Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2402989, 24 pages
https://doi.org/10.1155/2022/2402989

Research Article

Sensitivity Analysis of Stationarity Tests' Outcome to Time Series Facets and Test Parameters

Advait Amol Bawdekar,[1] **B. Rajanarayan Prusty**,[2] **and Kishore Bingi**[3]

[1]*School of Mechanical Engineering, Vellore Institute of Technology, Vellore, India*
[2]*Department of Electrical and Electronics Engineering, Alliance University, Bengaluru, India*
[3]*School of Electrical Engineering, Vellore Institute of Technology, Vellore, India*

Correspondence should be addressed to B. Rajanarayan Prusty; b.r.prusty@ieee.org

Received 22 July 2022; Accepted 27 September 2022; Published 7 October 2022

Academic Editor: Chao Huang

Time series stationarity is vital for the effective implementation of forecasting models. Time series of renewable generation rich power system input variables such as photovoltaic generations, wind power generations, load power, and ambient temperature have inherent time series facets such as trend, seasonality, and volatility. These inherent facets, combined with the length of the time series or time series clustering, have a propensity to bias the stationarity tests' outcome. This research conducts a rigorous comparative analysis to assess the tests' sensitivity to different time series facets. A seasonal, nonstationary load power time series and its derived time series with synthetically embedded trend and volatility effects are used to help the study capture tests' sensitivity to the above time series facets. This comprehensive analysis and discussion, via a set of well-delineated figures and tables, are expected to assist novice researchers in choosing a group of suitable tests for checking time series stationarity.

1. Introduction

Modern-day power systems are encouraged to accommodate renewable generations' high penetrations at transmission and distribution levels [1–4]. The renewable generations have major characterizing features such as (i) uncertainty (unexpected change), (ii) intermittency (unplanned unavailability), and (iii) uncontrollability (power output is not under the control of the system management) [2, 3]. The above features, alongside load variability, necessitate new methodologies to analyze system flexibility to accommodate higher renewable penetrations [5]. Developing prediction models for the above power system input variables is helpful to most power system studies [3]. On this note, stationarity is vital as many applicable forecasting models rely on stationary time series for easy modeling and obtaining accurate results [3, 6]. For a stationary time series, statistical properties do not change over time. The raw time series collected from sources are often nonstationary and methods such as differencing [7] and transformation [8], help yield a stationary time series. Stationarity tests check whether the application of these methods has succeeded in producing a stationary time series. ADF, KPSS, PP, Breitung, MK, Levene's, KW, KS, and SW tests are a few well-established stationarity tests used in the literature [9–16]. The outcome of these tests is solely based on the time series facets, such as trend, seasonality, and volatility effects, as well as the test's computational steps. Data analysts are always interested in a more effective test or a suitable combination of tests for a given application. Hence, a deliberate comparison study of the well-established tests is the need of the hour; therefore, it is carried out in this paper, taking into account various pertinent issues while applying to different time series.

The above tests are well-documented in the literature. These tests are often used to check data stationarity in power system analyses with renewable generations. ADF test was the first test establishing the unit root concept. The presence of a unit root can be detected using unit root tests. Because of its higher reliability, the ADF test is of choice in the applications such as solar irradiance/PV power [17–20] and

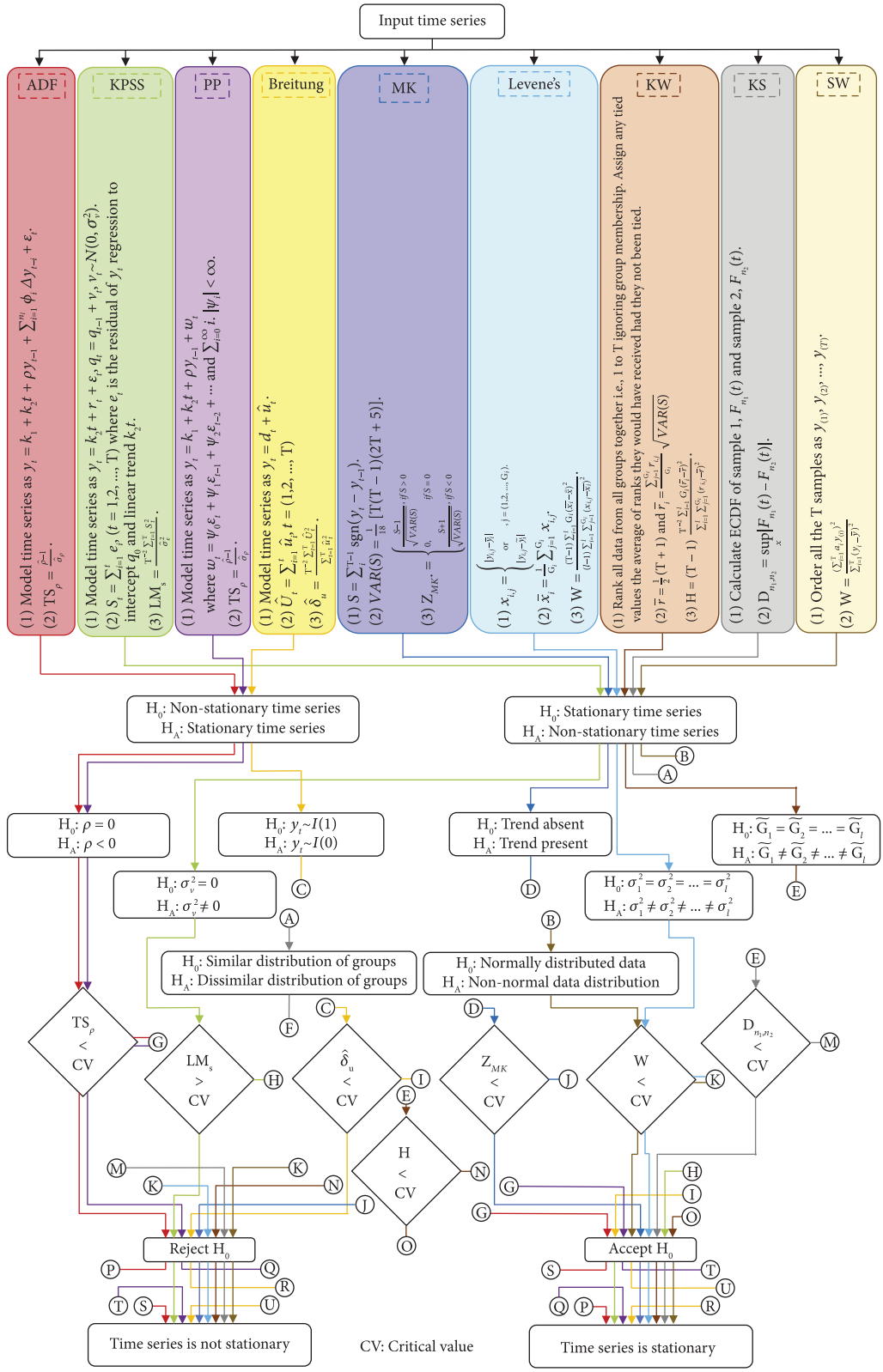

Figure 1: Pictorial representation of the thorough execution steps of well-established stationarity tests.

wind speed/wind power [21–23] time series to check data stationarity. The unit root concept was further used in the PP test, attempting to curb the disadvantages of the ADF test. The main distinction between the ADF and PP tests is how the serial correlation of error terms [10] is handled. The pairs of ADF and PP tests are commented on concrete data stationarity decisions in [24, 25]. In addition to this pair, various normality tests such as Jarque Bera, Liliefors,

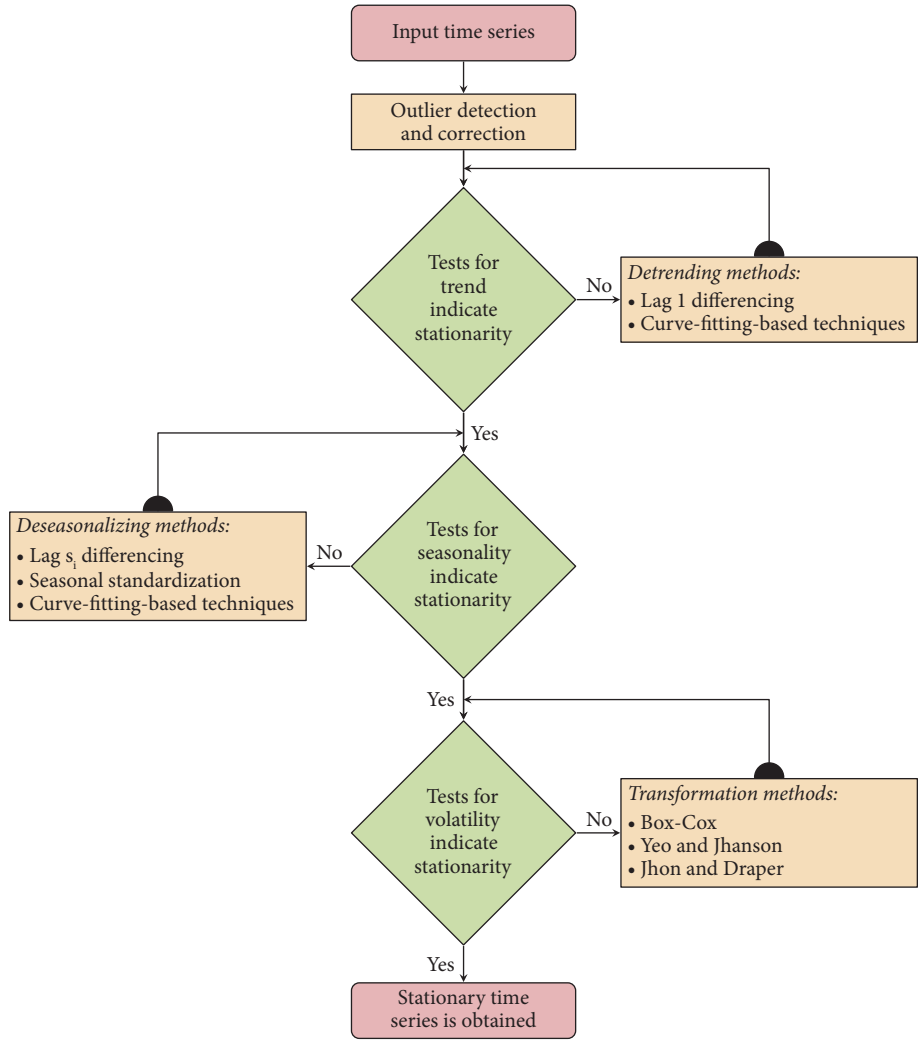

Figure 2: The flow diagram for effective stationarization of time series.

one-way KS, and SW tests were employed to assess stationarity by checking the normality of the time series [11]. The SW test is the most reliable of these tests as it has the highest power [26]. Levene's test was utilized to check the equality of variances to ensure time-series stationarity after using unit root tests such as ADF, PP, and DF-GLS tests [12]. Besides, the nonparametric tests, KPSS, and Breitung tests also assess a time series by determining the unit root. To further support the test results of the ADF, PP, or ADF-PP duo, the KPSS test (which checks the existence of stationarity around a deterministic trend) is readily chosen [3, 27–29]. MK test is applied along with the ADF-KPSS pair to analyze trend effects effectively. Besides, various normality tests discussed above are also chosen to confirm the effective stationarity of time series [14]. The KW test was used after applying ADF, PP, and KPSS tests in [30] to assess the seasonal behavior of the time series. Similarly, one-way ANOVA-based tests and the KPSS test are suggested to evaluate seasonal effects [13]. Breitung test [31] was also used to test data stationarity in wind power applications. Lastly, a two-way KS test is proposed to analyze time series stationarity, highlighting the limitations of ADF and KPSS tests [15]. Furthermore, a comparison study of these tests is also made, considering some of the crucial aspects [10]. The important information about the unit root tests was noted. The effect of time series length on the tests is analyzed in [9] with the help of the power of a test. The reliability of all the nine tests discussed in this paper is analyzed by calculating power [16]. Authors in [32, 33] compared and noted the behavior of specific stationarity tests. Further, the drawbacks of unit root tests were highlighted in [34].

Although the computational steps for tests and their characteristics are well documented in the literature, the question arises of which test or combination of tests to use and when. Further, a comprehensive comparison among the tests by examining their sensitivity towards various time series facets and the effect of other critical factors can better investigate the selection of particular tests for a specific application. Although a systematic summarization and comparison of unit root tests are made in [10], the impacts of time series length/time series clustering, parameters associated with various tests, and time series

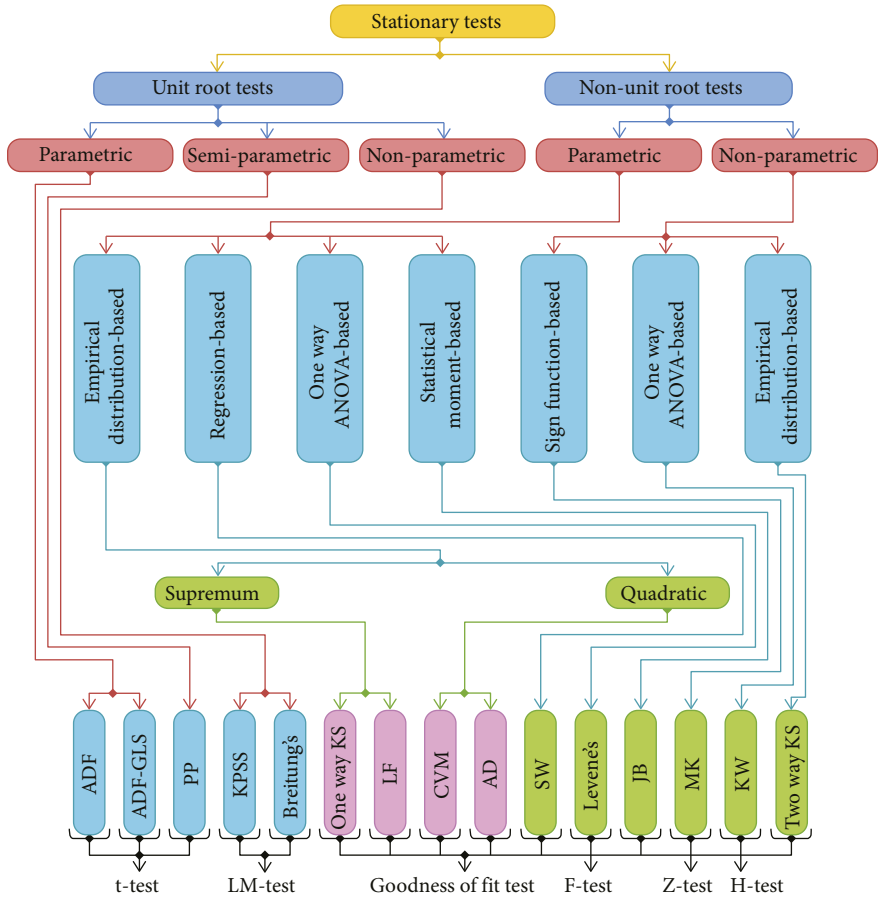

Figure 3: A novel classification of well-established stationarity tests.

