Hanck, Christoph

Working Paper

For Which Countries did PPP hold? A Multiple Testing Approach

Technical Report, No. 2006,47

Provided in Cooperation with:
Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475), University of Dortmund

Suggested Citation: Hanck, Christoph (2006) : For Which Countries did PPP hold? A Multiple Testing Approach, Technical Report, No. 2006,47, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:
https://hdl.handle.net/10419/22691

For Which Countries did PPP hold? A Multiple Testing Approach*

Christoph Hanck†

This version: November 13, 2006

Abstract

We use recent advances in multiple testing to identify the countries for which Purchasing Power Parity (PPP) held over the last century. The approach controls the multiplicity problem inherent in simultaneously testing for PPP on several time series, thereby avoiding spurious rejections. It has higher power than traditional multiple testing techniques by exploiting the dependence structure between the countries with a bootstrap approach. We use a sieve bootstrap approach to account for nonstationarity under the null hypothesis. Our empirical results show that, plausibly, controlling for multiplicity in this way leads to a number of rejections of the null of no PPP that is intermediate between that of traditional multiple testing techniques and that which results if one tests the null on each single time series at some level α.

Keywords: Multiple Testing, Bootstrap, PPP, Panel Data

JEL classification: C12, C23, F31

*I am grateful to Alan Taylor for providing the dataset used in this paper.

†Research supported by DFG under Sonderforschungsbereich 475. Universität Dortmund/Ruhr Graduate School in Economics, Vogelpothsweg 78, 44221 Dortmund, Germany. Contact Information: Tel. (+49) 0231-7553127, Fax (+49) 0231-7555284, christoph.hanck@uni-dortmund.de.

1 Introduction

Purchasing Power Parity (PPP) is among the most popular theories to explain the long run behaviour of exchange rates. Not least because it is ready-made for empirical implementation, it has been investigated by a host of econometric techniques. So-called "stage-two" tests [Froot and Rogoff, 1995] test the hypothesis that the real exchange rate follows a random walk. The alternative is that the real exchange rate is a stationary process, i.e. that PPP holds in the long run. Typically, researchers would obtain real exchange rate data over a certain time span for several countries and conduct appropriate unit root tests on each series [see, e.g., Taylor, 2002]. It is then argued that PPP holds for those countries for which the null is rejected.

Unfortunately, this simple and intuitive way of investigating the validity of PPP is problematic from a statistical point of view. Effectively, it ignores the issue of multiple testing. To illustrate the problem, consider the following artificial numerical example. Suppose one has exchange rate data on a panel of, say, $N = 20$ countries. Also assume for simplicity that the the units are independent and that PPP does not hold for any of the units. When conducting tests on each unit at the $\alpha = 0.05$ level, one might casually expect the probability to erroneously find evidence in favor of PPP in at most one case to equal 5%, because $1/20 = 0.05$. However, the event of a rejection is a Bernoulli random variable with "success" probability 0.05. Hence, P_k, the probability of finding k rejections in N tests, is the probability mass function of a Binomial random variable,

$$P_k = \left(\begin{array}{c} N \\ k \end{array} \right) \alpha^k (1-\alpha)^{N-k}.$$

Therefore, the probability of (at least) one erroneous rejection, also known as the

Familywise Error Rate[1] ($FWER$), equals

$$P_{k\geqslant 1} = \sum_{j=1}^{20} \left(\begin{array}{c} 20 \\ j \end{array} \right) 0.05^j (1-0.05)^{20-j} = 0.6415.$$

Even if PPP does not hold for any of the countries in the panel, one will falsely find some evidence of it with a rather high probability. Of course, the problem only worsens if one adds more units to the panel.

This so-called "multiplicity" problem, while not widely recognized in econometrics [Savin, 1984], has of course been realized long ago in the statistics literature [see Lehmann and Romano, 2005]. Several solutions to controlling the $FWER$ at some specified level α have been suggested. Among the most popular are the Bonferroni and the the Holm [1979] procedure. These procedures have however been less successful in econometric applications because ensuring $FWER \leqslant \alpha$ typically comes at the price of reducing the ability to identify false hypotheses. That is, the procedures are conservative or have low "power."[2] Hence, often quite reasonably, researchers have tended to ignore the issue of multiplicity.

