

**The Performance of Panel Unit Root and
Stationarity Tests:
Results from a Large Scale Simulation Study**

Jaroslava Hlouskova
Martin Wagner

05-03

March 2005

DISCUSSION PAPERS

The Performance of Panel Unit Root and Stationarity Tests: Results from a Large Scale Simulation Study*

Jaroslava Hlouskova
Institute for Advanced Studies
Department of Economics and Finance

Martin Wagner [†]
University of Bern
Department of Economics

March 15, 2005

Abstract

This paper presents results concerning the size and power of first generation panel unit root and stationarity tests obtained from a large scale simulation study, with in total about 290 million test statistics computed. The tests developed in the following papers are included: Levin, Lin and Chu (2002), Harris and Tzavalis (1999), Breitung (2000), Im, Pesaran and Shin (1997 and 2003), Maddala and Wu (1999), Hadri (2000), and Hadri and Larsson (2002). Our simulation set-up is designed to address i.a. the following issues. First, we assess the performance as a function of the time and the cross-section dimension. Second, we analyze the impact of *positive* MA roots on the test performance. Third, we investigate the power of the panel unit root tests (and the size of the stationarity tests) for a variety of first order autoregressive coefficients. Fourth, we consider both of the two usual specifications of deterministic variables in the unit root literature.

JEL Classification: C12, C15, C23

Keywords: Panel Unit Root Test, Panel Stationarity Test, Size, Power, Simulation Study

1 Introduction

Panel unit root and stationarity tests have become extremely popular and widely used over the last decade. Given that several such tests are now implemented in commercial software, their usage will most likely increase further. Thus, it is important to collect evidence on the size and power of these tests with large-scale simulation studies in order to provide practitioners with some guidelines for deciding which test to use (for a specific problem or sample size at hand).

*Financial support from the Jubiläumsfonds of the Österreichische Nationalbank under grant Nr. 9557 is gratefully acknowledged.

[†]Part of this work has been done whilst visiting Princeton University and the European University Institute. The hospitality of these institutions is gratefully acknowledged.

All tests included in this study are so called *first generation tests* that are designed for cross-sectionally independent panels. This admittedly very strong assumption simplifies the derivation of the asymptotic distributions of panel unit root and stationarity tests considerably. We include the panel unit root tests developed in the following papers: Levin, Lin and Chu (2002), Harris and Tzavalis (1999), Breitung (2000), Im, Pesaran and Shin (1997 and 2003), and Maddala and Wu (1999). We also include two panel stationarity tests, developed in Hadri (2000), and Hadri and Larsson (2002).

Note that in recent years several tests that avoid the assumption of cross-sectional independence have been developed, see e.g. Bai and Ng (2004), Chang (2002), Choi (2002), Moon and Perron (2004) or Pesaran (2003). These are, however, as of now not widely used and are also not yet available in commercial software. For these reasons a simulation performance analysis of such tests is not contained in this paper.

In our simulation study we are primarily interested in the following aspects.¹ First, we investigate the performance of the tests depending upon the time series and cross-sectional dimension. Since in the derivation of the asymptotic test statistics, different rates of divergence for the time series and the cross-sectional dimension are assumed for different tests (see Table 1), it is interesting to analyze the performance of the tests when varying the time and cross-sectional dimensions of the panel. We take for both the time dimension T and the cross-sectional dimension N all values in the set $\{10, 15, 20, 25, 50, 100, 200\}$. Thus, we investigate in total forty-nine different panel sizes. Second, we assess the performance of the tests for moving average roots tending to 1. It is well known from the time series unit root literature (e.g. Agiakloglou and Newbold, 1996) that unit root tests suffer from severe size distortions for large positive moving average roots. This is clear, since in the case of a moving average root at 1, the unit root is cancelled and the resultant process is stationary (see also the discussion in Section 3). In our study we consider moving average roots in the set $\{0.2, 0.4, 0.8, 0.9, 0.95, 0.99\}$ and also include the case of no moving average root. This latter case corresponds in our simulation design to serially uncorrelated errors, which is also the special case for which some of the tests listed above are developed (e.g. the test of Harris and Tzavalis, 1999, see the description in Section 2). Third, we study the performance as a

¹Our simulation study is based on ARMA(1,1) processes, respectively AR(1) processes if the MA coefficient is equal to 0, given by (ignoring deterministic components here for brevity): $y_{it} = \rho y_{it-1} + u_{it}$ with $u_{it} = \varepsilon_{it} + c\varepsilon_{it-1}$ and $\varepsilon_{it} \sim N(0, 1)$ and cross-sectionally independent. The parameter c is equal to *minus* the moving average root.

function of the first order autoregressive coefficient ρ . For the power analysis of the panel unit root tests we take ρ in the set $\{0.7, 0.8, 0.9, 0.95, 0.99\}$, and for the size analysis of the stationarity tests $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Fourth, we investigate the performance of the tests for the two most common, and arguably for economic time series most relevant, specifications of deterministic variables. These are intercepts in the data generating process (DGP) when stationary but no drifts when integrated (referred to as case 2), and intercepts and linear trends under stationarity and drifts when integrated (referred to as case 3).²

The number of simulations (i.e. the number of test statistics computed) is given by 290,080,000. This huge number is the product of all parameter choices (for ρ and c) and sample sizes with the number of replications for each choice, given by 10,000. The total set of results, comprising about 170 pages of tables and about 30 pages with multiple figures, is available from the authors upon request.

In Section 3 of the paper we discuss the main observations and display some representative results graphically. A brief outlook on some of the main findings is: The relative size of the panel (i.e. the size of T relative to N) has important influence on the performance of the tests. Especially for $T \leq 50$ the performance of all tests is strongly influenced by the cross-sectional dimension N . For increasingly negative MA coefficients, as expected, size distortions become more prominent and especially for large negative values of c the size diverges to 1 (even for $T, N \rightarrow 200$). The general impression concerning the size behavior is that the Levin, Lin and Chu (2002) and Breitung (2000) tests have their size closest to the nominal size. There are, however, exceptions (see the discussion in Section 3). Concerning power we observe that for case 2 either the Levin, Lin and Chu (2002) test or the Breitung (2000) test have the highest power, whereas in case 3 there exist parameter constellations and sample sizes such that each of the considered tests has highest power.

The stationarity tests show very poor performance. The tests essentially reject the null hypothesis of stationarity for all processes that are not ‘close to white noise’, for all but the smallest values of T . This finding is not inconsistent with the fact that empirical studies usually reject the null hypothesis of stationarity when using the tests of Hadri (2000) or Hadri and Larsson (2002).

The paper is organized as follows: Section 2 describes the implemented panel unit root

²A further issue that is studied is the effect of the choice of the autoregressive lag lengths, as explained in Section 2, on the performance of the tests.

and stationarity tests. Section 3 presents the simulation set-up and discusses the simulation results and Section 4 draws some conclusions. An appendix containing additional figures follows the main text.

2 The Panel Unit Root and Stationarity Tests

In this section we describe the implemented panel unit root and stationarity tests. We include a relatively detailed description here for two reasons. First, the detailed description allows the reader to see the differences and similarities across tests clearly at one place. Second, our description is intended to be detailed enough to allow the reader to implement the tests herself.

The data generating process (DGP) for which the considered tests are designed is in its most general form given by

$$y_{it} = \alpha_i + \beta_i t + \rho_i y_{it-1} + u_{it}, \quad i = 1, \dots, N \text{ and } t = 1, \dots, T \quad (1)$$

where $\alpha_i, \beta_i \in \mathbb{R}$ and $-1 < \rho_i \leq 1$.³ The noise processes u_{it} are stationary ARMA processes, i.e. the stationary solutions to $a_i(z)u_{it} = b_i(z)\varepsilon_{it}$, $a_i(z) = 1 + a_{i,1}z + \dots + a_{i,p_i}z^{p_i}$, $a_{i,p_i} \neq 0$, $b_i(z) = 1 + b_{i,1}z + \dots + b_{i,q_i}z^{q_i}$, $b_{i,q_i} \neq 0$, $a_i(z) \neq 0$ for all $|z| \leq 1$, $b_i(z) \neq 0$ for all $|z| \leq 1$ and with $a_i(z)$ and $b_i(z)$ relative prime. The innovation sequences ε_{it} are i.i.d. with variances σ_i^2 and finite fourth moments and are assumed to be cross-sectionally independent.

The above assumptions on the noise processes are stronger than required for the applicability of functional limit theorems. In particular the assumptions guarantee a finite *long-run variance* of the processes u_{it} , i.e. a bounded spectrum of u_{it} at frequency 0. The long-run variance of u_{it} is given by 2π times the spectrum of u_{it} at frequency 0. For stationary ARMA processes the long-run variance, $\sigma_{ui,LR}^2$ say, is immediately found to be $\sigma_i^2 b_i^2(1)/a_i^2(1)$.⁴

Some of the tests discussed below are designed for more restricted DGPs than the general DGP given in (1). In particular some tests are restricted to serially uncorrelated noise processes u_{it} .

³In all our simulations we restrict attention to balanced panels, i.e. panels where the number of observations is identical for all cross-sectional units. This is of course not required for all tests investigated. Some cross-sectional dependence can be handled with the tests discussed by including (random) time effects, θ_t say. We do not discuss this issue here.

⁴Solving the ARMA equation for the Wold representation $u_{it} = c_i(z)\varepsilon_t = \sum_{j=0}^{\infty} c_{ij}\varepsilon_{t-j}$, the (short-run) variance of u_{it} is given by $\sigma_{ui}^2 = \sigma_i^2 \sum_{j=0}^{\infty} c_{ij}^2$ and the long-run variance is given by $\sigma_{ui,LR}^2 = \sigma_i^2 (\sum_{j=0}^{\infty} c_{ij})^2$.

As in the time series unit root literature, three specifications for the deterministic components are considered in the panel unit root literature. These are DGPs with no deterministic component ($d_{1t} = \{\emptyset\}$), DGPs with intercept only ($d_{2t} = \{1\}$) and DGPs containing both intercept and linear trend ($d_{3t} = \{1, t\}$). Exactly as in the time series literature, three cases concerning the deterministic variables in the presence of a unit root and under stationarity are considered most relevant. Case 1 contains no deterministic components in both the stationary and the nonstationary case, case 2 allows for intercepts in the DGP when stationary but excludes a drift when integrated, and case 3 allows for intercepts and linear trends under stationarity and for a drift when a unit root is present.

2.1 Panel Unit Root Tests

Levin, Lin (and Chu): We start the description of the unit root tests with the Levin and Lin (1993) tests, abbreviated by *LL93* henceforth. Their results have only been recently published in Levin, Lin and Chu (2002).⁵ The null hypothesis of the *LL93* test is $H_0 : \rho_i = 1$ for $i = 1, \dots, N$, against the *homogenous* alternative $H_1^1 : -1 < \rho_i = \rho < 1$ for $i = 1, \dots, N$. Thus, under the homogenous alternative the first order serial correlation coefficient ρ is required to be identical in all units. This restriction stems from the fact that the test is pooled.

The approach is most easily described as a three-step procedure, with preliminary regressions and normalizations necessitated by cross-sectional heterogeneity. In the first step for each individual series an ADF type regression of the form⁶

$$\Delta y_{it} = (\rho_i - 1)y_{it-1} + \sum_{j=1}^{p_i} \gamma_{ij} \Delta y_{it-j} + \delta_{mi} d_{mt} + v_{it}, \quad m = 1, 2, 3 \quad (2)$$

is performed and v_{it} denotes the residual process of the AR equation. If the processes are AR processes and the AR orders p_i are specified correctly, then $v_{it} = u_{it}$ holds. Here and throughout the paper m indexes the case considered. The lag lengths in the autoregressive test equations have to be increased appropriately as a function of the time dimension of the panel to ensure consistency, if the processes u_{it} are indeed ARMA processes. More specifically

⁵Important foundations have already been laid in Levin and Lin (1992), where panel unit root tests have been developed for homogenous panels. These are panels where loosely speaking the ARMA coefficients are identical for all u_{it} .

