Review of International Economics, 9(3), 482–493, 2001

Misleading Inferences from Panel Unit-Root Tests with an Illustration from Purchasing Power Parity

*Janice Boucher Breuer, Robert McNown, and Myles S. Wallace**

Abstract

Simulations demonstrate that when unit-root behavior is rejected in a Levin and Lin panel test, it is incorrect to infer that all series are stationary. Recent tests proposed by Im, Pesaran and Shin, and by Sarno and Taylor, are also incapable of determining the mix of $I(0)$ and $I(1)$ series in a panel setting. This paper introduces a new unit-root test that allows the researcher to discern which series are $I(0)$ and which ones are $I(1)$. The test has double to triple the power of single-equation augmented Dickey–Fuller tests.

1. Introduction

Unit-root test methodology has undergone a major development due to Levin and Lin (1992) who advanced panel estimation over single-equation estimation in testing for a unit root. This development in testing methodology has also brought with it a dramatic change in conclusions regarding unit-root behavior. Panel unit-root tests typically produce much more evidence favorable to stationarity than when single-equation estimation is used. The dramatic change in results is mostly attributed to the substantial gain in power the panel studies afford when a researcher moves from a sample size of T (where T equals the number of time-series observations) to a sample size of $T \times N$ (where N equals the number of cross-section observations).

Levin and Lin's paper has spawned an industry of re-investigating the time-series properties of economic data. Examples of papers that use the panel estimation framework to test for a unit root are: for the real exchange rate, O'Connell (1998), Papell (1997), Lothian (1997), Frankel and Rose (1996), Jorion and Sweeney (1996), Oh (1996), and Wu (1996); for the law of one price, Parsley and Wei (1996); for nominal interest rates, Wu and Zhang (1996); for the permanent income hypothesis, Jin (1996); and for technology, Evans and Karras (1996).

In all of these papers, where unit-root behavior is rejected, the conclusion reached is that *all* members of the panel are stationary which is the alternative hypothesis put forth by Levin and Lin. For example, in commenting on one panel study that rejects the unit-root hypothesis, Jorion and Sweeney (1996, p. 538) state: "[t]he implication is that each of the real exchange rates has a mean value to which it reverts over time. . . ." This is a potentially erroneous inference. We demonstrate that while the Levin and Lin procedure has some statistical power to reject a false null of a unit root for all members of the panel, it lacks discriminatory power about the time-series properties of the individual members of a panel that mixes stationary and nonstationary processes.[1] That is, rejection of a unit root in a panel setting can arise when some, but not all, members of the panel are stationary.

* Breuer: Darla Moore School of Business, University of South Carolina, Columbia, SC 29208, USA. Tel: 803-777-7419; Fax: 803-777-6876; E-mail: boucher@darla.badm.sc.edu. McNown: Department of Economics, University of Colorado, Boulder, CO 80309, USA. Tel: 303-492-8295; Fax 303-492-8960; E-mail: mcnownr@colorado.edu. Wallace: Department of Economics, Clemson University, Clemson, SC 29634, USA. Tel: 864-656-6509; Fax: 864-656-4192; E-mail: myles@clemson.edu.

O'Connell (1998) modifies the Levin–Lin test to accommodate dependence across the cross-section, and Papell (1997) adapts the test for heterogeneous serial correlation. However, both of these generalized least-squares procedures impose a common value of the autoregressive parameter under both the null and alternative hypotheses. Although these are important extensions of the Levin–Lin tests, the all-or-nothing nature of these procedures may lead to serious misinterpretations of tests applied to mixed panels.

Recent methodological refinements of the Levin and Lin test fail to address the "all-or-nothing" nature of the test (and modifications to it) fully. Im et al. (1997), Maddala and Wu (1997), Sarno and Taylor (1998), and Taylor and Sarno (1998) develop tests that permit the autoregressive parameter to differ across the panel members under the stationary alternative. However, because these tests are joint tests of the null hypothesis, they are not informative about the number of series that are stationary processes when the null hypothesis is rejected.

2. The Problem with Panel Unit-Root Tests: An Example

As mentioned in the introduction, a number of the panel unit-root studies of the behavior of the real exchange rate find evidence supportive of stationarity. In contrast, surveys by Breuer (1994) and Froot and Rogoff (1995) document a lack of empirical support for stationarity of the real exchange rate when tests are conducted on an individual country basis, particularly for the floating rate period. In this section, we use a panel of real exchange rates as one example of how inference can potentially be affected by the mix of series in the panel.

