structural Change (SC-[OLS, RR]-[CUSUM, MOSUM, ME])	`efp(x ~ xlag)`	`strucchange`

We used the data generating model in Eq. (1) to generate these AR(1) time series for sample sizes $T = 32, 64, 128, 256, 512, 1024, 2048, 4096$ and $\phi = 0.99, 0.9, 0.8, 0.5, 0.2, 0, -0.2, -0.5, -0.8, -0.9, -0.99, 1, -1$. Parameter values specific to generating the stationary and each type of non-stationary time series are shown in Table 2; if not specified, the values used for these parameters are $\tau = 0$, $\omega = 0$, $\delta = 0$, $\theta = 0$, $\mu = 0$, $\sigma = 1$, $p = 1$ and $q = 0$.

The Trend Mean and Trend Variance used a constant value of ϕ_t (i.e., $\phi_t = \phi$) to generate the time series, while the Trend AC varies ϕ_t value in time, thus $\phi_t = \phi_1 + ((\phi_1 - \phi_2) \times t)/T)$. The structural break time series were generated using the constant values τ_1, ω_1 and ϕ_1 for the first half of the time series generation process and the constant values τ_2, ω_2 and ϕ_2 for the second half.

We ran 1,000 replications of each combination of ϕ and sample size with the parameter values listed in Table 2. For each generated time series, we applied the individual statistical tests shown in Table 1 assuming significance level of 0.05. Particularly for the SC tests using the SC-OLS-MOSUM and SC-OLS-ME methods, we also ran the tests for bandwidths from 10% to 90% in steps of 10%.

Depending on the individual test, the stationarity hypothesis is either the null hypothesis (i.e., KPSS, stationarity, serial correlation, homoscedasticity and structural change tests) or the alternative hypothesis (i.e., DF family unit root tests). Thus, instead of recording the outcome of the tests in terms of their hypothesis, the results were recorded in a binary format where 0 means the time series was identified as non-stationary and 1 as stationary. The results were summarized as the proportion of times out of 1,000 replications that the time series was identified as stationary independent of the time series real state. Summary values close to 1 indicate the test identified the time series mostly as stationary, while values close to 0 indicate the time series was mostly identified as non-stationary.

We used these values to calculate the *detection effectiveness* of each statistical test. *Detection effectiveness* refers to the correct detection of the actual time series state (i.e., stationary or non-stationary).

4 Results

In this section, we present a summary of the results of the simulation study conducted in Section 3. For the detailed results, we refer the reader to the *Supporting Information* (*SI*) material[1].

Although the results are discussed in terms of the *detection effectiveness* of the statistical tests, the plots show the percentage of times that the statistical test identified the time series as stationarity (i.e., % Detected Stationarity) irrespective of the real state of the times series. We opted for using this metric to standardize the presentation of the results since the assessed statistical tests adopt different null hypotheses, which renders confusing the comparison and discussion in terms of their empirical size and power. Thus, if the time series assessed is stationary (e.g., Figure 2 Panels A-D), an effective test should have the % Detected Stationarity value close to 100%. But, if the time series is non-stationary (e.g., Figure 2 Panels E-F), an effective test should have the % Detected Stationarity value close to 0%.

4.1 Stationary Time Series

Figure 2 Panels A-D shows the results for different stationary time series. Overall, the performance of the unit root tests and PSR test improved with the increase of sample size, while

[1]The Support Information material is available at `https://tinyurl.com/2p89spky`

