Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 5687518, 8 pages
https://doi.org/10.1155/2022/5687518

Research Article

Selection of Stationarity Tests for Time Series Forecasting Using Reliability Analysis

Advait Amol Bawdekar [1] **and B. Rajanarayan Prusty** [2]

[1]*School of Mechanical Engineering, Vellore Institue of Technology, Vellore, India*
[2]*School of Electrical Engineering, Vellore Institute of Technology, Vellore, India*

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

Received 17 April 2022; Accepted 5 May 2022; Published 27 May 2022

Academic Editor: Xiaoshuang Li

Stationarity is an essential concept in time series forecasting. A reliable stationarity test that yields unbiased test outcomes is vital as it is the gateway before a suitable forecasting model development. Renewable generation time series is inherently seasonal, comprising trend components, and often volatile. These characterizing facets alongside time series length tend to bias stationarity tests' outcomes. A critical comparison study to check the tests' reliability is carried out in this paper using different synthetic data required for the case-to-case analysis. Based on the tests' working, reliabilities are analyzed for different time series lengths and group sizes, varying from 200 to 1000 with an increment of 200. This provides information about changes in reliabilities of the tests for various time series lengths or group sizes. This comprehensive comparison report with a necessary set of well-illustrated figures, table, and thorough explanation of the obtained results is expected to help novice readers to select an apt combination of tests for stationarity check for renewable generation applications.

1. Introduction

In renewable generation forecasting, stationarity is a crucial notion [1]. As a result, knowing whether a renewable generation time series is effectively stationarized is important [2, 3]. A reliable stationarity test that can deliver impartial results for a particular application is necessary on this note. Therefore, a set of tests' reliability information would instill enough confidence in the user for the apt selection of tests. The calculation of the power of a test used for the reliability study confirms whether the test behavior is ideal for a set of parameters associated with the test. Any deviations from the ideal for specific parameters indicate that the test is unreliable [4]. Thus, a complete reliability record of tests by analyzing the plot of power vs. test parameters is crucial for the appropriate selection of tests for a given application.

Reliability analysis through power calculation is well-documented in the literature for various tests. Unit root tests examine time series stationarity using the concept of unit root, and power for some of these tests is calculated in [5] and is analyzed for various time series lengths. Similarly, the power calculations in [4, 6] for various data distributions, time series lengths, and significance levels expose the MK test's limitations. The power calculation of Levene's test is well explained in [7, 8], along with a comparison of type-I error probabilities enlightening the test's sensitivity for variance differences, various data distributions, and sample sizes. Besides, the power study and error analysis of the KW test considering sample sizes and data distribution notify the test's limitations [9, 10]. Power plots were also computed for SW and KS tests and are analyzed for various data distributions to study tests' behavior against nonnormal distributions [11–14].

Though reliability analysis of the above well-established tests is presented through detailed reasoning, a comparative analysis of the above tests in a common platform which is vital for assisting in the apt selection of tests for a particular application was not performed in the literature. Furthermore, in the above studies, the analysis of the effect of time

series lengths on tests' reliability was studied only for a few specific tests. Besides certain tests, for example, Levene's and KW tests whose work is based on dividing the data samples into various clusters open up opportunities to perform the reliability analysis for different selections of group sizes. However, such an analysis was never carried out in the literature. Furthermore, for KS and SW tests, no reliability analysis was done with respect to the tests' integral parameters, such as skewness and kurtosis. Lastly, a comparison table of the above tests' merits, demerits, and key application tips for identifying the best set of tests for a particular application is always of interest for novice researchers in time series forecasting.

The authors have considered all the above-highlighted research gaps to provide a unique reliability analysis, and the significant contributions are mentioned as follows. Firstly, the importance of power calculation for reliability study is enlightened, and then power calculation steps for the nine well-established tests are pictorially represented. The basis for stationarity outcome for the above tests alongside their ideal reliability plots is given special attention. Secondly, five different time series lengths/groups are considered for the above tests' reliability analysis to compare their reliability performance critically and to expose the reason for a test being reliable or not reliable for a particular case. Finally, considering their merits and demerits, a critical comparison of the above tests is tabulated and recommended with each test's key application tips for a better outcome.

The remainder of the paper includes a thorough discussion on the importance of reliability analysis, power calculation steps, and comparison of reliability plots.