What is responsible for the difference in conclusions and how reliable are the inferences drawn from the panel unit-root procedures? We first examine these questions using real exchange rates from 14 OECD countries to illustrate the sensitivity of panel unit-root tests to the possible mix of stationary and nonstationary series in the panel.

We construct panels of real exchange rates using combinations of up to 10 of the 14 countries. The data are from the IFS CD-ROM. The real exchange rate is defined as the natural logarithm of the number of foreign currency units per US dollar times the US CPI divided by the foreign CPI. All data are yearly and the period examined is 1950–95.[2] Prior to forming the panels, each country's real exchange was tested for a unit root using the augmented Dickey–Fuller (ADF) test. Because one lagged augmentation term was sufficient to clear up any problems of serial correlation and additional lags were insignificant, inferences are based on this specification.

Table 1 presents the results of the ADF test. The countries are listed in the order of their estimated autoregressive coefficients, starting with the value closest to 1.0. The ADF test indicates that the three real exchange rates of Finland, France, and Greece may be stationary with t-statistics significant at the 0.10, 0.10, and 0.05 levels, respectively. In the table, these three countries have implied autoregressive coefficients that are the farthest away from 1.0.

Next, we tested panels of the data using a Levin and Lin panel unit-root procedure that permits fixed effects; i.e., country-specific intercepts. We did this to explore how inference using the Levin and Lin test may change depending on the potential mix of stationary and nonstationary series. We chose the fixed-effects specification because the point estimates of the intercepts differ substantially across countries and there is no reason to believe that all the intercepts are identical. Also, based on the individual country results, all panel ADF tests included one lagged augmentation term. The

Table 1. Single-Equation Augmented Dickey–Fuller Test for a Unit Root, 1950–95

$$\Delta q_t = \alpha_0 + (\rho - 1)q_{t-1} + \delta\Delta q_{t-1} + \varepsilon_t$$

	$(\rho - 1)$	ρ	*t-statistic on* $(\rho - 1)$
1. Japan	–0.008	0.992	–0.28
2. Switzerland	–0.018	0.972	–0.63
3. Austria	–0.049	0.951	–0.90
4. Portugal	–0.051	0.949	–1.11
5. Norway	–0.054	0.946	–1.41
6. Netherlands	–0.056	0.944	–1.25
7. Denmark	–0.060	0.940	–1.36
8. Spain	–0.086	0.914	–1.64
9. Germany	–0.090	0.910	–1.51
10. Belgium	–0.139	0.861	–2.12
11. Luxembourg	–0.191	0.809	–2.49
12. Finland	–0.226	0.774	–2.76***
13. France	–0.251	0.749	–2.61***
14. Greece	–0.380	0.620	–3.60**

** and *** denote significance at the 0.05 and 0.10 levels, respectively.

critical values for the fixed-effects specification have been calculated by Levin and Lin (1992, Table 5). For a panel of 10 countries with 50 observations each, they are –5.13 at the 10% level and –5.43 at the 5% level of significance. While the usable number of time-series observations per country in our panel is 44, the critical values found by Levin and Lin are virtually identical for varying numbers of observations per panel.

The results from the Levin and Lin panel unit-root tests are shown in Table 2. First, we panel test the ten countries with implied autoregressive coefficients closest to 1.0 (and therefore much more apt to contain unit roots). In all cases, for these countries, the individual ADF tests did not reject a unit root. The panel test result (specification 1) is in agreement with the individual unit-root tests. A unit root cannot be rejected for the panel as a whole.

Next, we dropped Japan, the country with the autoregressive coefficient closest to 1.0 in the panel of ten countries, and replaced it with Greece, the country with the autoregressive coefficient farthest from 1.0. The new panel of ten countries (specification 2) still cannot reject a unit root although the implied autoregressive coefficient reported in Table 2 drops slightly from 0.952 to 0.930. In this case, the conclusion from the panel test would be that the ten real exchange rates are unit-root processes when, in fact, one of them (Greece) may not be.

