THE PERFORMANCE OF PANEL UNIT ROOT AND STATIONARITY TESTS: RESULTS FROM A LARGE SCALE SIMULATION STUDY

Jaroslava Hlouskova □ Department of Economics and Finance, Institute for Advanced Studies, Vienna, Austria
Martin Wagner □ Department of Economics and Finance, Institute for Advanced Studies, Vienna, Austria

This paper presents results on the size and power of first generation panel unit root and stationarity tests obtained from a large scale simulation study. The tests developed in the following papers are included: Levin et al. (2002), Harris and Tzavalis (1999), Breitung (2000), Im et al. (1997, 2003), Maddala and Wu (1999), Hadri (2000), and Hadri and Larsson (2005). Our simulation set-up is designed to address inter alia the following issues. First, we assess the performance as a function of the time and the cross-section dimensions. Second, we analyze the impact of serial correlation introduced by positive MA roots, known to have detrimental impact on time series unit root tests, on the 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.

Keywords Panel stationarity test; Panel unit root test; Power; Simulation study; Size.

JEL Classification C12; C15; C23.

1. INTRODUCTION

Panel unit root and stationarity tests have become extremely popular and widely used over the last decade. The fact that several such tests are now implemented in commercial software will lead to further increased usage. 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).

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 et al. (2002), Harris and Tzavalis (1999), Breitung (2000), Im et al. (1997, 2003), and Maddala and Wu (1999). We also include two panel stationarity tests, developed in Hadri (2000) and Hadri and Larsson (2005). We include also a discussion concerning the effect of two commonly used forms of cross-sectional covariance on the test performance (constant covariance and a covariance matrix in Toeplitz form; see the details in Section 3). It turns out that the performance comparison across tests is remarkably robust to these two covariance structures; some examples are displayed in the appendix. Therefore the presentation of the results focuses on cross-sectionally independent panels.

To overcome the cross-sectional independence restriction of first generation tests, in recent years several tests that allow for some form or another of cross-sectional dependence have been developed. These include Bai and Ng (2004), Chang (2002), Choi (2002), Moon and Perron (2004), and Pesaran (2003). The most general results are derived in Bai and Ng (2004) with a factor model approach, who allow for (multiple) common stationary and integrated components. The other papers mentioned generally allow only for stationary common components, which may be insufficient for many practical applications. For example in purchasing power parity studies, the base country price index is a potential candidate for a nonstationary common component (for a detailed discussion see Wagner, 2005). Except for the factor model approach, the theory for nonstationary cross-sectionally dependent panels appears to still be in an early stage, and no widely accepted modeling strategies for cross-sectional dependencies have emerged up to now. For example, for macroeconomic panels it may be necessary to consider dependence structures that are invariant to the ordering of the panel (see, e.g., Gregoir, 2004) or that include some notion of (economic) distance (see, e.g., Chen and Conley, 2001). A detailed simulation study of second generation panel unit root tests, including in particular also a discussion of the relative merits and limitations of approaches proposed in the literature, is undertaken in ongoing work.1

In our simulation study we are primarily interested in the following aspects.2 First, we investigate the performance of the tests depending upon

---

1Thus in a sense the present paper can be seen as the first stage in our simulation agenda. Note also that in applications mainly first generation tests are used, which makes a detailed understanding of their performance relevant.

2Our simulation study is based on ARMA(1,1) processes, respectively on AR(1) processes if the MA coefficient is equal to 0, given by (ignoring deterministic components here for brevity) $y_t = \rho y_{t-1} + u_t$ with $u_t = \varepsilon_t + c\varepsilon_{t-1}$, where $\varepsilon_t \sim N(0,1)$, and $\varepsilon_t$ is independent of $\varepsilon_{t'}$ for $i \neq j$. The parameter $c$ is equal to minus the moving average root.
TABLE 1 Summary of asymptotic behavior required for $T$ and $N$ for the derivation of the limiting distributions of the tests

| Asymptotics for the derivation of the test statistics          |
|---------------------------------------------------------------|
| $LL_{93}$          | $N \to \infty$ following $T \to \infty$, $N/T \to 0$ (for cases 2 and 3) |
| $HT$               | $N \to \infty$ and $T$ fixed                                         |
| $UB$               | $N \to \infty$ following $T \to \infty$                             |
| $IPS$              | White noise: $N \to \infty$ and $T$ fixed                            |
|                   | Serial correlation: $N \to \infty$ following $T \to \infty$, $N/T \to k > 0$ |
| $MW$               | $N$, $T$ fixed, approximation of ADF $p$-values for finite $T$        |
| $HLM$              | $N \to \infty$ following $T \to \infty$                             |
| $HT$               | $N \to \infty$ and $T$ fixed                                         |

the time series and cross-sectional dimensions. Since in the derivation of the asymptotic test statistics, differing 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 impact of serial correlation on the performance of the tests. We model serial correlation by simulating ARMA$(1, 1)$ processes and let the moving average roots tend toward 1. It is well known from the time series unit root literature (e.g., Agiakloglou and Newbold, 1996; Schwert, 1989) that unit root tests suffer from severe size distortions for large positive moving average roots. This is as expected, 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 more detailed discussion on this issue 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). Third, we study the performance as a function of the first-order autoregressive coefficient $\rho$. For the power analysis of the panel unit root tests we take $\rho$ 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 (macro)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). 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 $T, N \to 200$). The general impression concerning the size behavior is that the Levin et al. (2002) and Breitung (2000) tests have their size closest to the nominal size. There are, however, exceptions (see the discussion in Section 3.1). Concerning power we observe that for case 2 either the Levin et al. (2002) test or the Breitung (2000) test has 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 (2005).

The paper is organized as follows: Section 2 describes the implemented panel unit root and stationarity tests. Section 3 presents the simulation setup and discusses the simulation results, and Section 4 draws some conclusions. An appendix containing additional figures follows the main text.