Recently, panel econometric techniques have become popular to test for PPP. See, for instance, Wu [1996], Papell and Theodoridis [2001], Papell [2002] or Murray and Papell [2005]. Typically, these panel unit root tests formulate the null of the entire panel being nonstationary. The alternative quite often is that of a stationary panel [see, for instance, Harris and Tzavalis, 1999; Levin et al., 2002; Breitung, 2000]. These panel tests also have power against "mixed" panels, where only some fraction of the units is actually stationary [see Taylor and Sarno, 1998; Karlsson and Löthgren, 2000; Boucher Breuer et al., 2001]. Hence, erroneous conclusions on the number of countries for which PPP holds remain possible. (Concluding from a

[1] More generally, the j-$FWER$ is defined as $P_{k\geqslant j}$, the probability of j or more false rejections.

[2] For a discussion of "power" in a multiple testing framework see Romano and Wolf [2005], Sec. 2.2.

rejection of a panel unit test that *all* units are stationary is closely related to the erroneous inference that a rejection in an F test of the "significance of a regression" implies that *all* coefficients are nonzero.)

As a partial remedy, Maddala and Wu [1999] and Choi [2001] draw on the meta analytic literature [see Hedges and Olkin, 1985] to provide panel unit root tests having the more conservative alternative that some nonzero fraction of the panel is stationary. However, their approach neither allows to identify which nor how many of the countries in the panel have a stationary real exchange rate.

Recently, there has been substantial research on improving the ability of multiple testing approaches to detect false hypotheses while still controlling the *FWER*. Notably, Romano and Wolf [2005] have put forward a bootstrap scheme that exploits the dependence structure of the statistics in order to improve the power of the multiple test. In the present paper, we propose an adaptation of the Romano and Wolf [2005] approach to identify those countries of a panel of real exchange rate data for which the Purchasing Power Parity condition holds.

The plan of the paper is as follows. Section 2 offers a brief statement of the PPP condition and presents the general multiple testing approach of Romano and Wolf [2005]. Section 3 discusses the bootstrap approach employed in this paper. The empirical results are in Section 4. Section 5 concludes.

2 The Multiple Testing Approach

Our goal is to identify those countries of a panel for which the Purchasing Power Parity (PPP) relation held over the sample period. Let $p_{i,t}$ be the (log) price level in country i and period t, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$, p_t^* the "foreign" (log) price level of the reference country in the panel and $s_{i,t}$ the (log) nominal

exchange rate between the currencies of country i and the reference country. The real exchange rate is then given by

$$r_{i,t} = p_{i,t} - p_t^* - s_{i,t} \qquad (i = 1, \ldots, N)$$

Testing the strong PPP hypothesis is naturally formulated [see Rogoff, 1996] as a unit root test on the real exchange rate. A vast number of unit root tests have been suggested in the literature [see Phillips and Xiao, 1998, for a survey], many of which have been applied to the PPP question. We will use the standard augmented Dickey and Fuller [1979] test [see also Said and Dickey, 1984]. We do so because it is still the most popular unit root test and, more importantly, the bootstrap versions of the test required for the multiple testing scheme have desirable properties [Swensen, 2003; Chang and Park, 2003]. Accordingly, we investigate PPP by testing the individual hypotheses

$$H_i : \varrho_i = 0 \quad \text{vs.} \quad H_i' : \varrho_i < 0 \qquad (i = 1, \ldots, N) \tag{1}$$

where

$$\Delta r_{i,t} = \varrho_i r_{i,t-1} + \sum_{j=1}^{J_i} \nu_j \Delta r_{i,t-j} + \epsilon_{J_i,i,t}. \tag{2}$$

The number of lagged differences J_i required to capture serial correlation in $r_{i,t}$, is allowed to vary across i. Our test statistic is given by $\hat{\tau}_i = \hat{\varrho}_i / s.e.(\hat{\varrho}_i)$, the t-statistic of ϱ_i in (2), where $\hat{\varrho}_i$ is the usual OLS estimator and $s.e.(\hat{\varrho}_i)$ the associated standard error.