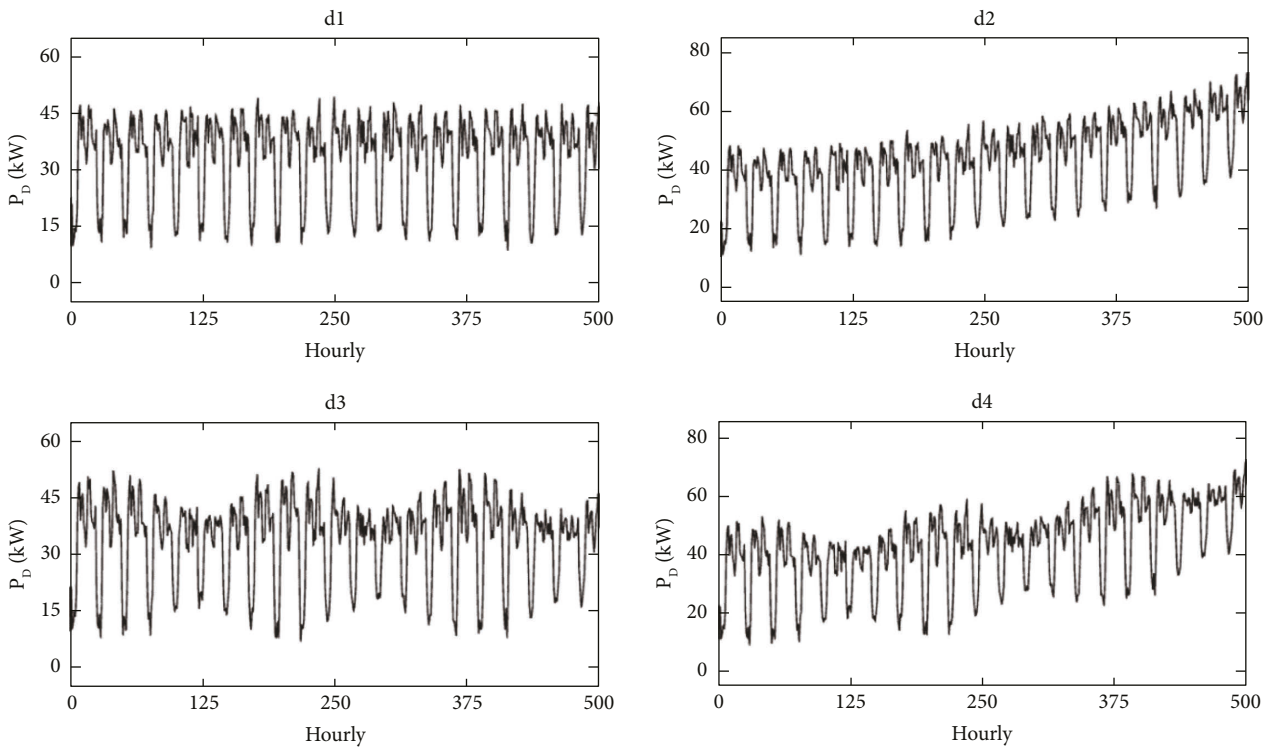

Figure 4: Time series of four synthetic datasets used for simulation.

Table 1: Datasets used for result analyses.

Type	Notation	Characteristics		Sample size	Time step
Load power, P_D (clean)	d1	Synthetic data	—	500	Hourly
	d2		Seasonal, non-linear trend	500	Hourly
	d3		Seasonal, volatile	500	Hourly
	d4		Seasonal, non-linear trend, volatile	500	Hourly

Table 2: Impact of drift and trend component consideration and time series facets on test results and critical values of ADF test.

Model equation (for ADF)	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
Without drift and trend components	−0.6790	−1.9412	−0.0399	−1.9412	−0.7335	−1.9412	−0.0200	−1.9412
With only drift component	−16.7321	−2.8683	−2.6612	−2.8683	−13.7298	−2.8683	−2.7889	−2.8683
With both drift and trend components	−16.7263	−3.4199	−10.6335	−3.4199	−13.7848	−3.4199	−9.8550	−3.4199

Table 3: Impact of trend component consideration and time series facets on test results and critical values of KPSS test.

Model equation (for KPSS)	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
With only drift component	0.0578	0.4630	15.9197	0.4630	0.0314	0.4630	14.8379	0.4630
With both drift and trend components	0.0242	0.1460	0.2979	0.1460	0.0293	0.1460	0.2922	0.1460

Table 4: Impact of drift and trend component consideration and time series facets on test results and critical values of PP test.

Model equation (for PP)	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
Without drift and trend components	−2.1086	−1.9411	−1.5034	−1.9411	−2.1505	−1.9411	−1.5446	−1.9411
With only drift component	−7.6076	−2.8680	−6.0787	−2.8680	−13.7298	−7.5627	−6.1654	−2.8680
With both drift and trend components	−7.6002	−3.4193	−7.4642	−3.4193	−13.7848	−7.5546	−7.4239	−3.4193

facets on test outcomes through a detailed analysis are overlooked. The time series have facets that alter the test results in many ways, as highlighted in [9, 16, 32, 33]. Still, only a few selective tests that belong to a specific category are analyzed and compared. Thus, a study examining these facets' impact on well-established stationarity tests becomes necessary.

The major contributions of the paper are as highlighted underneath

(i) Firstly, the working of the existing tests is pictorially portrayed in this paper, assisting readers in figuring out how different the approach of a particular test is from others.

(ii) Further, a stationarity test's attributes are highlighted, and the impacts of time series length/clustering and test parameters alongside time series facets on the test outcomes are examined.

(iii) The suitability of a test concerning these effects is also highlighted, and the capability of a stationarity test to characterize all properties of the time series and accordingly yield a fair outcome is thoroughly studied.

(iv) The appropriate incorporation of conclusions/suggestions at each analysis stage is expected to allow the readers to understand the tests' suitability for various power system applications.

The rest of the paper is organized as follows: Section 2 meticulously summarizes the nine well-established stationarity tests on a common platform. Their working steps are suitably explained, and an effective comparison is provided

Table 5: Impact of drift and trend component consideration and time series facets on test results and critical values of Breitung test.

Model equation (for breitung)	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
Without drift and trend components	0.3012	0.01986	0.2584	0.01986	0.3016	0.01986	0.2595	0.01986
With only drift component	0.0701	0.01046	0.0240	0.01046	0.0711	0.01046	0.0255	0.01046
With both drift and trend components	0.0603	0.00355	0.0230	0.00355	0.0611	0.00355	0.0244	0.00355

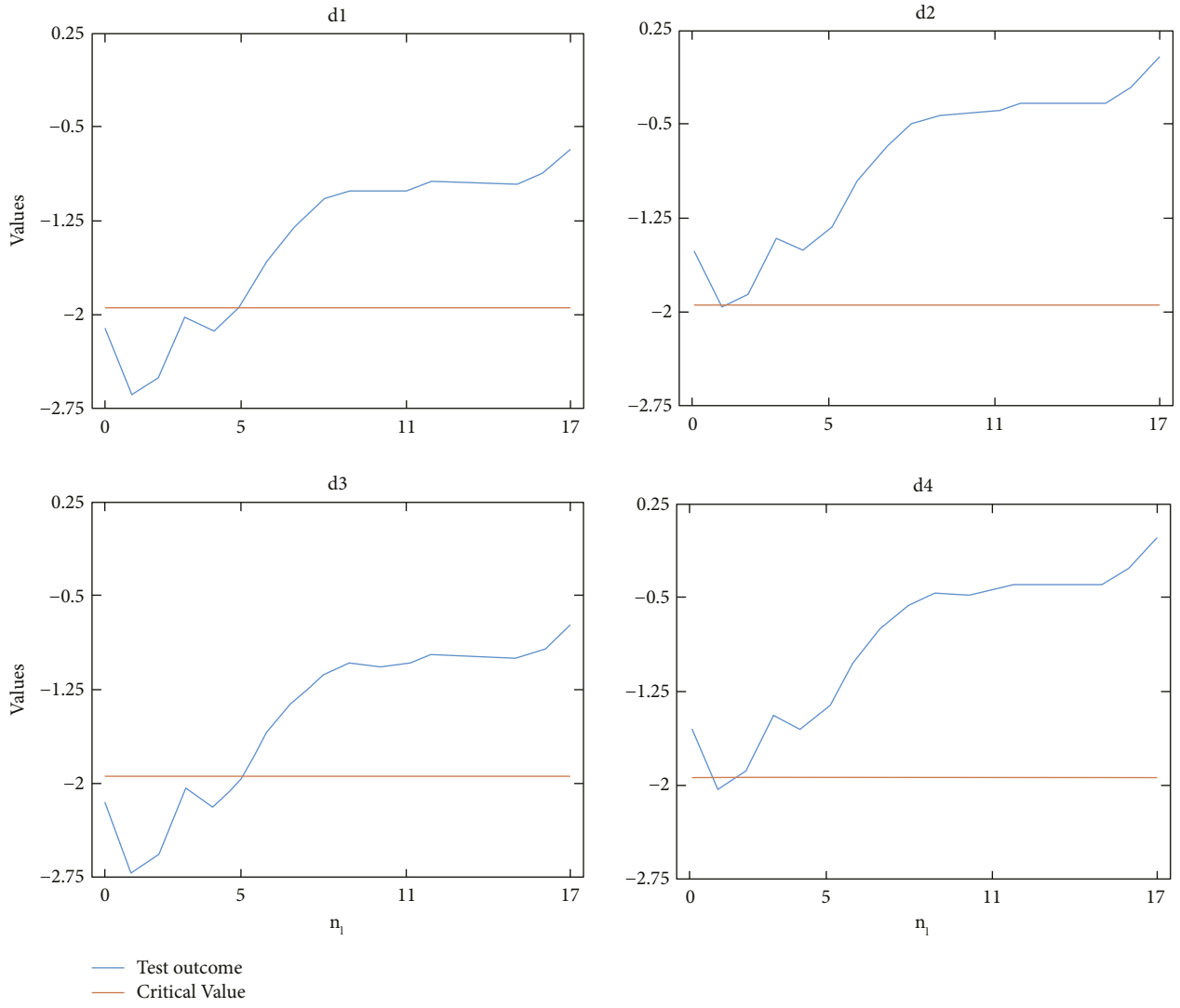

Figure 5: Impact of lag number of ADF test on test results and critical values.

for the tests. Various time series facets are highlighted, and the importance of understanding the effect of these facets on stationarity test outcomes is elaborated in Section 3. Comprehensive result analysis is carried out in Section 4 to compare the impacts of time series length/time series clustering, tests' parameters alongside the time series facets on test results to select tests' suitability for various applications astutely. Finally, the critical findings are concluded in Section 5 and provided with some future scopes.

2. Stationarity Tests

The stationarity tests analyze a time series and mark whether it is stationary through test statistics such as t-test, Z-test, LM

Table 6: Impact of seasonality consideration and time series effects on test results and critical values of MK test.

Parameter	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
Seasonality not considered	0.3480	1.9600	14.4893	1.9600	−0.6888	1.9600	13.6453	1.9600
Seasonality considered	11.7552	1.9600	19.3845	1.9600	11.7552	1.9600	19.7024	1.9600

Table 7: Impact of mean analysis and median analysis with time series effects on test results and critical values of Levene's test.

Parameter	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
Using mean	0.5836	0.0500	0.7196	0.0500	0.0152	0.0500	0.0223	0.0500
Using median	0.8327	0.0500	0.8846	0.0500	0.1307	0.0500	0.1238	0.0500

Table 8: Impact of time series effects on test results and critical values of KW, KS, and SW tests.

Test	d1		d2		d3		d4	
	Test result	Critical value	Test result	Critical value	Test result	Critical value	Test result	Critical value
KW	0.9168	0.0500	0.0000	0.0500	0.8173	0.0500	0.0000	0.0500
KS	0.5784	0.0500	0.0000	0.0500	0.1012	0.0500	0.0000	0.0500
SW	0.0000	0.0500	7.8276×10^{-8}	0.0500	1.1102×10^{-16}	0.0500	8.8698×10^{-9}	0.0500

test, and ANOVA. The test results for these tests are apportioned in various statistical distributions. For a 5% level of significance, the entire region under the distribution curve is divided into the acceptance region, constituting 95% of the whole region, and the critical region, making the rest 5%. When the result value lies in the acceptance region, the null hypothesis is accepted, whereas the null hypothesis is rejected when the test result value lies in the critical region. If the critical region lies at both ends of the curve, the test is two-tailed. When only one end contains the critical region, the test is one-tailed. The value of the test result for separating the critical region and acceptance region is called the critical value [35, 36]. The critical values mark the ends of the acceptance region, i.e., the region under which the null hypothesis is accepted. The tests are either one-tailed or two-tailed, meaning that the critical values signify one or both ends of the significance level, respectively. The working steps of nine well-established stationarity tests are portrayed in Figure 1.

The ADF test is a one-tailed t-test that checks the presence of a unit root [37]. Firstly, a model for the given time series is built (refer to Figure 1) and then tested for stationarity. The critical value of the test changes based on significance level and for cases whether k_1, k_2 are zero or nonzero. If k_1, k_2 components are nonzero, the respective components are eliminated using OLS detrending [10]. The value of n_l for the test is to be chosen according to the time series, which should not exceed a particular value given as, $n_{l,\max} = 12[T/100]^{1/4}$ [9]. The t-test here is conducted using the obtained ρ value. A t-test is an inferable statistic that evaluates the significance of the difference between the mean values of two groups and their relationship. Nextly, the KPSS test is also a one-tailed unit root test. Still, it uses LM statistics, which involves an examination of constraints on various statistical parameters using the gradient of the likelihood function. The critical value depends on the significance level based on constant and trend terms. When $r_0 = 0$ and $k_2 = 0$, e_t denotes y_t and when $r_0 = 0$ and $k_2 \neq 0$, e_t denotes residual of y_t regression to the nonzero intercept term [10]. Except for the difference in asymptotic theory and no need for a decision of lag number, the PP test is the same as the ADF test [10]. Asymptotic theory, or large sample theory, is a framework for evaluating the properties of estimators and statistical tests. The sample size is generally considered to grow indefinitely in this paradigm. The properties of estimators and tests are then examined under the limit of sample size tending to infinity. In reality, a limit analysis is thought to be roughly valid for large finite sample sizes. The Breitung test also follows a one-tailed LM statistic, and the critical values depend on significance level and length of time series [38].