### 2. 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 in 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

$$y_{it} = \alpha_i + \beta_t + \rho_i y_{i,t-1} + u_{it}, \quad i = 1, \ldots, N \quad \text{and} \quad t = 1, \ldots, T,$$

A further issue that is studied is the effect of the choice of the autoregressive lag lengths on the test performance, as explained in Section 2, on the performance of the tests.
where \( x_i, \beta_i \in \mathbb{R} \) and \(-1 < \rho_i \leq 1\).\(^4\) The noise processes \( u_{it} \) are stationary ARMA processes, i.e., the stationary solutions to \( a_i(z)u_{it} = b_i(z)e_{it} \), \( a_i(z) = 1 + a_{i,1}z + \cdots + a_{i,\beta_i}z^{\beta_i} \), \( a_{i,\beta_i} \neq 0 \), \( b_i(z) = 1 + b_{i,1}z + \cdots + b_{i,\gamma_i}z^{\gamma_i} \), \( b_{i,\gamma_i} \neq 0 \), \( a_i(z) \neq 0 \) for all \(|z| \leq 1\), \( b_i(z) \neq 0 \) for all \(|z| \leq 1\) and \( a_i(z) \) and \( b_i(z) \) relative prime. The innovation sequences \( e_{it} \) are i.i.d. with variances \( \sigma^2_{\epsilon_t} \) 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. For stationary ARMA processes the long-run variance, \( \sigma^2_{u_{it},LR} \) say, is immediately found to be \( \sigma^2_{\epsilon_t}b_t^2(1)/a_t^2(1) \).\(^5\)

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

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 cases, 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 \( LL.93 \) henceforth. Their results have only been recently published in Levin et al. (2002).\(^6\) The null hypothesis of the \( LL.93 \) test is \( H_0 : \rho_i = 1 \) for \( i = 1, \ldots, N \), against the homogeneous alternative \( H_1^1 : -1 < \rho < 1 \) for \( i = 1, \ldots, N \). Thus under the homogeneous alternative the first-order serial correlation coefficient

\[^4\]In all our simulations we restrict attention to balanced panels, i.e., to 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, \( \theta_t \) say. See the discussion above Section 3.1.

\[^5\]Solving the ARMA equation for the Wold representation \( u_{it} = \epsilon_i(z)e_t = \sum_{j=0}^{\infty} \epsilon_{ij}e_{t-j} \), the (short-run) variance of \( u_{it} \) is given by \( \sigma^2_{u_t} = \sigma^2_{\epsilon_t} \sum_{j=0}^{\infty} \epsilon^2_{ij} \), and the long-run variance is given by \( \sigma^2_{u_{it},LR} = \sigma^2_{\epsilon_t}(\sum_{j=0}^{\infty} \epsilon_{ij})^2 \).

\[^6\]Important foundations have already been laid in Levin and Lin (1992), where panel unit root tests have been developed for homogeneous panels. These are panels where, loosely speaking, the ARMA coefficients are identical for all \( u_{it} \).
\( \rho \) is required to be identical in all units. This restriction stems from the fact that the test statistic is computed in a pooled fashion.

The approach is most easily described as a three-step procedure, with preliminary regressions and normalizations necessitated by cross-sectional heterogeneity.\(^7\) In the first step for each individual series an ADF type regression of the form\(^8\)

\[
\Delta y_{it} = (\rho_i - 1)y_{it-1} + \sum_{j=1}^{p_i} \gamma_{ij} \Delta y_{it-j} + \delta_m d_{it} + v_{it}, \quad m = 1, 2, 3 \tag{2}
\]

is performed, where \( v_{it} \) 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, \( p_i(T) \sim T^{\kappa} \), with \( 0 < \kappa \leq 1/4 \) assumed in the ARMA case. It appears possible that this condition of Levin et al. (2002) may be relaxed, given the results of Chang and Park (2002) brought to our attention by a referee, who derive ADF test asymptotics for \( p = o(T^{1/2}) \). 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 chosen \( p_i \), orthogonalized residuals are obtained from two auxiliary (or partitioned) regressions: \( \tilde{\epsilon}_{it} \) say, from a regression of \( \Delta y_{it} \) on the lagged differences \( \Delta y_{it-j}, j = 1, \ldots, p_i \) and \( d_{it} \), and \( \tilde{f}_{it-1} \) say, from a regression of \( y_{it-1} \) on the same set of regressors. These residuals are standardized by the regression standard error from regressing \( \tilde{\epsilon}_{it} \) on \( \tilde{f}_{it-1}, \hat{\sigma}_{v_{it}} \) say, to obtain the standardized residuals \( \hat{\epsilon}_{it} \) and \( \hat{f}_{it-1} \). This step is necessary to correct for cross-sectionally heterogeneous

\(^7\)Up to the computation of correction factors to account for cross-sectional heterogeneity, the procedure consists essentially of the usual two regressions well known in unit root and cointegration testing. These two regressions are the regressions of both \( y_{it} \) and \( \Delta y_{it} \) on the lagged differences \( \Delta y_{it-j} \) and deterministic components. Then, the residuals of these two regressions are regressed onto each other to compute the first-order serial correlation coefficient, respectively, its \( t \)-statistic. These regressions can be performed in a pooled fashion if the panel is homogeneous. However, in heterogeneous panels, the optimal lag orders will in general differ across units. Furthermore, cross-sectional heterogeneity necessitates additional correction steps described in the text. A referee has pointed out to us that such an approach is known as a partitioned regression in microeconometrics. In the context of the standard regression model the feasibility of this approach is the content of the famous Frisch–Waugh theorem.

\(^8\)Actually, it is recommended by Levin et al. (2002) 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 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. See the discussion above Section 3.1 on cross-sectional dependence.
variances to allow for efficient pooled OLS estimation of \((\rho - 1)\) at a later stage; see (4) below.

The second step is to obtain an estimate of the ratio of the long-run variance to the short-run variance of \(\Delta y_{it}\), or equivalently of \(u_{it}\). This is required for the construction of mean (and variance) correction factors, since the \(t\)-statistic based on (4) diverges under the null hypothesis for cases 2 and 3. Therefore to obtain a nondegenerate limiting distribution, correction factors are required. The definition of the long-run variance, \(\sigma^2_{uLR} = \sigma^2_{ui} + 2 \sum_{j=1}^{\infty} \mathbb{E}(u_{it}u_{i,t-j})\), immediately leads to an estimator of the form

\[
\hat{\sigma}^2_{u_{i},LR} = \frac{1}{T} \sum_{i=1}^{T} \hat{u}_{it}^2 + \frac{2}{T} \sum_{j=1}^{L} w(j, L) \sum_{i=j+1}^{T} \hat{u}_{it} \hat{u}_{i,t-j},
\]

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 \(\hat{u}_{it} = \Delta y_{it} - \hat{\phi}_{i} d_{it}\). In our simulations the weights are given by \(w(j, L) = 1 - \frac{j}{L+1}\). This kernel is known as the Bartlett kernel. The estimated individual specific ratio of long-run to short-run variance is defined as \(\hat{s}_{i}^2 = \hat{\sigma}^2_{u_{i},LR} / \hat{\sigma}^2_{ui}\), with \(\hat{\sigma}^2_{ui} = \frac{1}{T} \sum_{i=1}^{T} \hat{u}_{it}^2\). Denote by \(\hat{S}_{NT} = 1/N \sum_{i=1}^{N} \hat{s}_{i}\), which 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 \(i = 1, \ldots, N\).