We aim to determine those countries $i \subset \{1, \ldots, N\}$ for which $r_{i,t}$ is a stationary process. As argued in the Introduction, in order to provide reliable statistical inference in the sense of controlling the *FWER*, it is important to take into account the multiplicity inherent in testing in a panel setting. We now present the general multiple testing framework used here, making suitable adjustments to adapt the procedure to the PPP testing case.

First, relabel the test statistics from smallest to largest, such that $\hat{\tau}_{r_1} \leqslant \hat{\tau}_{r_2} \leqslant \ldots \leqslant \hat{\tau}_{r_N}$. (The smaller a Dickey-Fuller test statistic, the stronger the evidence in favor of stationarity.) Form a joint rectangular confidence region for the vector $(\varrho_{r_1}, \ldots, \varrho_{r_N})^\top$. The region is of the form

$$(-\infty,\ \hat{\varrho}_{r_1} + s.e.(\hat{\varrho}_{r_1}) \cdot d_1] \times \cdots \times (-\infty,\ \hat{\varrho}_{r_N} + s.e.(\hat{\varrho}_{r_N}) \cdot d_1], \tag{3}$$

where one chooses d_1 so as to ensure a joint asymptotic coverage probability $1-\alpha$.[3] The bootstrap method to appropriately choose d_1 in the present problem will be discussed below. The decision rule is to reject a particular hypothesis H_{r_n} if the corresponding confidence interval satisfies $0 \notin (-\infty,\ \hat{\varrho}_{r_n} + s.e.(\hat{\varrho}_{r_n}) \cdot d_1]$. Romano and Wolf [2005] show that if the confidence region (3) has coverage probability $1-\alpha$, then this method asymptotically controls the *FWER* at level α, $\lim_T FWER \leqslant \alpha$. Crucially, the method does not stop there. In order to improve the ability of the method to detect false hypotheses, one can construct further confidence regions after having rejected, say, the first N_1 hypotheses. In a second step, one forms a confidence region for the remaining $N - N_1$ coefficients $(\varrho_{r_{N_1+1}}, \ldots, \varrho_{r_N})^\top$. This is again constructed to have nominal joint coverage probability $1-\alpha$ and is of the form

$$(-\infty,\ \hat{\varrho}_{r_{N_1+1}} + s.e.(\hat{\varrho}_{r_{N_1+1}}) \cdot d_2] \times \cdots \times (-\infty,\ \hat{\varrho}_{r_N} + s.e.(\hat{\varrho}_{r_N}) \cdot d_2],$$

potentially leading to the rejection of some further N_2 hypotheses. This step-down process can be repeated until no further hypotheses are rejected. Romano and Wolf [2005] show that the d_j should ideally be chosen as

$$d_j \equiv d_j(1-\alpha, P) = \inf\left\{x : \Pr_P\left[\max_{R_{j-1}+1 \leqslant n \leqslant N}\left(\frac{\hat{\varrho}_{r_n} - \varrho_{r_n}}{s.e.(\hat{\varrho}_{r_n})}\right) \leqslant x\right] \geqslant 1-\alpha\right\},$$

where $R_{j-1} = \sum_{k=0}^{j-1} N_k$ and $R_0 = 0$. In practice, however, P and hence d_j are unknown. Fortunately, Romano and Wolf [2005, Thms. 3.1 and 4.1] show that d_j

[3] As recommended by Romano and Wolf [2005] we use the studentized version of their method. For a discussion of the "basic" approach, see Sec. 3 of their paper.

can often be estimated consistently with the bootstrap without affecting asymptotic control of the *FWER*.

3 The Bootstrap Algorithm

We now outline the bootstrap approach to obtain an estimator $\hat{d}_j$ employed in this paper.