The MK test is a two-tailed test capable of detecting a trend in time series [14]. This test follows Z-statistics, and the critical values depend on the significance level. A Z-test is a statistical test to assess whether the mean values are different when the variances of two populations are known. Like the t-test, the Z-test is a univariate test based on the standard normal distribution; however, the population variance in the t-test is estimated from the sample with a known mean. Further, the seasonal MK test is applied if the time series has a seasonal component. Levene's test can be employed for testing the equality of variances [12, 39]. It is a one-tailed test following the F statistics; hence, critical values are dependent on the level of significance and corresponding degrees of freedom. As F-test determines whether the samples taken come from the same population, they assess whether the statistical properties of the samples considered are significantly similar. Besides, the KW test is a one-tailed H-test

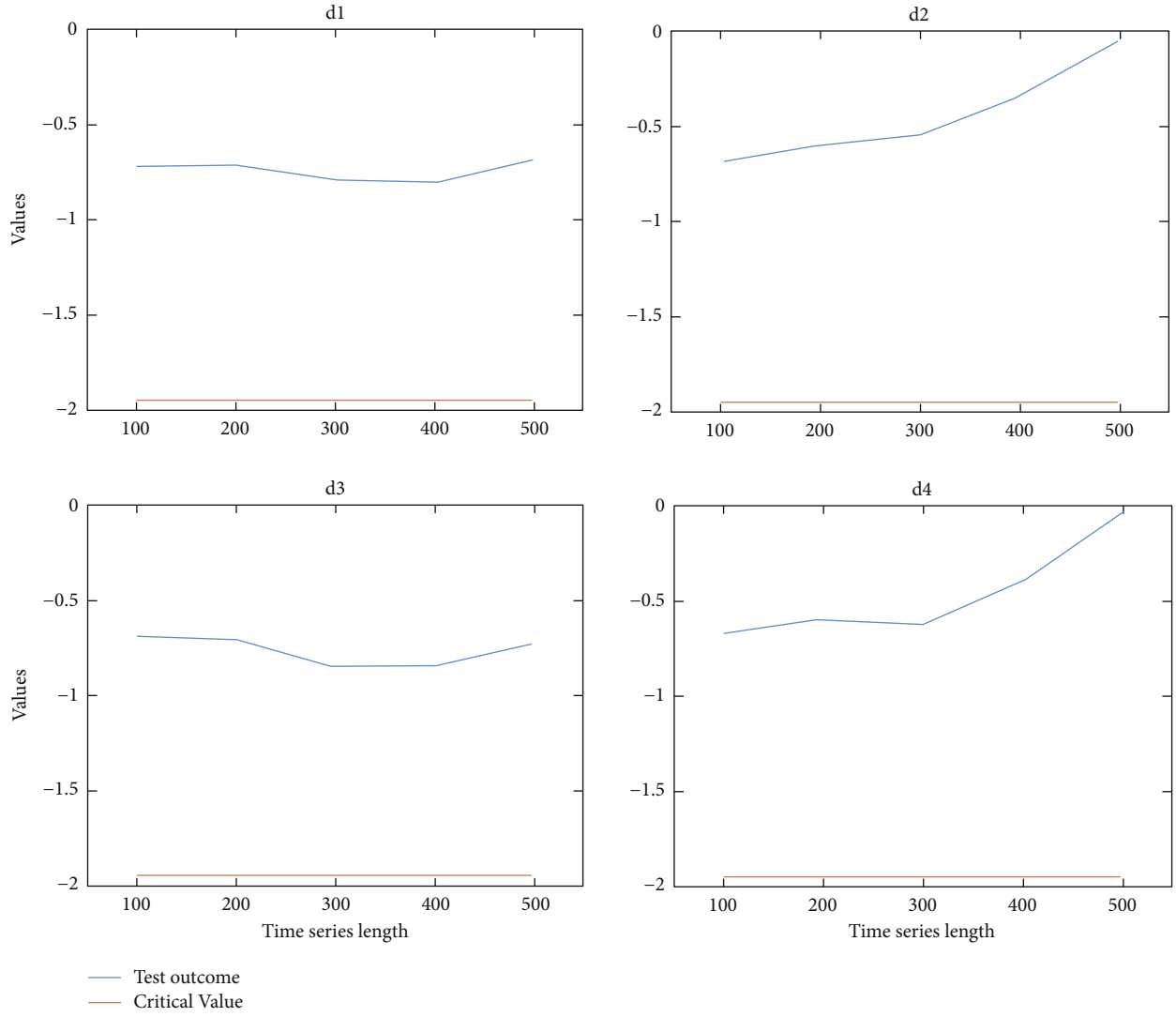

Figure 6: Impact of time series length and time series effects on test results and critical values of ADF test.

known as one-way ANOVA on ranks as it classifies its data into ranks and performs a one-way ANOVA test. The critical values for this test also depend on the level of significance and corresponding degrees of freedom [40]. Further, the two-way KS test is also one-tailed, comparing ECDFs effectively [15]. The critical value for this test is calculated as follows: $D_{m_1,m_2} > \sqrt{-\ln(c_\alpha/2)1 + m_2/m_1/2m_2}$ [15], and thus it is dependent on group size and level of significance. Lastly, the SW test is also based on one-way ANOVA. Therefore it is one-tailed with critical values dependent on the significance level and corresponding degrees of freedom [11, 26]. The formulations of $\widehat{\sigma}_\rho$ and $\widehat{\sigma}_\varepsilon^2$ are provided in [10], and those for calculating a_i are given in [26].

The existing test combos in the literature include, (i) ADF and KPSS [27], (ii) ADF and PP [24], (iii) KPSS and PP [13], (iv) ADF, KPSS, and PP [30], (v) ADF-GLS, PP, and normality tests [11], (vi) ADF, PP, DF-GLS, and Levene's tests [12], (vii) ADF, KPSS, PP, and KW tests,, and (viii) ADF, KPSS, seasonal MK and normality tests [14]. Breitung test [31] and two-way KS test [15] are not used with other tests. The unit root tests, i.e., ADF, KPSS, PP, and Breitung tests, and the MK test, can analyze the trend component in time series. The seasonal MK test is considered to avoid any discrepancies in the test results due to time series seasonality. Similarly, the Fisher, Levene, and KW tests help characterize seasonal behavior in time series. Two-way KS test analyses stationarity by comparing ECDFs in various time series fragments. Therefore, it should be capable of characterizing all the time series components, which account for its nonstationarity if the sizes of fragments are correctly chosen. Lastly, the SW normality test can analyze time series stationarity as normally distributed data are bound to be stationary. These stationarity tests cannot detect all the nonstationarizing time series factors such as trend, seasonality, and volatility. Further, the available stationarizing methods cannot remove all the nonstationary time series components. So a sequence of techniques must be employed for an effective stationarization. Therefore, deciding on a series of steps in which the stationarity components are required to be removed is essential and is suggested in

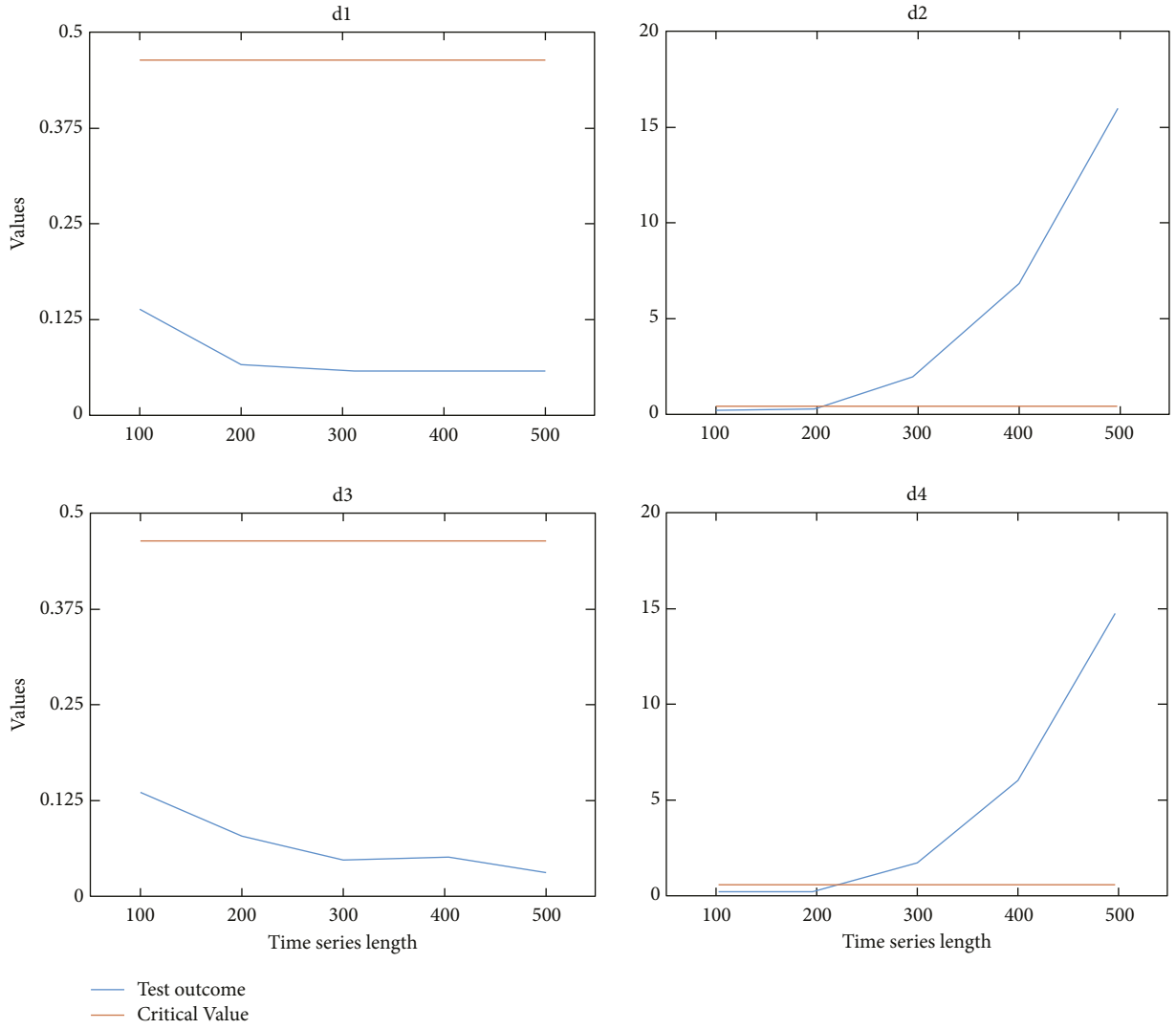

Figure 7: Impact of time series length and time series effects on test results and critical values of KPSS test.

Figure 2. The tests that can detect trend components can yield unbiased results without being concerned about the presence of seasonality or volatility in the time series. Tests useful for seasonality assessment provide biased results in the presence of trend components as they fail to detect differences between seasonal and trend variations. In contrast, the presence of volatility is not a prime concern. Lastly, as volatility is generally detected using normality tests or similar tests, the time series should be free from any trend or seasonality component for true evaluation of the time series volatility. In the presence of trend or seasonality, these tests will provide results pertaining to these facets, which means that these components overshadow the results as volatility remains unassessed.

The variations in the above-discussed tests necessitate a suitable classification for duly understanding the similarities and differences between all these tests. Thus, a novel standard classification tree for stationarity tests is proposed in Figure 3. A test's working principle being the main reason how a test is different from others is generally ignored in the existing stationarity test classifications; hence, after classifying the tests as unit root and nonunit root tests, the proposed classification is mainly prioritized on the working principles of the nonunit root tests. A detailed elaboration is provided underneath. The tests are classified as parametric, semiparametric, and nonparametric. Parametric tests assume the data to be normally distributed, whereas nonparametric tests do not have such assumptions. PP test is classified as semiparametric as initially, it follows a parametric model, but while constructing test statistics, a nonparametric approach is adopted. The tests are classified as EDF-based, regression and correlation-based statistical moment-based, sign function-based, and one-way ANOVA-based tests. Here, EDF-based tests function by computing the EDF of data. In contrast, regression and correlation-based tests use the ratio of normally distributed least squares estimates and the variance of the data sample. Similarly, statistical moment-based and sign function-based tests involve calculating moments and using sign function, respectively. One-way ANOVA-based tests mostly employ the

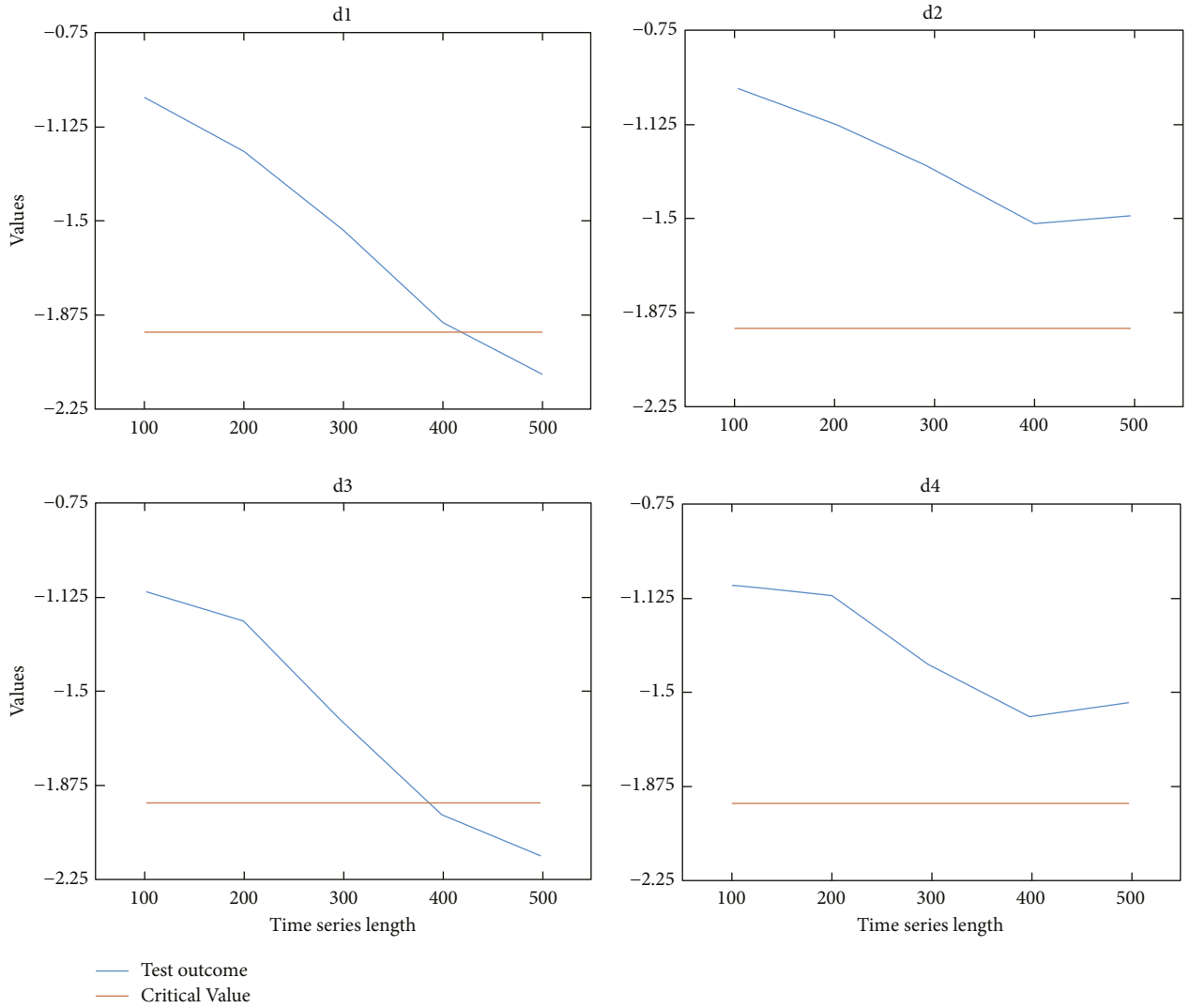

Figure 8: Impact of time series length and time series effects on test results and critical values of PP test.

same principles as classical one-way ANOVA tests; hence, they can also be called improvements of one-way ANOVA tests. Lastly, parametric EDF-based tests are further classified as supremum and quadratic based on the computation of test statistics by respective tests. Tests entailing supremum yield test results by calculating the largest difference between two quantities, whereas quadratic tests do the same by calculating the quadratic difference.