⁶Actually, it is recommended by the authors, that in a first step the cross-section average $\bar{y}_t = \frac{1}{N} \sum_{i=1}^N y_{it}$ is removed from the observations. This stems from the fact that the presence of time specific aggregate effects, θ_t , does not change the asymptotic properties, when the tests are performed on the transformed variables $y_{it} - \bar{y}_t$. Thus, as indicated already, a limited amount of dependence across the errors is allowed for, in a form that can easily be removed. For the panels we simulate this step is not required.

$p_i(T) \sim T^\kappa$, with $0 < \kappa \leq 1/4$ has to be assumed in the ARMA case. In practical applications some significance testing on the estimated $\hat{\gamma}_{ij}$, an information criterion or checking for no serial correlation in the estimated residuals \hat{v}_{it} is used to determine the lag lengths p_i . Then, for given p_i , orthogonalized residuals are obtained from two auxiliary regressions. \tilde{e}_{it} say, from a regression of Δy_{it} on the lagged differences Δy_{it-j} , $j = 1, \dots, p_i$ and d_{mt} , and \tilde{f}_{it-1} say, from a regression of y_{it-1} on the same set of regressors. These residuals are then standardized by the regression standard error from regressing \tilde{e}_{it} on \tilde{f}_{it-1} , $\hat{\sigma}_{vi}$ say, to obtain the standardized residuals $\hat{e}_{it}, \hat{f}_{it-1}$.

The second step is to obtain an estimate for the ratio of the long-run variance to the short-run variance of Δy_{it} , or equivalently of u_{it} . The definition of the long-run variance, $\sigma_{ui,LR}^2 = \sigma_{ui}^2 + 2 \sum_{j=1}^{\infty} \mathbb{E}(u_{it}u_{i,t-j})$ immediately leads to an estimator of the form

$$\hat{\sigma}_{ui,LR}^2 = \frac{1}{T} \sum_{t=1}^T \hat{u}_{it}^2 + \frac{2}{T} \sum_{j=1}^L w(j, L) \sum_{t=j+1}^T \hat{u}_{it} \hat{u}_{i,t-j+1} \quad (3)$$

where the lag truncation parameter L can be chosen e.g. according to Andrews (1991) or Newey and West (1994). In the above equation we choose as estimate for the unobserved noise $\hat{u}_{it} = \Delta y_{it} - \hat{\delta}_{mi} d_{mt}$.⁷ In our simulations the weights are given by $w(j, L) = 1 - \frac{j}{L+1}$. This kernel is known as Bartlett kernel. The estimated individual specific ratio of long-run to short-run variance is defined as $\hat{s}_i^2 = \hat{\sigma}_{ui,LR}^2 / \hat{\sigma}_{ui}^2$, with $\hat{\sigma}_{ui}^2 = \frac{1}{T} \sum_{t=1}^T \hat{u}_{it}^2$. Denote by $\hat{S}_{NT} = 1/N \sum_{i=1}^N \hat{s}_i$. The quantity \hat{S}_{NT} is used later for the construction of correction factors to adjust the t -statistics of the hypothesis that $\phi_i := (\rho_i - 1) = 0$ for $i = 1, \dots, N$.

The test statistic itself, which can be based on either the coefficient $\hat{\phi}$ itself or on the corresponding t -statistic, is given from the pooled regression of \hat{e}_{it} on \hat{f}_{it-1} ,

$$\hat{\phi} = \frac{\sum_{i=1}^N \sum_{t=p_i+2}^T \hat{e}_{it} \hat{f}_{it-1}}{\sum_{i=1}^N \sum_{t=p_i+2}^T \hat{f}_{it-1}^2} \quad (4)$$

The null hypothesis is $H_0 : \phi = 0$, and the test we use in the simulations is based on the corresponding t -statistic, t_ϕ say. The standard deviation of $\hat{\phi}$ can be straightforwardly computed from the regression, since due to the pre-filtering all the errors in this pooled regression have the same (asymptotic) variance.

⁷Note that a direct estimate for the long-run variance is given by $\hat{\sigma}_{ui}^2 (1 - \sum_{j=1}^{p_i} \hat{\gamma}_{ij})^{-2}$. Levin, Lin and Chu (2002) however indicate that a variance estimation based on the first differences is found to have smaller bias under the null hypothesis, which in turn should help to improve both (finite sample) size and power of the panel unit root test.

For case 1 and test $LL93_1$, Levin and Lin (1993) show that $t_\phi \Rightarrow N(0, 1)$. For cases 2 and 3 and tests $LL93_2$ and $LL93_3$ the t -statistic t_ϕ diverges to minus infinity, and thus has to be re-centered and normalized to induce convergence towards a well defined limiting distribution,

$$t_\phi^* = \frac{t_\phi - N\tilde{T}\hat{S}_{NT}STD(\hat{\phi})\mu_{mT}}{\sigma_{mT}} \quad (5)$$

Here μ_{mT} and σ_{mT} denote mean and variance correction factors, tabulated for various panel dimensions in Table 2 on page 14 of Levin, Lin and Chu (2002). \tilde{T} denotes the average effective sample size across the individual units and $STD(\hat{\phi})$ denotes the standard deviation of $\hat{\phi}$. The adjusted t -statistics t_ϕ^* converges to the standard normal distribution.

As a remark note that the relative rates of divergence for N and T required for the consistency proofs of the test statistics differ between case 1 and cases 2 and 3. For case 1, $\lim_{N,T} \sqrt{N}/T \rightarrow 0$ is required and for cases 2 and 3 $\lim_{N,T} N/T \rightarrow 0$ is imposed. For a detailed discussion of the relevant limit concepts for nonstationary panels (and their relations) see Phillips and Moon (1999).

Harris and Tzavalis: The test of Harris and Tzavalis (1999), labelled HT , augments the analysis of Levin and Lin (1993) by considering inference for fixed T and asymptotics only in the cross-section dimension N . They obtain their results (closed form correction factors as a function of T), however, only for serially uncorrelated errors. All three cases for the deterministic variables are considered. For fixed T , the authors derive asymptotic normality (for $N \rightarrow \infty$) of the appropriately normalized and centered coefficients $\hat{\phi}$ (which are for cases 2 and 3 inconsistent for $T \rightarrow \infty$, as can be seen from the above discussion). In particular the following results are shown:

$$\sqrt{N}(\hat{\rho} - 1 - B_m) \Rightarrow N(0, C_m) \quad (6)$$

with $B_1 = 0$, $C_1 = 2/T(T-1)$, $B_2 = -3/(T+1)$, $C_2 = 3(17T^2 - 20T + 17)/5(T-1)(T+1)^3$ and $B_3 = -7.5/(T+2)$ and $C_3 = 15(193T^2 - 728T + 1147)/112(T-2)(T+2)^3$.

The practical relevance of this result is to obtain improved tests for panels with small T and large N . E.g. for case 1 the variance scaling factor used for testing is – when the limit is taken only with respect to N – by a factor $T/(T-1)$ smaller than the $LL93$ scaling factor. This implies immediately that, compared to the fixed- T test, the $LL93$ test will be oversized, i.e. the test based on test statistics constructed by letting both T and N tend to infinity will reject the null hypothesis more often. The drawback of the Harris and Tzavalis results is the mentioned restriction to white noise errors.

Breitung: Breitung (2000) develops a pooled panel unit root test that does not require bias correction factors. This is achieved by an appropriate variable transformation. The Breitung test, *UB* henceforth, is also a pooled test against the homogenous alternative.

In the description we can build upon the above discussion. Suppose that the individual preliminary ADF regressions have already been performed and standardized residuals \hat{e}_{it} and \hat{f}_{it} are available. The following orthogonalization of the residuals renders the introduction of correction factors obsolete:⁸

$$e_{it}^* = \frac{T-t}{T-t+1} \left(\Delta \hat{e}_{it} - \frac{1}{T-t} (\Delta \hat{e}_{it+1} + \dots + \Delta \hat{e}_{iT}) \right) \quad (7)$$

$$f_{it}^* = \hat{f}_{it-1} - \hat{f}_{i0} + \frac{t-1}{T} \hat{f}_{iT-1} \quad (8)$$

Here we denote for notational simplicity by T also the sample size after the auxiliary regressions. Now the unit root test is performed in the pooled regression

$$e_{it}^* = \phi^* f_{it}^* + v_{it}^* \quad (9)$$

by testing the hypothesis $H_0 : \phi^* = 0$. Breitung shows that the t -statistic of this test has a standard Normal limiting distribution for a sequential limit of first $T \rightarrow \infty$ followed by $N \rightarrow \infty$ for cases 1 to 3

We now turn to panel unit root tests that are designed against the *heterogeneous* alternative $H_1^2 : -1 < \rho_i < 1$ for $i = 1, \dots, N_1$ and $\rho_i = 1$ for $i = N_1, \dots, N$. For asymptotic consistency (over N) of these tests, a non-vanishing fraction of the individual units has to be stationary under the alternative, i.e. $\lim_{N \rightarrow \infty} N_1/N > 0$. The tests are based on group-mean estimation and test statistics, i.e. on appropriately combined individual time series unit root tests.

Im, Pesaran and Shin: In two papers Im, Pesaran and Shin (1997 and 2003), henceforth abbreviated as IPS, the authors present two group-mean panel unit root tests designed against the heterogeneous alternative. IPS consider cases 2 and 3 and allow for individual specific autoregressive structures and individual specific variances.⁹

⁸See e.g. Breitung (2000) for a discussion of the underlying mathematical argument and required assumptions.

⁹The same arguments as used in Levin and Lin (1993) might cover the case of ARMA disturbances, with the lag lengths in autoregressive approximations increasing with the sample size at an appropriate rate. The authors seem to share this view given that one of the reported simulation experiments is based on moving average dynamics for the errors.

Note that in order to apply the tables with correction factors provided by the authors identical autoregressive lag lengths for all units and a balanced panel are required. The two tests are given by a t -test based on ADF regressions ($IPSt$) and a Lagrange multiplier (LM) test ($IPSLM$).

For the case of serially uncorrelated errors, the test statistics are derived for fixed T and asymptotic N . However, in that case the t -test is not exactly a usual t -test, since the applied variance estimator is taken from the restricted regression where the coefficient on the lagged level term is set equal to 0. IPS establish asymptotic normality (for $N \rightarrow \infty$) for case 2 when $T > 5$ and for case 3 when $T > 6$.

For serially correlated errors sequential limit theory is applied, with $T \rightarrow \infty$ followed by $N \rightarrow \infty$, with a particular relative rate restriction for the LM test, with $\lim N/T = k$ for some $k > 0$.