1. Fit an autoregressive process to $\Delta r_{i,t}$ $(i = 1, \ldots, N;\ t = 2, \ldots, T)$. It is natural to use the Yule-Walker procedure because it always yields an invertible representation [Brockwell and Davis, 1991, Secs. 8.1–2]. Letting $\overline{\Delta r_i} := (T_i - 1)^{-1} \sum_{t=2}^{T_i} \Delta r_{i,t}$, compute the empirical autocovariances of $\Delta r_{i,t}$ up to order q,

$$\hat{\gamma}_i(\ell) := \frac{1}{T_i - 1 - \ell} \sum_{t=2}^{T_i - \ell} (\Delta r_{i,t} - \overline{\Delta r_i})(\Delta r_{i,t+\ell} - \overline{\Delta r_i}),$$

where $i = 1, \ldots, N;\ \ell = 1, \ldots, q$.[4] Defining

$$\hat{\Gamma}_{i,q} := \begin{pmatrix} \hat{\gamma}_i(0) & \cdots & \hat{\gamma}_i(q-1) \\ \vdots & \ddots & \vdots \\ \hat{\gamma}_i(q-1) & \cdots & \hat{\gamma}_i(0) \end{pmatrix}$$

and $\hat{\gamma}_i := (\hat{\gamma}_i(1), \ldots, \hat{\gamma}_i(q))^\top$, obtain the AR coefficient vector as

$$(\hat{\phi}_{q,i,1}, \ldots, \hat{\phi}_{q,i,q})^\top := \hat{\Gamma}_{i,q}^{-1} \hat{\gamma}_i. \qquad (i = 1, \ldots, N)$$

2. The residuals are, as usual, given by

$$\hat{\epsilon}_{q,i,t} := \Delta r_{i,t} - \sum_{\ell=1}^{q} \hat{\phi}_{q,i,\ell} \Delta r_{i,t-\ell},$$

for $i = 1, \ldots, N$; $t = q + 2, \ldots, T$. Following Swensen [2003], center $\hat{\epsilon}_{q,i,t}$ to obtain

$$\tilde{\epsilon}_{q,i,t} := \hat{\epsilon}_{q,i,t} - \frac{1}{T_i - q - 1} \sum_{g=q+2}^{T_i} \hat{\epsilon}_{q,i,g}$$

for $i = 1, \ldots, N$; $t = q + 2, \ldots, T$.

[4] In practice, q can be chosen with a data-dependent criterion such as Akaike's.

3. Resample nonparametrically from $\tilde{\epsilon}_{q,i,t}$ to get $\epsilon^*_{q,i,t}$. To preserve the empirical cross-sectional dependence structure, jointly resample residual vectors

$$\tilde{\epsilon}_{q,\bullet,t} := (\tilde{\epsilon}_{q,1,t}, \ldots, \tilde{\epsilon}_{q,N,t}). \qquad (t = q+2; \ldots, T)$$

   See Hanck [2006] for evidence of the good performance of this step to account for cross-sectional dependence.

4. Recursively construct the bootstrap samples as[5]

$$\Delta r^*_{q,i,t} = \sum_{\ell=1}^{q} \hat{\phi}_{q,i,\ell} \Delta r^*_{q,i,t-\ell} + \epsilon^*_{q,i,t}$$

   for $i = 1, \ldots, N$, $t = q+2, \ldots, T$.

5. It is necessary to impose the null of a unit root when generating the artificial data in bootstrap unit root tests to achieve consistency [Basawa et al., 1991]. Accordingly, impose the null of nonstationarity by integrating $\Delta r^*_{i,t}$ to obtain $r^*_{i,t}$.

6. For each bootstrap sample $r^*_b := \big((r^*_{b,1,1}, \ldots, r^*_{b,1,T})^\top, \ldots, (r^*_{b,N,1}, \ldots, r^*_{b,N,T})^\top\big)$, compute the test statistics τ^*_{b,r_n}, and

$$\max^*_{b,j} := \max_{R_{j-1}+1 \leqslant n \leqslant N} (\tau^*_{b,r_n} - \hat{\tau}_{r_n}).$$

7. Repeat steps 3 to 6 many, say B, times.

8. Compute $\hat{d}_j$ as the $1-\alpha$ quantile of the B values $\max^*_{1,j}, \ldots, \max^*_{B,j}$.

Chang and Park [2003] and Swensen [2003] show that the above sieve bootstrap scheme yields asymptotically valid bootstrap ADF tests in the sense that using the α quantile of the bootstrap distribution of the τ^*_{b,r_n} as critical value asymptotically gives a test with size α. By a continuous mapping theorem argument, we expect the bootstrap to also consistently estimate the distribution of the $\max^*_{b,j}$ and hence $\hat{d}_j$.