3. Stationarity Test Attributes

The results of the tests for time series stationarity are often biased by time series facets, further affected by the time series length [38, 41, 42] as well as time series clustering [14, 39, 40, 43]. Various other parameters associated with the tests can also bias the test results. A practical stationarity test is expected to correctly indicate whether a given time series is stationary or not. Further, the test outcome should remain unbiased irrespective of the above facets.

3.1. Impact of Time Series Length on Test Results. The length of the time series significantly impacts the test results and critical values [38, 41, 44]. This capability varies differently for different tests. The time series length must not influence a test in providing unbiased results. For any given length, the tests should ideally yield an unbiased outcome; therefore, it is necessary to study the impact of length on the test results. This would provide colossal information about the tests' efficacy while dealing with time series of different lengths. The effect of time series length on test results can be noted by comparing the test results and critical values for various lengths, which help notice the significance of a test declaring a time series as stationary or

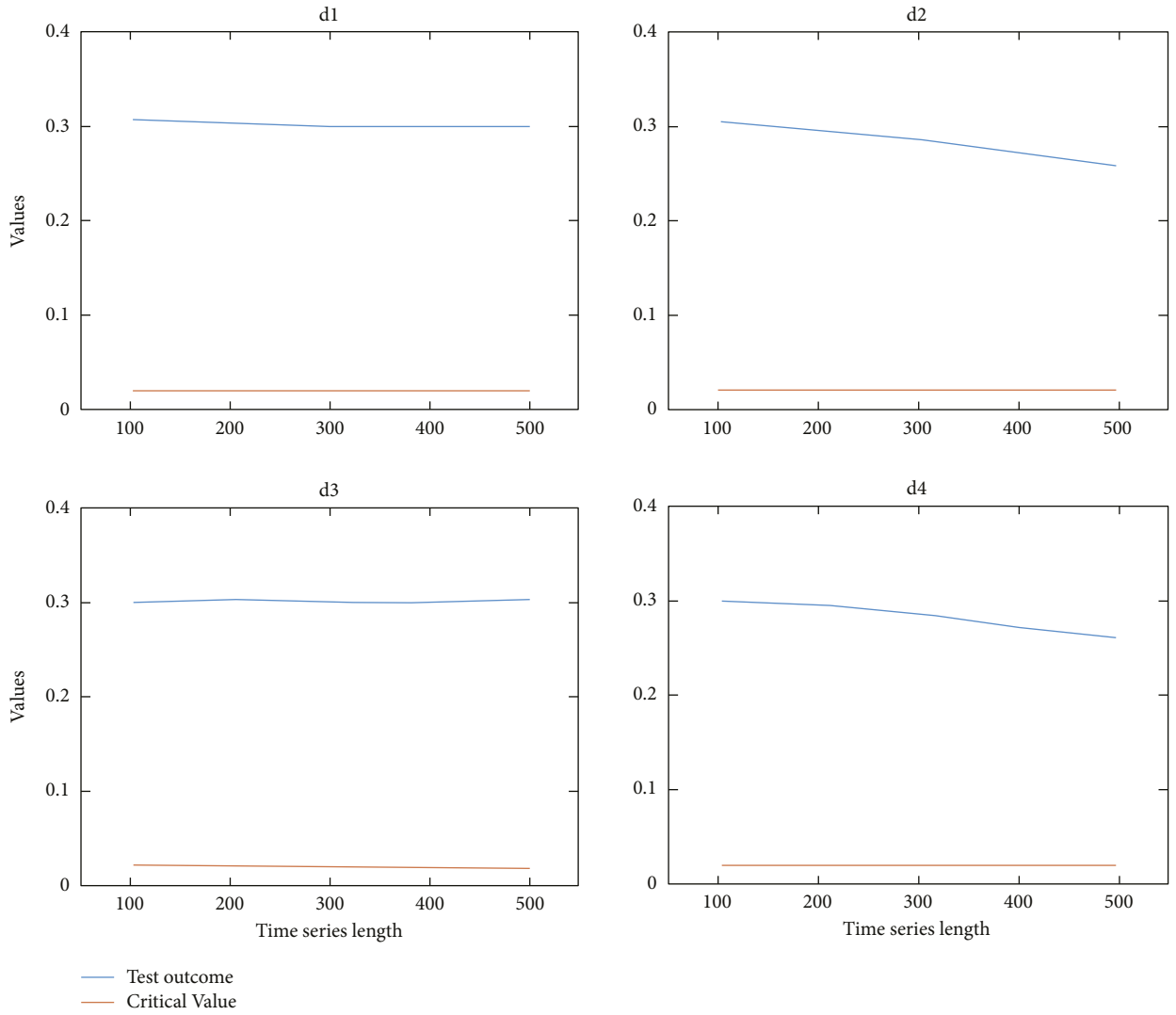

Figure 9: Impact of time series length and time series effects on test results and critical values of Breitung test.

nonstationary. Actually, the test result is compared with the critical value to accept or reject the null hypothesis. Thus, if the positive difference between the test result and the corresponding critical value is higher, the null hypothesis is accepted or rejected by a higher significance. Therefore, if a time series rejects the nonstationary hypothesis by a large factor, then the time series is more strongly stationary. A nonstationary component in the time series may be overlooked when the time series considered for testing has a high length. Perhaps for a shorter length, the presence of the component is highlighted through the results. The impact of length on test results can only be studied for ADF, PP, KPSS, Breitung, MK, and SW tests. The rest of the tests analyze time series by dividing them into groups or clusters; hence, the impact of length on these tests' outcomes cannot be characterized.

3.2. Impact of Time Series Clustering on Test Results. Some stationarity tests function by dividing the time series into various fragments [14, 39, 40, 43]. All these fragments are compared and analyzed, and test results are obtained. Among all the considered tests, Levene's, KW, and two-way KS tests are built to examine the time series by dividing them into various groups or clusters. These fragments, groups, or clusters are created by considering some parts of the time series without any intermixing of data. The size of the group can cause notable variations in test results. It is crucial to know the apt size of the groups for respective tests to obtain accurate and unbiased results. Considering a very small or very large group size may lead to significant discrepancies in test results. Further, cluster sizes can also be unequal for all these tests. KS test can only test two clusters at a time, while Levene's and KW tests can examine many groups at a time,

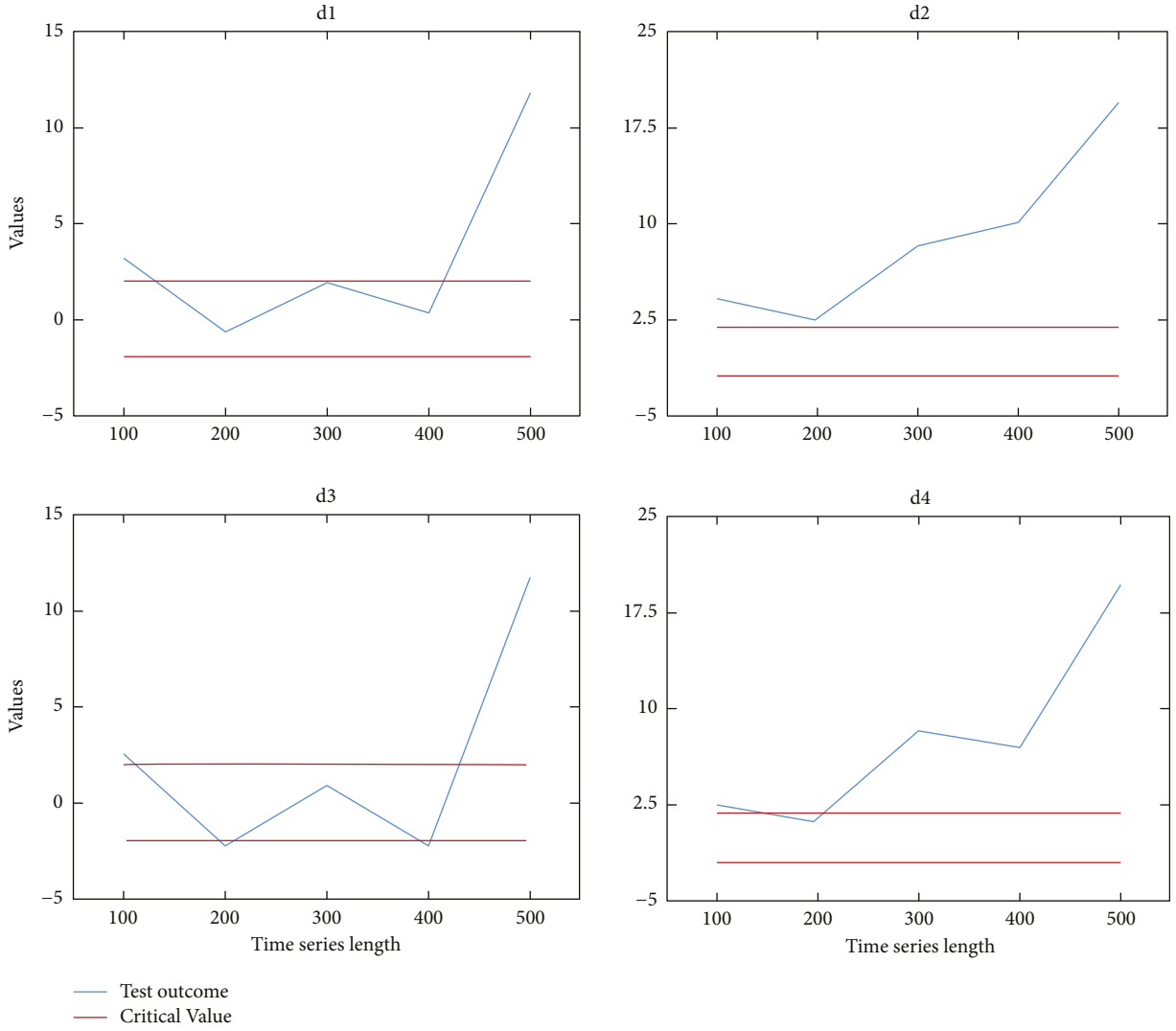

Figure 10: Impact of time series length and time series effects on test results and critical values of MK test.

and thus the number of groups also turns out to be an essential attribute. The size of clusters, the number of clusters, or the fact that they are equally or unequally sized, should not affect the outcome of an ideal test. But in reality, these elements have a substantial effect on the results, so it is necessary to note their behavior for all the variations in these elements.

3.3. Impact of Time Series Facets on Test Results. Trend and volatility effects in a seasonal time series account for its nonstationarity and might bias the test results [32]. It is also possible that the tests may overlook these effects and may fail to yield unbiased results. It is vital to notice the changes in test results and critical values due to the trend effect as the facet cause nonstationarity in time series [38, 41, 42, 44, 45]. Such an analysis can help understand how impactful a trend effect could be in making a test bias. Similar to the trend effect, it would also be interesting to notice the changes in test results due to volatility effects. The impact of volatility on test outcomes is often characterized based on the test results, as the volatility effect has less impact on changing critical values [29]. The presence of deterministic seasonality may inflate the size and reduce the power of the ADF test; as a result, the distribution of the test is shifted towards the left, and the dispersion is lower [44]. The presence of deterministic seasonality components can lead to discrepancies in the decision-making process of the KPSS test under the null hypothesis [44]. Possibly, other tests can also characterize seasonality through test results similarly.

3.4. Impact of Changes in Test Parameters on Test Results. Various test parameters characterize stationarity tests; they define the primary features of respective stationarity tests, e.g., the median is typically employed in Levene's test whenever the data distribution are not symmetrical, whereas the mean is appropriate when the data come from a symmetrical distribution [12]. Thus, significant impacts of these modifications could be observed through the test results. The parameters corresponding to various tests are listed as follows:

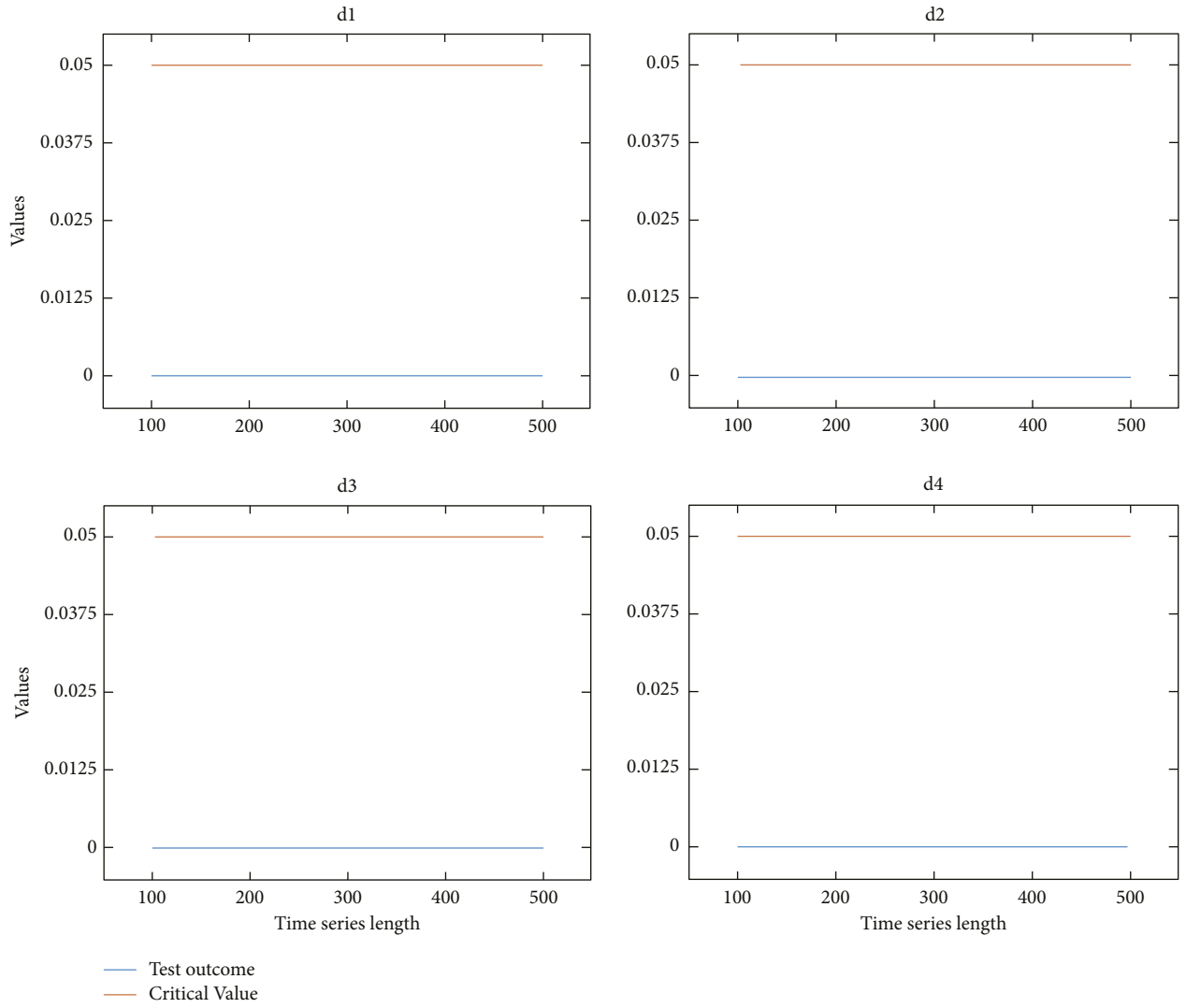

Figure 11: Impact of time series length and time series effects on test results and critical values of SW test.