We now describe the construction of the t -test for serially correlated errors. For the moment we focus on only one unit i . The errors u_{it} are assumed to follow an $AR(p_i + 1)$ process. Thus, the t -test statistic from the ADF regression (2) can be written as follows, with $m = 2, 3$ indicating again the deterministic terms present in the regression:

$$t_{iT,m}(p_i, \gamma_i) = \frac{\sqrt{T - p_i - m}(\mathbf{y}'_{i,-1} \mathbf{M}_{\mathbf{Q}_i} \Delta \mathbf{y}_i)}{(\mathbf{y}'_{i,-1} \mathbf{M}_{\mathbf{Q}_i} \mathbf{y}_{i,-1})^{1/2} (\Delta \mathbf{y}'_i \mathbf{M}_{\mathbf{X}_i} \Delta \mathbf{y}_i)^{1/2}}, \quad m = 2, 3 \quad (10)$$

using the notation $\gamma_i = (\gamma_{i1}, \dots, \gamma_{ip_i})'$, $\mathbf{y}_{i,-1} = [y_{i0}, \dots, y_{iT-1}]'$, $\Delta \mathbf{y}_{i,-s} = [\Delta y_{i1-s}, \dots, \Delta y_{iT-s}]'$, $s = 0, \dots, p_i$, $\Delta \mathbf{y}_i = \Delta \mathbf{y}_{i,-0}$, $\mathbf{d}_{2T} = [1, \dots, 1]'$, $\mathbf{t} = [1, \dots, T]'$, $\mathbf{d}_{3T} = [\mathbf{d}_{2T}, \mathbf{t}]$, $\mathbf{Q}_i = [\mathbf{d}_{mT}, \Delta \mathbf{y}_{i,-1}, \dots, \Delta \mathbf{y}_{i,-p_i}]$, $\mathbf{M}_{\mathbf{Q}_i} = \mathbf{I}_T - \mathbf{Q}_i(\mathbf{Q}'_i \mathbf{Q}_i)^{-1} \mathbf{Q}_i$, $\mathbf{X}_i = [\mathbf{y}_{i,-1}, \mathbf{Q}_i]$, $\mathbf{M}_{\mathbf{X}_i} = \mathbf{I}_T - \mathbf{X}_i(\mathbf{X}'_i \mathbf{X}_i)^{-1} \mathbf{X}_i$ (suppressing the index m in the matrix notation for \mathbf{Q}_i and \mathbf{X}_i). For finite values of T , the statistic t_{iT} depends upon the nuisance parameters γ_i . IPS show that this dependence vanishes for $T \rightarrow \infty$, but that the bias of the individual t -statistics under the null remains (under the null hypothesis convergence to the Dickey-Fuller distribution corresponding to the model prevails). Therefore mean and variance correction factors have to be introduced. The proposed test statistic itself is then the cross-section average of the corrected t -statistics:

$$IPSt,m(\mathbf{p}, \gamma) = \frac{\sqrt{N} \{ \bar{t}_m - \frac{1}{N} \sum_{i=1}^N \mathbb{E}(t_{iT,m}(p_i, \mathbf{0}) | \rho_i = 1) \}}{\sqrt{\frac{1}{N} \sum_{i=1}^N \text{Var}(t_{iT,m}(p_i, \mathbf{0}) | \rho_i = 1)}} \Rightarrow N(0, 1) \quad (11)$$

where $\bar{t}_m = \frac{1}{N} \sum_{i=1}^N t_{iT,m}(p_i, \gamma_i)$, $\mathbf{p} = [p_1, \dots, p_N]'$ and $\gamma = [\gamma'_1, \dots, \gamma'_N]'$. The correction factors $\mathbb{E}(t_{iT,m}(p_i, \mathbf{0}) | \rho_i = 1)$ and $\text{Var}(t_{iT,m}(p_i, \mathbf{0}) | \rho_i = 1)$ are simulated for $m = 2, 3$ for a

set of values for T and lag lengths p (see Table 3 in Im, Pesaran and Shin, 2003). Thus, without resorting to further *tailor made* Monte Carlo simulations, the applicability of the IPS tests is limited to balanced panels and identical lag lengths in all individual equations (and error processes). Simulating the mean and variance only as a function of the lag and setting the nuisance parameters $\gamma_i = 0$ introduces a bias of order $O_p(1/\sqrt{T})$, but still takes into account the finite sample effect of the different lag lengths chosen.¹⁰ Note that for $T \rightarrow \infty$ the t -statistics converge to the Dickey-Fuller distributions and thus the asymptotic correction factors are the mean and variance of the Dickey-Fuller statistic corresponding to the model. Thus, if one wants to avoid to use simulated critical values one can also refer to the asymptotic (for T) values.

Let us now turn to the Lagrange multiplier test. Using the Lagrange multiplier test principle implies that the alternative is actually given by $\rho_i \neq 1$ as opposed to $\rho_i < 1$, although the authors propose to use a 1-sided test nevertheless (see Im, Pesaran and Shin, 1997, Remark 3.2). For each individual unit the test statistic is given by

$$LM_{iT,m}(p_i, \gamma_i) = T \frac{(\Delta \mathbf{y}'_i \mathbf{M}_{\mathbf{Q}_i} \mathbf{y}_{i,-1})(\mathbf{y}'_{i,-1} \mathbf{M}_{\mathbf{Q}_i} \mathbf{y}_{i,-1})^{-1}(\mathbf{y}'_{i,-1} \mathbf{M}_{\mathbf{Q}_i} \Delta \mathbf{y}_i)}{(\mathbf{y}'_{i,-1} \mathbf{M}_{\mathbf{Q}_i} \mathbf{y}_{i,-1})} \quad (12)$$

As for the t -test above, for $T \rightarrow \infty$ the dependence upon the nuisance parameters disappears. Paralleling the above argument the Lagrange multiplier panel unit root test statistic is given by

$$\begin{aligned} IPS_{LM,m}(\mathbf{p}, \gamma) &= \frac{\sqrt{N} \{ \overline{LM}_m - \frac{1}{N} \sum_{i=1}^N \mathbb{E}(LM_{iT,m}(p_i, \mathbf{0}) | \rho_i = 1) \}}{\sqrt{\frac{1}{N} \sum_{i=1}^N \text{Var}(LM_{iT,m}(p_i, \mathbf{0}) | \rho_i = 1)}} \\ &\Rightarrow N(0, 1) \end{aligned} \quad (13)$$

where $\overline{LM}_m = \frac{1}{N} \sum_{i=1}^N LM_{iT,m}$. As indicated above, this result is developed for $\lim N/T = k$. The correction factors are available in Im, Pesaran and Shin (1997).

Maddala and Wu: Maddala and Wu (1999) tackle the panel unit root testing problem with a very elegant idea dating back to Fisher (1932).¹¹ The basic idea of Fisher can be explained based on the following simple observations that hold for any testing problem with continuous test statistics: Firstly, under the null-hypothesis the p -values, π say, of the test statistic are uniformly distributed on the interval $[0, 1]$. Secondly, $-2 \log \pi$ is therefore

¹⁰Simulation of these values for different values of T and p proceeds by generating $\Delta y_t = \varepsilon_t$, with ε_t i.i.d. $N(0, 1)$ for $t = 1, \dots, T$ and computing the t -statistic for $\rho = 1$ in the ADF regression (2) for $j = p$ and including d_{mt} .

¹¹Choi (2001) presents very similar tests that only differ in the scaling in order to obtain asymptotic normality for $N \rightarrow \infty$.

distributed as χ_2^2 , with \log denoting the natural logarithm. Thirdly, for a set of independent test statistics $-2 \sum_{i=1}^N \log \pi_i$ is consequently distributed as χ_{2N}^2 under the null hypothesis.

These basic observations can be very fruitfully applied to the panel unit root testing problem, provided that cross-sectional independence is assumed. Any unit root test with continuous test statistic performed on the individual units can be used to construct a Fisher type panel unit root test, provided that the p -values are available or can be simulated. We implement this idea by applying ADF tests on the individual units. For ADF tests estimated p -values for cases 1 to 3 can be obtained due to the extensive simulation work of James MacKinnon and his coauthors (see e.g. MacKinnon, 1994). Note as a further advantage that the Fisher test neither requires a balanced panel nor identical lag lengths in the individual equations. We have implemented the test for cases 1 to 3 based on individual ADF tests, they are labelled as MW_m for $m = 1, 2, 3$ (ignoring the dependence upon ADF in the notation).

2.2 Panel Stationarity Tests

Hadri: Hadri (2000) proposes a panel extension of the Kwiatkowski et al. (1991) test, labelled H_{LM} henceforth. Cases 2 and 3 are considered. The null hypothesis is stationarity in all units against the alternative of a unit root in all units. The alternative of a unit root in all cross-sectional units stems from the fact that this test is based on pooling. Individual specific variances and correlation patterns are allowed for. We start our discussion of the test statistics, however, assuming for the moment serially uncorrelated errors and only allow for individual specific variances σ_i^2 .

The test is constructed as a residual based Lagrange multiplier test with the residuals taken from the regressions

$$y_{it} = \delta_{mi}d_{mt} + \varepsilon_{it}, \quad m = 2, 3 \quad (14)$$

for $i = 1, \dots, N$. Denote the residuals of regression (14) by \hat{e}_{it} , and their partial sum by $S_{it} = \sum_{j=1}^t \hat{e}_{ij}$. The test statistic is then given by (m indexing again the case investigated)

$$H_{LM,m} = \frac{1}{NT^2} \sum_{i=1}^N \sum_{t=1}^T \frac{S_{it}^2}{\hat{\sigma}_{ei}^2} \quad (15)$$

with $\hat{\sigma}_{ei}^2 = 1/T \sum_{t=1}^T \hat{e}_{it}^2$.

Expression (15) can be disentangled to highlight the principle of the test. Under the null hypothesis of stationarity, the expressions in the numerator of the test statistics, $1/T^2 \sum_{t=1}^T S_{it}^2$

(for any fixed i) converge for $T \rightarrow \infty$ to an integral of a Brownian motion of the form $\hat{\sigma}_{ei}^2 \int V_m^2(r) dr$. This follows from the fact that this term is the appropriately scaled sum of squared partial sums of ε_{it} . The denominator scales the expression by the variance (see e.g. Phillips, 1987). For $m = 2$ it holds that $V_2(r) = W(r) - rW(1)$ (the so called Brownian bridge) and for $m = 3$, $V_3(r)$ is the so called second level Brownian bridge.¹² Recentering and rescaling the expressions by subtracting their mean and dividing by their standard deviation gives rise to asymptotic standard normality in the sequential limit (with now $N \rightarrow \infty$)

$$Z_{LM,m} = \frac{\sqrt{N}(H_{LM,m} - \xi_m)}{\zeta_m} \Rightarrow N(0,1) \quad (16)$$

Due to the simple shape of the correction terms, closed form solutions for the correction factors can be easily obtained. They are given by $\xi_2 = 1/6$, $\zeta_2 = \sqrt{1/45}$ and $\xi_3 = 1/15$, $\zeta_3 = \sqrt{11/6300}$. The extension to serially correlated errors is straightforward, the variance estimator $\hat{\sigma}_{ei}^2$ only has to be replaced by an estimator of the long-run variance of the noise processes in (14).

Hadri and Larsson: Hadri and Larsson (2002) extend the analysis of Hadri (2000) by considering the statistics for fixed T (the test is therefore abbreviated by H_T). The key ingredient for their result is the derivation of the exact finite sample mean and variance of the Kwiatkowski et al. (1991) test statistic that forms the individual unit building block for the Hadri type test statistic. For cases 2 and 3 they compute the exact mean and variance of $\eta_{iTm} = 1/T^2 \sum_{t=1}^T S_{iT}^2 / \hat{\sigma}_{ei}^2$, which is the core expression of the Hadri type test statistics, compare (15). Standard asymptotic theory for N then delivers asymptotic normality

$$H_{T,m} = \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\eta_{iTm} - \mathbb{E}\eta_{iTm}}{\sqrt{Var(\eta_{iTm})}} \right) \Rightarrow N(0,1) \quad (17)$$

with $\mathbb{E}\eta_{iT2} = (T+1)/6T$, $Var\eta_{iT2} = (T^2+1)/20T^2 - ((T+1)/6T)^2$ and $\mathbb{E}\eta_{iT3} = (T+2)/15T$, $Var\eta_{iT3} = (T+2)(13T+23)/2100T^3 - ((T+2)/15T)^2$.

