A Monte Carlo study on the size and power of panel unit root tests: Limitations in small data sets

Ivana Mravak[1,*],

[1] *Faculty of Economics, Business and Tourism, University of Split, Cvite Fiskovića 5, 21 000 Split, Croatia*
E-mail: ⟨*jerkovic@efst.hr*⟩

Abstract. The aim of this paper is to explore the properties of various panel unit root tests in terms of their power and size regarding different panel data structures, with a special focus on small data samples. In particular, different values of the cross-sectional and time dimensions for both heterogeneous and homogeneous data settings with intercept and with and without time trend are observed. We examine and compare the results of five commonly used first-generation panel unit root tests: Levin, Lin and Chu test, Im Pesaran and Shin test, Harris–Tzavalis test, Breitung test and a Fischer type test. The results are derived using Monte Carlo simulations and show that all the observed panel unit root tests suffer from a serious lack of power and tend to either over or under reject the null hypothesis when the time dimensions are small. It is evident that the results of conducting panel unit root tests for data with $T < 30$ are erroneous and unreliable, and it is therefore concluded that panel unit root tests should not be conducted for such samples in the first place. This is even more pronounced when there is a time trend or heterogeneity.

Keywords: first-generation panel unit root tests, homogeneous and heterogeneous panels, Monte Carlo simulations, power and size properties, small T panels

Received: September 14, 2023; accepted: November 24, 2023; available online: December 19, 2023

DOI: 10.17535/crorr.2023.0011

1. Introduction

Nowadays, the application of panel unit root tests has become a must-do as an initial step in performing panel data analysis and the wide availability of various user-friendly software solutions for performing these tests has facilitated their implementation, leading to their increased adoption and popularity. However, it is necessary to investigate the performance of different unit root tests on different data structures to ensure that they are used appropriately and that incorrect conclusions are not drawn. When assessing a panel unit root test, two key factors should be considered: the ability of a test to correctly detect stationary panels (power) and its control over type I errors (size). Consideration of these factors is essential when selecting the most appropriate panel unit root test for specific data structures, as they offer valuable insights into the performance of the test in those particular settings.

In this paper, the power and size properties of five commonly used panel unit root tests are discussed and compared based on the results obtained using Monte Carlo simulations. By systematically comparing the results of Monte Carlo simulations under different data structures, the strengths and limitations of each panel unit root test will be highlighted in this paper. The paper considers the following panel unit root tests: Levin, Lin and Chu test [16], Im Pesaran

[*] Corresponding author.

and Shin test [13], Harris–Tzavalis test [11], Breitung test [4] and a Fischer-type test [5]. These tests are first-generation panel unit root tests that assume cross-sectional independence and can all be used in the widely popular statistical software STATA.

Most of the existing literature dealing with the properties of panel unit root tests focuses on the comparison of a few most commonly used panel unit root tests, mainly the Im, Pesaran and Shin test and Levin, Lin and Chu test [9, 13, 15, 26]. In addition, asymptotical power and size are studied in most cases, especially in the papers introducing individual unit root tests, and the discussion of finite sample properties tends to focus on datasets with larger cross-sectional (N) and time (T) dimensions [20, 6]. In practice, however, many investigated data sets involve smaller N or T dimensions. Some examples are datasets related to EU, G-20, or G-8 countries, on yearly, quarterly or monthly frequencies for shorter time periods [8, 7]. In addition, smaller dimensions may result from the subdivision of datasets into subsets or subperiods [25]. As large values of N have been extensively researched in the existing literature, the focus of this paper is on values of $N \leq 30$.

Karlsson [15] discusses the power and size properties of panel unit root tests and concludes that researchers should be cautious when performing these tests on datasets with smaller T dimensions, as this can lead to false conclusions due to the low power of declaring a panel as non-stationary when in fact it is not. His conclusions are based on Monte Carlo simulations for datasets with $N, T \in \{5, 10, 25, 50, 100, 200\}$. However, he only investigates Levin, Lin and Chu and Im, Pesaran and Shin tests. Lopez [18] also notes that existing panel unit root tests are limited in their ability to accurately reject the unit root hypothesis when applied to data sets with a limited time span, but he discusses tests that account for cross-sectional dependence, i.e., the second-generation panel unit root tests. Moreover, Oh [23] shows how the low power of unit root tests can lead to erroneous conclusions. Moon et al. [21] investigate the power of the Breitung test and conclude that the power of the test is often lower than the predictions of the asymptotic theory, with the difference gradually decreasing as either N or T increases. These conclusions are derived from Monte Carlo simulations, but only for values of $T \in \{50, 100, 250\}$. Hlouskova and Wagner [12] explore a larger number of panel unit root tests and various combinations of N and T dimensions, but they only discuss the homogeneous case. Geppert et al. [9], on the other hand, thoroughly investigate the effects of panel heterogeneity on the power of a unit root test, but only for the Levin and Lin test and only for T=100. Maddala and Wu [19] have conducted the most extensive research to date on power and size properties of panel unit root tests. They compare Im, Pesaran and Shin, Fischer, and Levin, Lin and Chu tests on $N, T \in \{25, 50, 100\}$ for the homogeneous case and additionally discuss the effects of selecting incorrect lags. They also investigate the heterogeneous case, but only for a single combination of $N = 25$ and $T = 50$. Furthermore, some other studies discuss more recently developed second-generation unit root tests, which are not considered in this paper. [10, 17, 14].