(i) ADF – $k_1 = 0/k_1 \neq 0,\ k_1 = k_2 = 0\,/k_1 \neq 0, k_2 \neq 0,\ n_1$ [41].

(ii) KPSS – $k_2 = 0\,/k_2 \neq 0$ [44].

(iii) PP – $k_1 = 0/k_1 \neq 0,\ k_1 = k_2 = 0/\,k_1 \neq 0, k_2 \neq 0$ [45].

(iv) Breitung – $d_t = 0/d_t \neq 0$ [38].

(v) MK–Consideration of seasonality [14].

(vi) Levene's–Analysis using mean or median [12, 39].

4. Result Analysis

The selection of effective stationarity tests for a given application necessitates a thorough examination of well-established tests, investigating their efficacy when dealing with multifold seasonal datasets with trend and volatility effects. A comprehensive survey is usually helpful for exposing tests' weaknesses or failures when dealing with a particular time series facet. On this note, the performance of the nine well-established tests for four datasets (refer to Figure 4) is examined to check their ability to yield unbiased results. The analyses carried out in this section are fivefold. Firstly, the impact of test parameters on the test outcomes is studied considering four clean seasonal time series with synthetically embedded trend and volatility effects. Secondly, with the same datasets, the impact of time series length on test results is comprehensively studied. Furthermore, the sensitivity of stationarity tests' outcome to time series clustering considering cluster size, cluster ratio, and cluster numbers is examined. Lastly, deliberate discussions of the obtained results are carried out and provided with appropriate suggestions on the selection of tests. The non-stationary datasets used for the above analyses are summarized in Table 1. Considering the four WGN-assisted seasonal datasets with synthetically embedded time series facets helps make the study more generic and informative. The load power consumption of a restaurant in Anchorage,

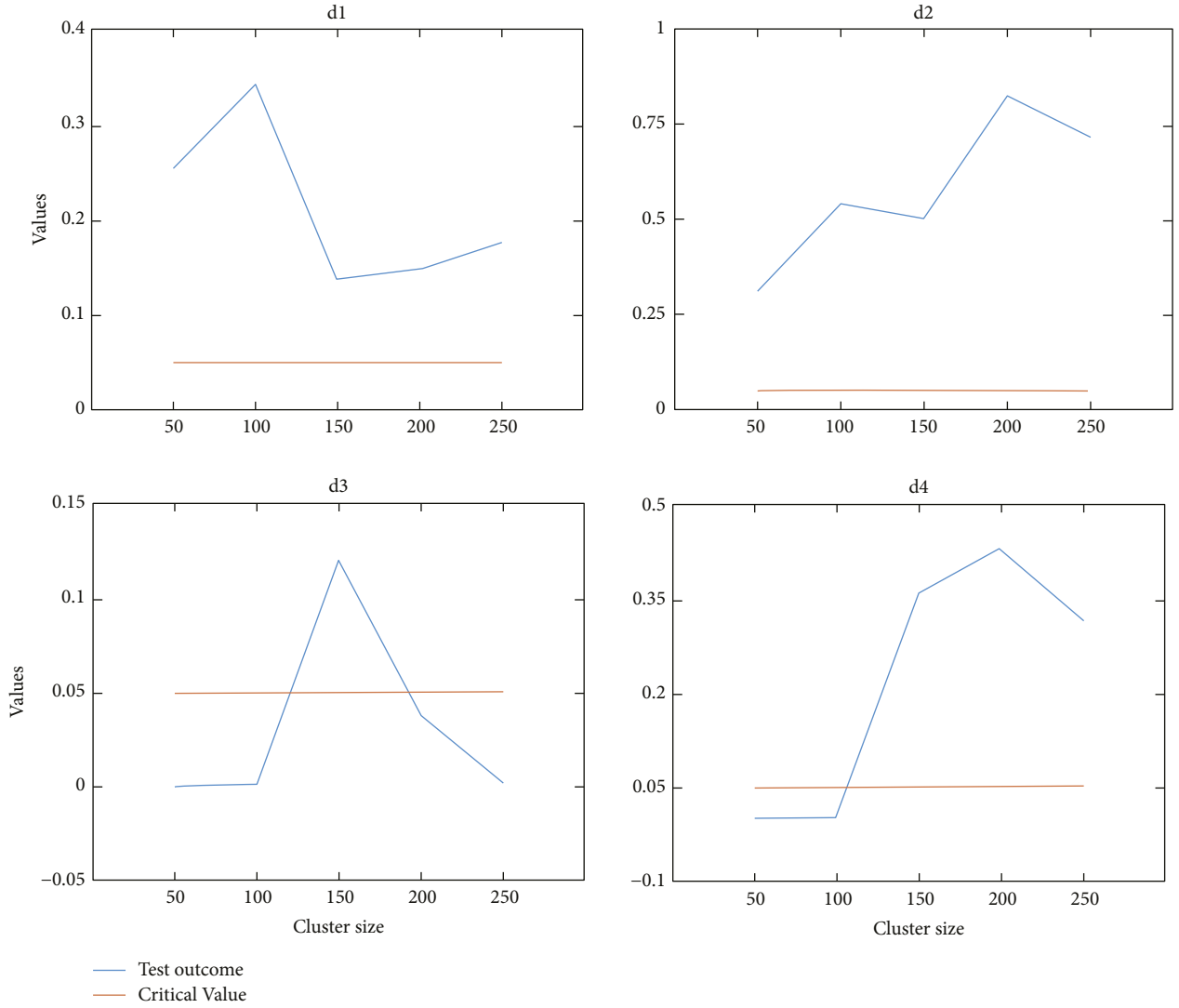

Figure 12: Impact of cluster size and time series effects on test results and critical values of Levene's test.

USA [46], denoted as d1, is the primary dataset for creating d2, d3, and d4.

4.1. Impact of Test Parameters on the Outcome of ADF, KPSS, PP, Breitung's, MK, and Levene's Tests

4.1.1. Impact of Test Parameters on the Outcome of Unit Root Tests, ADF, KPSS, PP, and Breitung. All unit root tests involve time series modeling (refer to Figure 1). While modeling time series, k_1 and k_2 values are essential parameters for tests such as ADF and PP. For the KPSS test, k_2 is only considered. The overall deterministic component is considered for the Breitung test, similar to ADF and PP tests. These tests model the time series with a drift and a time-varying trend component upon consideration of k_1 and k_2 values, respectively. These components are then removed from the time series using OLS detrending method in ADF, KPSS, and PP tests. In contrast, GLS detrending technique is used in the Breitung test. Finally, the residual is tested for stationarity. When time series is modeled under consideration of these parameters, provided these components exist in the time series, higher accuracy can be expected from results for conformance of time series stationarity. The observations from the obtained results are similar for all the unit root tests (refer to Table 2 through Table 5). For all these tests, consideration of k_2 or time-varying trend component in d_t does not have any impact on test outcome for data with no time-varying trend as in the case of d1 and d3. In contrast, due to the same trend component, notable changes are seen in results for d2 and d4 (refer to Table 2 through Table 5). Unit root tests effectively characterize trend components but cannot describe seasonality and volatility effects. Seasonality and volatility bias the test results of ADF, KPSS, and PP tests. Notably, the Breitung test did not render any biased results like all other unit root tests in the presence of the seasonality effect, thus showcasing its ability to detect the seasonality effect. Also, results without d_t indicate that it is better to consider the drift and trend components if they are present in time series for effectively detecting seasonal components. Volatility also has a minimal but positive impact on the test under all

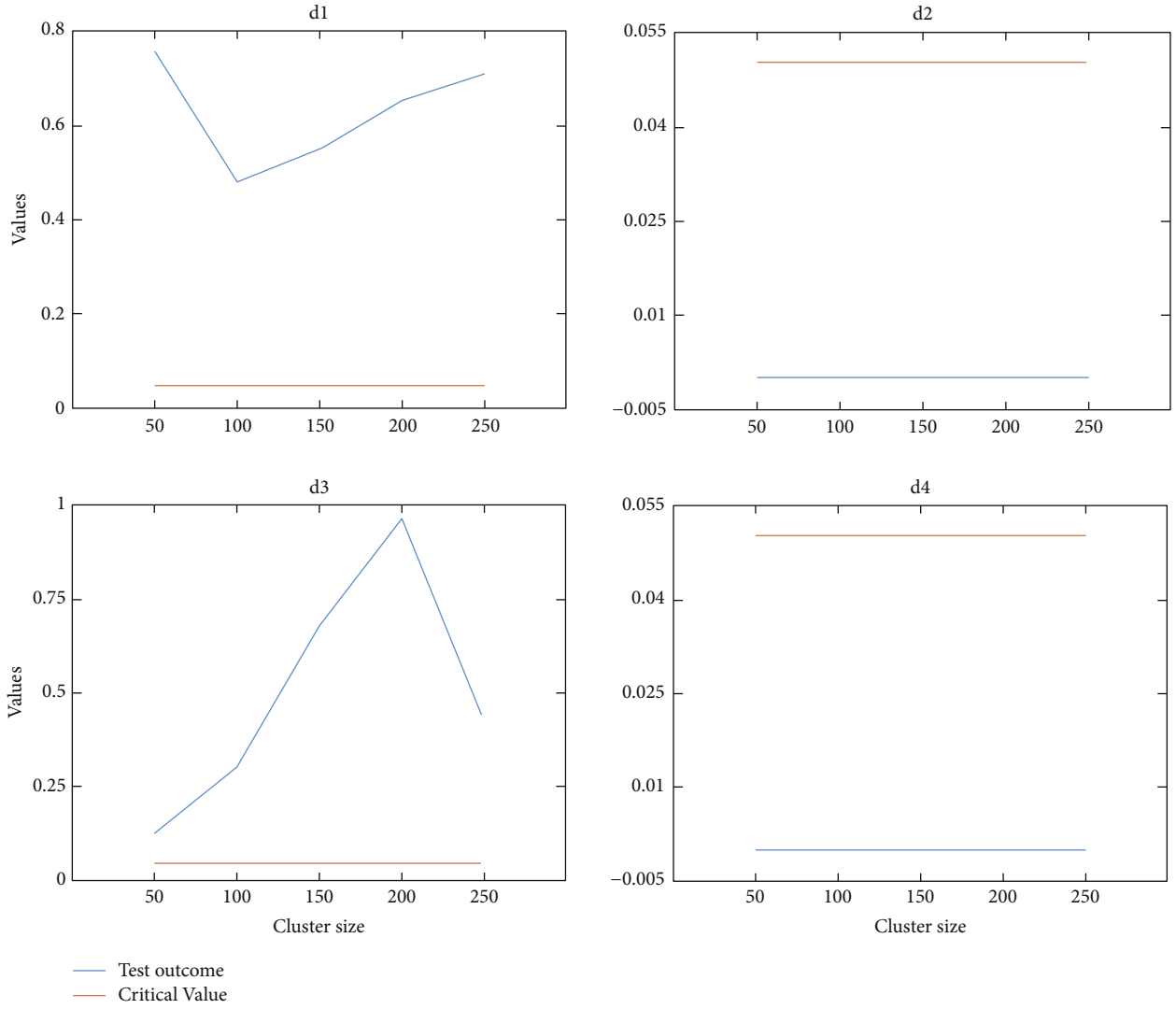

Figure 13: Impact of cluster size and time series effects on test results and critical values of KW test.

parameters. Further, critical values for all the tests are also reduced under consideration of k_1 and k_2 values or d_t. To check whether time series is not stationary according to the ADF and PP tests, $k_1 = k_2 = 0$ model is better for use. The results obtained under consideration of parameters $k_1 \neq 0$ with $k_2 = 0$ and $k_1 \neq 0$ with $k_2 \neq 0$ can be compared with of obtained with $k_1 = k_2 = 0$ as they indicate the necessary actions required for stationarizing the considered data. A similar approach can be used to check the presence of a trend with the KPSS test and the presence of drift and time-varying trends with the Breitung test.

4.1.2. Impact of Lag Number on ADF Test's Outcome. ADF test involves the determination of n_l values as it is an integral part of the model equation (refer to Figure 1). These lag numbers significantly impact ADF test results under various time-series effects. In all the cases, the test outcome decreases for $n_l = 1$, and then inconsistently rises as the n_l value increase to $n_{l,\max} = 17$. The critical values also decrease uniformly by a minimal factor with an increment in n_l value. Overall, it is seen that unbiased results are obtained for higher lag numbers. If the lag number is too small or too large, a loss in power of the ADF test is noticed [47]. This increases the chance of getting biased results. The trend component has notable effects on the test results as the entire plot shifts upward for d2 and d4, which can be visualized in Figure 5 as unit root tests can characterize the trend component effectively. The volatility effect is seen biasing the test outcome as the plots for cases d3 and d4 shift slightly down compared to d1 and d2 (refer to Figure 5).