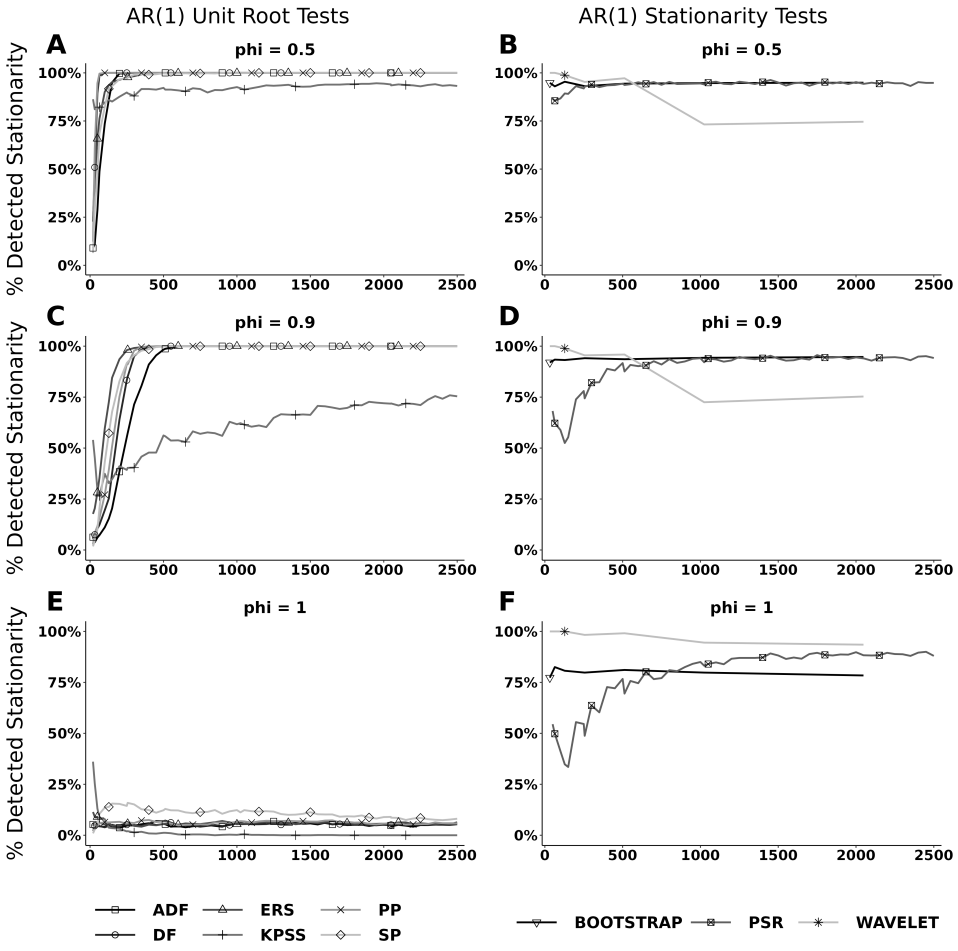

Figure 2. Detection effectiveness of the Unit Root and Stationarity tests for stationary and unit root time series.

Table 2. Parameter values used to generate the stationary and non-stationary time series.

Type	Parameter Values
Stationary	$\phi = \{0.99, 0.9, 0.8, 0.5, 0.2, 0, -0.2, -0.5, -0.8, -0.9, -0.99\}$
Unit Root	$\phi = \{1, -1\}$
Trend Mean	$\tau = \{0, 0.5, 1, 2, 3, 4\}$
Trend Variance	$\omega = \{0, 0.5, 1, 2, 3, 4\}$
Trend AC	$(\phi_1, \phi_2) = \{(0.9, -0.9), (0.45, -0.45)\}$
Break Mean	$(\tau_1, \tau_2) = \{(0,0), (0,0.5), (0,1), (0,2), (0,3), (0,4)\}$
Break Variance	$(\omega_1, \omega_2) = \{(0,0), (0,0.5), (0,1), (0,2), (0,3), (0,4)\}$
Break AC	$(\phi_1, \phi_2) = \{(0.9, 0.45), (0.9, 0), (0.9, -0.45), (0.9, -0.9), (0.45, 0),$ $(0.45, -0.45), (0.45, -0.9), (0, -0.45), (0, -0.9), (-0.45, -0.9)\}$

the performance of the WAVELET and BOOTSTRAP tests decreased with the increase of sample size.

The stationarity and unit root tests showed a high detection effectiveness. The performance of the unit root tests of the DF family (i.e., DF and ADF tests) improved with the increase of the sample size, while the KPSS test performance remained about the same regardless of the sample size for small values of ϕ. Although impervious to the sample size, the KPSS test had very low detection effectiveness for highly persistent autoregressive time series (see *SI Table 1 Stationarity*). The WAVELET and BOOTSTRAP tests outperformed the unit root tests for small sample sizes. The performance of the WAVELET test declined as the sample size increased above 512 observations. The BOOTSTRAP test had low detection power for highly persistent negative autoregressive values ($\phi \leq -0.9$), while unit root tests had lower detection power for highly persistent positive autoregressive values ($\phi \geq 0.9$) (see *SI Table 1 Stationarity*).

4.2 Non-Stationary Time Series: Unit Root

Figure 2 Panels E-F shows the results of non-stationarity unit root time series when $\phi = 1$. All unit root tests correctly detected the true state of the time series for $\phi = 1$, but they failed for $\phi = -1$ (see *SI Table 2 Unit Root*). The performance of the tests of the DF family and the KPSS test improved with the increase of the sample size, with the former outperforming the latter. The BOOTSTRAP and WAVELET tests detected incorrectly times series with $\phi = 1$ as stationary, but detected correctly time series with $\phi = -1$ as non-stationary for samples larger than 256 observations (see *SI Table 2 Unit Root*). The PSR test had a low detection effectiveness of unit root time series at small sample sizes, which also decreased as the sample size increased.

4.3 Non-Stationary Time Series: Trend and Break in Mean

Figure 3 shows selected results of non-stationarity time series with trends and breaks in the mean.