The potential advantage of these finite T statistics is, as noted above in the discussion of the Harris and Tzavalis (1999) test, to avoid oversized tests due to treating not only N but also T asymptotic.

Note finally that serial correlation can be handled again by appropriately computing the individual specific long-run variances as discussed several times in this section. Since for

¹²These are of course the well known limits known from the time series unit root literature. We have encountered related expressions already in the discussions of the Levin, Lin and Chu (2002) and the Im, Pesaran and Shin (1997) tests.

	The Asymptotics Used in the Derivation of the Test Statistics
<i>LL93</i>	$N \rightarrow \infty$ following $T \rightarrow \infty$, $N/T \rightarrow 0$
<i>HT</i>	$N \rightarrow \infty$ and T fix
<i>UB</i>	$N \rightarrow \infty$ following $T \rightarrow \infty$
<i>IPS</i>	White Noise: $N \rightarrow \infty$ and T fix
	Serial Correlation: $N \rightarrow \infty$ following $T \rightarrow \infty$, $N/T \rightarrow k > 0$
<i>MW</i>	N, T fix, approximation of ADF p -values for finite T
<i>H_{LM}</i>	$N \rightarrow \infty$ following $T \rightarrow \infty$
<i>H_T</i>	$N \rightarrow \infty$ and T fix

Table 1: Summary of asymptotic behavior required T and N for the derivation of the limiting distribution of the tests.

fixed T only estimates for the long-run variances are available it is not clear that the above result (17) holds exactly (for finite T and $N \rightarrow \infty$).

3 The Simulation Study

In this section we present a representative selection of results obtained from our large scale simulation study. Due to space constraints we only report a small subset of results and focus on some of the main observations that emerge. The full set of results (containing about 170 pages of tables and about 30 pages with multiple figures) is available from the authors upon request.

We only report results for cases 2 and 3, since case 1 is of hardly any empirical relevance for economic time series. The computations have been performed in `GAUSS`.¹³ The number of replications is 10,000 for each DGP and sample size. Both the time dimension T and the cross-sectional dimension N assume all values in the set $\{10, 15, 20, 25, 50, 100, 200\}$. Thus, we consider in total 49 different panel sizes. The performance of the tests in relation to the sample dimensions T and N is one aspect of interest in our simulations. Remember from the discussion in the previous section that the tests rely upon different divergence rates for T and N , summarized for convenience in Table 1. One question in this respect is whether the finite- T tests of Harris and Tzavalis (1999) and Hadri and Larsson (2002) exhibit less size distortions than their asymptotic- T counterparts for panels with T small (compared to N).

¹³The computations have been performed with a substantially extended, corrected and modified set of routines based originally on Chiang and Kao (2002). A description of major changes is available upon request.

The DGPs simulated for case 2 are of the following form

$$\begin{aligned} y_{it} &= \alpha_i(1 - \rho) + \rho y_{it-1} + u_{it} \\ u_{it} &= \varepsilon_{it} + c\varepsilon_{it-1} \end{aligned} \tag{18}$$

with $\varepsilon_{it} \sim N(0, 1)$. The parameters chosen in the simulations are $\alpha = [\alpha_1, \dots, \alpha_N]$, ρ and c . We summarize the dependency of the DGP upon these parameters notationally as $DGP_2(\alpha, \rho, c)$. Note for completeness that the formulation of the intercepts as $\alpha_i(1 - \rho)$ ensures that in the unit root case (when $\rho = 1$) no drift appears. Consequently, when $\rho = 1$ we set $\alpha = 0$ in the simulations. Otherwise, the coefficients α_i are chosen uniformly distributed over the interval 0 to 4, i.e. $\alpha_i \sim U[0, 4]$. We parameterize case 3, $DGP_3(\alpha, \rho, c)$, as

$$\begin{aligned} y_{it} &= \alpha_i + \alpha_i(1 - \rho)t + \rho y_{it-1} + u_{it} \\ u_{it} &= \varepsilon_{it} + c\varepsilon_{it-1} \end{aligned} \tag{19}$$

with $\varepsilon_{it} \sim N(0, 1)$. This formulation allows for a linear trend in the absence of a unit root and for a drift in the presence of a unit root. The coefficients α_i are, as for case 2, $U[0, 4]$ distributed.

For the unit root tests the following values are chosen for ρ : 0.7, 0.8, 0.9, 0.95, 0.99 and 1.¹⁴ The former five values are used to assess the power of the tests against the stationary alternative. For the stationarity tests we only report results for $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for the size analysis. These values are chosen because preliminary simulations have shown that the stationarity tests fail to deliver acceptable results for larger values, i.e. for $\rho \in \{0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$.

For the moving average parameter c we choose all values in the set $\{0, -0.2, -0.4, -0.8, -0.9, -0.95, -0.99\}$ for the size study of the panel unit root tests and the power study of the stationarity tests, and $c \in \{0, -0.2, -0.4\}$ for the power study of the panel unit root tests and the size study of the stationarity tests. Why do we choose 0 and negative values approaching -1? It is well known from the time series unit root literature that unit root tests suffer from severe size distortions in the presence of large *positive* MA roots. In the boundary case with the MA coefficient equal to -1, the unit root is cancelled and the resultant process is stationary. Thus, the closer the coefficient c is to -1, the larger the size distortions are expected to be for any given sample size.¹⁵ With our set-up we can analyze the extent of

¹⁴Preliminary simulations have shown that for values of ρ up to 0.6 all tests exhibit satisfactory power behavior already for medium sized panels. We thus focus here only on those cases for ρ , where differential results across tests can be observed widely across the simulation experiments. The case $\rho = 0$ is included as benchmark case and also to study the tests that are designed for serially uncorrelated errors.

¹⁵It is straightforward to show that the asymptotic bias for $T \rightarrow \infty$ of $\hat{\rho}$, estimated from an AR(1) equation when the errors are not white noise but MA(1), is linear in the MA coefficient c . This holds both in the stationary and the integrated case.

the size distortions as a function of both N and T . The value $c = 0$ serves as a benchmark case with no serial correlation and is also the special case for which the test of Harris and Tzavalis (1999) is designed for. For $c \neq 0$, the choice of the lag lengths in the autoregressive approximations that most of the tests are based on becomes potentially important. We try to assess the importance of this choice by running the panel unit root tests (in case of MA errors) for several choices for the autoregressive lag length. One of our choices is BIC. We, however, also compute the test statistics for $c \neq 0$ for autoregressive lag lengths varying from 0 to 2 (since 2 is for all values of $c \geq -0.4$ the maximum lag length according to BIC), to assess the influence of the lag length selection on the size behavior (see the discussion below on the effect of lag length selection).

The careful reader will have observed that our simulated DGPs all have a cross-sectionally identical coefficient ρ under both the null and the alternative. Thus, we are in effect in a situation where we analyze the test situation with the *homogenous* alternative. We do this, because only this more restrictive alternative can be used for all tests described in the previous section. This implies to a certain extent that we do not explore the additional degree of freedom that the tests against the *heterogeneous* alternative (IPS and MW) possess. Thus, to a certain extent the pooled tests are favored in our comparison, since the last step regression to estimate ρ , is for these tests one pooled regression with about $N(T - p)$ observations, and consists of N regressions with $T - p$ observations for the group-mean tests (denoting with p the autoregressive lag length). An analysis of group-mean tests and their performance under the heterogeneous alternative is not considered separately in this paper. The relative ranking of the group mean tests in our simulations, may however still serve as an indicator for the relative performance of these tests.¹⁶

3.1 The Size of the Panel Unit Root Tests

In this subsection we report the results of the analysis of the actual size of the panel unit root tests.¹⁷ The nominal critical level in the simulation study is 5%. As noted above, the Harris and Tzavalis (1999) test is only designed for serially uncorrelated errors. Thus, this test is only computed for $c = 0$. All other tests ($LL93$, UB , IPS_t , IPS_{LM} and MW) are computed for all values of c .

¹⁶Karlsson and Löthgren (2000) present some simulation results in this respect.

¹⁷In this study we use the word size to simply denote the type I error rate at the actual DGP. This is, of course, not the size as defined by the maximal type I error rate over all feasible DGPs under the null hypothesis, see Horowitz and Savin (2000) for an excellent discussion of this issue.

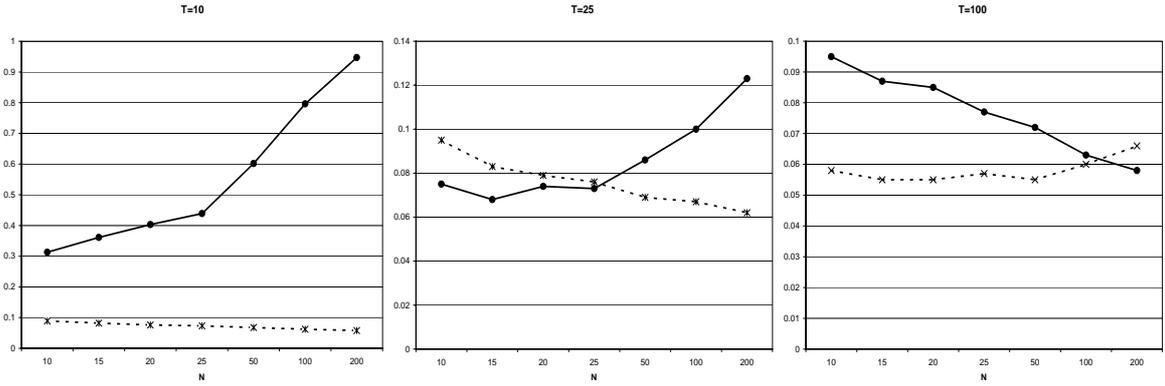


Figure 1: Comparison of the size of the Levin, Lin and Chu (2002) and the Harris and Tzavalis (1999) tests for case 2 with serially uncorrelated errors ($DGP_2(0, 1, 0)$). The $LL93_2$ results are displayed with solid lines with bullets, and the HT_2 results are displayed with dashed lines with stars.

We start with case 2 in Figures 1 and 2 and Figure 3 and 4 display results for case 3. For these and all other figures, it is always the cross-sectional dimension N that varies along the horizontal axis.¹⁸

Figure 1 displays for $c = 0$ a comparison of the size of the $LL93_2$ test and the HT_2 test, which is – as has been discussed – a fixed T version of the $LL93_2$ test (for serially uncorrelated errors). The graphs display the size for all values of N for $T \in \{10, 25, 100\}$. It becomes clearly visible that for small T like 10, the Harris and Tzavalis (1999) test has superior size performance. The difference in size performance increases with N , for both $T = 10$ and $T = 25$ (in the latter case for $N \geq 25$). This, of course, can be traced back to the fact that the asymptotic normality and the corresponding critical values of the $LL93_2$ test are based on sequential limit theory with $N \rightarrow \infty$ following $T \rightarrow \infty$ and furthermore with $\lim N/T \rightarrow 0$, see Table 1. For larger T , the improved performance of the ADF-type unit root of the $LL93_2$ test kicks in and starts to outweigh the performance deterioration with increasing N . For $T = 100$ the size of $LL93_2$ is monotonously decreasing towards 5% in the right graph of Figure 1.