[5] We run the recursion for 30 initial observations before using the $\Delta r^*_{q,i,t}$ to mitigate the effect of initial conditions.

Table I—Empirical Results

country	$\hat{\tau}_i$	p-value	Holm criterion
Mexico	-4.334	< 0.001	0.0026
Finland	-4.136	0.001	0.0028
Argentina	-3.632	0.006	0.0029
Italy	-3.344	0.015	0.0031
Norway	-3.285	0.018	0.0033
Sweden	-3.202	0.022	0.0036
UK	-2.996	0.038	0.0038
Belgium	-2.980	0.040	0.0042
Germany	-2.957	0.042	0.0046
France	-2.929	0.045	0.0050
Brazil	-2.561	0.104	0.0056
Australia	-2.544	0.108	0.0063
Netherlands	-2.498	0.119	0.0071
Portugal	-2.391	0.147	0.0083
Canada	-2.202	0.207	0.0100
Spain	-2.118	0.238	0.0125
Denmark	-2.058	0.262	0.0167
Switzerland	-1.349	0.604	0.0250
Japan	-1.323	0.617	0.0500

4 Results

We now present the empirical results of an application of the modified Romano and Wolf [2005] methodology to the PPP condition. We revisit the dataset used by Taylor [2002], which includes annual data for the nominal exchange rate, CPI and the GDP deflator. This dataset is particularly useful for our purposes because it covers a long period, ranging from 1892 through to 1996. The countries contained in our panel are given in Table I. We use the United States as the reference country throughout and report results using CPI price series. See Taylor [2002] for further details on data sources and definitions.

Using standard ADF unit root tests, we find rejections for 9 out of 19 countries at the 5% critical value -2.94. See the first column of Table I. (The entries are sorted for later use.) The number of lagged differences J_i in (2) is chosen with the data-

dependent criterion of Ng and Perron [2001]. The findings of Taylor [2002] are very similar.[6] Evidence in favor of PPP is therefore at best mixed. Taylor [2002] then argues that it may be possible to find more rejections in favor of PPP by employing more powerful techniques. Our goal, on the other hand, is to investigate whether some of the rejections are spurious in the sense that they would not occur when taking into account the multiplicity of the testing problem.

As a preliminary step, we report results for the more classical techniques to control the *FWER*, namely the Bonferroni and the Holm [1979] procedures. Recall that the former rejects H_i if the p-value $\hat{p}_i$ corresponding to the test statistic $\hat{\tau}_i$ satisfies $\hat{p}_i \leqslant \alpha/N$. The Holm [1979] procedure first sorts the p-values from smallest to largest, $\hat{p}_{r_1} \leqslant \ldots \leqslant \hat{p}_{r_N}$. Relabel the hypotheses accordingly as H_{r_n}. Then, reject H_{r_n} at level α if $\hat{p}_{r_j} \leqslant \alpha/(N-j+1)$ for all $r_j = 1, \ldots, r_n$.[7] The cutoff value for the first hypothesis is identical for both methods, but unlike the Bonferroni method, the Holm [1979] procedure uses gradually less challenging criteria for $H_{r_2}, \ldots, H_{r_N}$. Nevertheless, it often has low power because it also fails to exploit the dependence structure between the statistics.

The limit distribution of the ADF test statistics is a functional of Brownian motions that cannot be evaluated analytically to obtain p-values. We therefore rely on response surface regressions suggested by MacKinnon [1994, 1996] to obtain numerical distribution functions of the test statistics. We report results in columns 2 and 3 of Table I.

As expected, the number of rejections is now much lower. After controlling for multiplicity, we only observe rejections for Mexico and Finland for either method. These

[6] The small differences can be explained by different interpolation schemes for missing wartime data, other lag selection criteria as well as the fact that we balance our panel.