We continued the process by next eliminating the second, then third, and finally fourth countries with the autoregressive coefficients closest to 1.0 and replacing them respectively with the twelfth, thirteenth, and fourteenth countries with autoregressive coefficients farthest from 1.0. With three replacements, a unit root is rejected at the 0.10 level (specification 4), and with four replacements a unit root is rejected at the 0.05 level (specification 5). While the implied value of the autoregressive coefficient in the panel ADF test declines from 1.0 as the number of replacements increases, the

Table 2. Augmented Dickey–Fuller Test for a Unit Root in the Real Exchange Rate: Panel Results, 1950–95

$$\Delta q_t = \alpha_0 + (\rho - 1)q_{t-1} + \delta\Delta q_{t-1} + \varepsilon_t$$

Specification	$(\rho - 1)$	ρ	*t-statistic on* $(\rho - 1)$
1. Ten countries with value of $(\rho - 1)$ closest to 1.0 from Table 1	–0.048	0.952	–3.45
2. Japan replaced by Greece in spec. 1	–0.070	0.930	–4.32
3. Japan and Switzerland, replaced by Greece and France in spec. 1	–0.084	0.916	–4.85
4. Japan, Switzerland, and Austria replaced by Greece, France, and Finland in spec. 1	–0.095	0.905	–5.31***
5. Japan, Switzerland, Austria, and Portugal replaced by Greece, France, Finland, and Luxembourg in spec. 1	–0.105	0.895	–5.76**
6. Spec. 1 excluding Japan, Switzerland, Austria, and Portugal	–0.074	0.926	–3.70
7. Greece, France, Finland and Luxembourg	–0.252	0.748	–5.66*

Notes: Individual-specific effects were permitted for the intercept. Critical values are from Table 5 in Levin and Lin (1992). Specifications 1–5 use critical values $T = 50$, $N = 10$. Critical values for specifications 6 and 7 are based on $T = 50$, $N = 5$. See endnote 3.
*, **, and *** denote significance at the 0.01, 0.05, and 0.10 levels, respectively.

simulations demonstrate that changes in this coefficient are not particularly informative regarding the mix of stationary and nonstationary series.

Rejection of a unit root in specification 5 could be a result of replacement by the four smallest autoregressive coefficients and/or the elimination of the largest four autoregressive coefficients. These alternatives are examined in specifications 6 and 7, respectively. The Levin–Lin test from the panel of four countries with the smallest autoregressive coefficient rejects a unit root at the 0.01 significance level. The ADF test for the remaining group of six countries from the original ten is insignificant.[3] The rejection of the unit root in specification 5 is apparently due to the inclusion of the four countries with the smallest individual autoregressive parameters. If the panel actually consists of four stationary and six nonstationary real exchange rates, the overall inference of stationarity is clearly misleading. A researcher reporting these results as favorable to purchasing power parity would be basing the inference on a panel where the majority of real exchange rates may be nonstationary.

Of course, the results presented in this section suffer from pre-test bias and can only be considered illustrative of the problem. In the next section, we present simulations where, in the context of a controlled environment, the problems of inference in panel unit-root tests are clearly documented.

3. Simulation Experiments

To demonstrate how mixed data generation processes across the panel can affect the outcome of panel unit-root tests, we constructed a simulation environment similar to

Table 3. Rejection Frequencies and Estimates of Autoregressive Parameter in Fixed-Effects Model

	$T = 50$		$T = 100$	
Number of stationary cases	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.9$
$N = 5$				
1	0.120 (0.914)	0.108 (0.918)	0.163 (0.954)	0.134 (0.956)
2	0.257 (0.893)	0.151 (0.907)	0.301 (0.941)	0.264 (0.946)
3	0.458 (0.863)	0.250 (0.891)	0.576 (0.916)	0.447 (0.931)
$N = 10$				
1	0.092 (0.928)	0.079 (0.929)	0.128 (0.962)	0.096 (0.969)
2	0.140 (0.920)	0.108 (0.924)	0.181 (0.958)	0.154 (0.954)
3	0.223 (0.911)	0.143 (0.918)	0.297 (0.952)	0.219 (0.955)

Notes: Numbers reported are rejection frequencies and (in parentheses) estimates of the autoregressive parameter based on 1,000 replications of each experiment. Critical values for the tests come from Levin and Lin (1992, Table 5). The column labels are: N, for the number of individuals in the panel; T, the number of time-series observations; and ρ, the value of the autoregressive parameter for those series that are stationary.