2. Reliability Analysis of Well-Established Stationarity Tests

A reliable stationarity test is expected to indicate that a time series is stationary if the time series satisfies the conditions for stationarity. Nevertheless, in some instances, a test declares a stationary time series as nonstationary. The calculation of the power of a test, defined as "the probability with which a test detects a divergence from the null hypothesis conditional that the divergence exists," helps indicate the test's capability in yielding a fair outcome. For ADF, PP, Breitung, MK, Levene's, KW, KS, and SW tests, the power 1-β is the probability that the test rejects the null hypothesis conditional that it is actually false [4, 5, 8, 10, 13]. Here, β is the probability of accepting the null hypothesis when it is actually false whereas, for the KPSS test, the power calculation metric is different as the hypothesis in the KPSS test is reversed compared to other unit root tests [5]. Therefore, the power 1-α for this test indicates the probability that the test does not reject the null hypothesis conditional that it is true, where α is the probability of rejecting the null hypothesis when it is actually true. The plots of power of the test indicate the deviations in the test behavior compared to that of in the ideal case. The ideal plots of power for all the well-established tests for reliability analysis are presented in Figure 1. Furthermore, the basis for tests' outcome is highlighted, and various symbols used in Figure 1 are described underneath.

Power is calculated according to the test properties. For example, unit root tests are designed to detect the presence of a unit root. The unit root can be easily characterized using AR(1) process. When the AR(1) *parameter* (φ) is less than 1, the unit root is absent, whereas the unit root is present if $\varphi \geq 1$. In an ideal case, for $\varphi < 1$, the power values should be 1, and it is zero for $\varphi = 1$.

MK test confirms time series stationarity by detecting the presence of a monotonic trend. Hence, for any value of its *slope* (s) other than 0, the test should reject the null hypothesis. Thus, MK test power values are calculated against the slope of the added trend component. Ideally, for $s = 0$, the power value for the MK test should be 0, and for all other s values, it should be 1. Levene's test detects the equality of variances between equal/unequal sized groups created of the original time series. The test must reject the null hypothesis if inequality in standard deviation is present. Therefore, power values for Levene's test are calculated with respect to the difference in standard deviation (d) among the groups. Ideally, the power value should be 0 for $d = 0$, and the power values should be 1 for the rest of the values of d. KW test is used to test the equality of mean values of various groups. Thus, power values for this test are calculated against the difference in mean values (m). In an ideal case, the power value should be 0 for $m = 0$, and for all other values of m, it should be 1.

Lastly, SW and KS tests check whether a time series follows the normal distribution. For a normal distribution, the skewness (γ_1) and kurtosis (γ_2) values are 0, and thus, for any other values of γ_1 and γ_2, the null hypothesis must be rejected by the tests. Thus, power values are computed for these tests with respect to varying γ_1 and γ_2 values, keeping the mean and standard deviation constant for all the cases. Ideally, for 0 values of γ_1 and γ_2, power should be 0, and the power values should be 1 for all other values of γ_1 and γ_2.

3. Power Calculation for Stationarity Tests

Time series length "T" or group size "G" tends to affect stationarity tests' outcomes [5, 7]. Therefore, their impact on the well-established tests' reliability is always of interest. Power calculation being essential for reliability analysis, Figure 2 systematically elucidates power calculation steps for the well-established tests.

3.1. Power Calculation for Unit Root Tests. The power of a unit root test for a specific value of φ can be calculated by following the set of steps as suggested in Figure 2. Calculation of the same for different values of φ ranging from 0 to 1 [5], that is, 0.01, 0.02, ..., 0.99, yields a grid of values of power corresponding to the set of φ. The plots of power vs. φ for various values of "T" can help note how much a test is complying with the ideal behavior, confirming the test's reliability for a particular value of "T." This approach is suitable for unit root tests, and hence, ADF, PP, KPSS, and Breitung tests can be analyzed using this approach.