From the mentioned literature, it is evident that the power of panel unit root tests generally expectedly increases as the values of N and T increase, but there is no consensus on a threshold that would indicate a power level that would be considered as satisfactory. Similarly, the question arises as to when the power of the test becomes too low to justify its usage and reliance on the results. A rule of thumb often used in the literature is the 80% threshold [1, 24], although the cut-off point may be individualised for a researcher based on the data used and the consequences of a type II error, which is more likely when the power is low.

The aim of this paper is to investigate whether there are certain data specifications, particularly in relation to the N and T dimensions, for which the values for power and size are so distorted, regardless of the panel unit root test chosen, that the validity of conducting panel unit root tests is called into question. Therefore, the contribution of this paper lies in its comprehensiveness and the wide range of used panel unit root tests and data specifications used, which are more extensive than in the previous literature. Specifically, the properties of five different panel unit root tests are analysed under different data settings. Various time series and cross-sectional dimensions are analysed, with a particular focus on smaller samples, which have tended to be rather neglected in the literature to date. The values discussed values are $N \in \{10, 15, 25, 30\}$ and $T \in \{10, 20, 30, 50, 100, 200\}$. Both cases containing only the intercept and cases containing both the intercept and the linear time trend are considered. Moreover, both homogeneous and heterogeneous settings of the data in relation to different cross-sections are taken into consideration. For each combination of N and T and for each underlying data setting, 2000 replications of the data generating process are performed, and then the presence of unit roots is tested using panel unit root tests available in the STATA software. The results show that all panel unit root tests perform very poorly for data sets with smaller N and T values, especially for values of $T < 30$, to such an extent that the usage of unit root tests becomes redundant and the results obtained are erroneous. This is even more pronounced when the time trend is included.

The rest of the paper is organized as follows: Section 2 gives a brief overview of some assumptions underlying the panel unit root tests used in this paper. Section 3 demonstrates the basic framework for the data generating processes and describes the simulations conducted. Section 4 presents and discusses the results obtained. Finally, Section 5 provides a concluding summary.

2. Brief review of some unit root tests assumptions

All unit root tests compared in this paper do not have the same assumptions and the same formulations of the hypotheses. First of all, there is no general agreement on what it even means for a panel to be stationary, i.e., whether it is sufficient for some cross-sections to be stationary for the whole panel to be declared stationary or whether all cross-sections must be stationary. Existing panel unit root tests primarily test whether the null hypothesis of a unit root is true for each cross-section within a panel. However, the formulation of the alternative hypothesis is controversial and depends on the assumptions about the homogeneity or heterogeneity of the panel [22]. For this reason, two different versions of the alternative hypothesis are used in the literature. When interpreting and comparing the results of various tests, the reader should bear in mind that some of them cannot be directly compared due to these different assumptions of the alternative hypothesis.

Suppose we observe a general autoregressive AR(1) model:

$$y_{it} = (1 - \phi_i)\mu_i + \phi_i y_{i,t-1} + \epsilon_{it} \tag{1}$$

$i = 1, \ldots N, t = 1, \ldots, T$, where the inital values y_{i0} are given. Our focus is on investigating the presence of unit roots i.e., when $\phi_i = 1$, $\forall i \in 1, \ldots, N$. By subtracting $y_{i,t-1}$ from both sides of the equation, (1) can be rewritten as:

$$\Delta y_{it} = \alpha_i + \beta_i y_{i,t-1} + \epsilon_{it} \tag{2}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\alpha_i = (1 - \phi_i)\mu_i$ and $\beta_i = \phi_i - 1$. Now, testing for the presence of unit roots is equal to testing for $\beta_i = 0$, $\forall i \in 1, \ldots, N$. Therefore, the panel unit root tests assess the following null hypothesis:

$$H_0 : \beta_i = 0, \forall i = 1, \ldots, N \tag{3}$$

against the alternative:

$$H_1 : \beta_i < 0, \forall i = 1, \ldots, N \tag{4}$$

or

$$\begin{aligned} H_1 : \beta_i &< 0, i = 1, \ldots, N_1 \\ \beta_i &= 0, i = N_1 + 1, \ldots, N \end{aligned} \tag{5}$$