The BOOTSTRAP performed the best among the stationarity tests in detecting trends and breaks in the mean. The WAVELET test exhibited improved detection effectiveness as the sample size increased, but this might have been due to the sample size rather than the actual capacity of detecting trends in the mean. For time series with $\phi = 0$ and mean trends, all tests performed poorly for sample sizes smaller than 1,500, except for the CPM-STUDENT test that was able to detect the mean trends correctly for samples greater than 1,000. As the ϕ value changed (Figure 3 Panels C and E), the CPM-STUDENT test performance remained largely the same, but the SC and CPTM tests started to perform better. In general, no test detected the trend in the mean for very small samples. The SC tests had high detection effectiveness only for small samples and for larger samples. The CPM-STUDENT test had problems detecting breaks in the mean for negative ϕ autoregressive time series (see *SI Table 3 Trend in Mean*) because of the nature of the test (i.e., a two-sample t-test over sequentially increasing sample sizes) and its inability to distinguish between the high within-variability of highly persistent series from the between-samples variability (results not shown). Similar trends were present in the detection of trend and break in the mean, although, in general, break in the mean required smaller sample sizes than the trend in the mean to be detected (see *SI Table 3 Trend in Mean and Table 4 Break in Mean*).

4.4 Non-Stationary Time Series: Trend and Break in Variance

Figure 4 shows selected results of non-stationarity time series with trend and break in the variance.

The CPTV structural change tests increased their detection effectiveness with the increase of the sample size or the increase of the trend or break magnitudes. They also had, in general, the best detection effectiveness of trend and break in the variance for small sample sizes. The CPV-BARTLETT test performance was similar to the CPTV-PELT and the performance of the ARCH test for trend in the variance with sample sizes larger than 2048 and magnitude larger than three standard deviations, and for break in the variance with sample sizes larger than 1024 and magnitude larger than one standard deviation (see *SI Table 5 Trend in Variance* and *Table 6 Break in Variance*). The SC-OLS-ME tests and BOOTSRAP tests failed to detect most of the trend or break in the variance. Similar

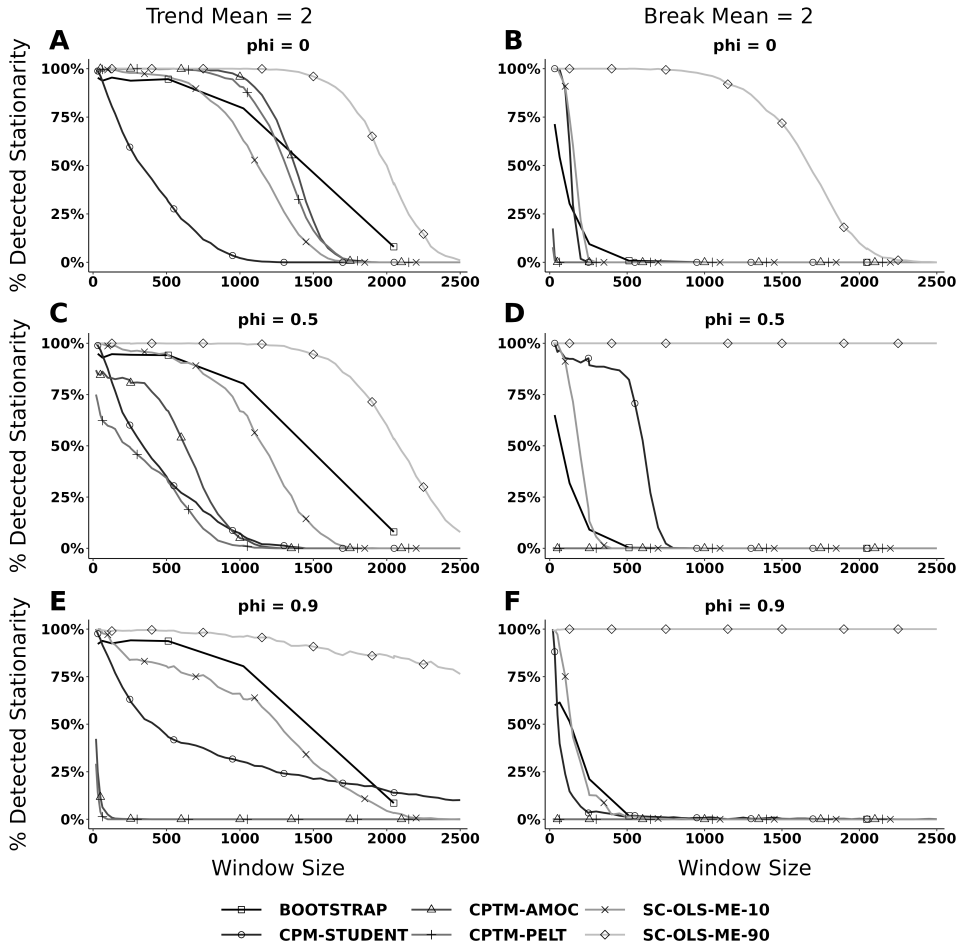

Figure 3. Detection effectiveness of statistical tests for non-stationarity time series with trend and break in the mean.