Thus, for panels with little or no serial correlation the HT test can be considered an interesting extension or implementation of the $LL93$ test. No serial error correlation is unfortunately a rare case for economic time series. We therefore turn next to study the size of the

¹⁸Please note that the vertical axis is not scaled identically across the sub-plots of the figures. This stems from the fact that for all the experiments we display, identical vertical scaling leads to closely bundled lines in some of the figures.

five panel unit root tests designed for serially correlated panels, see Figure 2. In this figure we display the size performance depending upon the MA parameter c for $T = 25$.

As a baseline case, and as a follow-up to the previous analysis, we include again the case $c = 0$ (the upper left graph of Figure 2). One sees that for short panels (similar results also hold for $T = 10, 15, 20$ not shown) in particular the $LL93_2$ test and also the MW_2 test are increasingly oversized with increasing N . The two tests of Im, Pesaran and Shin (1997) and the Breitung (2000) test exhibit satisfactory size behavior. In particular for these three tests the size is not increasing with N , but stays close to the nominal level of 5%. Note, however, that for medium length panels with $T = 50, 100$, both the $LL93_2$ test and the MW_2 test exhibit satisfactory size behavior as well (for $c = 0$). The general summary for the serially uncorrelated case is that for all T investigated the Im, Pesaran and Shin (1997) tests and the Breitung (2000) test have *comparably* acceptable size. The increase is slower for these tests than for the Levin, Lin and Chu (2002) and the Maddala and Wu (1999) test. Especially for T small relative to N an application of the Harris and Tzavalis (1999) test offers an improvement over Levin, Lin and Chu (2002).

For panels with increasingly negative serial correlation, i.e. with $c \rightarrow -0.99$, the size distortions become more prominent for any given T , as is illustrated for $T = 25$ in Figure 2. For this value of T , an MA coefficient of $c = -0.4$ is the ‘boundary’ case (amongst the values of c investigated) for which for some tests the size does not rise sharply (i.e. up to 0.2 or higher) as N is increased to 200. For the more negative values of c , the size diverges for all tests to 1 for $N \geq 100$. Somewhat surprising, also for the larger values of T , the ‘boundary’ value for the MA coefficient is still given by $c = -0.4$. For $T \geq 50$ and for $c \in \{-0.8, -0.9, -.95\}$, ‘size divergence’ occurs again for $N \geq 100$.¹⁹ This divergence can be partly mitigated by using smaller values for the autoregressive lags than suggested by BIC.²⁰ In light of Table 1, this divergence might not be too surprising, as most tests’ critical values are derived on the basis of sequential limit theory. There are, however, exceptions: The Maddala and Wu test is developed for finite given N , and uses an approximation of the p -values for the individual ADF tests. For serially uncorrelated errors furthermore Im, Pesaran and Shin (1997) provide

¹⁹Generally, for very small $T = 10, 15$ all tests exhibit smaller size distortions as a function of N than for larger T .

²⁰Surprisingly, performing *no correction* for serial correlation sometimes mitigates the ‘size-divergence’ for increasing N , in particular for c close to 0. For the values of c close to -1, including more lags is in general preferable. The values of c close to -1 also lead, as expected, to larger lag lengths suggested by BIC for $T \geq 100$. It is not clear whether these observations have practical implications or generalize beyond the MA(1) error processes simulated in this study. An investigation of this issue is left for future research.

critical values for the tests for finite T and only $N \rightarrow \infty$. Thus, we a priori expect the MW test (and the IPS tests for serially uncorrelated errors) to be less prone to the size distortions observed above. However, this is not observed throughout our simulations. The performance of the Maddala and Wu test as displayed in Figure 2 is quite representative. For $c = 0$ it shows the fastest size divergence for $N \rightarrow 200$ and for $c \neq 0$ its size performance is in the ball park of other tests. What about the two IPS tests? Both tests exhibit rather similar behavior and their size stays relatively stable close to the nominal value. Of course, for c becoming too negative some size distortions occur.

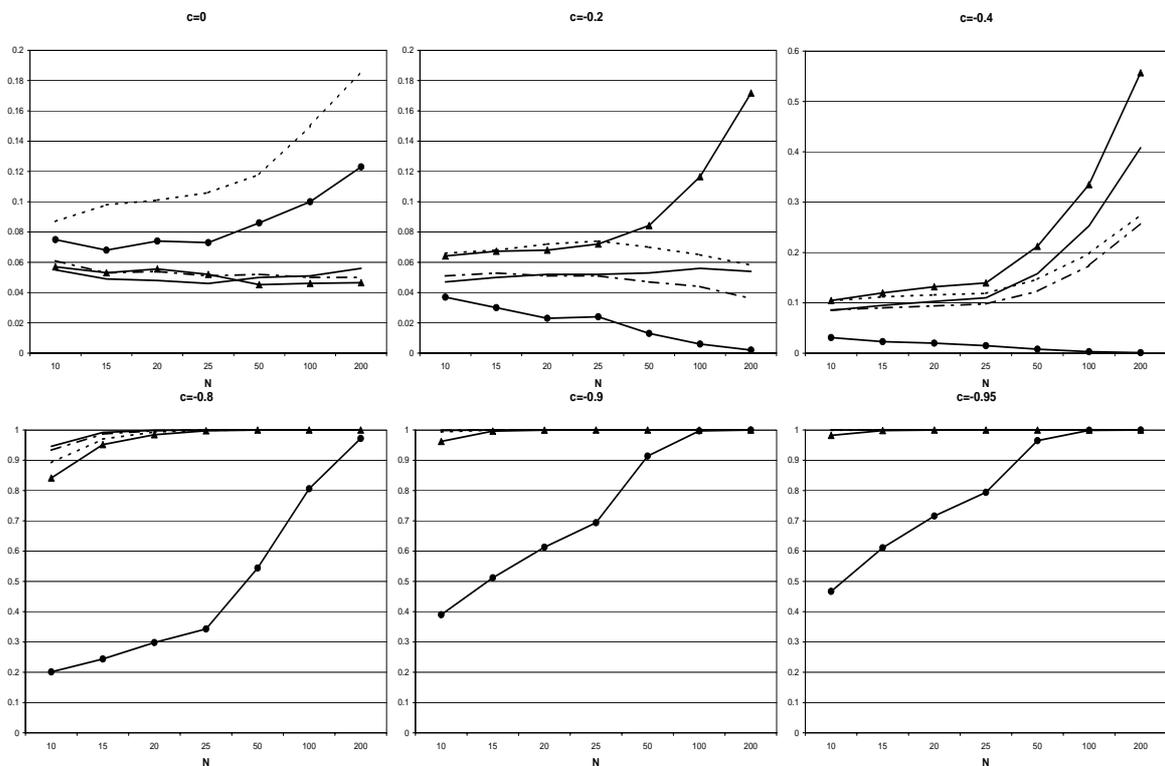


Figure 2: Size of panel unit root tests for case 2 ($DGP_2(0,1,c)$) with $c \in \{0, -0.2, -0.4, -0.8, -0.9, -0.95\}$ for $T = 25$. The solid line with bullets corresponds to $LL93_2$, the solid line with triangles corresponds to UB_2 , the solid line corresponds to $IPS_{t,2}$, the dash-dotted line corresponds to $IPS_{LM,2}$, and the dashed line corresponds to MW_2 .

The tests that exhibit in most cases the slowest divergence of the size as c is decreased towards -0.95 , are the $LL93_2$ and (usually second slowest) the UB_2 test. This behavior is the large T extension of the behavior observed for the Levin, Lin and Chu (2002) test for small $T \in \{10, 15, 25\}$. The nominal size of the $LL93_2$ test *even decreases* for fixed small $T \in \{10, 15, 25\}$ for N tending to 200 for certain values of c (e.g. for $T = 25$, this holds for

$c = -0.2, -0.4$). With increasing serial correlation, instead of being undersized this test has the slowest divergence of the size towards 1 for $N \rightarrow \infty$. For the UB_2 test the behavior is different, since it displays relatively fast size divergence for the smaller values of c (see for an example the center graph in the upper row of Figure 2). Thus, summarizing we find that for the panels with highly negative MA coefficients, the $LL93_2$ test is *grosso modo* the least distorted test, with in general a slight tendency for being undersized in small T and large N panels.

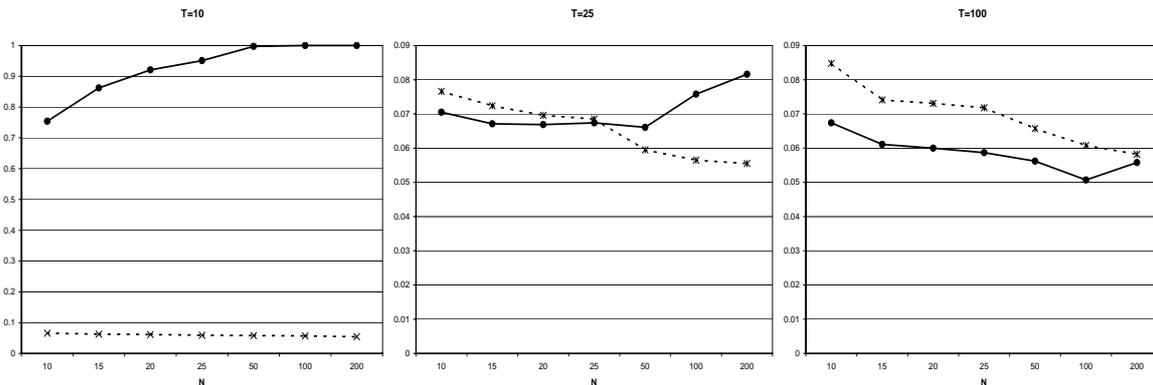


Figure 3: Comparison of size of the Levin, Lin and Chu (2002) and the Harris and Tzavalis (1999) tests for case 3 with serially uncorrelated errors ($DGP_3(\alpha, 1, 0)$). The $LL93_3$ results are displayed with solid lines with bullets, and the HT_3 results are displayed with dashed lines with stars.

We now turn to case 3 and start again with a comparison of the Levin, Lin and Chu (2002) and Harris and Tzavalis (1999) tests for $c = 0$. See Figure 3, where we display as above results for $T \in \{10, 25, 100\}$. As in the case of random walks without drift, substantially smaller size distortions are observed for the Harris and Tzavalis (1999) test (in particular again for small T). The differences for the larger values of T are slightly less pronounced than in case 2. For $T \geq 50$, the size performance is very satisfactory also for large values of N .

In case of no serial correlation in u_{it} size divergence only occurs for $T = 10, 15$ for the $LL93_3$ test and at a lesser rate for the MW_3 test. For $T = 25$ all tests except the MW_3 test exhibit satisfactory size performance for all N . Only the MW_3 test still has size distortions up to 0.3 when $N \rightarrow 200$ and $T = 25$. The two IPS tests have very similar performance. Thus, in case of no serial correlation, size divergence occurs only for the smallest values of T . The relative sample sizes are therefore not of great concern as soon as $T \geq 25$, and even for shorter panels three tests (UB_3 , $IPS_{t,3}$ and $IPS_{LM,3}$) show satisfactory size performance.