The panel unit root tests used in this article can be divided into two groups based on their alternative hypothesis. The Levin, Lin and Chu (LLC), Harris–Tzavalis (HT) and Breitung tests assume homogeneous coefficients, i.e., $\beta_1 = \beta_2 = \cdots = \beta_N = \beta$ and therefore use the alternative hypothesis shown in (3), while Im Pesaran and Shin (IPS) and the Fischer type test (Fischer) assume heterogeneous coefficients across different cross-sections and present the alternative hypothesis as shown in (5).

Apart from the differences with respect to the alternative hypothesis, the presented tests have different assumptions with respect to some other data settings and show different performances for different data set properties. Some tests require the panels to be balanced (LLC, HT, Breitung) while others allow for unbalanced panels (IPS, Fisher). All panel unit root tests compared in this article are the so-called first-generation tests and assume cross-sectional independence. In addition, all panel unit root tests are designed with certain asymptotic assumptions. For example, the LLC test assumes $T \to \infty$, followed by $N \to \infty$, but with $N/T \to 0$, which means that T increases at a faster rate than N. Therefore, the LLC test is expected to perform better for $T > N$ [3]. Similar assumptions hold for IPS test. On the other hand, Harris and Tzavalis [11] show that such an assumption leads to poor properties in terms of size and power of the test for panels with small T, and therefore assume that T is fixed while $N \to \infty$ and, therefore, their test should outperform the LLC test for smaller T sizes. Breitung [4] assumes that N, T approach infinity sequentially, i.e. $(N, T)_{seq} \to \infty$, and Choi [5], used as a Fischer type test, allows N to be finite, while $T \to \infty$. Therefore, due to the differences in the designs and assumptions of the various panel unit root tests, they are expected to exhibit different size and power properties for different values of N and T.

3. Simulation study

To examine panel unit root test properties in various data settings, three distinct panel data models are simulated. These models differ in terms of coefficient characteristics with respect to homogeneity and heterogeneity across cross-sections and the inclusion of intercepts and linear time trends. Model 1 assumes that the coefficients are homogeneous between all individuals in the panel, and it incorporates an intercept term. Model 2 also assumes homogeneous coefficients, but includes a linear time trend in addition to the intercept. Model 3 introduces heterogeneity by assuming that the coefficients vary across cross-sections and it includes an intercept term.

The data generating process (DGP) used to simulate the data for each model is described by the following equations:

$$\text{Model 1:}\quad y_{it} = (1-\phi)\mu_i + \phi y_{i,t-1} + \epsilon_{it} \tag{6}$$

$$\text{Model 2:}\quad y_{it} = \mu_i + \phi y_{i,t-1} + (1-\phi)\mu_i t + \epsilon_{it} \tag{7}$$

$$\text{Model 3:}\quad y_{it} = (1-\phi_i)\mu_i + \phi_i y_{i,t-1} + \epsilon_{it} \tag{8}$$

where $i = 1, \ldots, N$, $t = -51, -50, \ldots, T$, $y_{i,-51} = 0$, $\mu_i \sim N(0,1)$, $\epsilon_{it} \sim N(0, \sigma_i^2)$ with $\sigma_i^2 \sim U[0.5, 1.5]$. Cross-sectional specifics μ_i and σ_i^2 are generated once for each cross-section and are then fixed in all subsequent iterations. A total of 52 additional observations were created for all replications and later omitted to ensure the stabilization of the data. In homogeneous models (Model 1 and Model 2) the values $N \in \{10, 15, 25, 30\}$ and $T \in \{10, 20, 30, 50, 100, 200\}$ are used, while for the heterogeneous case (Model 3) the values $N \in \{15, 30\}$ and $T \in \{10, 30, 100\}$ are considered. The smaller number of N and T combinations is used for the heterogeneous case to reduce the total number of replications, as five different levels of heterogeneity are considered for each N and T combination in the hetereogeneous case.