Levin and Lin's where individual fixed effects are allowed. After confirming Levin and Lin's (1992, Table 5) critical values for a limited parameter set, we depart from their data generation process by allowing one or more individuals in the panel to be stationary processes. The panel-data unit-root tests are then applied to the data generated from these mixed panels, using Levin and Lin's 5% critical values, to observe the resulting rejection frequencies when the panel contains both stationary and nonstationary series. Our objective is to show that higher rejection rates of the null hypothesis of a unit root in *all* series can occur with as few as one stationary series in the panel.

The experiments were conducted using panels of 5 or 10 individuals (N), 50 or 100 time-series observations (T), with data for one, two, or three individuals generated to be stationary. The autoregressive coefficient for the $I(0)$ series was set equal to 0.9 or 0.8. In all cases the error terms for all processes were generated as independent normal random variables with mean zero and standard deviation of 0.1, a value consistent with the standard errors of the estimates from the individual unit-root tests. This set of parameter values combined to yield 24 experiments, showing the sensitivity of rejection frequencies to small departures from the environment found in Levin and Lin in which all members of the panel are $I(1)$ processes. Rejection frequencies and mean estimates of the pooled estimate of the autoregressive parameter for these experiments are reported in Table 3.

The primary results from these simulations are that rejection frequencies rise, as they should, with the proportion of $I(0)$ individuals and with the magnitude of the difference between their autoregressive parameter and unity. For example, in the panel with 10 members and 100 time-series observations, the rejection frequency rises from 0.096 to 0.219 as the number of processes with a stationary root of 0.90 increases from 1 to 3. In the same environments, but with a stationary root of 0.80, the corresponding rejection rates are 0.128 and 0.297. Also, note that for a given panel size, increasing the number of observations (from $T = 50$ to $T = 100$) while holding constant the value of the stationary root and the number of stationary series always increases the rejection frequency.

Table 3 also shows that the estimate of the autoregressive parameter unfortunately gives little indication of the number of stationary series in the panel. With $N = 5$, $T = 100$, and an autoregressive parameter of 0.9 for the stationary series, the estimate of ρ changes only slightly from 0.956 to 0.931 as the number of stationary series in the panel increases from one to three. Although the estimate of ρ changes very little, the rejection frequencies increase substantially, from 0.134 to 0.447, as the number of stationary series increases.

The problems that these experiments point out is that, while the tests correctly reject a false null (with moderate power), the rejections are uninformative about the number or proportion of series in the panel that are stationary. By construction, the alternative hypothesis for these tests is that all series are $I(0)$. However, in practice these tests are applied to panels containing mixtures of $I(0)$ and $I(1)$ series, and an all-or-nothing conclusion is not warranted. By analogy to simple regression, when an F-statistic rejects the null that a vector of coefficients is equal to zero, it does not follow that each coefficient is nonzero. Likewise, when the unit-root null hypothesis is rejected, it may be wrong to conclude that all series in the panel are stationary.

In the next section, we offer a test that accommodates a panel environment and the ability to test for a unit root in each member of the panel, individually.

4. The Seemingly Unrelated Regressions Augmented Dickey–Fuller Test

We introduce the "seemingly unrelated regressions augmented Dickey–Fuller" (SURADF) test which is an augmented Dickey–Fuller test based on the panel estimation method of seemingly unrelated regression (SUR). Equation (1) presents the system of ADF equations to be estimated:

$$\begin{aligned}
\Delta y_{1,t} &= \alpha_1 + (\rho_1 - 1)y_{1,t\text{-}1} + \sum\nolimits_{i=1} \delta_i \Delta y_{1,t\text{-}i} + u_{1,t} \\
\Delta y_{2,t} &= \alpha_2 + (\rho_2 - 1)y_{2,t\text{-}1} + \sum\nolimits_{i=1} \delta_i \Delta y_{2,t\text{-}i} + u_{2,t} \\
&\vdots \\
\Delta y_{N,t} &= \alpha_N + (\rho_N - 1)y_{N,t\text{-}1} + \sum\nolimits_{i=1} \delta_i \Delta y_{N,t\text{-}i} + u_{N,t},
\end{aligned} \tag{1}$$

where ρ_i is the autoregressive coefficient for series i. With the SURADF procedure this system is estimated by SUR, and the significance of each $(\rho_i - 1)$ is tested against critical values generated through simulation.