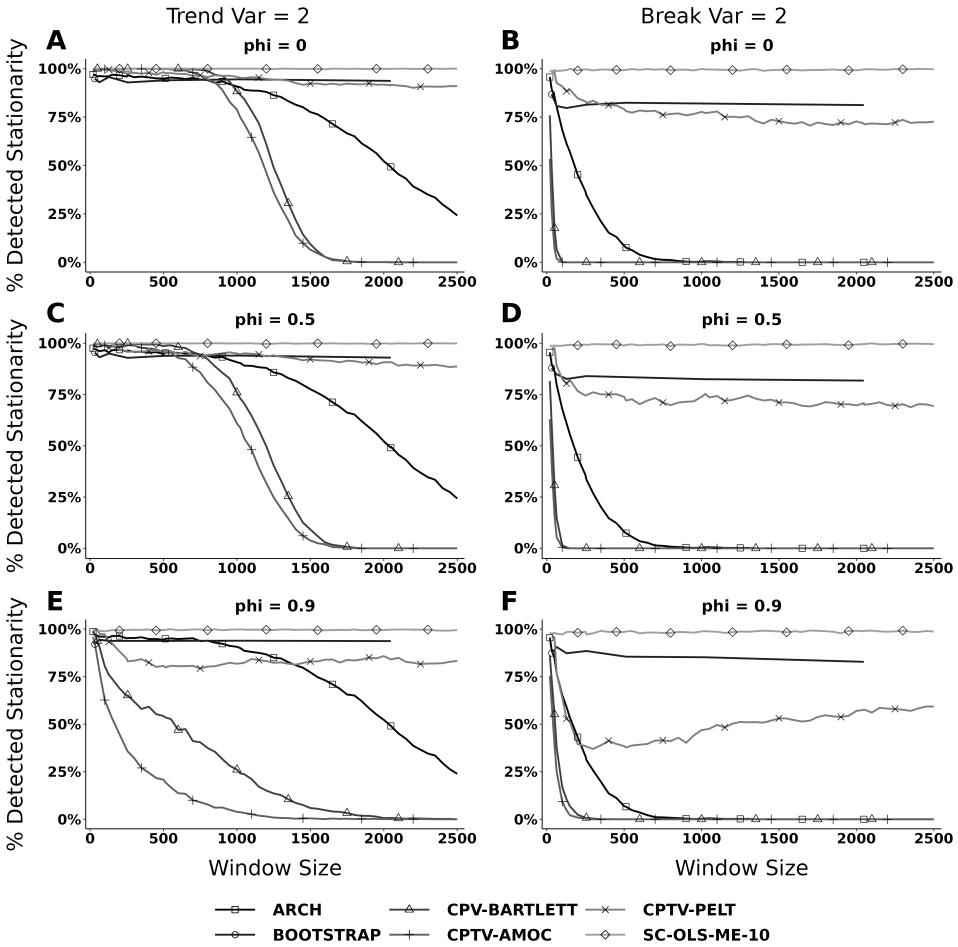

Figure 4. Detection effectiveness of statistical tests for non-stationary time series with trend and break in the variance.

patterns were present in the detection of trend and break in the variance, although, in general, break in the variance required smaller sample sizes than trend in the variance (see *SI Table 5 Trend in Variance* and *Table 6 Break in Variance*).

4.5 Non-Stationary Time Series: Trend and Break in Autocorrelation

Figure 5 shows selected results of non-stationarity time series with trend and break in the autocorrelation.

The BG test detection effectiveness increased with the increase of the sample size. The SC-OLS-ME structural change tests with small search windows had the highest detection effectiveness for smaller sample sizes. The LB and BOOTSRAP tests failed to detect most of the trend or break in the autocorrelation (see *SI Table 7 Trend in AC* and *Table 8 Break in AC*).

5 Discussion

Non-stationarity comes from potentially many sources. There is a wide variety of statistical tests designed for detecting specific sources of non-stationarity. Our simulation results show that these tests perform well in detecting the source of non-stationarity they have been designed for, especially for larger sample sizes. For example, DF family tests detect unit roots, and mean changes tests detect mean breaks and trends. But these tests show in our results low detection effectiveness outside their scope of operation. For example, unit root tests were not able to detect any other source of non-stationarity beside unit roots (for detailed results, we refer the reader to the *Supporting Information* (*SI*) material). Additionally, tests designed to detect the same type of non-stationarity may disagree. For example, Afriyie et al. [14] explored extensively the performance of unit root tests and they recommend the use of KPSS test when unit root tests disagree. Afriyie et al. [14], however, did not explore the case of other sources of non-stationarity. Our results corroborate their conclusions and extended them by comparing a larger variety of statistical tests and non-stationarity time series.

More general tests such as Stationarity tests and Structural Change (SC) tests are theoretically appealing since they claim to be able to detect whether a time series is stationary or not regardless of the source of non-stationarity. However, our results indicated that their performance are not consistent for all sources of non-stationarity. For example, the stationarity tests PSR, BOOTSTRAP and WAVELET were able to correctly detect stationary time series for small sample sizes, but they failed to detect positive unit roots, and trends and breaks in the variance and autocorrelation. Additionally, the SC tests failed to detect most trends and breaks in the variance.

Overall, the results indicate that the statistical tests have higher detection effectiveness for larger sample sizes and larger trend or break changes in the mean, variance or autocorrelation regardless of the source of non-stationarity.