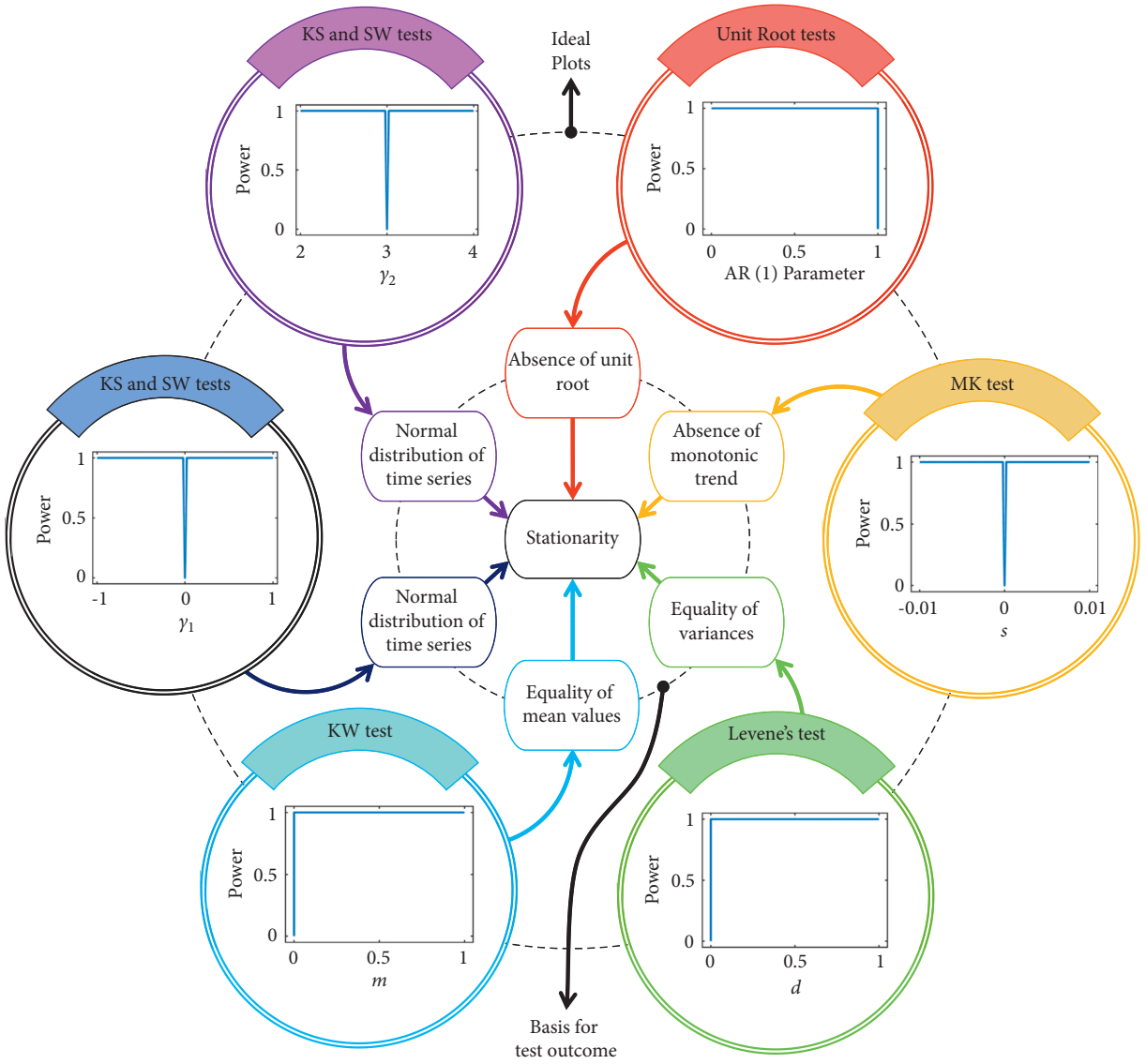

Figure 1: Ideal plots for reliability analysis of stationarity tests.

3.2. Power Calculation for MK Test. The power of the MK test can be calculated for a specific value of s. The values of s range from −0.01 to 0.01 [4], which will provide a set of power values corresponding to the s values. Therefore, the plots of power vs. s for various values of "T" can help comprehend the test's conformity with the ideal case.

3.3. Power Calculation for Levene's Test. The power for Levene's test is calculated for a specific value of d. The values for d range from 0.01 to 1 [7] yield a set of power values corresponding to the d values. Hence, the plots of power vs. d for various values of "G" can help visualize the test's conformity with the ideal case.

3.4. Power Calculation for KW Test. The power of the KW test is calculated for a specific value of m. The values for m range from 0.01 to 1 [10] yield a set of power values corresponding to the m values. The plot of power vs. m for different values of "G" can be envisaged for comprehending the test's conformity with its ideal case.

3.5. Power Calculation for KS and SW Tests. The power for KS and SW tests can be calculated for different time series lengths against various values of γ_1 and γ_2. A plot of power vs. γ_1 and power vs. γ_2 will clarify the behavior of the tests for normal and nonnormal distributions and further comment on tests' reliability.

4. Result Analysis

The selection of a suitable stationarity test for a specific application requires a critical study of well-established tests by revealing their efficacy while handling datasets with various lengths. A thorough survey is always of interest to expose a test's inability or failure to yield unbiased results with respect to critical parameters related to the basic working of the tests. To