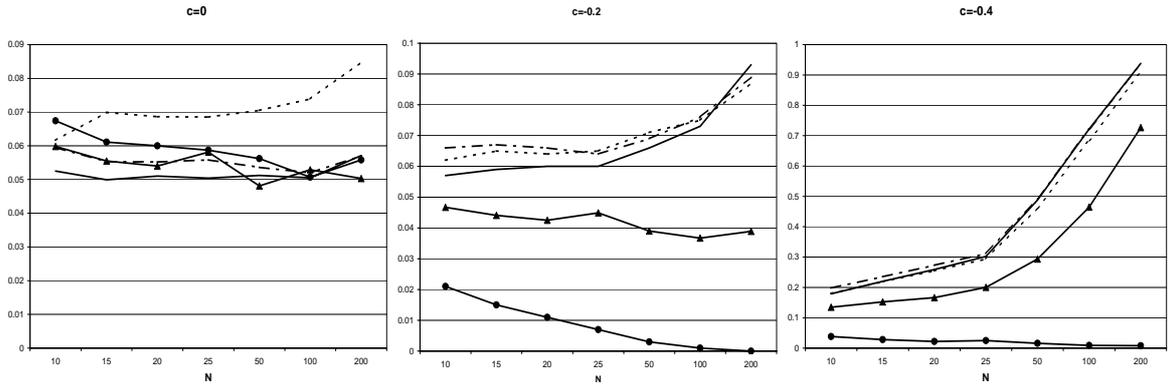


Figure 4: Size of panel unit root tests for case 3 ($DGP_3(\alpha, 1, c)$) with $c \in \{0, -0.2, -0.4\}$ for $T = 100$. The solid line with bullets corresponds to $LL93_3$, the solid line with triangles corresponds to UB_3 , the solid line corresponds to $IPSt_{,3}$, the dash-dotted line corresponds to $IPSLM_{,3}$ and the dashed line corresponds to MW_3 .

With serially correlated errors, as in case 2, the value $c = -0.4$ is the boundary value for which not all tests' size diverges to 1 for $T, N \rightarrow 200$. Two tests have substantially smaller size distortions (over a variety of combinations of T and N) than the other tests. These are the test of Levin, Lin and Chu (2002) and the Breitung (2000). For $c \geq -0.2$ these two tests have size below 0.1 for all combinations of T and N , whereas the other tests' size is diverging to at least 0.8 for $N \rightarrow 200$ and $T \geq 50$. The $LL93_3$ test is not subject to size divergence also for $c = -0.4$. The size divergence behavior of the $IPSLM_{,3}$, $IPSt_{,3}$ and MW_3 tests is very similar. Thus, for the case of random walks with drifts, the summary is that the $LL93_3$ and the UB_3 test outperform the other tests. The major exception to this general rule occurs for $T \in \{10, 15\}$, where the $IPSLM_{,3}$ tests shows good size properties and the $LL93_3$ test does not yet appear so favorable. This is to a certain extent surprising, given that the $LL93_3$ test is pooled and the $IPSLM_{,3}$ test is a group-mean test. The observation concerning the relative performance of the $IPSLM_{,3}$ and the $LL93_3$ tests, with the latter starting to outperform the former for $T \geq 15$ also holds for $c < -0.4$. As for case 2, it is worth noting that the divergence problem also occurs for the MW_3 test (see the upper right graph in Figure 4 for an example) despite being developed for fixed N inference. Performance improvements can be realized by varying the number of lagged differences included in the regressions, similar to case 2. We find again that the relative size of T and N has significant influence on the results obtained for small T . This observation has to be stressed again, although it is just a consequence of the construction of the tests, cf. Table 1.

3.2 The Power of the Panel Unit Root Tests

The discussion of the power of the panel unit root tests against the stationary alternative is *not* based on ‘so called’ *size-corrected* critical values. This follows from the fact, discussed in detail in Horowitz and Savin (2000), that ‘size correction’ based on arbitrary points in the set of feasible DGPs under the null hypothesis in general leads to empirically irrelevant critical values.²¹ The problem arises because the actual type I errors (of any of the unit root tests) vary substantially across integrated ARMA processes. Therefore, size corrections – which, hence, should correctly be labelled type I error corrections for given DGP – do not necessarily lead to insights that can be generalized.

Therefore our power analysis is based on the asymptotic critical values. Horowitz and Savin (2000) discuss situations when bootstrap based critical values lead to considerable power gains, this is not discussed here, but the interested reader will find bootstrap applications of the tests discussed in this paper in Wagner and Hlouskova (2004). The bootstrap based inference in that paper often leads to different conclusions than inference based on asymptotic critical values.

Before we next discuss the results briefly, let us start by summarizing a few general observations. First, maybe not surprising, power is monotonously increasing in N for all DGPs simulated under the stationary alternative for all values of T (see for example Figures 5 and 8). Note, however, that power does not increase monotonously in T for given N . This occurs for relatively small values of T and N , when ρ assumes values close to 1. For larger values of T power increases when T is increased further for any value of N . Most notably the $LL93$ test is subject to this non-monotonicity of power in T .

We start our discussion again with case 2, see Figure 5. In this figure we display the power of the panel unit root tests for $\rho = 0.8$ and $T \in \{10, 25, 100\}$. The upper row shows the case with serially uncorrelated errors and the lower row displays the case when $c = -0.4$. The figure displays one representative result clearly, namely the effect of the value of c on the ordering of the tests with respect to power. The highest power curve corresponds throughout to either the UB_2 or the $LL93_2$ test (also for parameter choices not displayed in figures here). For the larger values of T it is generally the UB_2 test, whereas the $LL93_2$ test has highest

²¹This occurs unless the test statistic is pivotal, which is not the case for finite samples for any of the unit root or stationarity tests discussed in this paper, as is illustrated by the large differences in size for different parameters c .

power in many cases for the smaller or smallest values of T . Corresponding to the sensitivity discussed in the previous subsection, altering the lag lengths in the ADF regressions can be used to improve the power performance of the Levin, Lin and Chu (2002) test. The most variable power performance is observed for the MW_2 test, who is ranked from second to last place without any detectable dependence upon sample size or parameters (see Figure 5). For the two group-mean tests of Im, Pesaran and Shin, power is comparatively low for small values of T (this is most likely a consequence of the group-mean construction of the test statistic), but is in general quite appealing for larger values of T . However, the UB_2 test is for those large panels the most powerful test. Note also that for $T \geq 100$ even for $N = 10$ all tests have power equal to 1, for $\rho \leq 0.9$. For even larger values of $\rho \in \{0.95, 0.99\}$, $N \geq 50$ is required to have power tending to 1 for $T \geq 100$. The previous observations holds both for $c = 0$ and $c \neq 0$.

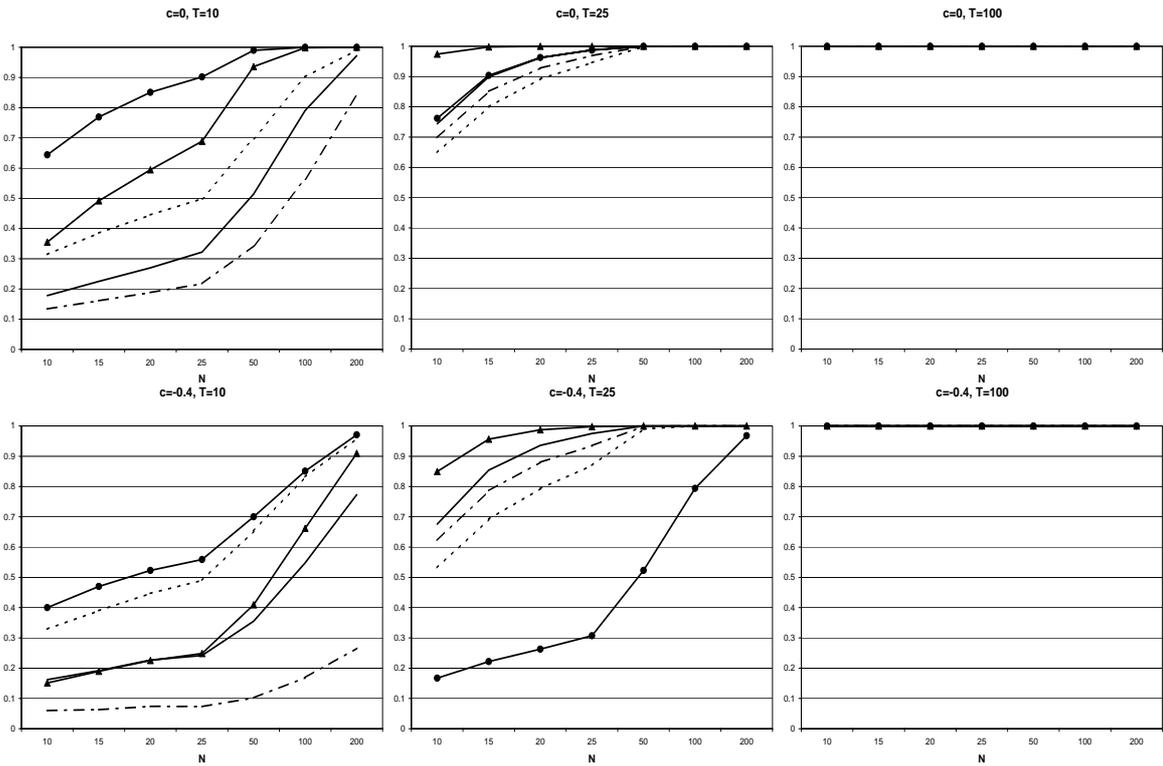


Figure 5: Power of panel unit root tests for case 2 with $\rho = 0.8$ ($DGP_2(\alpha, 0.8, c)$) for $c \in \{0, -0.4\}$ and $T \in \{10, 25, 100\}$. The solid line with bullets corresponds to $LL93_2$, the solid line with triangles corresponds to UB_2 , the solid line corresponds to $IPSt,2$, the dash-dotted line corresponds to $IPSLM,2$ and the dashed line corresponds to MW_2 .

For case 3 the results are less clear than for case 2. There are panel dimensions (T, N)

and parameter values (ρ, c) for which each of the five tests has highest power. Some clear observations emerge only for $T = 10$, where the $LL93_3$ test is the most powerful test for $c = 0$ and the MW_3 test is the most powerful test for $c \neq 0$. The latter is the second most powerful test when $T = 10$ and $c = 0$. This is a bit surprising, since the Maddala and Wu (1999) test is a group-mean test. It seems that the finite N inference mitigates this disadvantage, which does not occur for the Im, Pesaran and Shin tests. The UB_3 test is performing relatively well, but not as good compared to the other tests as in case 2. Also for case 3 power is basically equal to 1 for all tests for all values of N for values of ρ up to 0.9 for $T \geq 100$. Some graphical results of the power of the panel unit root tests for case 3 are displayed in Figure 8 in the appendix.

3.3 The Size of the Panel Stationarity Tests

Some representative results for the size behavior of the stationarity tests of Hadri (2000) and Hadri and Larsson (2002) are displayed in Figure 6 for case 2. The figure displays the size of both tests as a function of $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $c \in \{0, -0.4\}$ for $T = 25$. Remember from the discussion in Section 2.2 that the H_T test of Hadri and Larsson (2002) is based on finite T inference.

One of the aspects we want to compare is the relative performance of the H_T and the H_{LM} test of Hadri (2000). Focusing on this aspect first, we find that only for the case $\rho = 0$ and $c = 0$ substantial differences between these two tests occur. This holds not only for $T = 25$ shown, but also for smaller and larger values of T . As expected, for larger values of T , the differences become smaller. The explanation for this result is that the non-parametric estimation to allow for serial correlation is too imprecise to allow for improved size performance, since advantages of the H_T test only materialize in the single case where no serial correlation corrections are required.