Note that in the case where the unit root is present, i.e. $\phi = 1$, the term $(1 - \phi)\mu_i$ ensures that there is no drift. The situation is simmilar for the version with a linear time trend and the presence of a unit root: the term $(1 - \phi)\mu_i t$ implies that the time trend will disappear while the drift remains. When simulating the data to calculate the power of the test for the homogeneous case (Model 1 and Model 2), $\phi = 0.9$ is set and to calculate the size, $\phi = 1$. For the heterogeneous case (Model 3), some portions of the simulated data are created with different ϕ coefficients, which now vary between different cross-sections. For each N and T combination, five different proportions of cross-sections that exhibit unit root are considered. Stationary portions of the data in this model are generated using $\phi_i \sim U[0.7, 0.9]$.

Balanced data are used In all our simulations and all tests are performed on the basis of the 5% nominal level. For each combination and each underlying data setting, 2000 replications of the data are simulated and for each replication a panel unit root test is conducted in the STATA computer software with the appropriate settings (regarding the intercept and time trend) using the *xtunitroot* command. For the Fischer type test, the version proposed by Choi [5] which performs the Augmented Dickey Fuller test on each cross-section's time series and then combines the obtained p-values is used. The power is calculated as the proportion of correctly rejected null hypotheses and the size is the proportion of incorrectly rejected nulls.

4. Results

This section presents the results of the described Monte Carlo simulations. First, the panel unit root tests results for the homogeneous case with intercept are discussed (Model 1 (6)). The results of power and size for different combinations of N and T are presented in Table 1.

		LLC		IPS		HT		Breitung		Fisher	
N	T	size	power	size	power	size	power	size	power	size	power
30	200	0.106	1	0.049	1	0.085	1	0.058	1	0.055	1
	100	0.164	1	0.05	1	0.092	1	0.049	1	0.055	1
	50	0.255	0.932	0.04	0.99	0.091	1	0.063	1	0.044	0.99
	30	0.39	0.761	0.051	0.67	0.1	0.998	0.05	0.999	0.058	0.69
	20	0.628	0.872	0.046	0.334	0.127	0.911	0.044	0.896	0.058	0.388
	10	0.946	0.975	0.075	0.161	0.195	0.617	0.032	0.295	0.106	0.211
25	200	0.098	1	0.04	1	0.083	1	0.06	1	0.044	1
	100	0.152	1	0.048	1	0.088	1	0.062	1	0.052	1
	50	0.243	0.884	0.054	0.978	0.099	1	0.049	1	0.059	0.977
	30	0.346	0.709	0.048	0.603	0.106	0.987	0.055	0.992	0.056	0.626
	20	0.584	0.809	0.046	0.3	0.119	0.87	0.049	0.828	0.061	0.346
	10	0.924	0.959	0.079	0.159	0.17	0.56	0.028	0.262	0.114	0.214
15	200	0.084	1	0.052	1	0.1	1	0.072	1	0.057	1
	100	0.108	0.983	0.042	1	0.103	1	0.078	1	0.05	1
	50	0.18	0.687	0.047	0.86	0.113	0.999	0.06	0.999	0.055	0.877
	30	0.23	0.486	0.048	0.407	0.115	0.928	0.051	0.918	0.059	0.443
	20	0.423	0.6	0.049	0.196	0.118	0.694	0.057	0.608	0.066	0.242
	10	0.79	0.872	0.08	0.131	0.16	0.444	0.033	0.188	0.115	0.192
10	200	0.081	1	0.049	1	0.114	1	0.067	1	0.054	1
	100	0.102	0.884	0.061	0.999	0.112	1	0.07	1	0.067	1
	50	0.133	0.474	0.046	0.684	0.119	0.996	0.062	0.991	0.06	0.714
	30	0.187	0.355	0.053	0.29	0.112	0.833	0.061	0.774	0.065	0.333
	20	0.341	0.469	0.044	0.148	0.136	0.6	0.056	0.473	0.065	0.192
	10	0.695	0.772	0.079	0.119	0.153	0.388	0.036	0.148	0.121	0.174

Table 1: *Model 1 results.*

The results reveal that the power of all observed tests significantly decreases when the values of N and T are small. In the LLC, IPS and Fischer tests, this decline is serious even with a moderate value of $T = 50$ when $N = 10$. At first glance, the LLC test seems to exhibit excellent power properties for small T values. Although, as expected, the power starts decreasing as T decreases, this decrease stops at about $T = 30$, and then the power starts to increase again, reaching a value of 0.975 for $N = 30, T = 10$ and 0.959 for $N = 25, T = 10$. However, a closer look at the size of this test shows that the LLC test drastically over-rejects the null as T becomes small, with sizes of 0.946 and 0.924 for $T = 10$ and $N = 30$ and 25, respectively. This means that LLC test rejects the null more than 92% of the time even when it should not be rejected. The 95% of correctly rejected nulls no longer seem so impressive. The high power values are therefore explained by the fact that the LLC test over rejects the null hypothesis. This shows that it is necessary to observe the powers and sizes of the tests simultaneously and in conjunction with each other. The poor properties of the LLC test for data sets with smaller T were somewhat expected, as the LLC test is designed to perform best when $N < T$, and the authors of the test themselves suggest using their test for panels with moderate N and T sizes, with N between 10 and 250 and T between 25 and 250 [16].