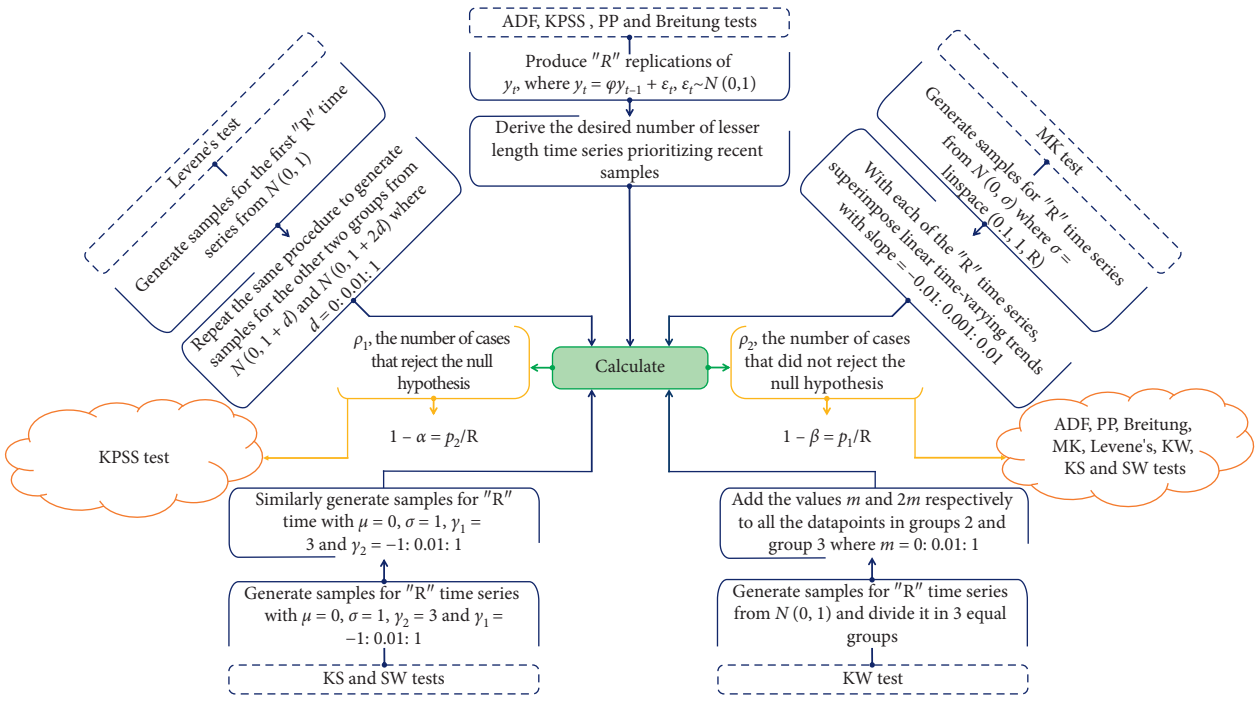

Figure 2: Step-by-step procedure for power calculation of stationarity tests.

provide a detailed comparison report of the reliability of well-established stationarity tests and to reveal the pertinent issues, the analyses carried out in this section are twofold, as listed underneath.

(i) Firstly, the powers of all the well-established tests for different time series lengths/group sizes are compared and critically analyzed

(ii) Next, the merits and demerits of the well-established tests are compared, further suggested with tips for better test outcomes

4.1. Analysis of Power Values of Stationarity Tests. To inspect the tests' reliability, the primary task is to determine the number of replications, that is, the value of "R" (refer to Figure 2). For all the analyses in the simulation study, the value of "R" is set to 3000. The steps as suggested in Figure 2 are followed to construct reliability plots. For randomly chosen five different values of "T/G," the comparison of reliability plots using the nine tests is portrayed in Figures 3–6.

Authors in [5] have carried out a reliability analysis considering only a few selective stationarity tests. The time series lengths chosen for analysis were also very small for real-time applications. Therefore, this work includes other established unit root tests like the Breitung test, and the time series lengths considered for the analysis are suitable for real-time applications. Similar problems are associated with the MK test too. Furthermore, Levene's and KW tests are analyzed with relatively low and high-sized groups to notice any vital changes in the results. The KS and SW tests have been analyzed previously for various data distributions in the literature [12–14]. But this approach does not provide any useful information about the tests' reliabilities based on their working; instead, it provides information on the usability of these tests for various distributions. Therefore, a novel approach is used for KS and SW tests where reliability is analyzed for changing skewness and kurtosis values that notify about any possibilities of discrepancies in outputs of the tests based on their basic working of detecting the normal distribution of data. The ensuing paragraphs critically elucidate the reliability performance of unit root and nonunit root tests.

The power values for the ADF test can be seen to start approaching zero at a lower value of φ for smaller lengths. And, for the increase in length, the power of the ADF test increases (refer to Figure 3). However, no such pattern of change in power is seen for the KPSS test. The KPSS test power value begins to fall to zero for $T = 800$ at the lowest value of φ compared to other lengths. Furthermore, the power plot for $T = 200$ begins to fall at the highest φ value. It is to highlight that the overall performance of the test is better for shorter lengths. Hence, the KPSS test is suggested to be used along with other tests due to its disadvantage of frequently committing a type-I error, leading to discrepancies in the obtained power values. PP test power values follow a similar trend to that of the ADF test, but, for shorter lengths, the PP test is significantly reliable compared to the ADF test. This is because the former uses nonparametrically adjusted test statistics. For the Breitung test, the output is the same for all lengths except for $T = 1000$. It is noticed that the power for $T = 1000$ falls by a small amount for $\varphi = 0.98$ while, for other lengths, the values do not approach zero even for $\varphi = 1$. Here, the test statistic is based on that of the KPSS test, with specific changes made to counter the disadvantages of the KPSS test. Although the power problem of the KPSS test is solved, the