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

\[
\hat{\phi} = \frac{\sum_{i=1}^{N} \sum_{t=p+2}^{T} \hat{e}_{it} \hat{f}_{i,t-1}}{\sum_{i=1}^{N} \sum_{t=p+2}^{T} \hat{f}_{i,t-1}^2}.
\]

The null hypothesis is \(H_0: \phi = 0\), and the test we use in the simulations is based on the corresponding \(t\)-statistic, \(t_{\phi}\) say. The standard deviation of \(\hat{\phi}\) as given in (4), \(STD(\hat{\phi})\) say, can be straightforwardly computed from the pooled regression residuals, since due to the prefiltering all the errors in this pooled regression have the same (asymptotic) variance.

For case 1, Levin and Lin (1993) show that \(t_{\phi} \Rightarrow N(0, 1)\). For cases 2 and 3, the \(t\)-statistic \(t_{\phi}\) diverges to minus infinity and thus has to be recentered and normalized to induce convergence toward a well-defined

\footnote{Note that a direct estimate for the long-run variance is given by \(\hat{\sigma}^2_{ui}(1 - \sum_{j=1}^{h} \hat{\gamma}_{j})^{-2}\). Levin et al. (2002) indicate that variance estimation based on the first differences is found to have a 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.}
limiting distribution

\[ t_\phi^* = \frac{t_\phi - N \hat{T} \hat{S}_{NT} STD(\hat{\phi}) \mu_{mT}}{\sigma_{mT}}. \]  

(5)

Here \( \mu_{mT} \) and \( \sigma_{mT} \) denote mean and variance correction factors, tabulated for various panel dimensions in Table 2 on page 14 of Levin et al. (2002). \( \hat{T} \) denotes the average effective sample size across the individual units. The adjusted \( t \)-statistics \( t_\phi^* \) converge to the standard normal distribution for cases 2 and 3.

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

\[ \sqrt{N}(\hat{\rho} - 1 - B_m) \Rightarrow N(0, C_m), \]

(6)

with \( B_1 = 0 \), \( C_1 = 2/(T-1) \), \( B_2 = -3/(T+1) \), \( C_2 = 3(17T^2 - 20T + 17)/5(T-1)(T+1)^3 \), 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 using correction factors based on asymptotics in both \( T \) and \( N \) 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, which is achieved by appropriate (depending upon case considered) variable transformations. Due to its pooled construction, also the Breitung test, \( UB \) henceforth, is a test against the homogeneous alternative.

In case 1, this test coincides exactly with the Levin et al. (2002) test, since in this case no bias corrections are required. For case 2, bias correction factors are avoided by subtracting the initial observation. Subtracting the initial observation instead of the mean circumvents the Nickell bias.
Thus case 2 is equal to case 1 of LL93 on the transformed variables \( \tilde{y}_{it} = y_{it} - y_{i0} \). In both cases the asymptotic distribution of the test statistic is standard normal without the need of resorting to correction factors.

For case 3, slightly more complicated transformations have to be applied, after serial correlation has been removed with first step regressions. There are two ways of removing the serial correlation: the first is resorting to preliminary regressions as in the description of the Levin et al. (2002) test, and the second, suggested by Breitung and Das (2005) to have better small sample performance, is prewhitening. Prewhitening involves in the first step the regressions (for each \( i \))

\[
\Delta y_{it} = z_i + \sum_{j=1}^{p_i} \gamma_j \Delta y_{it-j} + v_{it}, \tag{7}
\]

from which the prewhitened variables \( \tilde{e}_{it} \) and \( \tilde{f}_{it-1} \) are computed as

\[
\tilde{e}_{it} = \Delta y_{it} - \sum_{j=1}^{p_i} \hat{\gamma}_j \Delta y_{it-j}, \tag{8}
\]

\[
\tilde{f}_{it-1} = y_{it-1} - \sum_{j=1}^{p_i} \hat{\gamma}_j y_{it-j-1}. \tag{9}
\]

Note that at this step no correction for the mean or trend is performed. The prewhitened variables are next standardized by the regression standard error of (7) to obtain \( \hat{e}_{it} \) and \( \hat{f}_{it-1} \). Here we use for simplicity the same notation as for the residuals obtained via auxiliary regressions. We do this for notational simplicity and also because the two approaches are asymptotically equivalent. Finally \( \hat{e}_{it} \) and \( \hat{f}_{it-1} \) are transformed as

\[
e^*_it = \sqrt{\frac{T - t}{T - t + 1}} \left( \hat{e}_{it} - \frac{1}{T - t} (\hat{e}_{it+1} + \cdots + \hat{e}_{iT}) \right), \tag{10}
\]

\[
f^*_it-1 = (\hat{f}_{it-1} - \hat{f}_{i1}) - \frac{t - 1}{T} (\hat{f}_{iT} - \hat{f}_{i1}). \tag{11}
\]

The above transformations demean \( e^*_it \) and demean and detrend \( f^*_it-1 \). 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

\[
e^*_it = \phi^* f^*_it-1 + v^*_it \tag{12}
\]

Prewhitening is based on the idea of deriving an estimator of the nuisance parameters under the null hypothesis. As has been pointed out by a referee it is equivalently possible to perform the correction for the short-run dynamics in a similar way as for the Levin et al. (2002) test. However, from personal communication with Jörg Breitung, we have learned that prewhitening results in better small sample properties in his simulation experiments.
by testing the hypothesis $H_0 : \phi^* = 0$. Breitung (2000) shows that the $t$-statistic of this test has a standard normal limiting distribution.

We now turn to panel unit root tests that are designed to test against the heterogeneous alternative $H^2_0$: $-1 < \rho_i < 1$ for $i = 1, \ldots, N_i$ and $\rho_i = 1$ for $i = N_i + 1, \ldots, N$. For asymptotic consistency (in $N$) of these tests, a nonvanishing fraction of the individual units has to be stationary under the alternative, i.e., $\lim_{N \to \infty} N_i / N > 0$. The tests are based on group-mean computations, i.e., on appropriately combined individual time series unit root tests.

**Im, Pesaran, and Shin.** In two papers, Im et al. (1997, 2003), henceforth abbreviated as IPS, present two group-mean panel unit root tests designed against the heterogeneous alternative. IPS consider only cases 2 and 3 and allow for individual specific autoregressive structures and individual specific variances. The 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. Im et al. 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 Im et al., 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 $(IP_{SM})$ and a Lagrange multiplier $(LM)$ test $(IP_{ML})$.