The specification of model (1) affords several advantages over the Levin and Lin panel unit-root test. First, since SUR estimation takes account of contemporaneous cross-correlation of the error terms, it has an information advantage over single-equation augmented Dickey–Fuller tests and the Levin and Lin (1992, 1993) tests. Our strategy departs from O'Connell (1998) since he used SUR but retained the Levin and Lin restriction that $(\rho_1 - 1) = (\rho_2 - 1) = \ldots = (\rho_N - 1)$. Second, the estimation allows for heterogeneity of the lag structure across panel members as suggested by Papell (1997). Allowing member-specific lag structures ensures that misspecification is not introduced into the equations, and thereby ensures that each of the error terms are white noise. Imposition of an identical lag structure across panel members could potentially bias the test statistic. However, in our case, it turned out that one lag was sufficient to remove evidence of serial correlation for each and every member of the panel. Third, the specification permits the magnitude of the autoregressive coefficient to differ across the panel members. That is, we relax the restriction of $(\rho_1 - 1) = (\rho_2 - 1) = \ldots = (\rho_N - 1)$ and thereby avoid the joint null hypothesis that all series contain a unit root and the corresponding alternative hypothesis that all series are stationary with the same autoregressive coefficient.

The MADF test of Sarno and Taylor (1998) and Taylor and Sarno (1998) shares some of these advantages of the SURADF test. Both procedures employ SUR estimation that accounts for the error correlations across the panel, and both permit differing lag structures for the various panel members. One implication of these extensions is that critical values depend upon the specific properties of the errors in the system of equations and must be generated by Monte Carlo methods for each individual application. (Copies of the RATS program used to generate research-specific critical values for both MADF and SURADF tests are freely available on request.) The important difference between the MADF and SURADF tests derives from the formulation of the null hypothesis. The MADF test is constructed as a joint test of the null hypothesis $[(\rho_1 - 1) = (\rho_2 - 1) = \ldots = (\rho_N - 1)] = 0$ on equation (1). Thus, while the MADF provides a joint test of a unit root for all members of the panel, the SURADF tests a separate unit-root null hypothesis for each individual panel member. The SURADF test therefore indicates how many series in the panel, and which ones, are stationary processes.

If the MADF test rejects its null hypothesis, then the only inference that can be drawn is that *at least one panel member has a stationary root*. The test does not determine how many or which ones are stationary. Sarno and Taylor (1998) and Taylor and Sarno (1998) advocate combining the MADF test with their Johansen likelihood ratio (JLR) test, in which the null hypothesis is that at least one series in the panel is nonstationary and the alternative is that all series are stationary. In the special case that both MADF and JLR null hypotheses are rejected, one can conclude that all series are stationary. However, in the more likely case of mixed panels, neither of these two tests will be informative about the stationary/nonstationary composition of the panel.[4]

5. Results from the SURADF Test

We apply the SURADF test to the panel of ten real exchange rates of specification 5 in Table 2. This is the panel for which the Levin and Lin test had previously rejected a unit root. We must simulate critical values that are specific to the lag structure and covariance matrix estimated for this panel of real exchange rates. In the data generation phase of the simulation, the "true" values of both the intercept and the coefficient on the lagged level for each series were set equal to zero to create $I(1)$ series without drift. The error series were generated to be normally distributed with the variance–covariance matrix obtained from the SUR estimation on the actual real exchange rate data. Also, the coefficients on the lagged differences were set equal to those obtained from the actual data under SUR estimation. In the data generation all parameters were fixed for all 10,000 replications: the error covariance matrix, the coefficients on the lagged differences, and the zero values for the coefficients on the lagged levels and the constant term.

In simulating the critical values, the restriction that the true value of the intercept (drift) is zero is not innocuous. The critical values would be asymptotically normal, and much lower in absolute value than the results presented below, if the intercept were included as the only deterministic term under both the null and the estimating equation (West, 1988).