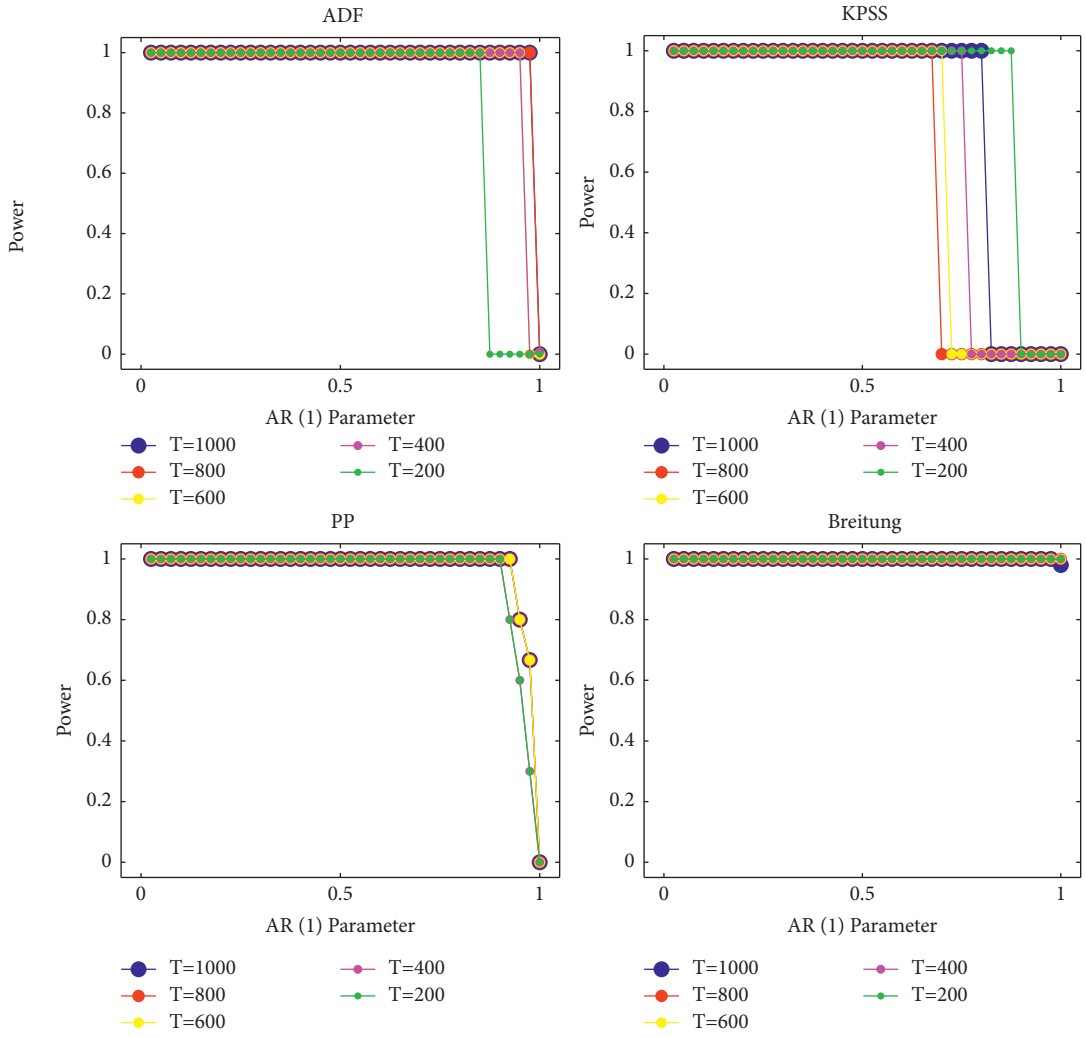

Figure 3: Comparison of power vs. φ for different values of "T" using unit root tests.

issue of power value not approaching to zero for $\varphi = 1$ arose, that is, while solving the type-I error of the KPSS test, the test statistics of the Breitung test became prone to type-II error.

MK test has very low reliability (refer to Figure 4) due to frequent type-I errors. The test formulations cannot differentiate between trend effect and general data highs and lows. Levene's test is completely reliable as the plots obtained are analogous to the ideal plot (refer to Figure 5). KW test fails to detect a very small difference in mean values between groups, and hence, biased results are seen. This biased nature is prominent for lower group sizes. For $G = 200$, power values begin to rise for higher m value as compared to that of $G = 1000$ (refer to Figure 5). Reliability for SW and KS tests is checked against various γ_1 and γ_2 values. SW test performs better than the KS test in both aspects, but a very high rate of committing type-I errors in the SW test results in biased power values with respect to γ_2. The KS test is designed to be sensitive to every form of difference between two distributions leading to low power.