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:

$$t_{iT,m}(p_i, \gamma) = \frac{\sqrt{T - p_i - m(y'_{i,T-1}M_Q \Delta y_i)}}{(y'_{i,T-1}M_Q, y_{i,T-1})^{1/2} (\Delta y_i'M_M \Delta y_i)^{1/2}}, \quad m = 2, 3,$$

(13)

using the notation $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{ip_i})'$, $y_{i,T-1} = [y_{i0}, \ldots, y_{iT-1}]'$, $\Delta y_{i,T-1} = [\Delta y_{i1}, \ldots, \Delta y_{iT-1}]'$, $s = 0, \ldots, p_i$, $\Delta y_i = \Delta y_i - 0$, $d_{2T} = [1, \ldots, 1]'$, $t = [1, \ldots, T]'$, $d_{3T} = [d_{2T}, t]$, $Q_i = [d_{3T}, \Delta y_i, \ldots, \Delta y_{i,T-1}]$, $M_Q = I_T - Q_i(Q_i'Q_i)^{-1}Q_i$, $X_i = [y_{i1}, \ldots, y_{iT}]$, and $M_{X_i} = I_T - X_i(X_iX_i)^{-1}X_i$ (suppressing the index $m$ in the matrix notation for $Q_i$ and $X_i$). For finite values of $T$, the statistics $t_{iT,m}$ depend upon the nuisance parameters $\gamma_i$. IPS show that this dependence vanishes for $T \to \infty$, but that the bias of the individual $t$-statistics under the null hypothesis remains. This follows from the fact that under the null hypothesis convergence to the Dickey–Fuller distribution corresponding to the DGP and model prevails. Therefore mean and
variance correction factors have to be introduced. The proposed test statistic itself is the cross-sectional average of the corrected \( t \)-statistics:

\[
IPS_{t,m}(\mathbf{p}, \gamma) = \frac{\sqrt{N} \left\{ \tilde{t}_m - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(t_{iT,m}(p_i, 0) \mid \rho_i = 1) \right\}}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \text{Var}(t_{iT,m}(p_i, 0) \mid \rho_i = 1)}} \Rightarrow N(0, 1), \tag{14}
\]

where \( \tilde{t}_m = \frac{1}{N} \sum_{i=1}^{N} t_{iT,m}(p_i, \gamma_i) \), \( \mathbf{p} = [p_1, \ldots, p_N]' \), and \( \gamma = [\gamma_1', \ldots, \gamma_N']' \). The correction factors \( \mathbb{E}(t_{iT,m}(p_i, 0) \mid \rho_i = 1) \) and \( \text{Var}(t_{iT,m}(p_i, 0) \mid \rho_i = 1) \) are simulated for \( m = 2,3 \) for a set of values for \( T \) and lag lengths \( p \) (see Table 3 in Im et al., 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 length 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.\(^{11} \) Note that for \( T \to \infty \) 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 using simulated critical values one can also refer to the asymptotic values for \( T \to \infty \) (which has the additional advantage of allowing the use of cross-section-specific lag lengths \( p_i \)).

Let us now turn to the Lagrange multiplier test. Using the Lagrange multiplier test principle implies that the alternative hypothesis 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 et al., 1997, Remark 3.2). For each individual unit the test statistic is given by

\[
LM_{t,m}(p_i, \gamma_i) = T \frac{(\Delta y_i'y_i'y_i')^{-1}(y_i'y_i'M_Q'y_i')}{(y_i'y_i'M_Q'y_i')} \tag{15}
\]

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

\[
IPS_{LM,m}(\mathbf{p}, \gamma) = \frac{\sqrt{N} \left\{ LM_m - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(LM_{t,m}(p_i, 0) \mid \rho_i = 1) \right\}}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \text{Var}(LM_{t,m}(p_i, 0) \mid \rho_i = 1)}} \Rightarrow N(0, 1), \tag{16}
\]

\(^{11} \)Simulation of these values for different values of \( T \) and \( p \) proceeds by generating \( \Delta y_t = e_t \), with \( e_t \) i.i.d. \( N(0,1) \) for \( t = 1, \ldots, T \), and computing the \( t \)-statistic for \( p = 1 \) in the ADF regression (2) for \( j = p \) where \( d_{au} \) is included.
where $\overline{LM}_m = \frac{1}{N} \sum_{i=1}^{N} LM_{it,m}$. The correction factors are available in Im et al. (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). Note that Choi (2001) presents very similar tests that only differ in the scaling in order to obtain asymptotic normality for $N \to \infty$. The basic idea of Fisher can be explained with the following simple observations that hold for any testing problem with continuous test statistics: First, under the null hypothesis the $p$-values, $\pi$ say, of the test statistic are uniformly distributed on the interval $[0, 1]$. Second, $-2 \log \pi$ is therefore distributed as $\chi^2_2$, with log denoting the natural logarithm. Third, for a set of independent test statistics, $-2 \sum_{i=1}^{N} \log \pi_i$ is consequently distributed as $\chi^2_{2N}$ 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 owing to the extensive simulation work of James MacKinnon and his coauthors (see, for one example, MacKinnon, 1994). Note as a further advantage that the Fisher test requires neither 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 labeled 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. (1992) test, labeled $H_{LM}$ 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 hypothesis 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 $\sigma^2_i$.

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

$$y_{it} = \delta_{mt} d_{mt} + \varepsilon_{it}, \quad m = 2, 3$$  \hspace{1cm} (17)
for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Denote the OLS residuals of equation (17) by $\hat{e}_i$ and their partial sum by $S_d = \sum_{t=1}^{T} \hat{e}_i$. The test statistic is then given by ($m$ indexing again the case investigated)

$$H_{LM,m} = \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{S_d^2}{\hat{\sigma}_n^2},$$

(18)

with $\hat{\sigma}_n^2 = 1/T \sum_{t=1}^{T} \hat{e}_i^2$.

Recentering and rescaling the expressions by subtracting their mean and dividing by their standard deviation gives rise to asymptotic standard normality

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

(19)

Owing to the simple shape of the correction terms, closed form solutions for the correction factors can be easily obtained. They are given by $\xi_2 = 1/6$, $\xi_2 = \sqrt{1/45}$, and $\xi_3 = 1/15$, $\xi_3 = \sqrt{11/6300}$. The extension to serially correlated errors is straightforward; the variance estimator $\hat{\sigma}_n^2$ only has to be replaced by an estimator of the long-run variance of the noise processes in (17).