Within each replication, samples of 94 observations on each series were generated, with initial values for each simulated exchange rate set to zero. To minimize the sen-

Table 4. SURADF Tests and Critical Values

		Critical values		
Country panel label	*SURADF*	*0.01*	*0.05*	*0.10*
1. Norway	–5.452***	–6.263	–5.523	–5.116
2. Netherlands	–5.123	–6.589	–5.911	–5.497
3. Denmark	–5.824***	–6.731	–5.994	–5.609
4. Spain	–3.435	–5.551	–4.801	–4.407
5. Germany	–5.003	–6.620	–5.917	–5.526
6. Belgium	–5.340	–6.701	–6.040	–5.668
7. Greece	–4.716**	–5.042	–4.165	–3.743
8. France	–4.339	–6.067	–5.292	–4.860
9. Finland	–3.920	–5.361	–4.530	–4.106
10. Luxembourg	–5.253	–6.538	–5.896	–5.533

Notes: The column "SURADF" presents the estimated augmented Dickey–Fuller statistics based on SUR estimation. The three right-hand-side columns report the estimated critical values obtained from simulations based on 44 observations for each series and 10,000 replications using the lag and covariance structure from the panel of real exchange rates.
** and *** denote significance at the 0.05 and 0.10 levels, respectively.

sitivity of the results to the initial conditions, the first 50 observations on each series were omitted. This left 44 usable observations for each series, which corresponds exactly to the number of observations available in the real data.

In the estimation phase of the simulation, the intercepts, the coefficients on the lagged levels and lagged differences, and the covariance matrix were all estimated freely. The SURADF test statistic for each of the ten series was computed as the *t*-statistic for the coefficient on the lagged level for that equation.

The experiment was replicated 10,000 times, and the test statistics determining the lowest α-percentile defined the α-percent critical value. The 1%, 5%, and 10% critical values for each of the ten panel members were tabulated in this manner and are reported in Table 4. Since the critical values are also affected by the error covariance and other equation parameters, country labels are attached to the simulated series corresponding to the country-specific parameters imposed on each of the ten simulated series. For example, in the row labeled "Norway," the critical values for series 1 are specific to the Norwegian parameters since these were imposed on series 1 in the simulation. The results of the SURADF test using these critical values are also presented in Table 4.

Table 4 shows that of the ten countries, three reject a unit root at the 10% significance level or better. Norway and Denmark reject at the 10% level (with implied autoregressive coefficients of 0.89 and 0.886) and Greece rejects at the 5% level (with an implied autoregressive coefficient of 0.615).

These results differ from Table 1 where the single-equation ADF test picked up Greece, France, and Finland as rejecting a unit root at the 10% level or better. The difference in results may arise for several reasons. First, a nonzero covariance matrix

Table 5. Power Analysis of Ten-country Panel: 5,000 Replications, 44 Observations

	$\rho = 0.95$		$\rho = 0.90$		$\rho = 0.80$	
	ADF	*SURADF*	*ADF*	*SURADF*	*ADF*	*SURADF*
1. Norway	8.4	13.0	13.9	31.8	35.0	74.4
2. Netherlands	7.5	12.4	12.8	33.2	30.0	79.7
3. Denmark	8.2	18.1	13.8	46.1	33.6	90.7
4. Spain	8.2	12.9	15.0	26.2	39.5	61.1
5. Germany	7.7	13.9	13.0	37.4	31.1	84.5
6. Belgium	8.3	17.8	14.0	48.2	35.1	92.6
7. Greece	7.4	7.7	13.4	15.0	31.8	35.1
8. France	7.4	11.4	13.3	26.6	33.2	65.6
9. Finland	8.2	10.9	15.6	23.0	39.4	53.3
10. Luxembourg	8.1	17.6	13.9	44.5	34.3	89.4

Notes: This table reports rejection rates of the unit-root null hypothesis for tests applied to ten simulated series based on the lag and covariance structure of the countries listed. Rejection rates are reported for the augmented Dickey–Fuller test (ADF) and the augmented Dickey–Fuller test based on the seemingly unrelated regressions estimator (SURADF). All ten series are generated as stationary processes with the common value of the largest autoregressive root given by ρ.

introduces more information into the estimation and results in lower standard errors. This should increase the chance of rejection. Second, the $(\rho_i - 1)$ may change, moving either closer to or farther away from zero, and this, too, could result in a change in inference. Third, the critical values for SURADF change. They are higher in absolute value than for the single-equation ADF test. The rise in critical values will make it more difficult to reject a unit root even if the standard errors are reduced and the implied autoregressive coefficient departs from 1 (toward zero).