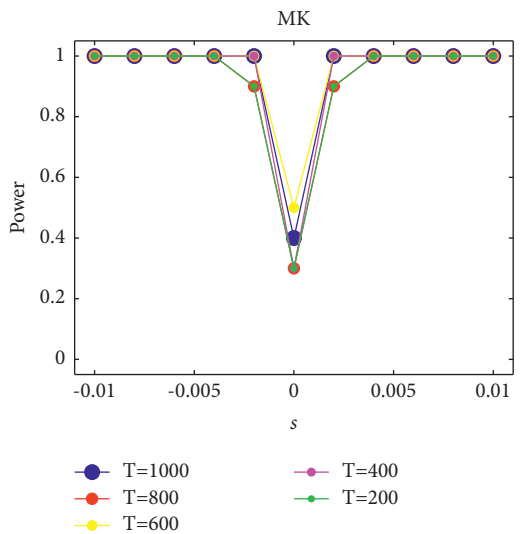

Figure 4: Comparison of power vs. s for different values of "T" using MK test.

4.2. Performance Comparison of Well-Established Stationarity Tests. It is quite clear from the results that all the stationarity tests are not perfectly reliable. Also, the

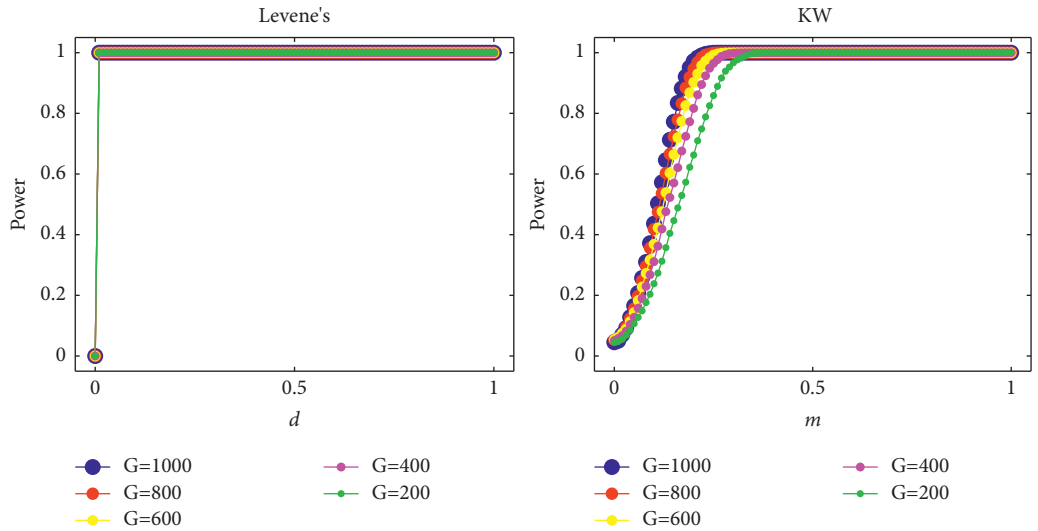

Figure 5: Comparison of power vs. d using Levene's test and power vs. m using the KW test for different values of "G."

Table 1: Performance comparison of stationarity tests.

ADF	Merit	(i) Test is reliable for high time series length.
	Demerits	(i) Calculates low power for a smaller time series length, often resulting in unit root conclusions even for a stationary time series (ii) Inappropriate choice of lag number adversely affects the test results
	Suggestion	(i) Reliable for apt selection of lag number
KPSS	Merits	(i) Test is nonparametric (ii) Test outcome indicates stationary if the time series is strongly stationary
	Demerit	(i) The test statistic is vulnerable to type-I errors lowering the test's reliability
	Suggestion	(i) Reliable for low time series lengths and recommended to be used along with another test
PP	Merits	(i) Test is reliable for high time series lengths (ii) Test is nonparametric
	Demerit	(i) Low reliability for small and moderately large time series lengths due to severe size distortions
	Suggestion	(i) Reliable for higher time series lengths and shorter time series lengths having low parameter value
Breitung	Merits	(i) Reliable for any time series length (ii) Test result is unbiased by any time series characteristics
	Demerit	(i) Fails to detect the presence of unit root for $\varphi = 1$ in absence of other nonstationary components for lower lengths
	Suggestion	(i) The test is helpful in accurately understanding the impacts of trend, seasonality, and volatility effects through test results
MK	Merit	(i) Presence of the slightest trend component can be effectively detected
	Demerit	(i) Test is not reliable, particularly for higher time series length
	Suggestion	(i) Test cannot be solely used for assessing stationarity
Levene's	Merit	(i) Completely reliable
	Demerits	(i) Test is parametric (ii) Cannot detect the presence of trend component if the variance is constant throughout
	Suggestion	(i) Test is very effective for assessing variance and is recommended for use with some other tests for trend assessment
KW	Merits	(i) Test is nonparametric (ii)Test is fairly reliable for higher time series lengths
	Demerits	(i) Cannot detect small differences in mean values (ii) Low reliability for lower time series lengths
	Suggestion	(i) Recommended for use with higher time series lengths
SW	Merits	(i) Test is reliable (ii) Best performing test in all normality tests
	Demerits	(i) Nonstationary outcome does not mean that the time series is not stationary (ii) Low reliability with respect to skewness (iii) Significant type-I error for 0 kurtosis
	Suggestion	(i) This test is to be used first. If the test outcome is nonstationary, then other tests are to be used
KS	Merit	(i) Test is nonparametric (two-way) and low type-I errors
	Demerit	(i) Very low reliability with respect to skewness and kurtosis
	Suggestion	(i) Two-way KS test is useful for stationarity assessment, but the use of other tests for confirmation of results is recommended