**Hadri and Larsson.** Hadri and Larsson (2005) 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. (1992) 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 $\eta_{1Tm} = 1/T^2 \sum_{t=1}^{T} S_d^2 / \hat{\sigma}_n^2$, which is the core expression of the Hadri type test statistics, compare (18). Standard asymptotic theory for $N$ then delivers asymptotic normality

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

(20)

with $\mathbb{E}\eta_{1T2} = (T + 1)/6T$, $\text{Var}\eta_{1T2} = (T^2 + 1)/20T^2 - ((T + 1)/6T)^2$, and $\mathbb{E}\eta_{1T3} = (T + 2)/15T$, $\text{Var}\eta_{1T3} = (T + 2)(13T + 23)/2100T^3 - (T + 2)/15T^2$. 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 computing the individual specific long-run variances as discussed several times in this section. However, since the long-run variance generally has to be estimated,
the corresponding test will not have exactly the same distribution as in the case of serially uncorrelated errors. In other words, result (20) does not hold exactly in case of serially correlated errors for finite $T$ if a long-run variance estimator is used. The resultant distortions in the test distribution depend upon the unknown long-run variances and thus cannot be quantified in applications. This implies that the usefulness of the Hadri and Larsson (2005) finite $T$ test for serially correlated errors is hard to assess.

The tests discussed in this section are based on different limit arguments. The most widely used concept is that of a sequential limit where first $T \to \infty$ followed by $N \to \infty$, employed in the tests of Levin et al. (2002), Breitung (2000), Im et al. (1997, 2003), and Hadri (2000). Some of the tests require furthermore a relative rate restriction, e.g., $N/T \to 0$ (Levin et al., 2002) or $N/T \to k$ (Im et al., 1997, 2003). As has been seen above, inference for fixed $T$ and only $N \to \infty$ is only developed for the case of serially uncorrelated errors. This is done in Harris and Tzavalis (1999) for the Levin et al. tests and in Hadri and Larsson (2005) for the Hadri tests. Thus the performance of such tests will depend upon the magnitude of both $T$ and $N$ and may also depend upon the relative magnitude of the time and cross-section dimension of the panel. This is one of the issues to be analyzed with the simulation study.

The tests following the Fisher principle developed in Maddala and Wu (1999) are the only ones derived for fixed $N$ and $T$. However, the critical values for the ADF tests have to be approximated for finite $T$. We summarize the asymptotics used in the derivation of the test statistics in Table 1. A detailed discussion of the relevant limit concepts for nonstationary panels and the relations among the different limit concepts is contained in Phillips and Moon (1999).

### 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 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 with a substantially extended, corrected, and modified set of routines based originally on Chiang and Kao (2002). A list containing the major changes is available upon request. 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 \( N \); compare again Table 1. One question in this respect is whether the finite-\( T \) tests of Harris and Tzavalis (1999) and Hadri and Larsson (2005) exhibit smaller size distortions than their asymptotic-\( T \) counterparts for panels with \( T \) small (compared to \( N \)).

The DGPs simulated for case 2 are of the form

\[
y_{it} = \alpha_i (1 - \rho) + \rho y_{i,t-1} + u_{it} \\
u_{it} = \varepsilon_{it} + c \varepsilon_{i,t-1} 
\]

with \( \varepsilon_{it} \sim N(0, 1) \). The parameters chosen in the simulations are \( \alpha = [\alpha_1, \ldots, \alpha_N] \), \( \rho \), 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 for computational efficiency. Otherwise, the coefficients \( \alpha_i \) 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

\[
y_{it} = \alpha_i + \alpha_i (1 - \rho) t + \rho y_{i,t-1} + u_{it} \\
u_{it} = \varepsilon_{it} + c \varepsilon_{i,t-1} 
\]

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 \( \alpha_i \) are, as for case 2, \( U[0, 4] \) distributed.

For the unit root tests the following values are chosen for \( \rho \): 0.7, 0.8, 0.9, 0.95, 0.99, and 1.\textsuperscript{12} 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?

\textsuperscript{12}Preliminary simulations have shown that for values of \( \rho \) up to 0.6 all tests exhibit satisfactory power behavior already for medium sized panels. We thus focus here only on those cases for \( \rho \), where differential results across tests can be observed widely across the simulation experiments. The case \( \rho = 0 \) is included as benchmark case. Also the case \( \rho = 0 \) is the single case for which the tests designed for serially uncorrelated errors can be applied with all assumptions fulfilled.
approaching $-1$? It is well known from the time series unit root literature, compare Schwert (1989) or Agiakloglou and Newbold (1996), 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.\footnote{It is straightforward to show that the asymptotic bias for $T \to \infty$ of $\hat{p}$, 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. Note that in case that $c = -1$, Equations (21) and (22) are unidentified ARMA systems that allow for stationary solutions. The lack of identifiability for $c = -1$ stems from the fact that the autoregressive and the moving average polynomial are not left coprime or in other words contain a common factor.} These observations are rooted in the asymptotic theory of autoregressive approximations for rather general process classes (for the multivariate case see Bauer and Wagner, 2005). Such results show that the approximation quality of autoregressive approximations depends (in case of ARMA processes) upon the MA root closest to the unit circle in absolute values. Therefore with ARMA(1,1) processes we can control directly the relevant dimension of the approximation quality of autoregressive models and at the same time allow for higher order serial correlation with only one parameter. This is the main reason for choosing ARMA processes and not AR processes with higher lag lengths as DGPs. Concerning the question of the relevance of ARMA processes for econometric modeling, we want only briefly to mention two motivations here. First, Zellner and Palm (1974) show that any subset of variables of a vector autoregressive process generally follows a (vector) ARMA process. Since (panel) unit root tests are often used as an individual variable preliminary step for (panel) vector autoregressive modeling, this shows that the robustness of the performance with respect to ARMA processes is very important. Second, more structurally, Campbell (1994) shows within the real business cycle paradigm that the exactly linearized solution processes to dynamic stochastic general equilibrium models are typically ARMA processes.

With our setup we can analyze the extent of 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 $c \neq 0$, 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).\textsuperscript{14} Note that we choose the value of $c$ identical for all cross-section members. We do this to study “cleanly” the effect of the moving average coefficient approaching $-1$, which is harder to assess when the MA coefficients are drawn randomly for the cross-section units.