4.1.3. Impact of Test Parameters on MK Test's Outcome. For the MK test, considering seasonality is an essential parameter, as seasonal variations can lead to biased test outcomes. Consideration of seasonality enables the test to detect trend components in time series in the presence of seasonality by ignoring the seasonal variations, as it is possible that these variations can bias the test results because

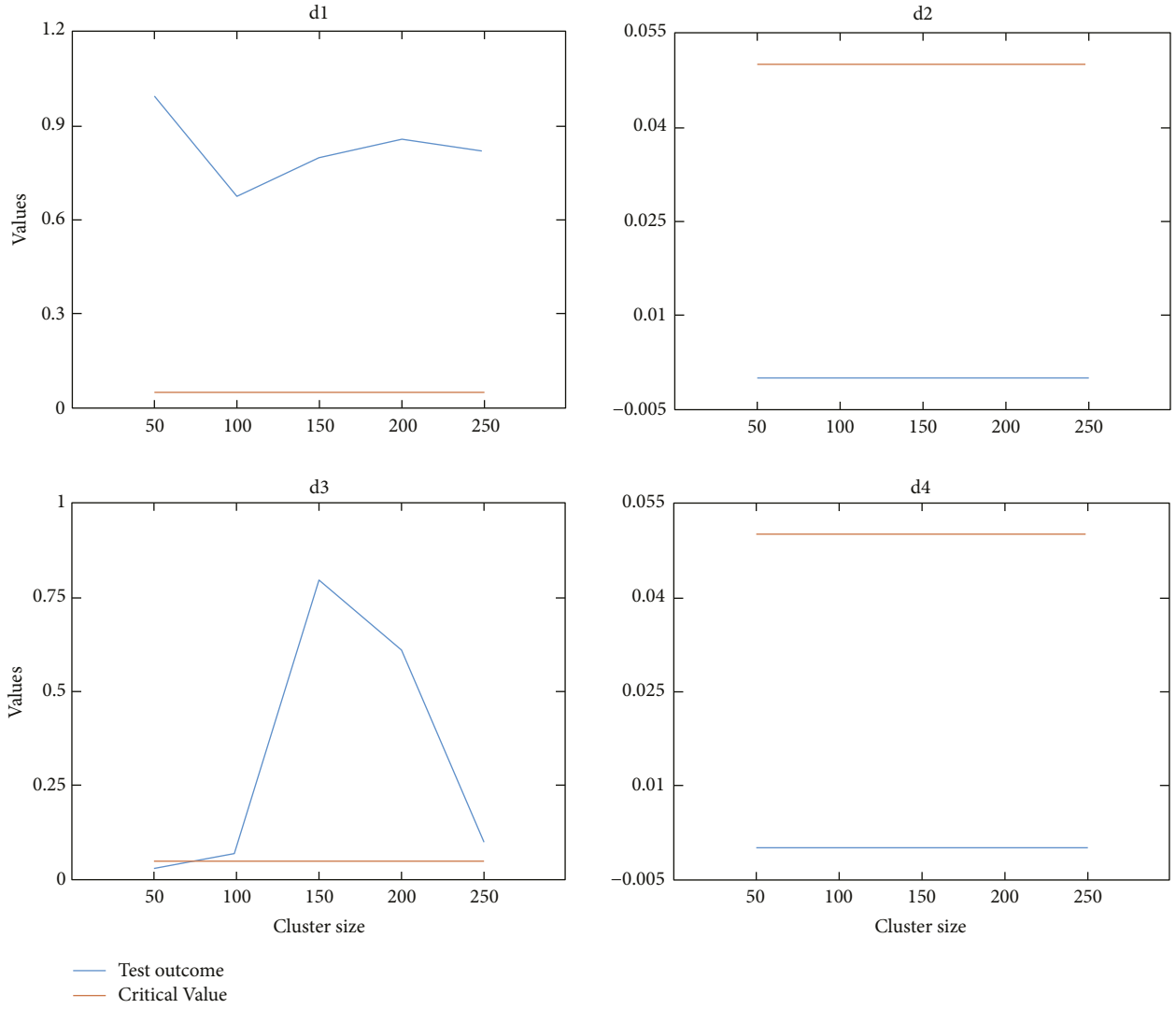

Figure 14: Impact of cluster size and time series effects on test results and critical values of KS test.

the test detects a monotonic trend by examining a gradual difference between data over time. MK test is designed to detect monotonic trends; hence, it effectively detects trend components in all the data (refer to Table 6). Biased results are obtained for the MK test, even considering seasonality due to the test's high frequency of committing type-II errors. Thus, the overall outcomes are better when no seasonality is considered. Also, volatility has minimal or no effects on test outcomes of the MK test.

4.1.4. Impact of Test Parameters on Levene's Test's Outcome. Levene's test is designed to test the stationarity of a time series by dividing it into clusters, and the test formulations require the calculation of the mean or median of the groups (refer to Figure 1). Therefore, consideration of the mean or median for time series analysis is a vital parameter for the test. In this analysis, the value of "G" is taken as 100. It is noticed that Levene's test using mean has an overall better performance than that using the median as the results obtained are more biased for analysis using median than that using mean (refer to Table 7). The volatility effect is evidently characterized by Levene's test, which can be seen in the results for d3 and d4. The test is seen as giving biased results for seasonality and trend components as the test is only capable of detecting a difference in variance. There is no significant difference in the variances of different groups, so biased results are obtained.

4.2. Impact of Time Series Effects on the Outcome of KW, KS, and SW Tests. KW test detects a difference in mean values between groups, whereas the two-way KS test checks the difference between the distributions of two groups. Thus, analyzing the time series is divided into five equal groups with $G = 100$ for the KW test and two groups with $G = 250$ each for the KW test. Both tests significantly characterized the trend's impact, as seen in Table 7. The test outcomes are returned as 0, indicating that the stationary null is being rejected with utmost confidence. Biased results are obtained with both the tests for data in the absence of trend components, e.g., d1 and d3, due to the incapability of these tests to detect seasonality. But overall, the KS test performed better than the KW test as the results for KW tests are very

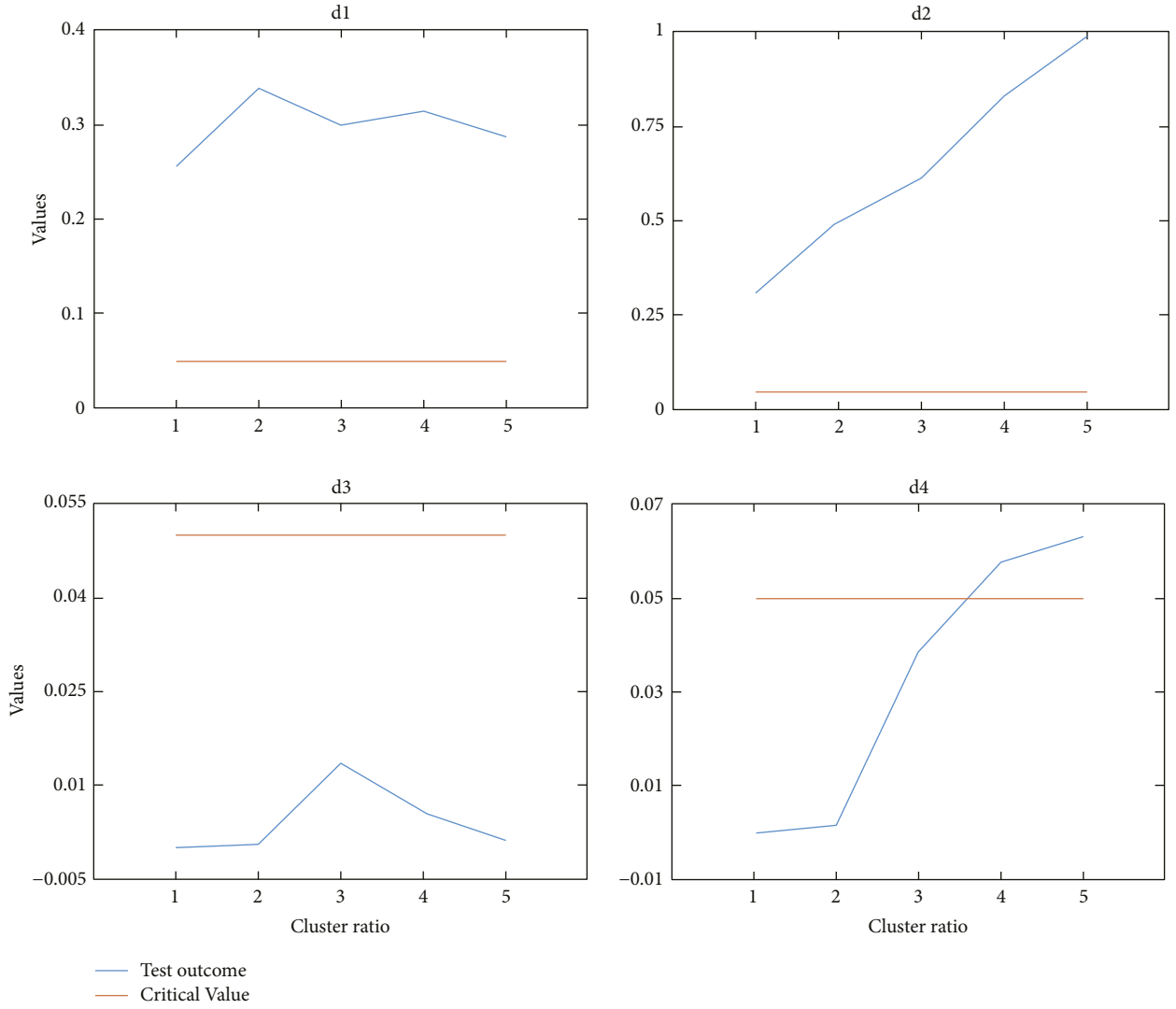

Figure 15: Impact of cluster ratio and time series effects on test results and critical values of Levene's test.

significantly biased compared to that of the KS test. Also, the volatility effect is quite effectively characterized by the KS test, as inferred through the results. Lastly, the SW test has given unbiased results. It is clear from the results that the test can detect seasonality (refer to Table 8). For d2 and d4, which have a trend component, an increase is seen in values of test results though insignificant. Similarly, the test result value rose for d3, having a volatility effect, but the difference is trivial. Thus, trend and volatility effects have minimal impacts on the SW test outcomes.

4.3. Impact of Time Series Lengths on Stationarity Tests' Outcomes. Time series length is an essential feature of a time series. A higher number of data points can assist in building an effective time series model. Therefore, for analysis of test results of ADF, KPSS, PP, Breitung, MK, and SW tests for different time series lengths, the same four sets of time series are considered, with lengths 100, 200, 300, 400, and 500. The variations in test results and critical values are visualized for all the given datasets and are thoroughly analyzed (refer to Figures 6 through Figure 11). For ADF and PP tests, the critical values rise by a small factor as the length increases. In contrast, the Breitung test's critical values decrease as length increases. There is no change in the critical values for KPSS, MK, and SW tests. ADF test result values show slight variations for length with d1, primarily due to seasonal variations. Further, if the same plot is seen for d2, test outcome values are increased as length increases due to the increasing impact of the trend effect. As the ADF test can characterize trend components effectively, the test results indicate higher nonstationarity. Considering the plots in d3 and d4, it is noticeable that they are similar to d1 and d3 but with more significant variations due to the volatility effect. KPSS test results are biased as the length increases for time series with no trend effect. The result values keep decreasing, implying that the null hypothesis is being retained with a higher significance indicating stronger stationarity. While d2 and d4 datasets have an inherent trend component, the outcome values increase with length. The outcome is stationary for

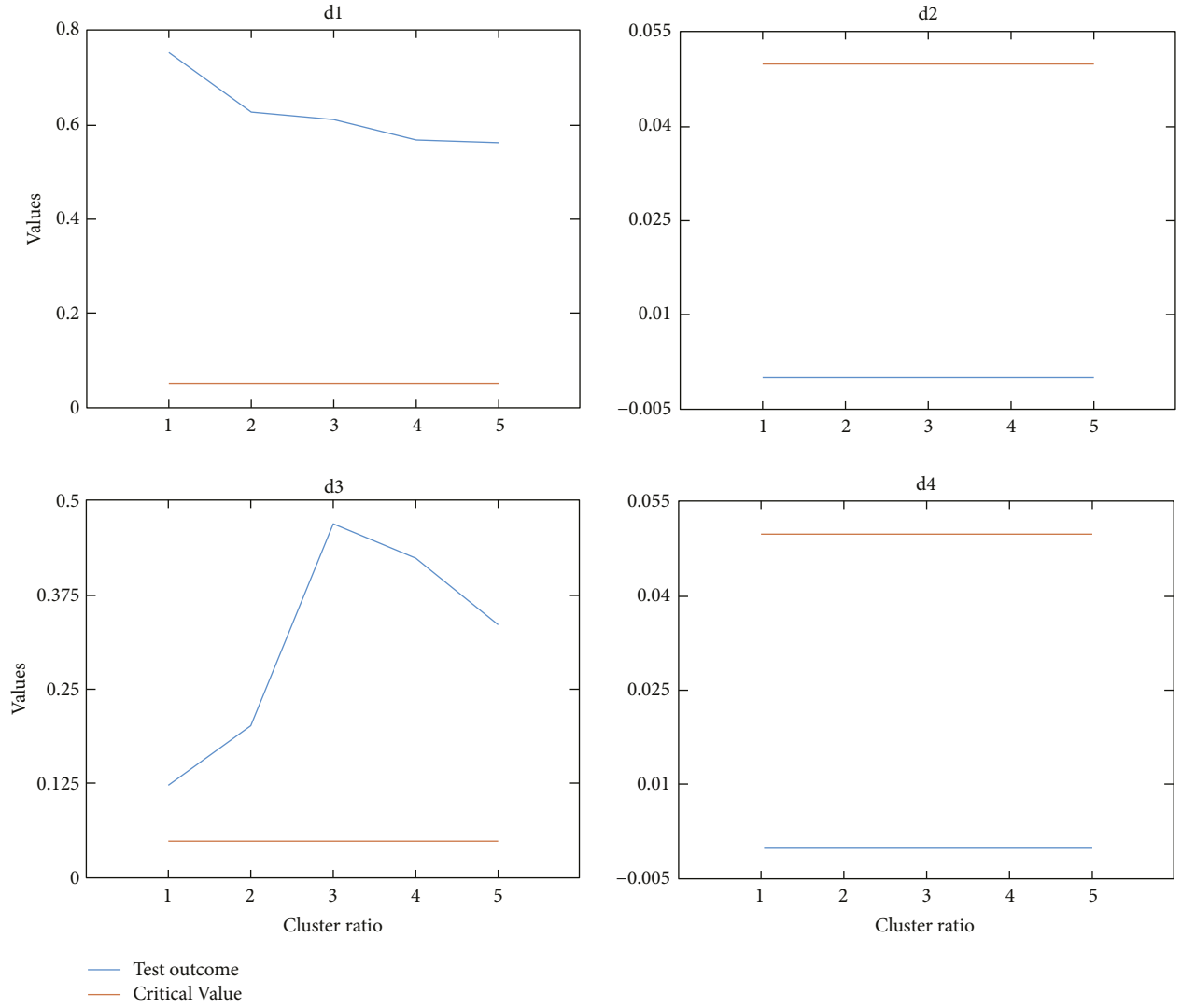

Figure 16: Impact of cluster ratio and time series effects on test results and critical values of KW test.

lengths 200 and 400, as there is no significant trend effect in these lengths. As the length increases, the stationarity null gets rejected with increasing significance due to the trend component. Volatility biases the test results overall as it can be noticed that for d3 and d4, the results indicate stronger stationarity or weaker nonstationarity than those of d1 and d2. Further, PP test result values decrease with an increase in length.