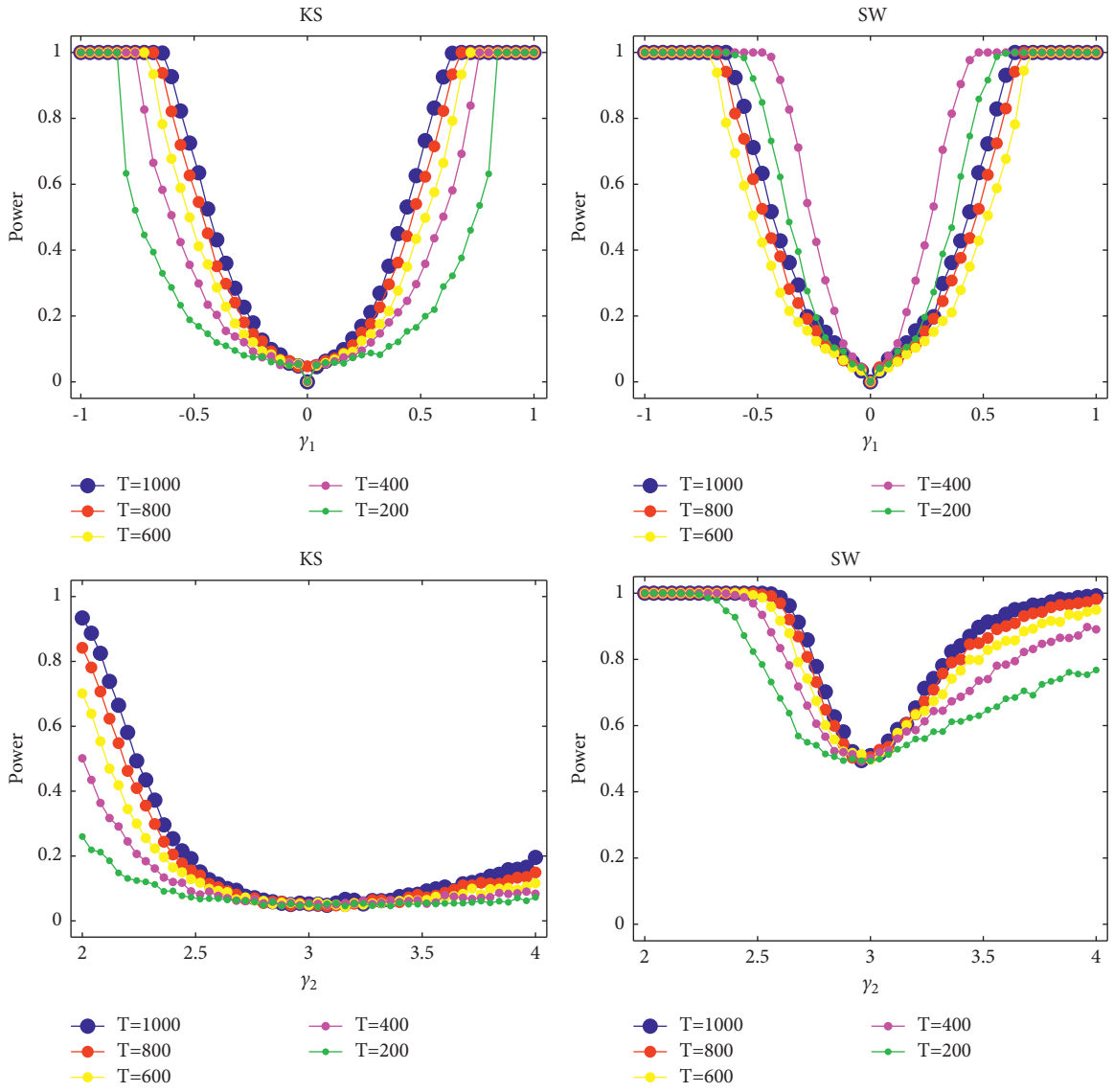

Figure 6: Comparison of power vs. γ_1 and power vs. γ_2 for different values of "T" using KS and SW tests.