Since there is no single designed test capable of detecting all sources of non-stationarity, time series practitioners must employ several tests to reach a conclusion about the state of a time series. The selection of these tests, however, must not be done solely on the basis of the tests assumptions (i.e., unit root tests used for unit root time series), but also based on the time series characteristics for which these statistical tests provide erroneous or questionable results. For example, we observed that the tests of the DF family performed well in detecting stationary time series and non-stationary unit root time series, but how do they perform in case of non-stationary time series with other sources of non-stationarity?

Figure 6 summarizes in a heatmap the percentage of times that time series were detected as stationary for all tests against stationary and all sources of non-stationarity (unit roots, and trend and break in the mean, variance and autocorrelation). For example, the DF test (first row) identified the ARMA ($\phi = 0.99$) time series (column 1) as stationary between 50% – 75% of the time, identified ARMA ($\phi = 1$) time series (column 4) as stationary between 0% – 25% of the time, and identified ARMA ($\phi = 0$, trend in the mean = 1), a non-stationary time series with trend in the mean (column 6), as stationary between 75% –100% of the time.

Figure 6 still shows that the tests of the DF family perform very well in the case of stationary series and positive unit root time series. But they are not able to detect any other source of non-stationarity. This suggests that the actual interpretation of the results generated by these tests should be that the time series is either (1) non-stationary with positive unit root or (2) stationary, or non-stationary with negative unit root, or non-stationary with trend in the mean, variance, or autocorrelation. Hence, if the time series has any source of non-stationarity different than unit root, the test cannot detect and wrongly assume the time series is stationary.

On the other extreme, tests as the CPM-STUDENT and the CPV-BARLETT detect all the time series as either having a trend or break in the mean or variance regardless of their source of non-stationarity and their real state. The CPM-STUDENT test is designed to detect differences in mean, while the CPV-BARLETT test is designed to detect differences in variance. When assessing these tests only against the trend and break in the mean and variance (Figures 3 and 4), we might conclude that they perform well. But when assessing them against time series with other sources of non-stationarity, we saw that they perform poorly and misclassify most time series. Thus, the use of these tests in time series assessments are detrimental and might lead to erroneous conclusions.

All tests suffer from misclassification problems. The ARCH test, for example, performs well in detecting trend and break in the variance, but it also detects trends and breaks in the autocorrelation. If used alone, the actual binary result of the ARCH test should be interpreted that the time series (a) is stationary, or it is non-stationary with a unit root or trend or break in the mean, or (b) has trend or break in the variance or autocorrelation. Thus, based on the results of just an individual test, it is challenging to identify the source of non-stationarity or even accurately detect if the time series is non-stationary.

Another important problem is the misclassification of stationary time series as non-stationary. Such an example can be observed in the results of the CPTV tests. These tests identify the trend and break in the variance as intended, although