A comparison of inference based on the SURADF test with those from the Levin and Lin test on the same ten countries reveals the weakness of the Levin and Lin test.[5] The inference from the Levin and Lin test would be that all series in this panel are stationary while the SURADF test indicates that only three of the series are stationary.

We next demonstrate that the SURADF test affords a substantial gain in power when compared with the single-equation augmented Dickey–Fuller test. The power of both tests is examined by constructing a panel of series that are simulated to be stationary with an autoregressive coefficient of 0.95. 0.90, or 0.80. Rejection rates are produced from 5,000 replications with a sample size of 44 for each series, using the same ten-country panel that was used in the simulation of critical values. In the generation of the ten series, the intercepts, coefficients on lagged differences, and the error covariance matrix were set equal to their estimates from the empirical data. Least squares and SUR estimation were both applied to the simulated data in order to compute the ADF and SURADF test statistics for each series in the panel. These computed test statistics were compared with 5% critical values, and the frequencies of rejection of the false null were recorded for every case.

Table 5 presents the results of the power analysis for the two tests. Reported for each series are the rejection rates across each of the three assumed autoregressive coefficients. For the SURADF test, the ten series are given the country names that correspond to the country-specific parameters used in the simulation. For example, in the row labeled "Norway," the SURADF entries present the rejection frequencies of the null hypothesis of a unit root when the simulated series has an autoregressive coefficient of 0.95, 0.90, or 0.80, and the parameters used in the simulation are those estimated for the Norwegian real exchange rate.

With the exception of Greece, the power of the SURADF test is substantially higher than that of the ADF in every environment. When the autoregressive coefficient is equal to 0.95, for example, the rejection frequency of the SURADF test ranges from 33% (Finland) to 120% (Denmark) higher than that of the ADF test. As the autoregressive coefficient departs further from 1.0, the power to reject increases for both the ADF and SURADF test. However, the SURADF test has nearly double to triple the power of the single-equation ADF test. The efficiency gain associated with SUR estimation has contributed to the improved power of the ADF test for a unit root. We conclude that in settings where a researcher admits the possibility that there may be a mix of stationary and nonstationary series in the panel, the SURADF test be used for investigating unit-root behavior. Indeed, in a working paper we find the power of the SURADF to substantially exceed that of the ADF in a number of simulated environments; see Breuer et al. (1999).

6. Conclusion

The panel unit-root test of Levin and Lin (1992, 1993) and extensions of their procedure have been widely used in empirical investigations of whether panels contain series that are nonstationary or stationary. However, these tests and advances by Im et al. (1997), Maddala and Wu (1997), Sarno and Taylor (1998), and Taylor and Sarno (1998) are joint tests of the null hypothesis that all members of the panel contain a unit root. Consequently, the tests are not able to provide information about the specific time-series behavior of each member of the panel. Moreover, the test outcomes can be misinterpreted in the plausible case where the panel contains a mix of stationary and unit-root processes. Our simulations demonstrate that rejection can be produced with as few as one stationary member of the panel, and therefore that it would be inappropriate to infer that all members of the panel are stationary.

We propose a panel unit-root test that allows testing for unit-root behavior of each individual panel member. Our test uses the method of seemingly unrelated regressions (SUR) applied to a panel of augmented Dickey–Fuller equations in which the autoregressive coefficients across the panel members are not restricted to be the same. Since the SURADF test is based on a panel environment, the increased sample size and efficiency gains would be expected to produce a more powerful test than the single-equation ADF test. Indeed, our simulations show that the SURADF test has double to triple the power of the single equation augmented Dickey–Fuller test in rejecting a false null hypothesis.

References

Breuer, Janice Boucher, "An Assessment of the Evidence on Purchasing Power Parity," in John Williamson (ed.), *Estimating Equilibrium Exchange Rates*, Washington: Institute for International Economics (1994).

Breuer, Janice Bouncher, Robert McNown, and Myles Wallace, "Series Specific tests for a Unit Root in a Panel Setting with an Application to Real Exchange Rates," Working Paper, University of South Carolina (1999).

Evans, Paul and Georgios Karras, "Do Economies Converge? Evidence from a Panel of US States," *Review of Economics and Statistics* 78 (1996):384–8.

Frankel, Jeffrey A. and Andrew K. Rose, "A Panel Project on Purchasing Power Parity: Mean Reversion Within and Between Countries," *Journal of International Economics* 40 (1996):209–24.