The power of the IPS test also becomes critical and falls below 70% for $T \leq 30$ regardless of the size of N, while the size shows less bias than the LLC test, but becomes increasingly extensive for $T = 10$. These results are consistent with the simulations performed by Im et al. [13], which show that both IPS and LLC tests exhibit distortions in size when N becomes significantly larger than T. The power values of the HT and Breitung tests remain above 80% for $N \geq 15, T \geq 30$. These two panel unit root tests have the best properties in terms of power and size, but again small values of T create distortions. The power of the HT test falls below the 80% threshold when $T \leq 20, N \leq 15$, although the size distortions start increasing as T decreases and begin to over-reject the null at around $T = 30$. Breitung test shows size distortions in the opposite direction as T decreases. It exhibits under rejection of the null hypothesis and declares the panel to be non-stationary, although in reality it is not. Severe fall in power occurs for smaller N values. With $N = 10$, $T = 30$, the power drops below 80%.

The outperformance of the HT test over the LLC and IPS tests was expected since the design of the tests. Harris and Tzavalis [11] found that the assumption that T increases faster than N towards infinity leads to poor power properties, especially for smaller values of T. To address this issue, in their design of the panel unit root test, they derived it under the assumption that T is fixed, which is a typical scenario in micro-panel studies [2]. Breitung [4] also addressed the same issue of LLC and IPS tests by suggesting a test statistic that avoids the use of a bias adjustment. This approach yielded significantly higher power compared to the LLC and IPS tests [2], as demonstrated by Monte Carlo experiments. As far as the Fischer type test is concerned, the power falls below 70% at $T \leq 30$, while for $N = 10$, even $T = 50$ exhibits poor power properties of 71.4%. Size values increase gradually as T deacreases.

Table 2 presents the results for data settings with homogeneous coefficients which now containing both the intercept and the linear time trend as shown in Model 2 (7). The pattern of results is simillar to Model 1, but with an even faster decline in power. Namely, now all tests exhibit power values less than 80% for $N = 15, 10$ for $T \leq 50$, while for moderate N values, $N = 25, 30$, power drops below 80% for $T \leq 30$ for all tests. The same conclusions as before are also drawn for the size properties, but with stronger distortions, so that the observed panel unit root tests now prove to be misleading even for higher T values. For the LLC test, the same remark regarding the high power values applies as for the intercept only version.

Although all tests show poor performances for $T \leq 30$, the HT and Breitung tests show the best properties for $T = 50$. Breitung [4] found that LLC and IPS tests lose a lot of power when individual-specific time trends are included, so he addressed this issue in the design of his test. Although the time trends in our simulations are not individual-specific but common, the Breitung test still shows better results than the LLC and IPS tests when the time trend is included.

		LLC		IPS		HT		Breitung		Fisher	
N	T	size	power	size	power	size	power	size	power	size	power
30	200	0.12	1	0.054	1	0.089	1	0.064	1	0.058	1
	100	0.258	1	0.046	1	0.079	1	0.056	1	0.05	1
	50	0.368	0.848	0.048	0.576	0.098	0.941	0.05	0.873	0.058	0.62
	30	0.476	0.657	0.052	0.169	0.106	0.536	0.05	0.171	0.071	0.209
	20	0.914	0.954	0.055	0.086	0.137	0.324	0.046	0.158	0.087	0.134
	10	0.999	1	0.114	0.126	0.247	0.317	0.035	0.058	0.196	0.225
25	200	0.117	1	0.043	1	0.067	1	0.055	1	0.049	1
	100	0.208	0.997	0.05	1	0.087	1	0.06	1	0.054	1
	50	0.335	0.769	0.052	0.5	0.103	0.893	0.054	0.825	0.062	0.548
	30	0.437	0.578	0.052	0.171	0.106	0.475	0.057	0.323	0.077	0.21
	20	0.884	0.921	0.053	0.081	0.136	0.305	0.049	0.137	0.08	0.123
	10	0.998	0.997	0.117	0.124	0.24	0.316	0.037	0.054	0.202	0.218
15	200	0.105	1	0.05	1	0.083	1	0.057	1	0.057	1
	100	0.193	0.956	0.054	0.984	0.093	1	0.06	0.998	0.06	0.987
	50	0.254	0.548	0.046	0.327	0.093	0.745	0.051	0.608	0.063	0.375
	30	0.283	0.436	0.047	0.142	0.12	0.373	0.056	0.243	0.067	0.1855
	20	0.71	0.765	0.054	0.087	0.132	0.248	0.053	0.1225	0.087	0.13
	10	0.986	0.984	0.094	0.113	0.196	0.248	0.036	0.055	0.036	0.21
10	200	0.086	1	0.046	1	0.093	1	0.072	1	0.051	1
	100	0.145	0.847	0.05	0.912	0.099	1	0.07	0.965	0.053	0.92
	50	0.196	0.429	0.056	0.252	0.105	0.625	0.068	0.487	0.067	0.306
	30	0.226	0.301	0.055	0.101	0.133	0.277	0.055	0.185	0.076	0.142
	20	0.555	0.619	0.056	0.064	0.117	0.194	0.055	0.098	0.079	0.107
	10	0.94	0.94	0.1	0.114	0.174	0.184	0.044	0.049	0.183	0.2