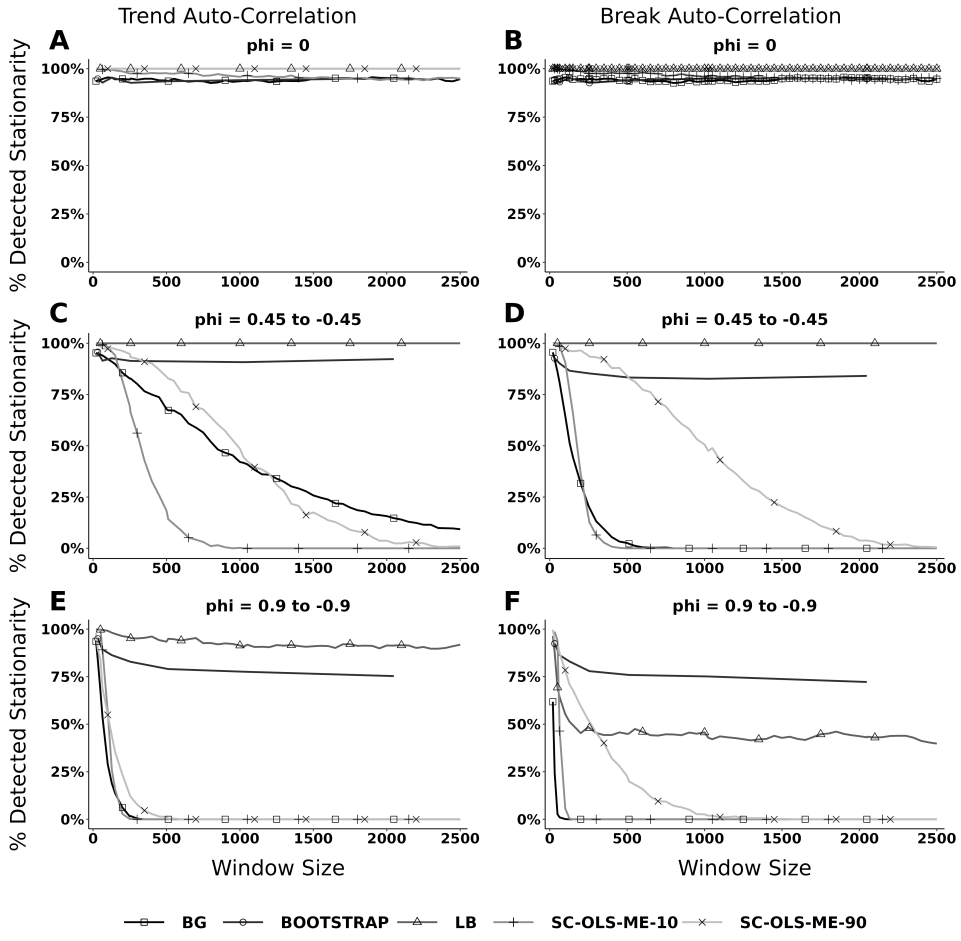

Figure 5. Detection effectiveness of statistical tests for non-stationary time series with trend and break in the autocorrelation.

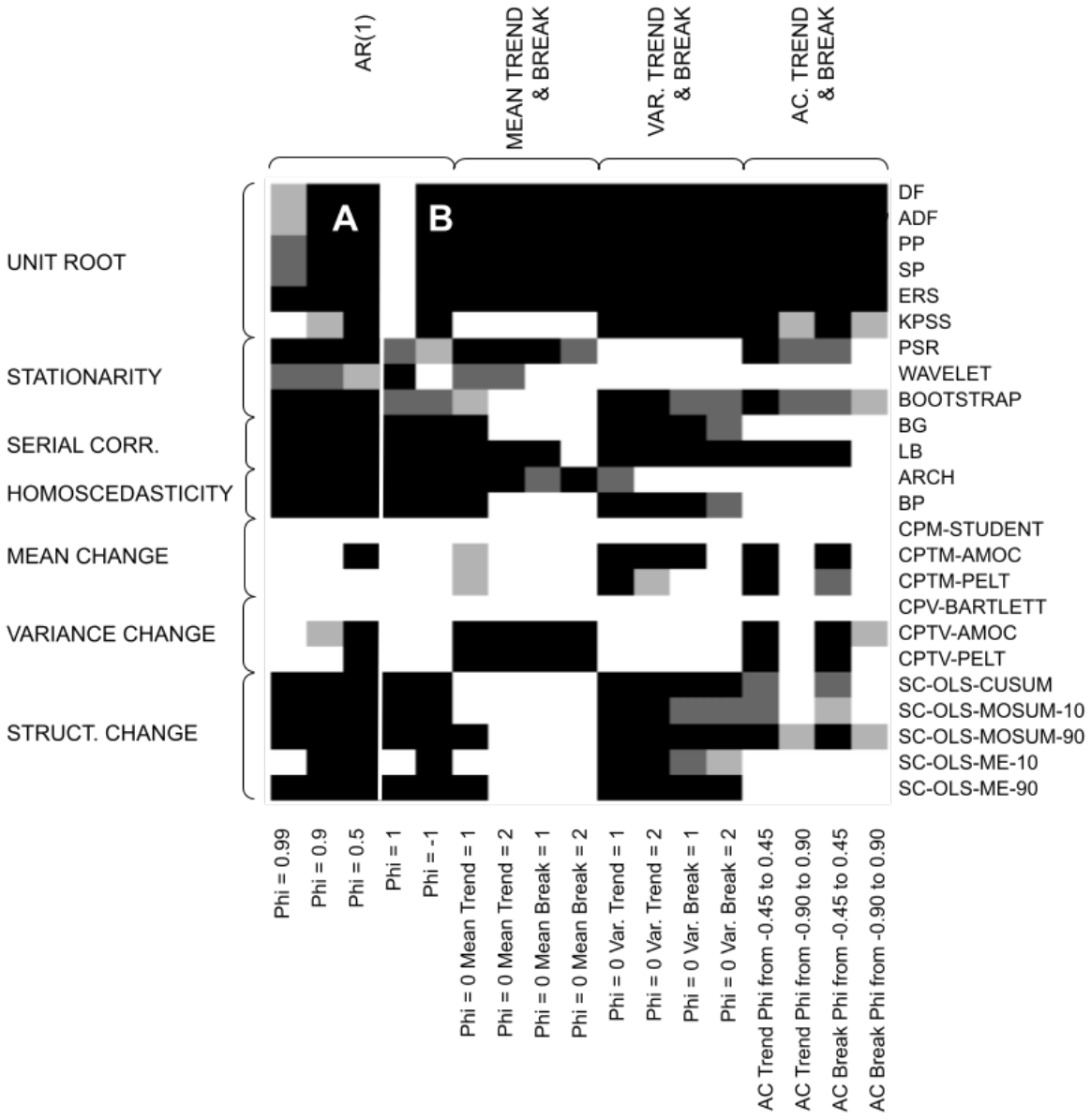

Figure 6. Heatmap of the percentage of time series detected as stationary for sample size 2048. Panel A shows stationary time series. Panel B shows non-stationary time series. The bottom labels show a subset of time series assessed. The top labels indicate the type of the time series assessed. The right labels show a subset of the tests assessed. The left labels indicate the type of the tests. The color represents the percentage of times out of 1,000 that the time series has been detected as stationary: Black $\geq$ 90%, Dark-Gray $\geq$ 75% and $<$ 90%, Light-Gray $\geq$ 50% and $<$ 75%, and White $<$ 50%.

misclassified more severe trend and break in the autocorrelation also as variance trends and breaks. Likewise, they classify stationary time series as non-stationary with trend and break in the variance. These misclassifications can really have very negative consequences in time series modeling since they can direct practitioners on the wrong analytical path or make them draw the wrong conclusions.