Considering that all the datasets under analysis are nonstationary, a decrease in outcome value with an increment in length indicates that the PP test performs better for time series with lower lengths. The test can also detect seasonal variations for lower lengths, possibly because the test characterizes the variation as a form of the stochastic trend for lower lengths. The test again effectively detects the trend component, as shown in Figure 8, as it is a unit root test. Similar to that of the KPSS test, the volatility effect biases the test results of the PP test too. In the case of the Breitung test, the test result values also decrease minimally as the length increases for d1 and d3 with no trend effect. The difference in the test results and critical values is more or less maintained. But, for d2 and d4, the test results decreased significantly with increased length. The trend effect becomes more prominent as the length increases, and it slightly biases the test results. Volatility has no significant impact on the test results, which can be noticed by comparing plots for d1 and d2 with d3 and d4, respectively. Further, the MK test has two lines of critical values as it is a two-tailed test, and seasonality is considered in this analysis. Trend component is detected for very low and very high lengths in all cases due to high Type-II error rates. For d2 and d4, as the length increases, the test rejects stationary null with higher significance as the trend effect becomes more and more prominent as the length increases. There are changes in plots of d3 and d4 to d1 and d2, respectively, indicating differences in results due to the volatility effect. The MK test possibly detects the volatility effect as a form of trend, and thus, nonstationary outcomes are yielded. Lastly, for the SW test, no significant changes are noticed for time series length for all the datasets. The test is seen behaving optimally against all

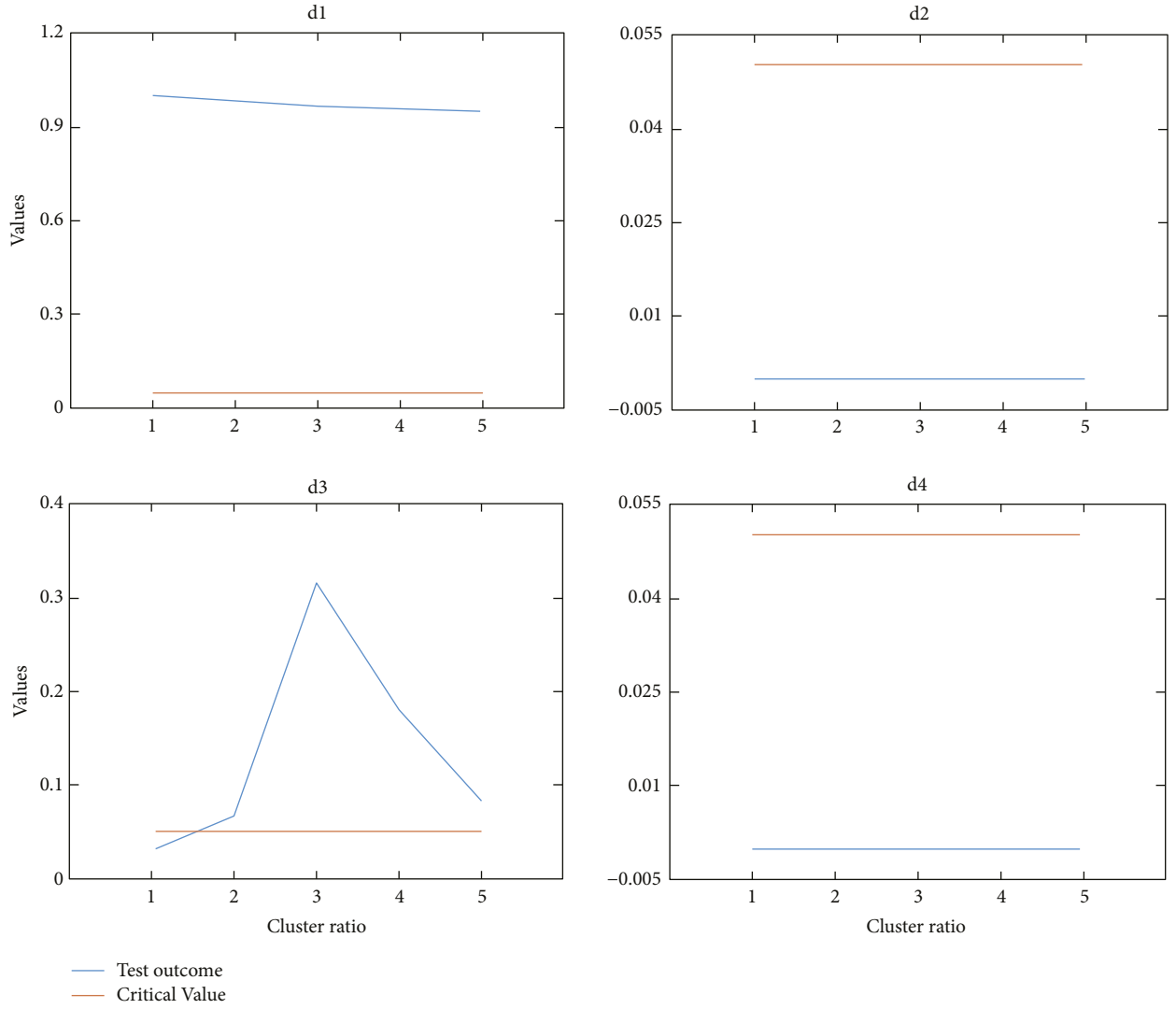

Figure 17: Impact of cluster ratio and time series effects on test results and critical values of KS test.

the considered lengths for all the time series effects, further stating that these components have no significant impact on the test.

4.4. Impact of Time Series Clustering on Stationarity Tests' Outcome. A threefold analysis of impacts of time series clustering considering various time series effects is executed. Levene's, KW, and KS tests can be used with equal and unequal-sized clusters. For equal-sized group analysis, two groups are taken with sizes varying from 50 to 250 with a difference of 50. Next, the term cluster ratio is employed to conduct a uniform analysis of these tests for unequal-sized groups. It is defined as the ratio of the size of group 2 to that of group 1. The cluster size of group 1 is taken as 50 throughout the cluster ratio analysis, whereas the size of group 2 increases to 50, 100, 150, 200, and 250. Therefore, to study the impacts of increasing inequality among groups, the test results are plotted against cluster ratio. Further, Levene's and KW tests can be employed with any clusters, while for the KS test; the number of clusters is always two. Thus, the variations in the test results for all three can be studied for varying cluster sizes (equal cluster sizes) and increasing cluster ratios (unequal cluster sizes), keeping the number of clusters as two. Hence, the impacts of the overall size of the clusters and the difference in sizes of the clusters on the test results can be visualized. For Levene's and KW's tests, the impacts of the number of clusters are also studied with a constant group size of 50. After analyzing the effects of changing sizes of two equal clusters, it can be noted that Levene's test performs better with a larger cluster size in the presence of only the seasonality effect. From all other plots in Figure 12, it can be marked that the test yields better results for a small sample size. Trend effect yet continues to bias results while the test effectually detects volatility. Actually, using a smaller cluster size would be beneficial for obtaining more precise results. Still, the value of "G" should not be minimal as it could call a stationary time series nonstationary. Further, from Figures 13 and 14, KW and KS tests have similar variations in test results. A drop is seen in result

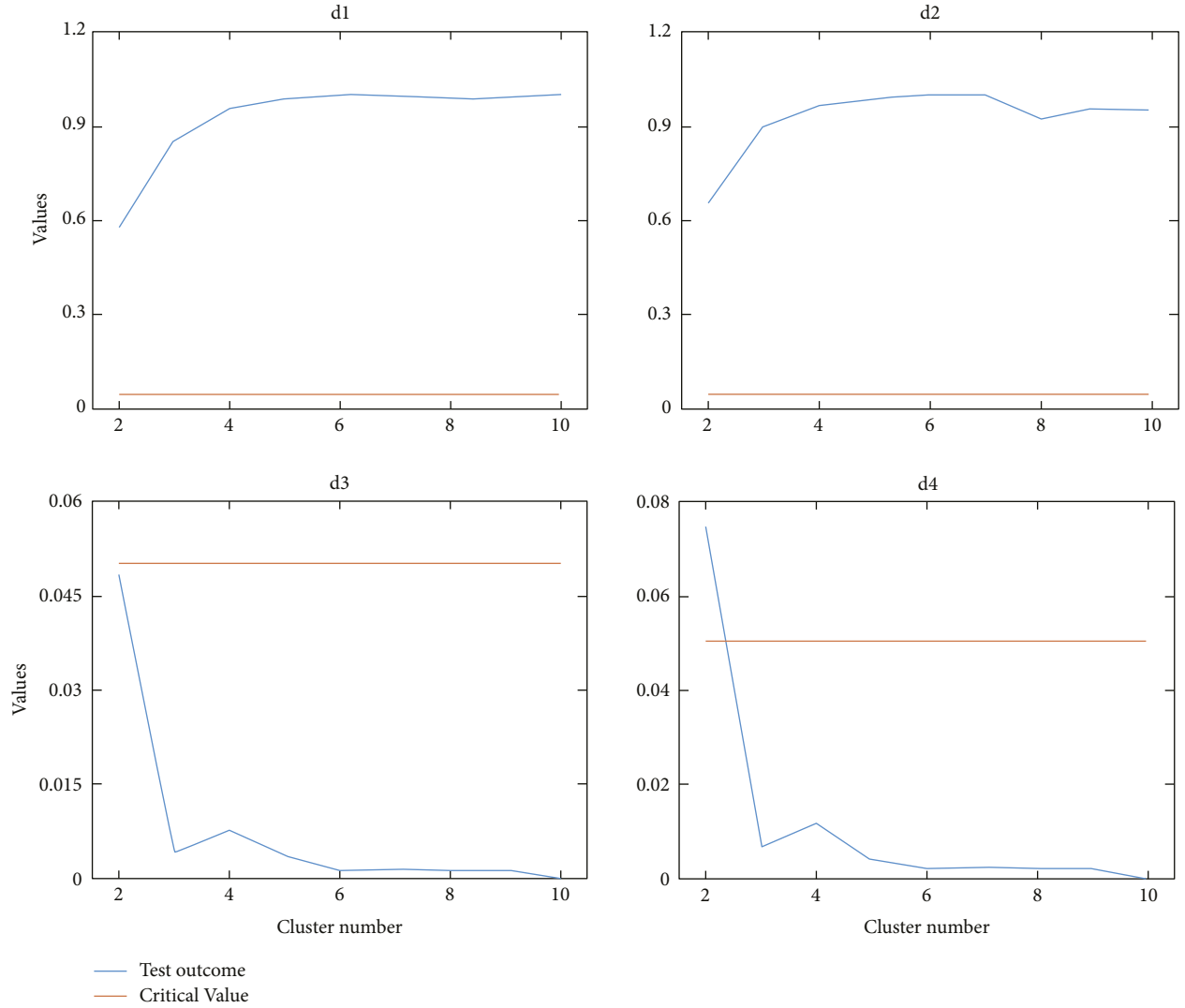

Figure 18: Impact of the number of clusters and time series effects on test results and critical values of Levene's test.

values for $G = 100$ and a further increase is seen as "G" values increments for d1. Likewise, for d3, the result values rise as "G" increases to 200 for the KW test and 150 for the KS test, and then a drop is seen. Overall, the performance of the tests is better for smaller cluster sizes for seasonal and volatile data. In addition to a trend component, as in d2 and d4, perfect results are obtained as these tests can effectively characterize the trend component.

Further, on studying the impacts of changes in cluster ratio on the test results, it can be seen that cluster ratio difference does not significantly affect Levene's test's results for detection of seasonality solely, which can be seen from results for d1 in Figure 15. For d3, overall performance is better for the lower cluster ratio, while an improvement is seen in the results again for a very high cluster ratio. For d2 and d4, having a trend effect, better results are obtained for a lower cluster ratio. Increasing cluster ratio also escalates the biased nature of the test's results for d4. KW test results for d1 indicate that a higher cluster ratio is preferable for detecting seasonality, whereas, for d3, overall performance is improved for a lower cluster ratio. The results show enhancement for a very high cluster ratio (refer to Figure 16). Further considering the KS test, somewhat similar observations are noticed. Variations in results are detected for d3 in Figure 17, considering all these tests due to the strong volatility effect. Lastly, considering d2 and d4, perfect results are obtained for KW and KS tests as both these tests are capable of noticing trend components.

Considering the number of clusters, Figure 18 shows that Levene's test yields better outcomes for lower cluster numbers for d1 and d2. In contrast, for d3 and d4, unbiased and more robust results are obtained for a higher number of clusters. The test is well-suited for analyzing volatility effects. Therefore, the test can notice more variations in data if the number of groups increases, as for d3 and d4. As the data are not volatile for d1 and d2, the variations noticed for the amount of data under testing dominate lower cluster numbers. The test's biased output becomes stronger. Lastly, the KW test fails to detect seasonality in d1 (refer to Figure 19), but comparatively better results are obtained when the cluster number is lower. Further, for d3, vicissitudes are seen in the plots due to the volatility effect. Considering the

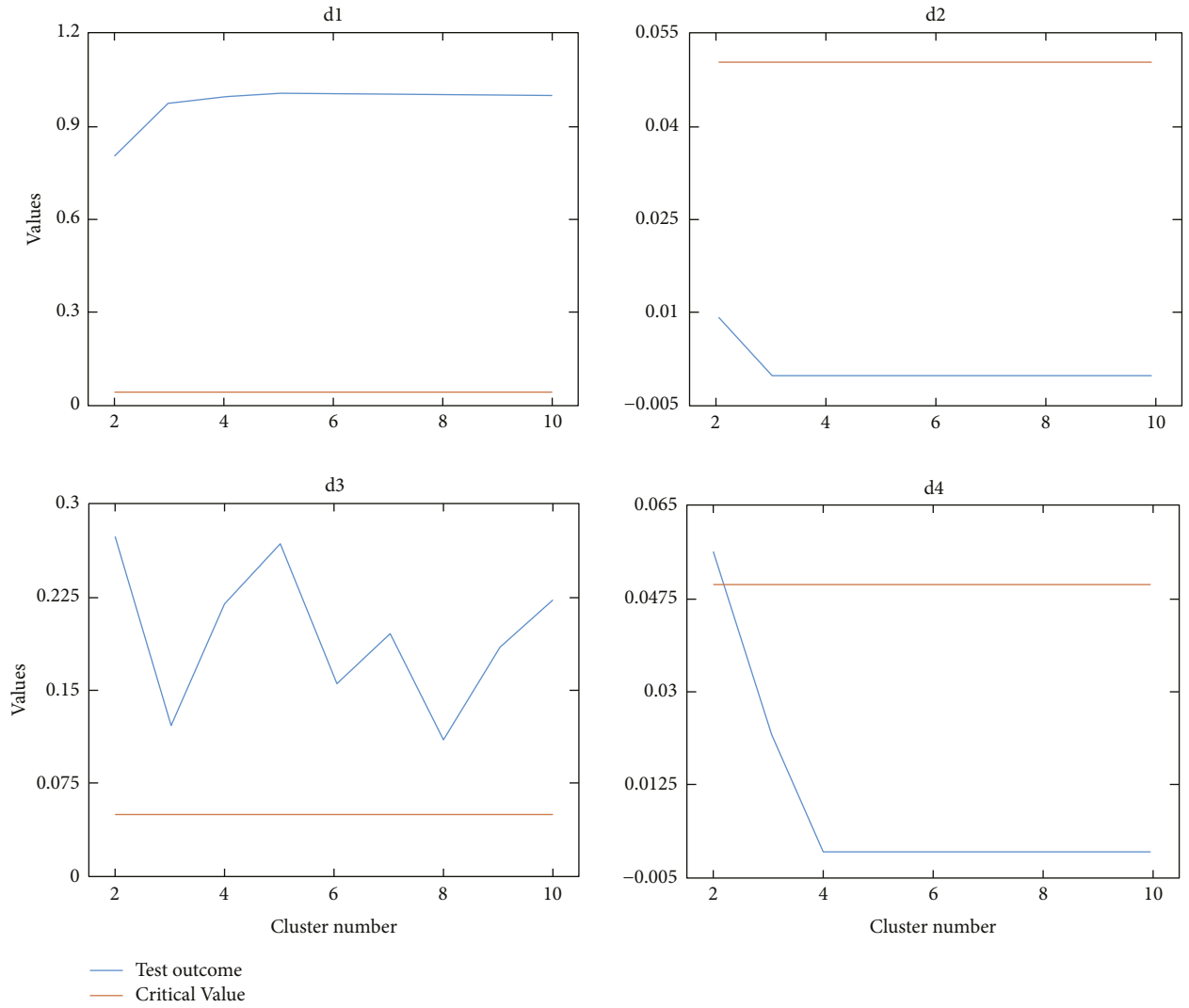

Figure 19: Impact of the number of clusters and time series effects on test results and critical values of Levene's test.

plots for d2 and d4, better and unbiased results are noted for a higher number of clusters as the trend effect is effectively highlighted by increasing the number of clusters.