Table 2: *Model 2 results.*

Overall, for the homogeneous cases, both with and without time trend, HT and Breitung tests outperform the other panel unit root tests, both for small and large values of T. This is to be expected since HT and Breitung tests are designed for homogeneous coefficients, whereas IPS and Fischer tests allow for heterogeneity across cross-sections. The LLC test, despite its design for homogeneous cases, has the worst size and power characteristics compared to the other tests. Although this comparison of panel unit root tests applies to all observed combinations of N and T, it can be generally inferred from our simulations that for values of $T < 30$ all observed tests lack in power and/or size properties and mainly lead to incorrect conclusions.

The results for the heterogeneous settings of the underlying data, as described in Model 3 (8), are shown in Table 3. In the case of these heterogeneous scenarios, only the powers are calculated, as the concept of size is not meaningful. The reason for this is that in each panel certain portions of the data exhibit stationarity, so the tests used in this paper cannot falsely reject the null hypothesis. In line with our previous observations and the existing research, the inclusion of the time trend has been shown to further degrade the performance of the panel unit root tests. Therefore, for this particular case of heterogeneity, only the scenario with intercept is presented.

The results show that the primacy of the HT and Breitung tests is now lost, as these tests assume homogeneity, just like the LLC test. The IPS and Fisher tests also show better results, as expected, as they are designed for heterogeneous data settings. However, good power properties are obtained only for larger T values. When T becomes smaller, the unit root tests specifically designed for heterogeneous panels do not outperform other ones anymore. The threshold value of T at which the power falls below 80% depends on the number of cross-sections that contain unit roots. It is also important to note that in this Model, the ϕ_i values are no longer homogeneously equal to 0.9, as in previous models, but follow a uniform distribution $\phi_i \sim U[0.7, 0.9]$. These values of ϕ_i make it easier for the unit root test to detect stationarity as the distance from 1 is greater.

N	T	Number of cross-sections containing unit root	LLC	IPS	HT	BR	FSH
30	100	1	1	1	1	1	1
		5	0.997	1	1	0.999	1
		15	0.76	1	0.898	0.743	1
		25	0.249	0.416	0.246	0.136	0.447
		29	0.162	0.042	0.086	0.059	0.046
	30	1	0.976	0.996	1	1	0.996
		5	0.93	0.976	0.997	0.981	0.984
		15	0.649	0.513	0.773	0.55	0.552
		25	0.9635	0.162	0.555	0.206	0.207
		29	0.37	0.056	0.104	0.053	0.065
	10	1	0.986	0.235	0.889	0.623	0.311
		5	0.98	0.243	0.811	0.504	0.305
		15	0.988	0.277	0.894	0.619	0.354
		25	0.958	0.097	0.268	0.062	0.138
		29	0.943	0.067	0.183	0.02	0.1
15	100	1	0.99	1	0.996	0.98	1
		3	0.915	1	0.981	0.901	1
		8	0.45	0.975	0.692	0.438	0.983
		12	0.167	0.191	0.207	0.135	0.215
		14	0.121	0.044	0.084	0.064	0.046
	30	1	0.728	0.799	0.976	0.922	0.827
		3	0.583	0.582	0.848	0.752	0.624
		8	0.4	0.301	0.564	0.314	0.338
		12	0.297	0.099	0.217	0.11	0.119
		14	0.297	0.099	0.217	0.11	0.119
	10	1	0.892	0.175	0.594	0.285	0.237
		3	0.887	0.173	0.562	0.265	0.236
		8	0.844	0.114	0.376	0.126	0.172
		12	0.826	0.087	0.24	0.055	0.129
		14	0.806	0.081	0.167	0.04	0.13

Table 3: *Model 3 results.*

Additionally, when the proportion of cross sections containing a unit root is above 80%, the heterogeneity of the test assumptions does not yield improvements and all tests exhibit poor properties in detecting the existence of stationary units. Similarly to previous models, we observe remarkably high power values for the LLC test when T is small. As noted in Table 1, the LLC test tends to over-reject the null hypothesis in such data sets, which explains this outcome.