Overall, our results suggest that it is not possible to definitively determine the exact source of non-stationarity. Moreover, they also suggest that certain tests can actually lead to even more erroneous conclusions.

Therefore, the use of multiple tests is the best approach, but the statistical tests should be selected accordingly to complement each other's results. For example, because the DF family tests performance leads to the interpretation that time series is either stationary or non-stationary with positive unit root, any other test used in tandem should really focus in discerning if the time series is in fact stationary, or if it has any other source of non-stationarity like negative unit root, or trend or break in the mean, variance or autocorrelation. The identification of the source of non-stationarity from this perspective should be viewed as a bonus, with the acknowledgment that there is no individual test that is not misclassified. From this perspective, the first criteria for selecting additional tests to be used beside the tests of the DF family should be to pick those that do not misclassify stationary time series as non-stationary. Based on this criterion, the WAVELET, the CPM and CPV tests (e.g., CPM-STUDENT and CPV-BARLETT) are not appropriate to be used in tandem with tests of the DF family. For the test to meet this criterion, it will be useful to choose those that at least help to identify the type of non-stationarity to some extent. For example, the BG test that identifies trend and break in autocorrelation, the ARCH test that identifies trend and break in variance, and the SC tests that identify the trend and break in the mean and autocorrelation.

6 Conclusions

Non-stationarity potentially comes from many sources. There is a wide variety of statistical tests for checking specific departures from stationarity. Our study shows that these tests have a low statistical power outside their scope of operation by comparing a larger variety of statistical tests and sources of non-stationarity. While tests capable of identifying if a time series is stationary exist, the same cannot be said about tests that can detect sources of non-stationarity. However, there are several tests that can detect non-stationarity different from unit root, but these tests are not always reliable for identifying the actual source of non-stationarity.

It is a common strategy among the time series practitioners to run several tests on a given time series and deem it stationary only if it passes all or a set of these tests (depending on the decision rules established by the practitioner). Our results also corroborate with this practice and are useful to inform the strategies of time series practitioners.

Finally, statistical tests can be important support tools in the decision-making process. Yet our results showed that they should not be used exclusively in the decision-making process, but rather as one information source in a more deliberative process, together with other tools like a visual exploration of the data, in addition to a profound understanding of the process that generated the data.

Acknowledgements

This work was supported by the NIGMS of the NIH under Grant P20GM104420; EPSCoR Program, National Science Foundation under Grant IIA-1301792. This research made use of the resources of the IBEST Computational Resources Core sponsored by the NIH grant number P30GM103324 and the High Performance Computing Center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the US DoE under Contract No. DE-AC07-05ID14517. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] H. J. Bierens, "Unit roots," in *A Companion to Theoretical Econometrics*, B. H. Baltagi, Ed. Wiley, 2003, ch. 29, pp. 610–633.

[2] S. Bandyopadhyay and S. Subba Rao, "A test for stationarity for irregularly spaced spatial data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, pp. 95–123, 2017.

[3] A. Cardinali and G. P. Nason, "Practical powerful wavelet packet tests for second-order stationarity," *Applied and Computational Harmonic Analysis*, vol. 44, no. 3, pp. 558–583, 2018.

[4] Y. Dwivedi and S. Subba Rao, "A test for second-order stationarity of a time series based on the discrete fourier transform," *Journal of Time Series Analysis*, vol. 32, no. 1, pp. 68–91, 2011.

[5] G. P. Nason, "A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 5, pp. 879–904, 2013.

[6] M. B. Priestley and T. Subba Rao, "A test for non-stationarity of time-series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 31, no. 1, pp. 140–149, 1969. [Online]. Available: http://www.jstor.org/stable/2984336

[7] R. von Sachs and M. H. Neumann, "A wavelet-based test for stationarity," *Journal of Time Series Analysis*, vol. 21, no. 5, pp. 597–613, 2000.

[8] A. Aue and L. Horváth, "Structural breaks in time series," *Journal of Time Series Analysis*, vol. 34, no. 1, pp. 1–16, 2013.

[9] J. Glynn, N. Perera, and R. Verma, "Unit root tests and structural breaks: A survey with applications," *Revista de Métodos Cuantitatívos para la Economía y la Empresa*, vol. 3, no. 1, pp. 63–79, 2007.

[10] P. C. B. Phillips and Z. Xiao, "A primer on unit root testing," *Journal of Economic Surveys*, vol. 12, no. 5, pp. 423–470, 1998.

[11] J. H. Cochrane, "A critique of the application of unit root tests," *Journal of Economic Dynamics and Control*, vol. 15, no. 2, pp. 275–284, 1991.