[7] See Lehmann and Romano [2005] for a proof that the Bonferroni and the Holm method control the *FWER* at level α.

results indeed suggest that the Bonferroni and Holm procedures are conservative.

We therefore now turn to the results of the Romano and Wolf [2005] approach. The algorithm presented in Section 2 yields $\hat{d}_1 = 4.050$, leading to a rejection for Mexico and Finland. In the second round, we obtain $\hat{d}_2 = 3.429$, implying evidence in favor of PPP for Argentina. Next, we find $\hat{d}_3 = 3.252$ such that we reject for Italy and Norway. Finally, $\hat{d}_4 = 3.075$ means that we also reject the null in the case of Sweden.

Observe that the number of rejections is intermediate between the results for the Holm and Bonferroni methods and that of the individual country results. In view of the above discussion, we find that this result is rather plausible. Furthermore, the ability of the Romano and Wolf [2005] method to detect several false hypotheses in a stepwise fashion proved instrumental in improving upon the more traditional multiple testing methods.

5 Conclusion

We have used recent advances in the multiple testing literature to make an attempt to identify those countries for which Purchasing Power Parity (PPP) held over the last century. The approach controls the multiplicity problem inherent in simultaneously testing for PPP on several time series, thereby avoiding spurious rejections. It has higher power than traditional multiple testing techniques by exploiting the dependence structure between the countries with a bootstrap approach. We use a sieve bootstrap approach to account for nonstationarity under the null hypothesis. On the other hand, our empirical results show that, plausibly, controlling for multiplicity leads to fewer rejections of the null of no PPP than if one tests the null on each single time series at some level α. Specifically, we find rejections of the null of no PPP for Mexico, Finland, Argentina, Italy, Norway and Sweden.

Several open issues remain. Hlouskova and Wagner [2006] point out that bootstrapping in a nonstationary framework is a "delicate issue." It would therefore be interesting to investigate the performance of other resampling techniques in the present problem. Consider, for instance, block bootstrapping as in Psaradakis [2006].

Obviously, the present framework is fairly general and could be applied to other macroeconomic questions such as savings-investment correlation or spot and forward exchange rates [Mark et al., 2005] that have hitherto been dealt with using panel techniques. Similarly, it is possible to accommodate problems that imply testing for cointegration.

References

Basawa, I., A. Mallik, W. McCormick, J. Reeves, and R. Taylor: 1991, 'Bootstrapping Unstable First-Order Autoregressive Processes'. *Annals of Statistics* **19**(2), 1098–1101.

Boucher Breuer, J., R. McNown, and M. S. Wallace: 2001, 'Misleading Inferences from Panel Unit Root Tests with an Illustration from Purchasing Power Parity'. *Review of International Economics* **9**(3), 482–493.

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

Brockwell, P. J. and R. A. Davis: 1991, *Time Series: Theory and Methods*. New York: Springer, 2nd edition.

Chang, Y. and J. Y. Park: 2003, 'A Sieve Bootstrap for the Test of a Unit Root'. *Journal of Time Series Analysis* **24**(4), 379–400.

Choi, I.: 2001, 'Unit Root Tests for Panel Data'. *Journal of International Money and Finance* **20**(2), 249–272.

Dickey, D. and W. Fuller: 1979, 'Distribution of the Estimators for Autoregressive Time Series with a Unit Root'. *Journal of the American Statistical Association* **74**(366), 427–431.

Froot, K. A. and K. Rogoff: 1995, 'Perspectives on PPP and Long-Run Real Exchange Rates'. In: G. Grossman and K. Rogoff (eds.): *The Handbook of International Economics*, Vol. 3. Amsterdam: Elsevier, Chapt. 32, pp. 1647–1688.

Hanck, C.: 2006, 'Cross-Sectional Correlation Robust Tests for Panel Cointegration'. *SFB 475 Technical Report Series*.

Harris, R. D. and E. Tzavalis: 1999, 'Inference for unit roots in dynamic panels where the time dimension is fixed'. *Journal of Econometrics* **91**(2), 201–226.