The careful reader will have observed that our simulated DGPs all have a cross-sectionally identical coefficient $\rho$ under both the null and the alternative. Thus we are in effect in a situation where we generate data either under the null hypothesis or under the \textit{homogeneous} alternative. We do this because only the more restrictive homogeneous 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 \textit{heterogeneous} alternative hypothesis (IPS and MW) possess. Thus, to a certain extent, the pooled tests are favored in our comparison, since the last step regression to estimate $\rho$ 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.\textsuperscript{15}

As indicated in the introduction we also simulate DGPs that allow for cross-sectional correlation. Denote with $\Omega = [\omega_{ij}]_{i,j=1,\ldots,N}$ the covariance matrix of $\epsilon_i$. Then we allow for two different forms of covariance, labeled constant covariance ($\Omega_{CC}$) and a covariance matrix in Toeplitz form ($\Omega_{TP}$):

$$
\Omega_{CC} = \begin{bmatrix}
1 & \omega & \cdots & \omega \\
\omega & 1 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \omega \\
\omega & \cdots & \omega & 1
\end{bmatrix},
\Omega_{TP} = \begin{bmatrix}
1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\
\omega & 1 & \omega & \cdots & \vdots \\
\omega^2 & \omega & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\omega^{N-1} & \cdots & \cdots & 1 & \omega
\end{bmatrix}
$$

Note that in our simulation setup, owing to the unit variances of $\epsilon_i$, these coincide with the correlation matrices. The first of the two covariance matrices has, e.g., been used in O’Connell (1998), and the second

\textsuperscript{14}Some (unreported) simulation results also suggest that choosing lag lengths by starting with a large initial number of lags and sequential truncation of one lag until the coefficient to the largest lag is significant can improve performance sometimes. However, we do not have systematic evidence on the effects of lag length selection on the test performance (for different sample sizes).

\textsuperscript{15}Karlsson and Löthgren (2000) present some simulation results in this respect.
corresponds to a spatial autoregression of order 1 (interpreting the cross-section dimension spatially). In our simulations we take the (correlation) coefficient $\omega$ in the set $\{0.3, 0.6, 0.9\}$, where of course $\omega = 0$ is the cross-sectionally uncorrelated case. The major insight we have obtained from these additional simulations with cross-sectional correlation is that the performance rankings of the tests (for both size and power) are essentially unchanged compared to the cross-sectionally uncorrelated case. For $\Omega_{cc}$ this is not so surprising, since by applying cross-sectional demeaning to such a process asymptotically for $N \to \infty$ removes the cross-sectional correlation. To be precise, the covariance matrix of the cross-sectionally demeaned innovations, $\tilde{\varepsilon}_{it} = \varepsilon_{it} - 1/N \sum_{i=1}^{N} \varepsilon_{it}$, is given by

$$\tilde{\Omega}_{cc} = \frac{1 - \omega}{N} \begin{bmatrix} N - 1 & -1 & \cdots & -1 \\ -1 & N - 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \cdots & -1 & N - 1 \end{bmatrix}$$

Thus, in the case of constant covariance, cross-sectional demeaning decreases the cross-sectional covariance to $(\omega - 1)/N$. This explains why for such processes cross-sectional demeaning leads to comparable results as in the cross-sectionally uncorrelated case. However, even when abstaining from cross-sectional demeaning the rankings are very robust. Some authors, e.g., Levin et al. (2002), suggest to cross-sectionally demean as a first step in any case (which as mentioned above also removes for $N \to \infty$ time specific aggregate effects $\theta_t$). From this perspective therefore this case of cross-sectional covariance is seen to be not a great problem.

The second studied case is a bit more complicated, since cross-sectional demeaning does not lead to monotonic reductions of the correlations, not even when $N \to \infty$. Therefore for the Toeplitz case the results without cross-sectional demeaning may be more relevant. The finding—in its extent surprising—is that the orderings across tests are extremely robust with respect to this form of cross-sectional dependence. Of course, all tests become increasingly distorted with increasing $\omega$. This holds both for size and power and also for panel unit root and stationarity tests; see Figure 8 for panel unit root tests and Figure 9 for panel stationarity tests in the Appendix. These findings (with more detailed results upon request) lead us to report here only the results for cross-sectionally independent panels. We only want to stress again that for the cross-sectional correlations investigated, the differences to the results obtained for cross-sectionally uncorrelated panels remain small for $\omega$ up to 0.6 and the rankings across tests remain almost unchanged throughout.
**3.1. Size of Panel Unit Root Tests**

In this subsection we report the results of the analysis of the actual size of the panel unit root tests. In this study we use the word size to denote the type I error rate at the actual DGP. This is 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. 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 (\( LL_{93}, UB, IPS_t, IPS_{LM} \), and \( MW \)) are computed for all values of \( c \).

We start with case 2 in Figures 1 and 2 and display results for case 3 in Figures 3 and 4. For these and all other figures, it is always the cross-sectional dimension \( N \) that varies along the horizontal axis.\(^{16}\)

Figure 1 displays for \( c = 0 \) a comparison of the size of the \( LL_{93} \) test and the \( HT_2 \) test, which is—as has been discussed—a fixed \( T \) version of the \( LL_{93} \) 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 \( LL_{93} \) test are based on sequential limit theory with \( N \to \infty \) following \( T \to \infty \) and furthermore with \( \lim N/T \to 0 \); see Table 1. For larger \( T \), the improved performance

\(^{16}\)Please note that the vertical axis is not scaled identically across the subplots of the figures. This stems from the fact that for all the experiments we display, identical vertical scaling would lead to closely bundled lines in some of the figures.
of the ADF-type test performed in 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 monotonically decreasing toward 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 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 et al. (1997, 2003) and the Breitung (2000) test exhibit satisfactory size behavior. In particular, for these three tests, 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 $LL_{932}$ 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 et al. (1997, 2003) tests and the Breitung (2000) test have comparably acceptable size. The increase is slower for these tests than for the Levin et al. (2002) and the Maddala and Wu (1999) tests. Especially for $T$ small relative to $N$, an application of the Harris and Tzavalis (1999) test offers an improvement over Levin et al. (2002).

For panels with increasingly negative serial correlation, i.e., with $c \to -0.99$, 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 (among 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 surprisingly, 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 $c \in \{-0.8, -0.9, -0.95\}$, “size divergence” occurs again for $N \geq 100$.17 This divergence can be partly mitigated by using smaller values for the autoregressive lags than suggested by BIC.18 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 $N$ and uses a finite $T$ approximation of the $p$-values for the individual ADF tests. For serially uncorrelated errors, furthermore, Im et al. (1997) provide critical values for the tests for finite $T$ and only $N \to \infty$. 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 $N \to 200$ 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.