[12] G. Elliott, T. J. Rothenberg, and J. H. Stock, "Efficient tests for an autoregressive unit root," *Econometrica*, vol. 64, no. 4, pp. 813–836, 1996.

[13] U. K. Müller, "Size and power of tests of stationarity in highly autocorrelated time series," *Journal of Econometrics*, vol. 128, no. 2, pp. 195–213, 2005.

[14] J. K. Afriyie, S. Twumasi-Ankrah, K. B. Gyamfi, D. Arthur, and W. A. Pels, "Evaluating the performance of unit root tests in single time series processes," *Mathematics and Statistics*, vol. 8, no. 6, pp. 656–664, 2020.

[15] T. E. Bartlett, A. M. Sykulski, S. C. Olhede, J. M. Lilly, and J. J. Early, "A power variance test for nonstationarity in complex-valued signals," in *IEEE 14th International Conference on Machine Learning and Applications*, 2015, pp. 911–916.

[16] O. Darné and C. Diebolt, "Non-stationarity tests in macroeconomic time series," in *New Trends in Macroeconomics*, C. Diebolt and C. Kyrtsou, Eds. Springer, 2005, pp. 173–194.

[17] F. X. Diebold and G. D. Rudebusch, "On the power of Dickey-Fuller tests against fractional alternatives," *Economics Letters*, vol. 35, no. 2, pp. 155–160, 1991.

[18] J. Lee, "On the power of stationarity tests using optimal bandwidth estimates," *Economics Letters*, vol. 51, no. 2, pp. 131–137, 1996.

[19] D. V. Metes, "Visual, unit root and stationarity tests and their power and accuracy," University of Alberta, Department of Mathematical and Statistical Sciences, Tech. Rep., 2005.

[20] N. Mohd Razali and B. Y. Wah, "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, pp. 21–33, 2011.

[21] M. Pourahmadi, *Foundations of time series analysis and prediction theory*, ser. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley, 2001.

[22] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: Forecasting and control*, 5th ed. Hoboken, NJ: Wiley, 2015.

[23] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–432, 1979.

[24] ——, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica*, vol. 49, no. 4, 1981.

[25] S. E. Said and D. A. Dickey, "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, vol. 71, no. 3, pp. 599–607, 1984.

[26] E. Paparoditis and D. N. Politis, "The asymptotic size and power of the augmented dickey–fuller test for a unit root," *Econometric Reviews*, vol. 37, no. 9, pp. 955–973, 2018.

[27] P. C. B. Phillips and P. Perron, "Testing for a unit root in time series regression," *Biometrika*, vol. 75, no. 2, pp. 335–346, 1988.

[28] P. Schmidt and P. C. B. Phillips, "Lm test for a unit root in the presence of deterministic trends," *Oxford Bulletin of Economics and Statistics*, vol. 54, no. 3, pp. 257–287, 1992.

[29] D. Kwiatkowski, P. C. B. Philips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, no. 1–3, pp. 159–178, 1992.

[30] M. Caner and L. Kilian, "Size distortions of tests of the null hypothesis of stationarity: Evidence and implications for the PPP debate," *Journal of International Money and Finance*, vol. 20, no. 5, pp. 639–657, 2001.

[31] S. Ng and P. Perron, "LAG length selection and the construction of unit root tests with good size and power," *Econometrica*, vol. 69, no. 6, pp. 1519–1554, 2001.

[32] M. L. King, *Serial Correlation*. Wiley, 2003, ch. 3, pp. 62–81.

[33] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression. iii," *Biometrika*, vol. 58, no. 1, pp. 1–19, 1971.

[34] T. S. Breusch, "Testing for autocorrelation in dynamic linear models," *Australian Economic Papers*, vol. 17, no. 31, pp. 334–355, 1978.

[35] L. G. Godfrey, "Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables," *Econometrica*, vol. 46, no. 6, pp. 1293–1301, 1978.

[36] G. M. Ljung and G. E. P. Box, "On a measure of a lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.

[37] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979.

[38] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.

[39] G. J. Ross, "Parametric and nonparametric sequential change detection in R: The cpm package," *Journal of Statistical Software*, vol. 66, no. 1, pp. 1–20, 2015.

[40] R. Killick and I. A. Eckley, "changepoint: An R package for changepoint analysis," *Journal of Statistical Software*, vol. 58, no. 1, pp. 1–19, 2014.

[41] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[42] W. Ploberger and W. Krämer, "The CUSUM test with OLS residuals," *Econometrica*, vol. 60, no. 2, pp. 271–285, 1992.

[43] A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber, "strucchange: An R package for testing for structural change in linear regression models," *Journal of Statistical Software*, vol. 7, no. 1, pp. 1–38, 2002.

[44] A. Zeileis, C. Kleiber, W. Krämer, and K. Hornik, "Testing and dating of structural changes in practice," *Computational Statistics & Data Analysis*, vol. 44, no. 1–2, pp. 109–123, 2003.

[45] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/