Hedges, L. and I. Olkin: 1985, *Statistical Methods for Meta-Analysis*. San Diego: Academic Press.

Hlouskova, J. and M. Wagner: 2006, 'The Performance of Panel Unit Root and Stationarity Tests: Results from a Large Scale Simulation Study'. *Econometric Reviews* **25**(1), 85–116.

Holm, S.: 1979, 'A simple sequentially rejective multiple test procedure'. *Scandinavian Journal of Statistics* **6**(1), 65–70.

Karlsson, S. and M. Löthgren: 2000, 'On the power and interpretation of panel unit root tests'. *Economics Letters* **66**(3), 249–255.

Lehmann, E. L. and J. P. Romano: 2005, *Testing Statistical Hypotheses*. New York: Springer, 3rd edition.

Levin, A., C. Lin, and C.-S. J. Chu: 2002, 'Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties'. *Journal of Econometrics* **108**(1), 1–24.

MacKinnon, J. G.: 1994, 'Approximate Asymptotic Distribution Functions for Unit-Root and Cointegration Tests'. *Journal of Business & Economic Statistics* **12**(2), 167–176.

MacKinnon, J. G.: 1996, 'Numerical Distribution Functions for Unit Root and Cointegration Tests'. *Journal of Applied Econometrics* **11**(6), 601–618.

Maddala, G. and S. Wu: 1999, 'A Comparative Study of Unit Root Tests with Panel Data

and a New Simple Test'. *Oxford Bulletin of Economics and Statistics* **61**(S1), 631–652.

Mark, N. C., M. Ogaki, and D. Sul: 2005, 'Dynamic Seemingly Unrelated Cointegrating Regression'. *Review of Economic Studies* **72**(3), 797–820.

Murray, C. J. and D. H. Papell: 2005, 'Do Panels Help Solve the Purchasing Power Parity Puzzle?'. *Journal of Business & Economic Statistics* **23**(4), 410–415.

Ng, S. and P. Perron: 2001, 'Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power'. *Econometrica* **69**(6), 1519–1554.

Papell, D. H.: 2002, 'The great appreciation, the great depreciation, and the purchasing power parity hypothesis'. *Journal of International Economics* **57**(1), 51–82.

Papell, D. H. and H. Theodoridis: 2001, 'The Choice of Numeraire Currency in Panel Tests of Purchasing Power Parity'. *Journal of Money, Credit and Banking* **33**(3), 790–803.

Phillips, P. C. and Z. Xiao: 1998, 'A Primer on Unit Root Testing'. *Journal of Economic Surveys* **12**(5), 423–469.

Psaradakis, Z.: 2006, 'Blockwise bootstrap testing for stationarity'. *Statistics & Probability Letters* **76**(6), 562–570.

Rogoff, K.: 1996, 'The Purchasing Power Parity Puzzle'. *Journal of Economic Literature* **34**(2), 647–668.

Romano, J. P. and M. Wolf: 2005, 'Stepwise Multiple Testing as Formalized Data Snooping'. *Econometrica* **73**(4), 1237–1282.

Said, S. E. and D. A. Dickey: 1984, 'Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order'. *Biometrika* **71**(3), 599–607.

Savin, N. E.: 1984, 'Multiple Hypothesis Testing'. In: Z. Griliches and M. Intriligator (eds.): *Handbook of Econometrics*, Vol. 2. Amsterdam: North-Holland Publishing, Chapt. 14, pp. 827–879.

Swensen, A. R.: 2003, 'Bootstrapping Unit Root Tests for Integrated Processes'. *Journal of Time Series Analysis* **24**(1), 99–126.

Taylor, A. M.: 2002, 'A Century of Purchasing-Power Parity'. *The Review of Economics and Statistics* **84**(1), 139–150.

Taylor, M. P. and L. Sarno: 1998, 'The behavior of real exchange rates during the post-Bretton Woods period'. *Journal of International Economics* **46**(2), 281–312.

Wu, Y.: 1996, 'Are Real Exchange Rates Stationary? Evidence from a Panel Data Test'. *Journal of Money, Credit and Banking* **28**(1), 54–63.