The tests that exhibit in most cases the slowest divergence of size as $c$ is decreased towards $-0.95$ are the $LL_{932}$ and (usually second slowest) the
The behavior is the large $T$ extension of the behavior observed for the Levin et al. (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 $N \to \infty$. 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 grossly the least distorted test, with in general a slight tendency for being undersized in small $T$ and large $N$ panels.

We now turn to case 3 and start again with a comparison of the Levin et al. (2002) and Harris and Tzavalis (1999) tests for $c = 0$. In Figure 3 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 $u_t$, 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 $N \to 200$ and $T = 25$. The two IPS tests have very similar performance. Thus, in case of no serial correlation, size divergence for $N \to 200$ occurs only for the smallest values of $T$. The relative sample sizes are therefore not of great

**FIGURE 3** Size of the Levin et al. (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.
concern as soon as \( T \geq 25 \), and even for shorter panels three tests \((UB_3, IPS_{1,3}, \text{and} IPS_{LM,3})\) show satisfactory size performance.

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 \( T, N \to 200 \). Two tests have substantially smaller size distortions (over a variety of combinations of \( T \) and \( N \)) than the other tests. These are the Levin et al. (2002) and Breitung (2000) tests. 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 \( N \to 200 \) and \( T \geq 50 \). Also for \( c = -0.4 \) the \( LL_{933} \) test is not subject to size divergence. The size divergence behavior of the \( IPS_{LM,3}, IPS_{1,3}, \text{and} MW_3 \) tests is very similar. Thus for the case of random walks with drifts, the summary is that \( LL_{933} \) and \( UB_3 \) outperform the other tests. The major exception to this general rule occurs for \( T \in \{10, 15\} \), where the \( IPS_{LM,3} \) test shows good size properties and the \( LL_{933} \) test does not yet appear so favorable. This is to a certain extent surprising, given that the \( LL_{933} \) test is a pooled test and the \( IPS_{LM,3} \) test is a group-mean test. The observation concerning the relative performance of the \( IPS_{LM,3} \) and the \( LL_{933} \) 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 essentially a direct consequence of the construction of the tests; cf., Table 1.
3.2. Power of 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. 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 labeled type I error corrections for given DGP—do not necessarily lead to insights that can be generalized.

Consequently 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 from inference based on asymptotic critical values. A precise theoretical analysis concerning the validity of bootstrap inference in nonstationary panels is yet to be provided.

Before we discuss the results, let us start by summarizing a few general observations. First, which may be not surprising, power is monotonically increasing in \( N \) for all DGPs simulated under the stationary alternative hypothesis for all values of \( T \) (see, for example, Figure 5). Note, however, that power does not increase monotonically in \( T \) for given \( N \). This occurs for relatively small values of \( T \) and \( N \), when \( \rho \) 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 nonmonotonicity 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 \( c = -0.4 \). The figure clearly displays one representative result, 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 that has highest power, whereas the \( LL93_2 \) test has highest power in many cases for smaller values of \( T \). Corresponding to

---

19This occurs unless the test statistic is pivotal, which is not the case in finite samples for any of the unit root or stationarity tests discussed in this paper, as is illustrated, e.g., by the large differences in size for different parameters \( c \).
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 et al. (2002) test. The most variable power performance is observed for the \( MW_2 \) test, 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 et al., 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 when \( T \geq 100 \). The previous observations hold both for \( c = 0 \) and \( c \neq 0 \).

For case 3, the results are less clear than for case 2. There are panel dimensions \((T, N)\) and parameter values \((\rho, c)\) 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

FIGURE 5 Power of panel unit root tests for case 2 with \( \rho = 0.8 \) (\( DGP_2(x, 0.8, c) \)) for \( c \in \{0, -0.4\} \) and \( T \in \{10, 25, 100\} \). The solid lines with bullets correspond to \( LL93_2 \), the solid lines with triangles correspond to \( UB_2 \), the solid lines correspond to \( IPS_{t,2} \), the dash-dotted lines correspond to \( IPS_{LM,2} \), and the dashed lines correspond to \( MW_2 \).
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. The \( UB_3 \) test performs relatively well, but not as well 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 \( \rho \) up to 0.9 for \( T \geq 100 \). Detailed graphical results of the power of the panel unit root tests for case 3 are available from the authors upon request.

### 3.3. Size of Panel Stationarity Tests

Some representative results for the size behavior of the stationarity tests of Hadri (2000) and Hadri and Larsson (2005) 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 (2005) is based on finite \( T \) inference.

![Figure 6](image_url)

**FIGURE 6** Size of Hadri (2000) and Hadri and Larsson (2005) 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 \( H_{LM.2} \) 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.
One of the aspects we want to compare is the relative performance of the \( H_{LM} \) test and the \( H_T \) test. Focusing on this aspect first, we find that only for the case \( \rho = 0 \) and \( c = 0 \) do substantial differences between these two tests occur. This holds not only for \( T = 25 \) shown but for all values of \( T \). As expected, for larger values of \( T \), the differences become smaller. The explanation for this result is that the nonparametric estimation to correct for serial correlation is too imprecise to result in 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 is the size of the tests acceptable.\(^{20}\) This is 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 à 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 (2005) test can be “used to find unit roots” (although, of course, strictly speaking a rejection of the null hypothesis does not imply acceptance of the alternative hypothesis).

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

### 3.4. Power of Panel Stationarity Tests

We finally briefly discuss the power of the panel stationarity tests. The size results (rejection of stationarity for many cases) already allow 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 that show exactly what has just been discussed. Note that similar results are available for case 3 upon request. Summing up: The high power stems from the fact that the Hadri (2000) and Hadri

\(^{20}\)The difference between 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 nonparametric correction. For \( c \) close to \(-1\), as expected, large size distortions occur throughout.
FIGURE 7 Power of the Hadri (2000) test (upper row) and of the Hadri and Larsson (2005) test (lower row) for case 2, \((DGP_2(0, 1, c))\) against the alternative of a unit root for \(c \in \{-0.2, -0.4, -0.8, -0.9, -0.95, 0\}\) 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.

and Larsson (2005) tests tend to reject stationarity most of the time 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 \(T \in \{10, 15\}\).