4.5. Discussions. It can be noted from the obtained results that in some cases, the test outcome is contradictory, i.e., a nonstationary time series is represented as stationary or vice-versa. This can be due to the limitations in tests' working because they cannot characterize the nonstationary facets. If the test's working is entirely suitable for the desired application, calculating the power of a test would be beneficial to note how reliable the tests are in assisting the tests' selection [16]. Calculation of power of a test provides information on whether the test is completely efficient in distinguishing the given data as per its null and alternative hypotheses. A test cannot be considered reliable if it does not have good power. Also, using two tests capable of characterizing the same component in time series is not helpful. Further, it can be seen that all the tests cannot characterize all the nonstationary time series facets. Few of these tests also have biased results over changes in other attributes such as time series length or clustering and tests' parameters. For detecting trend components, all the tests except for Levene's test are capable of varying strength in the unbiased examination, but the best among these are KW and KS tests. After that, only Breitung and SW tests could detect seasonality in all datasets. Lastly, Levene's and SW's tests could only characterize the volatility effect. SW test can characterize all components effectively, and its results do not bias by changes in time series length. But, if the SW test generates a nonstationary outcome, it does not always mean that the time series is not stationary. The time series could have a stationary non-normal distribution. In that case, a combo of Breitung, Levene's, and KS/KW tests could be instrumental. The Breitung test can be considered with no deterministic components based on the results. Levene's test can be employed using small cluster sizes, more or less equal group

sizes, and higher cluster numbers. KW and KS tests can be used with any cluster size and cluster ratio, but a higher cluster number is preferable for the KW test.

Thus, the detailed result analysis and corresponding discussions would help a novice reader understand.

(1) Working models of various stationarity tests and the similarities and differences between them.

(2) Impacts of time series length and other time series facets such as trend, volatility, and seasonality on the test outcomes that assist in the selection of tests.

(3) The correct usage of clustering and various other test parameters helps achieve good efficiency in acquiring unbiased test outcomes.

(4) The best combination of tests that can be used to attain accurate information on the stationarity of a time series.

5. Conclusion

This research aimed to compare and examine time series stationarity tests, considering the effects of various time series facets to help the readers choose the best test for a given application. Nine well-established tests for stationarity were compared and analyzed thoroughly. The impacts of various time series effects on the test results with respect to various test parameters, time series lengths, and various types of time series clustering were taken into account. These impacts on the tests' performance were studied in detail by applying them to clean data and the same data with synthetically embedded trend and volatility effects. Test results and critical values of the tests were compared for all the above analyses that highlighted their capabilities in characterizing all these effects for various test parameters. The variations in capabilities of tests in portraying all these effects through their results for different test parameters helped in choosing these parameters for additional analysis. Based on the results and discussions, it is suggested to use the SW test first, and if a nonstationary outcome is generated by the test, then Breitung, Levene's, and KW/KS test is used. It is to mention here that a test's performance should be analyzed considering all possible parameters' settings to ascertain the optimal/appropriate parameter value to yield an unbiased test outcome. All the attributes considered in this paper are within certain limits. Possibly, more attributes exist that are not considered in this study. Further, the obtained test results may be biased. Critical research establishing the above future scope would add one more dimension to the judicious selection of stationarity tests for a particular application.

Nomenclature

ADF: Augmented Dickey–Fuller
ANOVA: Analysis of variance
AR: Autoregressive
ECDF: Empirical cumulative distribution function
GLS: Generalized least squares
KPSS: Kwiatkowski Phillips Schmidt Shin
KS: Kolmogorov Smirnov
KW: Kruskal Wallis
LM: Lagrange multiplier
OLS: Ordinary least squares
PP: Phillips Perron
PV: Photovoltaic
SW: Shapiro–Wilk
VR: Variance ratio
WGN: White Gaussian noise
$\widehat{\delta}_u$: Test statistic for Breitung test
ε: White noise
ζ , ϕ, ψ: Model parameters
$\theta(q)$: Variance term associated with Z-statistic for VR test
ρ: Parameter of lag 1 term σ
$\widehat{\rho}$: OLS estimator of parameter ρ
σ: Standard deviation
$\widehat{\sigma}_{\rho}$: OLS estimator of σ_{ρ}
σ^2: Variance
$\widehat{\sigma}_{\varepsilon}^2$: Long run estimate of variance σ_{ε}^2
a_i: Coefficient for i^{th} ordered data sample in SW test
c_{α}: Level of significance
d_t: Deterministic part of time series
i, j, l: Counting variable
k_1, k_2: Constant/intercept term and linear trend parameter respectively
r_i: i^{th} rank of data in KW test
n: Integer part of T/q in VR test; q is the period of VR test
n_l: Lag number for ADF test
$n_{l,\max}$: Maximum value of n_l
s_i: Seasonal interval or seasonal period
sgn(.): sign function used to extract sign of a given real number
$\widehat{u}$: Residual part of time series
$\overline{\overline{x}}$: Overall mean of all the groups in Levene's test
y: Time series data
$\overline{y}_i$: Mean of data in i^{th} group
$\widetilde{y}_i$: Median of data in i^{th} group
D_{n_1,n_2}: Test statistic for KS test
$E(.)$: Calculates expected value
G: Group size
H: Test statistic for KW test
H_0: Null hypothesis
H_A: Alternate hypothesis
$I(L)$: "L" order of integration (differencing of lag "L")
LM_s: LM statistic for KPSS test
P_D: Power consumption
T: Time series length
TS_{ρ}: Test statistic for ADF and PP tests
W: Test statistic for Levene's and SW tests
Z_{MK}: Test statistic for MK test

Data Availability

Data are openly available in a public repository [46] that does not issue DOIs.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Fan, V. Vittal, G. T. Heydt, and R. Ayyanar, "Probabilistic power flow studies for transmission systems with photovoltaic generation using cumulants," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2251–2261, 2012.

[2] B. R. Prusty and D. Jena, "A critical review on probabilistic load flow studies in uncertainty constrained power systems with photovoltaic generation and a new approach," *Renewable and Sustainable Energy Reviews*, vol. 69, pp. 1286–1302, 2017.

[3] D. D. Le, G. Gross, and A. Berizzi, "Probabilistic modeling of multisite wind farm production for scenario-based applications," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 3, pp. 748–758, 2015.

[4] B. R. Prusty and D. Jena, "A sensitivity matrix-based temperature-augmented probabilistic load flow study," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 2506–2516, 2017.

[5] B. R. Prusty and D. Jena, "A spatiotemporal probabilistic model-based temperature-augmented probabilistic load flow considering PV generations," *International Transactions on Electrical Energy Systems*, vol. 29, no. 5, pp. 28199–e2913, 2019.

[6] C. Chatfield, *Time-series Forecasting*, CRC Press, Boca Raton, Florida, 2000.

[7] S. Porter-Hudak, "An application of the seasonal fractionally differenced model to the monetary aggregates," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 338–344, 1990.

[8] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *Journal of Climate and Applied Meteorology*, vol. 23, no. 8, pp. 1184–1195, 1984.

[9] D. Fedorová, "Selection of unit root test on the basis of length of the time series and value of AR (1) parameter," *Statistika*, vol. 96, p. 3, 2016.

[10] X. Zuo, "Several important unit root tests," in *Proceedings of the 2019 2nd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 10–14, IEEE, Weihai, China, September 2019.

[11] A. Lojowska, D. Kurowicka, G. Papaefthymiou, and L. van der Sluis, "Advantages of ARMA-GARCH wind speed time series modeling," in *Proceedings of the 2010 11th International Conference on Probabilistic Methods Applied to Power Systems*, pp. 83–88, IEEE, Singapore, June 2010.

[12] N. Masseran, A. M. Razali, K. Ibrahim, and W. Wan Zin, "Evaluating the wind speed persistence for several wind stations in Peninsular Malaysia," *Energy*, vol. 37, no. 1, pp. 649–656, 2012.

[13] C. Modin, "Short-term wind power forecasting," *Doctor, Economy and Society*, Dalarna University, Falun, Sweden, 2009.

[14] I. Ebtehaj, H. Bonakdari, M. Zeynoddin, B. Gharabaghi, and A. Azari, "Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models," *International journal of Environmental Science and Technology*, vol. 17, no. 1, pp. 505–524, 2020.

[15] D. Yang, "Choice of clear-sky model in solar forecasting," *Journal of Renewable and Sustainable Energy*, vol. 12, no. 2, Article ID 026101, 2020.

[16] A. A. Bawdekar and B. R. Prusty, "Selection of stationarity tests for time series forecasting using reliability analysis," *Mathematical Problems in Engineering*, vol. 20228 pages, Article ID 5687518, 2022.

[17] F. N. Melzi, T. Touati, A. Same, and L. Oukhellou, "Hourly solar irradiance forecasting based on machine learning models," in *Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 441–446, IEEE, Anaheim, CA, USA, December 2016.

[18] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, 2013.

[19] S. Atique, S. Noureen, V. Roy, V. Subburaj, S. Bayne, and J. Macfie, "Forecasting of total daily solar energy generation using ARIMA: a case study," in *Proceedings of the 2019 9th annual computing and communication workshop and conference (CCWC)*, pp. 0114–0119, IEEE, Las Vegas, NV, USA, January 2019.

[20] R. P. Silva, B. B. Zarpelão, A. Cano, and S. B. Junior, "Time series segmentation based on stationarity analysis to improve new samples prediction," *Sensors*, vol. 21, p. 7333, 2021.

[21] L. Lijuan, L. Hongliang, W. Jun, and B. Hai, "A novel model for wind power forecasting based on Markov residual correction," in *Proceedings of the IREC2015 6th International Renewable Energy Congress*, pp. 1–5, IEEE, Sousse, Tunisia, March 2015.

[22] J. Wang, Q. Zhou, and X. Zhang, "Wind power forecasting based on time series ARMA model," in *Proceedings of the 2018 IOP Conference Series: Earth and Environmental Science*, Banda Aceh, Indonesia, September 2018.

[23] S. Gao, Y. He, and H. Chen, "Wind speed forecast for wind farms based on ARMA-ARCH model," in *Proceedings of the 2009 International Conference on Sustainable Power Generation and Supply*, pp. 1–4, IEEE, Nanjing, China, April 2009.

[24] H. Chen, Q. Wan, F. Li, and Y. Wang, "GARCH in mean type models for wind power forecasting," in *Proceedings of the 2013 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, Vancouver, BC, July 2013.

[25] X. Hu, J. Jaraite, and A. Kažukauskas, "The effects of wind power on electricity markets: an evaluation using the Swedish electricity market data," Department of Economics, Umeå University, Umeå, Sweden, 2020.

[26] N. M. Razali and Y. B. Wah, "Power comparisons of shapiro-wilk, Kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of statistical modeling and analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[27] J. Li, G. Lyu, and H. Zhang, "Characteristics and forecast of short-term wind speed series in the Donghai Bridge wind farm," *SCIENTIA SINICA Physica, Mechanica & Astronomica*, vol. 46, no. 12, Article ID 124713, 2016.

[28] N. Bokde, A. Feijóo, N. Al-Ansari, S. Tao, and Z. M. Yaseen, "The hybridization of ensemble empirical mode decomposition with forecasting models: application of short-term wind speed and power modeling," *Energies*, vol. 13, no. 7, p. 1666, 2020.

[29] N. H. Hussin, F. Yusof, R. Jamaludin, and S. M. Norrulashikin, "Forecasting wind speed in peninsular Malaysia: an application of ARIMA and ARIMA-GARCH

models," *Pertanika Journal of Science and Technology*, vol. 29, no. 1, 2021.

[30] H. Wilms, M. Cupelli, and A. Monti, "On the necessity of exogenous variables for load, pv and wind day-ahead forecasts using recurrent neural networks," in *Proceedings of the 2018 Electrical Power and Energy Conference (EPEC)*, pp. 1–6, IEEE, Toronto, ON, Canada, October 2018.

[31] N. Odam and F. P. de Vries, "Innovation modelling and multi-factor learning in wind energy technology," *Energy Economics*, vol. 85, Article ID 104594, 2020.

[32] J. L. Carrion-i-Silvestre and A. Sansó, "A guide to the computation of stationarity tests," *Empirical Economics*, vol. 31, no. 2, pp. 433–448, 2006.

[33] R. Gimeno, B. Manchado, and R. Minguez, "Stationarity tests for financial time series," *Physica A: Statistical Mechanics and Its Applications*, vol. 269, no. 1, pp. 72–78, 1999.

[34] 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.

[35] T. K. Kim, "T test as a parametric statistic," *Korean Journal of Anesthesiology*, vol. 68, no. 6, p. 540, 2015.

[36] R. C. Sprinthall, *Basic Statistical Analysis*, Pearson Education, London, UK, 9th. Edition, 2012.

[37] Y. W. Cheung and K. S. Lai, "Lag order and critical values of the augmented Dickey–Fuller test," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 277–280, 1995.

[38] J. Breitung, "Nonparametric tests for unit roots and cointegration," *Journal of Econometrics*, vol. 108, no. 2, pp. 343–363, 2002.

[39] H. Levene, "Robust tests for equality of variances." contributions to probability and statistics," *Essays in honor of Harold Hotelling*, pp. 278–292, Stanford University Press, Palo Alto, 1960.

[40] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, pp. 583–621, 1952.

[41] 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. 427–431, 1979.

[42] G. Elliott, T. J. Rothenberg, and J. H. Stock, "Efficient tests for an autoregressive unit root," *NBER technical working papers series*, 1992.

[43] R. Simard and P. L'Ecuyer, "Computing the two-sided Kolmogorov-Smirnov distribution," *Journal of Statistical Software*, vol. 39, no. 11, pp. 1–18, 2011.

[44] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root?" *Journal of Econometrics*, vol. 54, no. 1-3, pp. 159–178, 1992.

[45] P. C. B. Phillips, "Time series regression with a unit root," *Econometrica*, vol. 55, no. 2, pp. 277–301, 1987.

[46] "Hourly load consumption," 2022, https://openei.org/datasets/files/961/pub.

[47] G. W. Schwert, "Tests for unit roots: a Monte Carlo investigation," *Journal of Business & Economic Statistics*, vol. 20, no. 1, pp. 5–17, 2002.