Overall, all the observed panel unit root tests demonstrate efficient performances in heterogeneous settings only for larger values of T and small proportions of cross sections containing a unit root. Considering the case where about 50% of the cross-sections are exhibiting a unit root, the power values of all the tests remain below 80% for $T \leq 30$. For large values of T, the IPS and Fischer tests show the best power properties.

5. Conclusion

In this paper, the power and size properties of five different panel unit root tests are compared and discussed. Three distinct underlying data generating processes were used for modelling: homogeneous panels containing an intercept; homogeneous panels containing both an intercept and a linear time trend; and heterogeneous panels with an intercept. Various combinations of N and T values were considered but the focus was on smaller data samples, namely for $N \leq 30$.

Monte Carlo simulations conducted reveal drastic power and size distortions for data sets with small T values. Depending on the underlying data structure and the combination of N and T, different tests exhibit different properties. However, a generally derived observation is that the vast majority of panel unit root tests for data sets with T size smaller than 30 exhibit such low power and/or size properties that their performance is essentially meaningless. The aforementioned threshold of $T = 30$ is derived as the point at which the distortions become apparent and the power mainly falls below 80%, while the size values show an under- or over-rejection of the null hypothesis, depending on the test. This is particularly pronounced when the time trend is included. When the T value drops further, below $T = 30$, the distortions mentioned are particularly severe and there is no doubt that the results obtained cannot be trusted as they are erroneous. The best performance for both smaller and larger T values in homogeneous data settings is shown by HT and Breitung tests, both of which are homogeneous in construct. Thus, when heterogeneous data are observed, their advantage disappears and they are outperformed by IPS and Fischer tests. In the heterogeneous case, the power properties deteriorate with both the decrease in panel dimensions and the number of cross-sections containing unit roots. If the proportion of non-stationary units is more than 50%, panel unit root tests for $T \leq 30$ become inefficient again.

The simulations carried out in this paper show that panel unit root tests for $T < 30$ generally yield incorrect results. There is no general rule in the literature for the threshold of acceptable values for the power of a unit root test, but a common rule of thumb is 80% [1, 24] Researchers can consider the impact of errors related to the presence of unit roots for their individual research purpose and set their own threshold accordingly, higher or lower. However, it is advisable not to set the threshold much closer to 50%, as success at this level would depend more on luck than on a meaningful result.

In this paper, only first-generation panel unit root tests are discussed. Consequently, the cross-sectional dependence of the data was not considered. This remains a possible topic for future research to conduct similar small data sets analysis but for second-generation panel unit root tests that do account for cross-sectional dependence.

Data availability

The Python code used to simulate the data in this paper is available at: https://figshare.com/s/1b591c879f4dc5ef4acb

References

[1] Arltová, M. and Fedorová, D. (2016). Selection of Unit Root Test on the Basis of Length of the Time Series and Value of AR(1) Parameter. Statistika - Statistics and Economics Journal. 96. 47-64. Retrieved from: https://www.czso.cz/csu/czso/2-statistika

[2] Baltagi, B. H. (2021). Econometric Analysis of Panel Data. Springer. Retrieved from: link.springer.com

[3] Barbieri, L. (2005). Panel Unit Root Tests: A Review. Department of Business and Social Sciences. Faculty of Economics and Law. Retrieved from: publires.unicatt.it

[4] Breitung, J. (2001), The local power of some unit root tests for panel data. In Baltagi, B. H., Fomby, T. B. and Carter Hill, R. (Eds.) Nonstationary Panels, Panel Cointegration, and Dynamic Panels (pp. 161-177). Emerald Group Publishing Limited. doi: 10.1016/S0731-9053(00)15006-6

[5] Choi, I. (2001). Unit root tests for panel data. Journal of International Money and Finance, 20(2), 249–272. doi: 10.1016/S0261-5606(00)00048-6

[6] Choi, I. (2019). Unit Root Tests for Dependent Micropanels. The Japanese Economic Review 70, 145–167. doi: 10.1111/jere.12170