The second general observation that is exemplified in Figure 6 is that $c = 0$ leads to *larger* size distortions than $c < 0$, as shown with the example $c = -0.4$ in the figure. This finding can be explained by noting that our generated processes are for $\rho = 0.4$ and $c = -0.4$ white noise processes, since the AR and the MA root are cancelled in this case. Thus, it seems that only for processes ‘close to white noise’ the size of the tests is acceptable.²² This is

²²The difference to the white noise case displayed in the upper left graph and again in the middle graph in the lower row (also white noise) is entirely attributable to the non-parametric correction. For c close to -1, as expected large size distortions occur throughout.

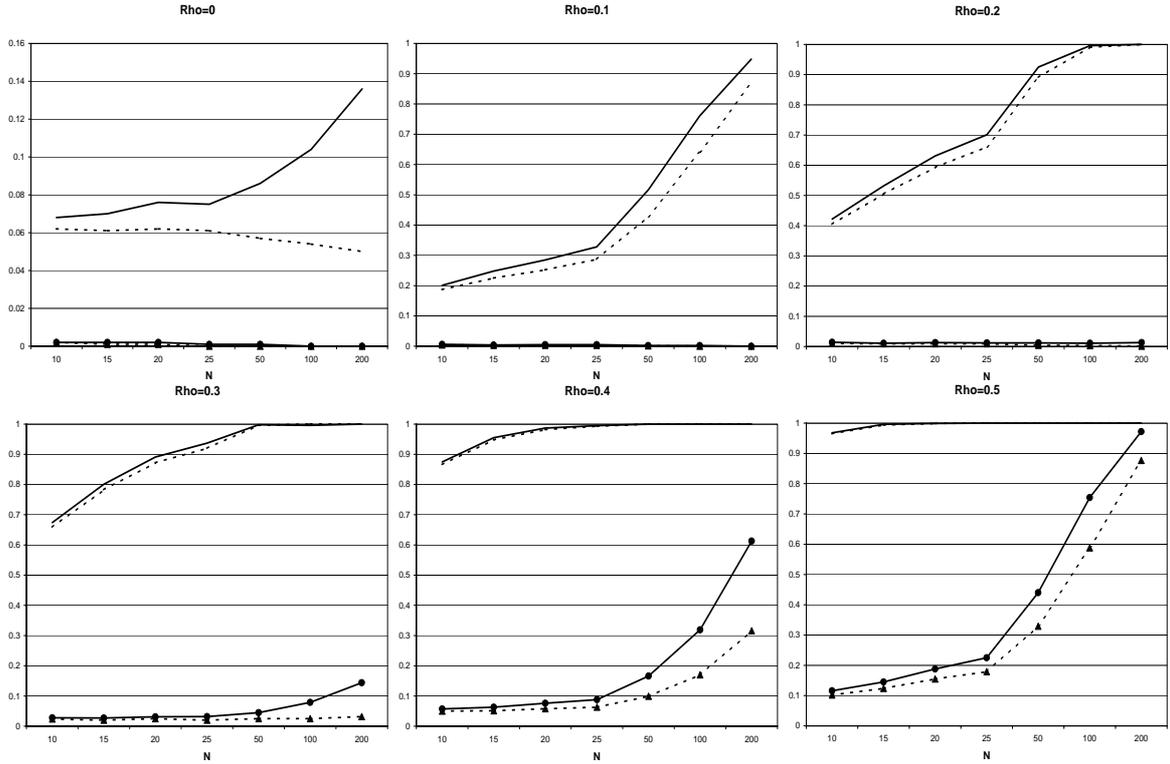


Figure 6: Size of Hadri (2000) and Hadri and Larsson (2002) stationarity tests for $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, $c \in \{0, -0.4\}$, and $T = 25$. The $H_{LM,2}$ results are displayed by solid lines for $c = 0$ and by solid lines with bullets for $c = -0.4$. The $H_{T,2}$ results are displayed by dashed lines for $c = 0$ and by dashed lines with triangles for $c = -0.4$. For $c = -0.4$ the window-size of the Bartlett kernel is chosen according to the Newey-West criterion.

bad news since, in case of stationary autoregressive time series with strong serial first order autocorrelation, the tests have basically size 1. This finding also holds for larger values of T . Our observation, however, can explain to a certain extent the fact that an application of the panel stationarity tests a la Hadri (2000), often leads to a *rejection of the null hypothesis*. Even for ‘highly stationary panels’ as displayed in the figure, the null is rejected in almost all replications (unless the AR and the MA root are nearly or exactly cancelled). In other words, the Hadri (2000) and the Hadri and Larsson (2002) test can be ‘used to find unit roots’ (although, of course, strictly speaking a rejection of the null does *not imply acceptance* of the alternative).

Note that qualitatively entirely similar findings are obtained for case 3, which we therefore do not discuss separately.

3.4 The Power of the Panel Stationarity Tests

We finally briefly discuss the power of the panel stationarity tests. The size results (rejection of stationarity for many cases) already allows for predictions concerning the behavior of the power function. First, power will be low for small T and processes ‘close’ to white noise. ‘Close’ to white noise here means that the MA coefficient is close to -1, so that the unit root is nearly cancelled. This is exactly what happens, see the graphical results for case 2 in Figure 7 and for case 3 in Figure 9 in the appendix that show exactly what has just been discussed. Summing up: The high power stems from the fact that the Hadri (2000) and Hadri and Larsson (2002) tests tend to reject stationarity most of the times even for highly stationary series. It is thus not a surprise that stationarity is also rejected for unit root series. It is only the general observation that it is hard to detect nonstationarity in short time series, that *reduces* power (and size) of the tests for small $T \in \{10, 15\}$.

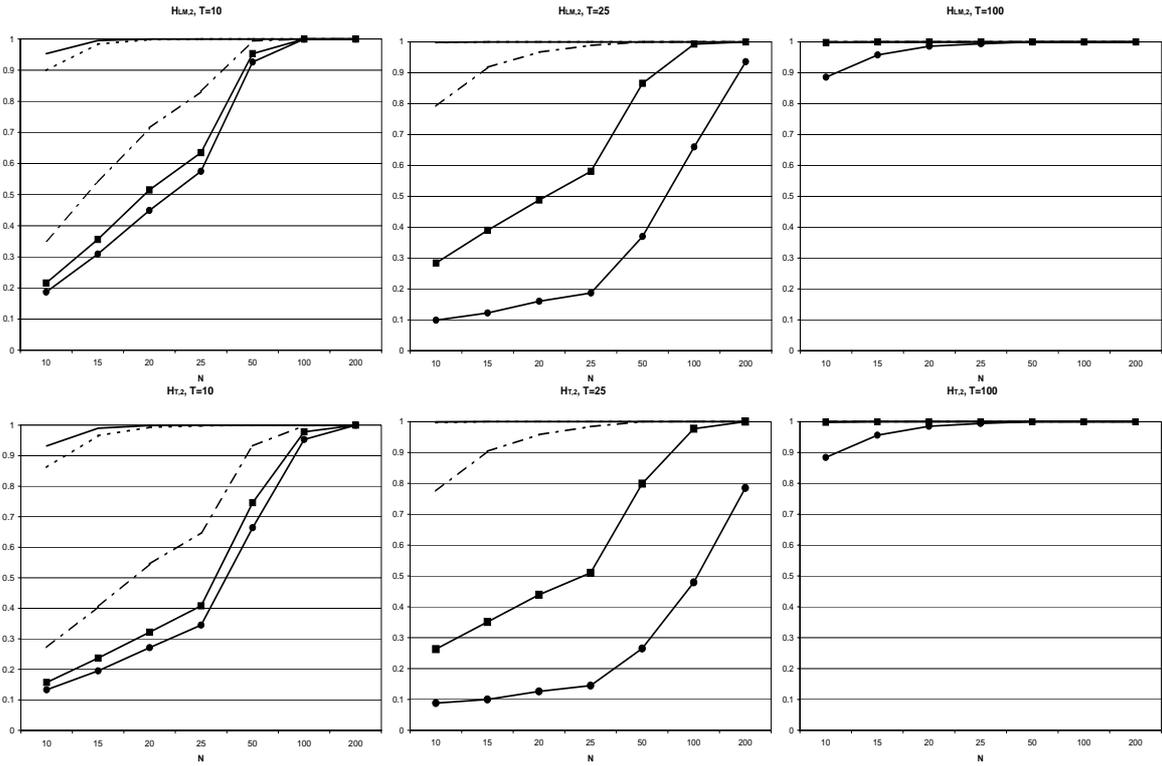


Figure 7: Power of the Hadri (2000) test (upper row) and of the Hadri and Larsson (2002) test (lower row) for case 2 ($DGP_2(0, 1, c)$) against the alternative of a unit root for $c \in \{-0.2, -0.4, -0.8, -0.9, -0.95\}$ and $T \in \{10, 25, 100\}$. The results for $c = -0.2$ are displayed with solid lines, for $c = -0.4$ with dashed lines, for $c = -0.8$ with dash-dotted lines, for $c = -0.9$ with solid lines with squares, and for $c = -0.95$ with solid lines with bullets.

4 Conclusions

The strongest and most unequivocal conclusion from our simulations is that the panel stationarity tests of Hadri (2000) and Hadri and Larsson (2002) perform very poorly. This is to a certain extent similar to the often observed poor performance of the Kwiatkowski et al. (1991) test, which is the time series building block of these tests. The null of stationarity is rejected as soon as sizeable serial correlation of either the autoregressive or the moving average type is present.

The picture that emerges for the panel unit root tests is much more differentiated and hardly any clear-cut patterns emerge (which is itself an interesting observation). There are, however, a few exceptions. First, for case 2 (with intercepts under stationarity) the best power behavior is displayed by either the Levin, Lin and Chu (2002) test or by the Breitung (2000) test. Second, for serially uncorrelated panels the Harris and Tzavalis (1999) implementation of the Levin and Lin test offers substantial improvements for short panels. The third clear message that emerges from the simulations is that for short panels size and power problems emerge when the cross-sectional dimension is too large, i.e. when T is too small compared to N . This finding is in line with the fact that most test statistics are based on sequential limits with first $T \rightarrow \infty$ followed by $N \rightarrow \infty$. However, the test of Maddala and Wu (1999) developed for fixed- N inference does not show superior performance with respect to variations of N , e.g. concerning size divergence as a function of N . Fourth, as expected for the moving average coefficient $c \rightarrow -1$ the size distortions become stronger. Across our simulations the value of $c = -0.4$ has emerged as a ‘boundary’ case for which at least some tests exhibit satisfactory behavior (for $T \geq 25$ and all values of N). Taking a rough average over all experiments the Levin, Lin and Chu (2002) and Breitung (2000) tests have the smallest size distortions. However, there is large variance around this result and there are constellations where e.g. the Levin, Lin and Chu (2002) has very rapid size divergence. Combined with the good power performance (notably for case 2) those two tests appear *grosso modo* quite favorable.

At this point, however, we have to note again that the *group-mean* tests of Im, Pesaran and Shin (1997 and 2003) and of Maddala and Wu (1999) are to a certain extent disadvantaged in our simulation study. This stems from the fact that we simulate (up to the intercepts and trend slopes) homogenous panels under both the null and the alternative. For such panels

pooling is apparently both advantageous and straightforward. When comparing only the group-mean tests, we do not find a stable ranking over parameter values and sample sizes, neither with respect to size nor with respect to power. However, only a detailed analysis with more heterogeneous panels will allow to understand the relative performance of these tests for situations where the additional degree of freedom they offer (the heterogeneous alternative) is utilized

The impact of lag length selection in the ADF type regressions, which has found to be ‘non-monotonous’ in c , is an open issue for future research. By non-monotonicity we mean the observation that for c close to 0 smaller lag lengths than suggested by BIC lead in many cases to improved performance, whereas for values of c close to -1 a larger number of lagged differences than suggested by BIC often leads to improvements. A priori such behavior is not expected. In this respect also the influence of the time dimension of the panel on this observation has to be investigated further.