Froot, Kenneth A. and Kenneth Rogoff, "Perspectives on PPP and Long-run Real Exchange Rates," in Gene Grossman and Kenneth Rogoff (eds.), *The Handbook of International Economics*, Vol. 4, New York: North-Holland (1995).

Im, Kyung So, M. Hashem Pesaran, and Yongcheol Shin, "Testing for Unit Roots in Heterogeneous Panels," Manuscript, Department of Applied Economics, University of Cambridge (1997).

Jin, Fuchin, "Cointegration of Consumption and Disposable Income: Evidence from Twelve OECD Countries," *Southern Economic Journal* 62 (1996):77–88.

Jorion, Phillipe and Richard Sweeney, "Mean Reversion in Real Exchange Rates: Evidence and Implications for Forecasting," *Journal of International Money and Finance* 15 (1996):535–50.

Levin, Andrew and Chien-fu Lin, "Unit Root Tests in Panel Data: Asymptotic and Finite-sample Properties," University of California-San Diego Discussion Paper 92-23 (1992).

———, "Unit Root Tests in Panel Data: New Results," University of California-San Diego Discussion Paper 93-56 (1993).

Lothian, James R, "Multi-country Evidence on the Behavior of Purchasing Power Parity under the Current Float," *Journal of International Money and Finance* 16 (1997):19–35.

Maddala, G. S. and Shaowen Wu, "A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test," manuscript, Ohio State University (1997).

O'Connell, Paul G. J., "The Overvaluation of Purchasing Power Parity," *Journal of International Economics* 44 (1998):19–35.

Oh, Keun-yeob, "Purchasing Power Parity and Unit Root Tests using Panel Data," *Journal of International Money and Finance* 15 (1996):405–18.

Papell, David, "Searching for Stationarity: Purchasing Power Parity under the Current Float," *Journal of International Economics* 43 (1997):313–32.

Parsley, David and Shang-Jin Wei, "Convergence to the Law of One Price without Trade Barriers or Currency Fluctuations," *Quarterly Journal of Economics* 111 (1996):1211–36.

Sarno, Lucio and Mark Taylor, "Real Exchange Rates Under the Recent Float: Unequivocal Evidence of Mean Reversion," *Economics Letters* 60 (1998):131–7.

Taylor, Mark and Lucio Sarno, "The Behavior of Real Exchange Rates during the Post Bretton Woods Period," *Journal of International Economics* 46 (1998):281–312.

West, Kenneth D., "Asymptotic Normality when Regressors Have a Unit Root," *Econometrica* 56 (1988):1397–418.

Wu, Yangru, "Are Real Exchange Rates Nonstationary? Evidence from a Panel-Data Test," *Journal of Money, Credit, and Banking* 28 (1996):54–63.

Wu, Yangru and Hua Zhang, "Mean Reversion in Interest Rates: New Evidence from a Panel of OECD Countries," *Journal of Money, Credit, and Banking* 28 (1996):604–21.

Notes

1. Levin and Lin (1993) emphasize that their test requires all members of the panel be identical in terms of the presence or absence of a unit root. We believe this caution has been disregarded in applications of their test.
2. We do not confront the issue of whether the data should be broken into the two separate regimes of fixed and floating exchange rates. Our purpose is to demonstrate the sensitivity of panel unit-root inference to the mix of nonstationary and stationary series included in the panel.
3. Levin and Lin (1992) do not report critical values for panels of $N = 4$ or 6. However, the critical values increase with N, and this allow us to make inferences. Since the panel of four countries is significant using the critical values for $N = 5$ (given) it is significant for $N = 4$. Similarly, the panel of six countries is insignificant for $N = 5$ and must also be insignificant for $N = 6$.

4. In addition, because the JLR is derived from an error-correction framework, degrees of freedom are rapidly used up as the panel expands in size. As a consequence, the test cannot be used in large panels as can the SURADF.
5. We do not evaluate the relative power of the Levin and Lin, MADF, and SURADF tests because of the differences in the null and alternative hypotheses of these three tests. The MADF and Levin–Lin procedures test the null of a common unit root for all series in the panel. In addition, these two tests are not informative about the stationary/nonstationary composition of the panel. Even though these two tests may have more power to reject a false null than the SURADF, the different information contained in the test results undermines meaningful power comparisons.