[7] Cota, B., Erjavec, N. and Jakšić, S. (2023). Economic complexity and income inequality in EU countries. Croatian Operational Research Review, 14(1), 77-86. doi: 10.17535/crorr.2023.0007

[8] Ganic, M. Hodzic, L. and Ridjic O. (2021). A Test of the validity of Crowding-out (or- in) hypothesis: A new examination of link between public borrowing and private investment in Emerging Europe. Croatian Operational Research Review, 12(1), 91-103. doi: 10.17535/crorr.2021.0008

[9] Geppert, J. M., Jares, T. E. and Lavin, A. M. (2002). The Effect of Time-Series and Cross-Sectional Heterogeneity on Panel Unit Root Test Power. Journal of Financial Research, 25(3), 321–335. doi: 10.1111/1475-6803.00021

[10] Gutierrez, L. (2006). Panel Unit-root Tests for Cross-sectionally Correlated Panels: A Monte Carlo Comparison. Oxford Bulletin of Economics and Statistics, 68(4), 519–540. doi: 10.1111/j.1468-0084.2006.00176.x

[11] Harris, R. D. F. and Tzavalis, E. (1999). Inference for unit roots in dynamic panels where the time dimension is fixed. Journal of Econometrics, 91(2), 201–226. doi: 10.1016/S0304-4076(98)00076-1

[12] Hlouskova, J. and Wagner, M. (2006). The Performance of Panel Unit Root and Stationarity Tests: Results from a Large Scale Simulation Study. Econometric Reviews, 25(1), 85–116. doi: 10.1080/07474930500545504

[13] Im, K. S., Pesaran, M. H. and Shin, Y. (2003). Testing for unit roots in heterogeneous panels. Journal of Econometrics, 115(1), 53–74. doi: 10.1016/S0304-4076(03)00092-7

[14] Kappler, M. (2006). Panel Tests for Unit Roots in Hours Worked. ZEW - Centre for European Economic Research. Discussion Paper No. 06-022. doi: 10.2139/ssrn.896066

[15] Karlsson, S. and Löthgren, M. (2000). On the Power and Interpretation of Panel Unit Root Tests. Economics Letters. 66. 249-255. doi: 10.1080/13504851.2012.695067

[16] Levin, A., Lin, C.-F. and James Chu, C.-S. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. Journal of Econometrics, 108(1), 1–24. 10.1016/S0304-4076(01)00098-7

[17] Lin, J., Hu, Y., Wang, M. and Xia, X. (2013). A Monte Carlo comparison of panel unit root tests under factor structure. Applied Economics Letters. 20(3), 288-291. doi: 10.1080/13504851.2012.695067

[18] Lopez, C. (2009). A Panel Unit Root Test with Good Power in Small Samples. Econometric Reviews. 28(4), 295-313. doi: 10.1080/07474930802458620

[19] Maddala, G. S., and Wu, S. (1999). A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. Oxford Bulletin of Economics and Statistics, 61(S1), 631–652. doi: 10.1111/1468-0084.0610s1631

[20] Madsen, E. (2010). Unit root inference in panel data models where the time-series dimension is fixed: a comparison of different tests. The Econometrics Journal. 13(1), 63–94. doi: 10.1111/j.1368-423X.2009.00302.x

[21] Moon, H. R., Perron, B. and Phillips, P. C. B., (2006). On The Breitung Test For Panel Unit Roots And Local Asymptotic Power. Econometric Theory. 22(6), 1179-1190. doi: 10.1017/S0266466606060555

[22] Pesaran, M. H. (2012). On the interpretation of panel unit root tests. Economics Letters, 116(3), 545–546. doi: 10.1016/j.econlet.2012.04.049

[23] Oh, K.-Y. (1996) Purchasing power parity and unit root tests using panel data. Journal of International Money and Finance. 15(3), 405-418. doi: 10.1016/0261-5606(96)00012-5

[24] Serdar, C. C., Cihan M., Yücel D. and Serdar M. A. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia Med-

ica. 13(1), 27-53. doi: 10.11613/BM.2021.010502

[25] Škrabić Perić, B., Sorić, P., Jerković, I. (2023). Behavioural antecedents of Bitcoin trading volume: A panel Granger causality test. Croatian Operational Research Review, 14(1), 87-97. doi: 10.17535/crorr.2023.0008

[26] Westerlund J. and Breitung S. (2002). Lessons from a Decade of IPS and LLC. Econometric Reviews. 32(5-6), 547-591. doi: 10.1080/07474938.2013.741023