limitations present in the working of these tests make it evident that a single test is not entirely sufficient to prove the stationarity of a test accurately. Some tests can be weak only in areas like high or low time series lengths or group sizes. It is also possible that other stationarity tests in the study prove advantageous in that particular area. Therefore, a thorough comparison of the performance of stationarity tests for their reliability and working would help novice readers select a group of tests for their respective applications. Using the detailed analysis and comparison of tests' reliability as carried out in Section 4.1, the merits, demerits, and suggestions for yielding the best outcome for all the tests are portrayed in Table 1. The best functioning tests for time series stationarity are identified and summarized in Section 5 based on the above-tabulated data.

5. Conclusion

The objective of this paper was to effectively compare and critically study nine well-established time series stationarity tests taking reliability into account and assisting the reader in selecting tests for a given application. The tests' reliability was characterized using a metric known as power, and the inferences from the reliability plots were examined. Furthermore, the merits and demerits of the tests were compared. And suggestions for the tests' direct application with apt setting(s) and information on other pertinent aspects are expected to help novice readers build accurate forecasting models.

Based on the obtained results, using PP, Breitung, and Levene's tests combined is recommended as the combo is

highly reliable in handling inherent renewable generation time series components, such as trend, seasonality, and volatility. The Breitung test suitably solves the reliability problem of the PP test for lower time series lengths, while the latter solves the same problem for higher time series lengths with the former. But, the above two unit root tests suffer from the incapability of detecting seasonality and volatility effects in a time series. Therefore, their suitable hybridization with a nonunit root test, such as Levene's test, can solve the above limitations. Levene's test's inability to detect the trend component can be further resolved by the above two unit root tests.

Data Availability

The figures and tables used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] 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, Article ID e2819, 2019.

[2] 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 IEEE Electrical Power and Energy Conference (EPEC)*, pp. 1–6, Toronto, ON, Canada, October 2018.

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

[4] S. Yue, P. Pilon, and G. Cavadias, "Power of the mann–kendall and spearman's rho tests for detecting monotonic trends in hydrological series," *Journal of Hydrology*, vol. 259, no. 1-4, pp. 254–271, 2002.

[5] D. Fedorová and M. Arltová, "Selection of unit root test on the basis of length of the time series and value of ar (1) parameter," *Statistika*, vol. 96, no. 3, pp. 47–64, 2016.

[6] A. Kulkarni and H. von Storch, "Simulationsexperimente zur Wirkung serieller Korrelation auf den Mann-Kendall Trend test," *Meteorologische Zeitschrift*, vol. 4, no. 2, pp. 82–85, 1992.

[7] T. Vorapongsathorn, S. Taejaroenkul, and C. Viwatwongkasem, "A comparison of type i error and power of bartlett's test levene's test and cochran's test under violation of assumptions," *Songklanakarin Journal of Science and Technology*, vol. 26, no. 4, pp. 537–547, 2004.

[8] D. Sharma and B. M. G. Kibria, "On some test statistics for testing homogeneity of variances: a comparative study," *Journal of Statistical Computation and Simulation*, vol. 83, no. 10, pp. 1944–1963, 2013.

[9] J. Gleason, *Comparative Power of the Anova Randomization Anova and Kruskal-wallis Test*, Wayne State University Dissertations, 2013.

[10] T. V. Hecke, "Power study of anova versus kruskal-wallis test," *Journal of Statistics & Management Systems*, vol. 15, no. 2-3, pp. 241–247, 2012.

[11] M. Mendes and A. Pala, "Type i error rate and power of three normality tests," *Information Technology Journal*, vol. 2, no. 2, pp. 135–139, 2003.

[12] B. M. Boyerinas, *Determining the Statistical Power of the Kolmogorov-Smirnov and anderson-darling Goodness-Of-Fit Tests via Monte Carlo Simulation*, Center of Naval Analyses Arlington United States Tech Rep, Arlington County, Virginia, USA, 2016.

[13] 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.

[14] S. binti Yusoff and Y. B. Wah, "Comparison of conventional measures of skewness and kurtosis for small sample size," in *Proceedings of the 2012 International Conference on Statistics in Science Business and Engineering (ICSSBE)*, pp. 1–6, IEEE, Langkawi, Malaysia, September 2012.