4. CONCLUSIONS

The strongest and most unequivocal conclusion from our simulations is that the panel stationarity tests of Hadri (2000) and Hadri and Larsson (2005) perform very poorly. This is to a certain extent similar to the often observed poor performance of the Kwiatkowski et al. (1992) test, which is the time series building block of these tests. The null hypothesis 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 only a few clear-cut patterns emerge (which is itself an interesting observation). Some of the main findings are: First, for case 2 (with intercepts under stationarity) the best power behavior is displayed by either the Levin et al. (2002) test or by the Breitung (2000) test. Second, for serially uncorrelated panels the Harris and Tzavalis (1999) implementation of the Levin et al. (2002) 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 $T \to \infty$ followed by $N \to \infty$. 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, the size distortions become larger when the moving average coefficient $c \to -1$. 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 et al. (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 et al. (2002) has very rapid size divergence. Combined with the good power performance (notably for case 2), these two tests appear *grosso modo* quite favorable. All our results generalize to the simulation experiments performed with cross-sectionally dependent panels. The extent of robustness of the performance ranking across tests to the two discussed cross-sectional covariance structures is remarkable.

At this point, however, we have to note again that the *group-mean* tests of Im et al. (1997, 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) homogeneous 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 heterogeneous panels will allow us 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 “nonmonotonic” in $c$, is an open issue for future research. By nonmonotonicity 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 consistent bootstrap inference. However, it is well known from the time series literature that bootstrap consistency is a delicate issue in unit root situations. Similar problems will arise in panels, in particular also in panels with cross-sectional dependencies. This is probably one of the most important problems to be solved in the panel unit root test literature for practical purposes.

**APPENDIX: ADDITIONAL FIGURES**

![Figure 8](image_url)

**FIGURE 8** Size (upper panel) and power (lower panel) of panel unit root tests with Toeplitz cross-sectionally correlated DGP's for $T = 100$. In the upper panel the errors are serially correlated ($c = -0.4$), in the lower panel $\rho = 0.95$ and the errors are uncorrelated. The solid lines with bullets correspond to $LL932$, the solid lines with triangles correspond to $UB2$, the solid lines correspond to $IPSt2$, the dash-dotted lines correspond to $IPSLM2$, and the dashed lines correspond to $MW2$. The left figures display the results for $\omega = 0$, i.e., for the cross-sectionally uncorrelated case, the figures in the center display the results for $\omega = 0.6$, and the right figures display the results for $\omega = 0.9$. 
FIGURE 9 Size of Hadri (2000) and Hadri and Larsson (2005) stationarity tests for case 2 with Toeplitz cross-sectional correlation. The errors are serially correlated with $\epsilon = -0.4$ and $T = 25$ and $\rho \in [0, 0.5]$. The results for $\omega = 0$ are displayed with solid lines, for $\omega = 0.3$ with dashed lines with stars, for $\omega = 0.6$ with dash-dotted lines, and for $\omega = 0.9$ with dashed lines.

ACKNOWLEDGMENTS

Financial support from the Jubiläumsfonds of the Oesterreichische Nationalbank under grant Nr. 9557 is gratefully acknowledged. We furthermore thank Jörg Breitung, Robert Kunst, Klaus Neusser, Chiara Osbat, Peter Pedroni, and especially two anonymous referees for very helpful comments. The usual disclaimer applies.

Part of this work has been done while the second author visited Princeton University and the European University Institute. The hospitality of these institutions is gratefully acknowledged.

REFERENCES

Agiakloglou, C., Newbold, P. (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., Ng, S. (2004). A PANIC attack on unit roots and cointegration. *Econometrica* 72:1127–1178.

Bauer, D., Wagner, M. (2005). Autoregressive approximations of multiple frequency I(1) processes. Institute for Advanced Studies: Vienna, Working Paper: Economics Series No. 174.

Breitung, J. (2000). The local power of some unit root tests for panel data. In: Baltagi, B. H., ed. *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*. Amsterdam: Elsevier, pp. 161–177.

Breitung, J., Das, S. (2005). Panel unit root tests under cross sectional dependence. *Statistica Neerlandica* 59:1–20.

Campbell, J. Y. (1994). Inspecting the mechanism: an analytical approach to the stochastic growth model. *Journal of Monetary Economics* 33:463–506.
Chang, Y. (2002). Nonlinear IV unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* 110:261–292.

Chang, Y., Park, J. Y. (2002). On the asymptotics of ADF tests for unit roots. *Econometric Reviews* 21:431–447.

Chen, X., Conley, T. G. (2001). A new semiparametric spatial model for panel time series. *Journal of Econometrics* 105:59–83.

Chiang, M.-H., Kao, C. (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., Durlauf, S., Hansen, B., 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*. Edinburgh: Oliver and Boyd.

Gregoir, S. (2004). Representation and estimation of weakly linearly exchangeable dynamic panels. Mimeo.

Hadri, K. (2000). Testing for stationarity in heterogeneous panel data. *Econometrics Journal* 3:148–161.

Hadri, K., Larsson, R. (2005). Testing for stationarity in heterogeneous panel data where the time dimension is fixed. *Economics Letters* 8:55–69.

Harris, R. D. F., Tzavalis, E. (1999). Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics* 90:1–44.

Horowitz, J. L., Savin, E. N. (2000). Empirically relevant critical values for hypothesis tests: a bootstrap approach. *Journal of Econometrics* 95:375–389.

Im, K. S., Pesaran, M. H., Shin, Y. (1997). *Testing for Unit Roots in Heterogeneous Panels*. Mimeo: Cambridge University.

Im, K. S., Pesaran, M. H., Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics* 115:53–74.

Karlsson, S., Löthgren, M. (2000). On the power and interpretation of panel unit root tests. *Economics Letters* 66:249–255.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* 54:91–115.

Levin, A., Lin, C. F. (1992). Unit root tests in panel data: asymptotic and finite sample properties. UC San Diego Working Paper 92–23.

Levin, A., Lin, C. F. (1993). Unit root tests in panel data: new results. UC San Diego Working Paper 93–56.

Levin, A., Lin, C. F., Chu, C.-S. J. (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., Wu, S. (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., Perron, B. (2004). Testing for a unit root in panels with dynamic factors. *Journal of Econometrics* 122:81–126.

Newey, W., West, K. D. (1994). Automatic lag selection for covariance matrix estimation. *Review of Economic Studies* 61:631–653.

O’Connell, P. G. J. (1998). The overvaluation of purchasing power parity. *Journal of International Economics* 44:1–19.

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., Moon, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica* 67:1057–1111.

Schwert, W. G. (1989). Tests for unit roots: a Monte Carlo investigation. *Journal of Business and Economic Statistics* 7:147–159.

Wagner, M. (2005). On PPP, unit roots and panels. Vienna: Institute for Advanced Studies, Working Paper: Economics Series No. 176.

Wagner, M., Hlouskova, J. (2004). What’s really the story with this Balassa-Samuelson effect in the CEECs? Department of Economics, University of Bern, Discussion Paper 04/16.

Zellner, A., Palm, F. C. (1974). Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* 2:17–54.