Finally, the variability of the results over the parameters, observed not only for small but also for large panels, suggests that substantial performance improvements might be realized by relying upon bootstrap inference.

Appendix: Additional Figures

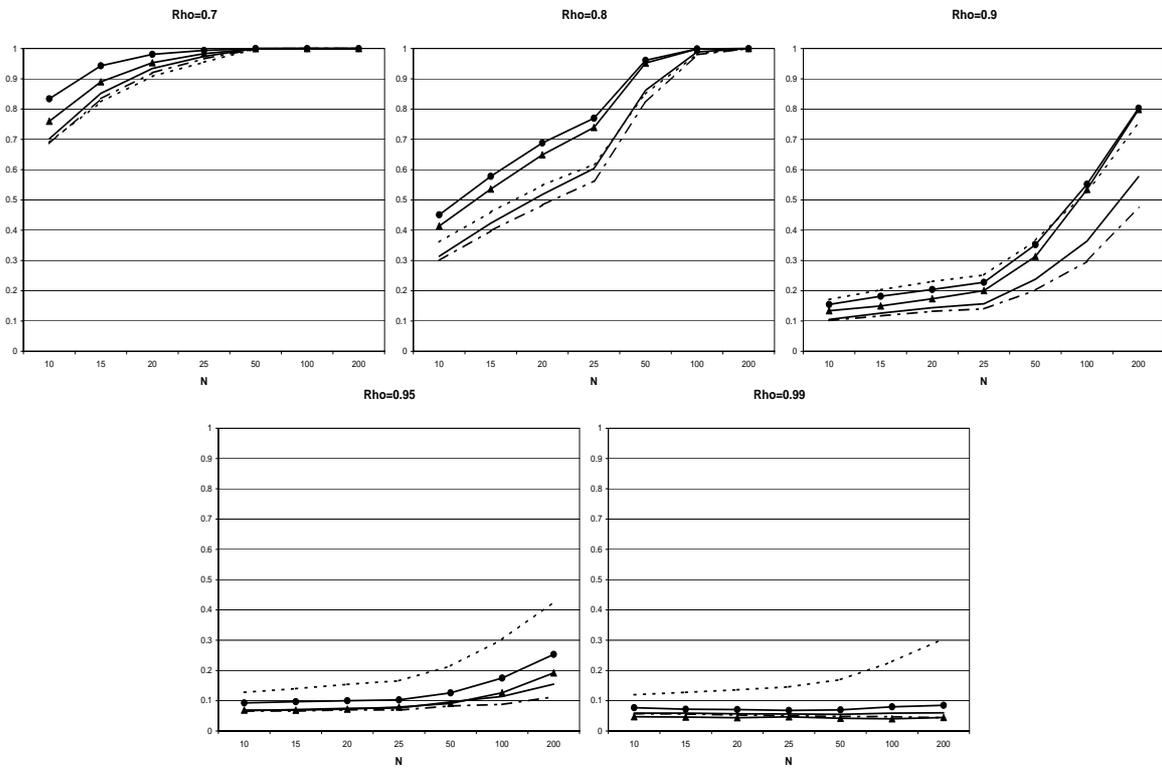


Figure 8: Power of panel of unit root tests for case 3 with $\rho \in \{0.7, 0.8, 0.9, 0.95, 0.99\}$ for $T = 25$ and $c = 0$ ($DGP_3(\alpha, \rho, 0)$). The solid line with bullets corresponds to $LL93_3$, the solid line with triangles corresponds to UB_3 , the solid line corresponds to $IPS_{t,3}$, the dash-dotted line corresponds to $IPS_{LM,3}$ and the dashed line corresponds to MW_3 .

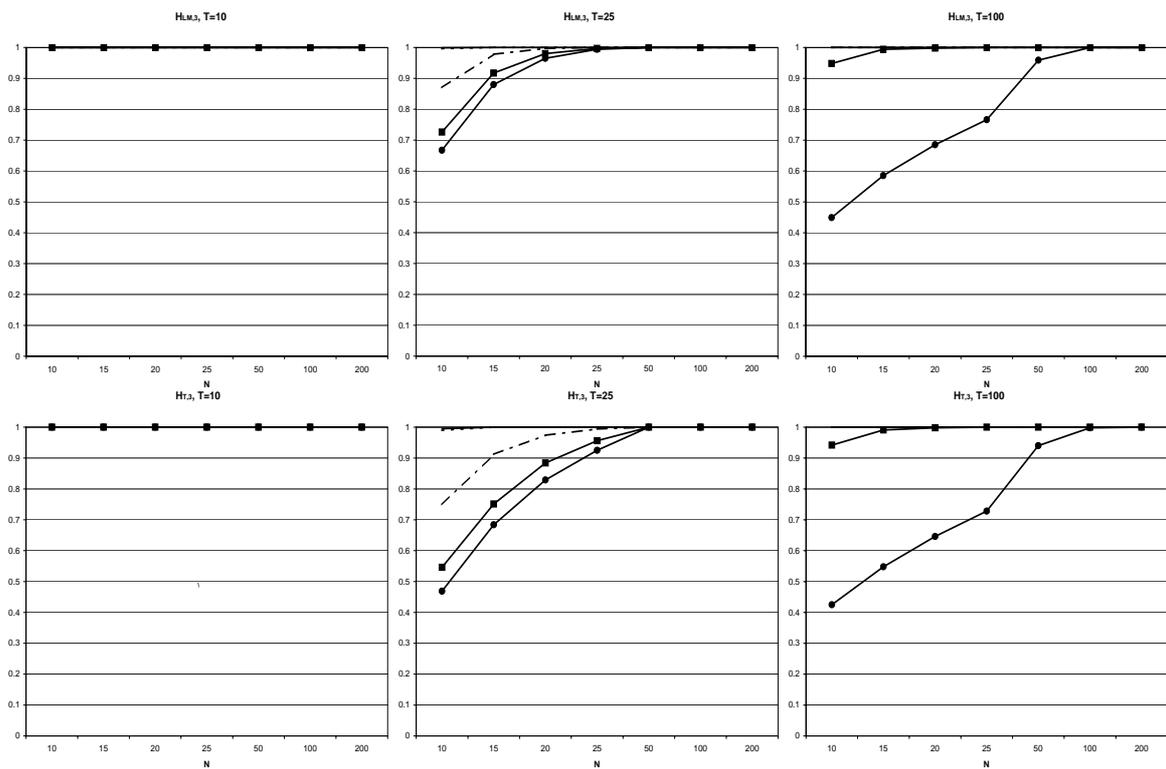


Figure 9: Power of the Hadri (2000) test (upper row) and of the Hadri and Larsson (2002) test (lower row) for case 3 ($DGP_3(\alpha, 1, c)$) against the alternative of a unit root for $c \in \{-0.2, -0.4, -0.8, -0.9, -0.95\}$ and $T \in \{10, 25, 100\}$. The results for $c = -0.2$ are displayed with solid lines, for $c = -0.4$ with dashed lines, for $c = -0.8$ with dash-dotted lines, for $c = -0.9$ with solid lines with squares, and for $c = -0.95$ with solid lines with bullets.

References

- Agiakloglou, C. and P. Newbold (1996). The Balance between Size and Power in Dickey-Fuller Tests with Data-dependent Rules for the Choice of Truncation Lag. *Economics Letters* **56**, 229–234.
- Andrews, D.W.K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica* **59**, 817–858.
- Bai, J. and S. Ng (2004). A PANIC Attack on Unit Roots and Cointegration. *Econometrica* **72**, 1127–1178.
- Breitung, J. (2000). The Local Power of some Unit Root Tests for Panel Data, 161–177. In Baltagi, B.H. (Ed.) *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, Elsevier, Amsterdam.
- Chang, Y. (2002). Nonlinear IV Unit Root Tests in Panels with Cross-Sectional Dependency. *Journal of Econometrics* **110**, 261–292.
- Chiang, M-H. and C. Kao (2002). Nonstationary Panel Time Series Using NPT: A User Guide. Center for Policy Research, Syracuse University.
- Choi, I. (2001). Unit Root Tests for Panel Data. *Journal of International Money and Finance* **20**, 249–272.
- Choi, I. (2002). Combination Unit Root Tests for Cross-Sectionally Correlated Panels. Forthcoming in: Corbae, D., S. Durlauf and B. Hansen (Eds.) *Econometric Theory and Practice: Frontiers of Analysis and Applied Research: Essays in Honor of Peter C. B. Phillips*, Cambridge University Press.
- Fisher, R.A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
- Hadri, K. (2000). Testing for Stationarity in Heterogeneous Panel Data. *Econometrics Journal* **3**, 148–161.
- Hadri, K. and R. Larsson (2002). Testing for Stationarity in Heterogeneous Panel Date Where the Time Dimension is Fixed. Mimeo.

- Harris, R.D.F. and E. Tzavalis (1999). Inference for Unit Roots in Dynamic Panels Where the Time Dimension is Fixed. *Journal of Econometrics* **90**, 1–44.
- Horowitz, J.L. and E.N. Savin (2000). Empirically Relevant Critical Values for Hypothesis Tests: A Bootstrap Approach. *Journal of Econometrics* **95**, 375–389.
- Im, K.S., M.H. Pesaran and Y. Shin (1997). Testing for Unit Roots in Heterogeneous Panels. Mimeo.
- Im, K.S., M.H. Pesaran and Y. Shin (2003). Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics* **115**, 53–74.
- Karlsson, S. and M. Löthgren (2000). On the Power and Interpretation of Panel Unit Root Tests. *Economics Letters* **66**, 249–255.
- Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Y. Shin (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. *Journal of Econometrics* **54**, 91–115.
- Levin, A. and C.F. Lin (1992). Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties. UC San Diego Working Paper 92-23.
- Levin, A. and C.F. Lin (1993). Unit Root Tests in Panel Data: New Results. UC San Diego Working Paper 93-56.
- Levin, A., C.F. Lin and C-S.J. Chu (2002). Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties. *Journal of Econometrics* **108**, 1–22.
- MacKinnon, J.G. (1994). Approximate Asymptotic Distribution Functions for Unit Root and Cointegration Tests. *Journal of Applied Econometrics* **11**, 601–618.
- Maddala, G.S. and S. Wu (1999). A Comparative Study of Unit Root Tests with Panel Data and a Simple New Test. *Oxford Bulletin of Economics and Statistics* **61**, 631–652.
- Moon, H.R. and B. Perron (2004). Testing for a Unit Root in Panels with Dynamic Factors. *Journal of Econometrics* **122**, 81–126.
- Newey, W. and K.D. West (1994). Automatic Lag Selection for Covariance Matrix Estimation. *Review of Economic Studies* **61**, 631 – 653.

- Pesaran, M.H. (2003). A Simple Panel Unit Root Test in the Presence of Cross Section Dependence. Mimeo, University of Cambridge.
- Phillips, P.C.B. (1987). Time Series Regression with a Unit Root. *Econometrica* **55**, 277-301.
- Phillips, P.C.B. and H.R. Moon (1999). Linear Regression Limit Theory for Nonstationary Panel Data. *Econometrica* **67**, 1057–1111.
- Wagner, M. and J. Hlouskova (2004). What's Really the Story with this Balassa-Samuelson Effect in the CEECs? University of Bern, Department of Economics, Discussion Paper 04/16.