# Non-convergent incomes with a new DF-Fourier test: most likely you go your way (and I'll go mine)

Silva Lopes, Artur

ULisboa

2021

# Non-convergent incomes with a new DF-Fourier test: most likely you go your way (and I'll go mine)*

Artur Silva Lopes †
Universidade de Lisboa

This version: October 9, 2023

**Abstract**

Motivated by the purpose to assess the income convergence hypothesis, a simple new Fourier-type unit root test of the Dickey-Fuller family is introduced and analysed. In spite of a few shortcomings that it shares with rival tests, the proposed test generally improves upon them in terms of power performance in small samples.

The empirical results that it produces for a recent and updated sample of data for 25 countries clearly contrast with previous evidence produced by the Fourier approach and, more generally, they also contradict a recent wave of optimism concerning income convergence, as they are mostly unfavourable to it. This evidence appears to be particularly robust to the possibility of undetected convergence.

Keywords: income convergence; unit root tests; structural breaks.
JEL codes: O47, C22, F43.

*This is a revised version of a paper which previously circulated under the titles of *Revisiting income convergence with DF-Fourier tests: old evidence with a new test* and *Most likely you go your way (and I'll go mine): non-convergent incomes with a new DF-Fourier test.*

†Email: artursl(at)meo.pt.

# 1 Introduction

According to the income convergence hypothesis, the diminishing marginal product of capital of the neoclassical growth model implies that, in the long run, initial conditions — namely the physical and the human stocks of capital — should play no role in determining a country's *per capita* income. Therefore, in the long run, independently of those initial conditions, the *per capita* incomes of different countries should converge to an identical level.

Adopting a time series framework, this paper contains two contributions to the empirical assessment of this hypothesis: a) a simple new unit root test, of the Dickey-Fuller family, Fourier-type variety, is proposed and analysed; b) it is applied to a sample of 25 countries whose data were recently released in an updated and improved version of the Maddison database (see Bolt et al., 2018).

Robustness to general non-linearities and, in particular, to breaks in level (and/or trend) is an attractive feature of some unit root tests, specially when a long span of time is involved[1]. Fourier-type unit root tests are considered to possess this property and, when combined with the Dickey-Fuller approach, they become also simple to implement. They encompass any type of non-linearity in the deterministic component of the series, they contain the damages on power resulting from any kind of break and they are particularly suited to smooth breaks. Moreover, they do not require any knowledge about neither the nature nor the number and the date(s) of the break(s). The breaks can be left unspecified and it is not even necessary to estimate them. An important source of leaks in power (see, e.g., Lee and Strazicich, 2001) is therefore avoided.

The test proposed in this paper adds to this flexibility some further benefits: besides particularly adapted to the income convergence testing problem and with improved power properties, two shortcomings of previous Fourier-type tests are also overcome. These concern the disregard for the endogeneity of the selection process for the frequency parameter and for the (pre-testing) nature of those versions, both liable to contribute to size distortion problems. These benefits are partially supported at the price of a simple null hypothesis that does not allow the presence

[1]Greasley and Oxley (1997) and Li and Papell (1999) pioneered this approach in the income convergence testing literature and they have found evidence restoring some credibility to the hypothesis.

of breaks, i.e., consisting only on the presence of a unit root, as in the tests by Zivot and Andrews (1992).

As regards the power properties, they are improved mostly through the (*min*) form adopted for the test statistic, which appears particularly useful for testing the null hypothesis of a unit root against non-linear alternatives (see, e.g., Kiliç, 2011). Together with the simple no-break null hypothesis, this form is also useful to overcome the previously mentioned endogeneity problem, and particularly to circumvent the so-called "Davies problem" concerning the frequency parameter[2] , i.e., its lack of identification under the null hypothesis. A narrowing of the length of the interval for the set of admissible values for the frequency parameter also contributes to an improvement in power.

Compared to a close surrogate of the Enders and Lee (2012a) Dickey-Fuller Fourier-type test (FDF), the new test has better size and power properties. In terms of power, the new test clearly dominates the standard DF and the FDF tests, particularly when the sample size is relatively large (and provided there are really breaks in the DGP). However, it is not free from a few shortcomings, which it shares with its rivals: a) low power in some cases when the sample size is small; b) over-rejections of the null hypothesis in some cases where breaks coexist with it, i.e., a few observations of the "converse Perron disease" [3].

This incomplete reliability is a small cost to bear for the containment of the generalized catastrophic power losses incurred by DF tests when breaks are present. Moreover, the "over-rejections" problem appears only in a few cases, almost coinciding with those where DF tests are also affected by the same problem [4]. The main objective of the new test is achieved: its general power performance is very good, power frequently becoming (much) larger than the one of the DF test for

[2]The "Davies problem" refers to the impossibility of applying usual asymptotic theory to derive the distribution of a test statistic because there is a nuisance parameter that is not identified — since it does not even exist — under the null hypothesis.

[3]The loss in power that usual unit root tests incur when there is a break in a stationary process is now well known due to the work of Pierre Perron (1989). Spurious rejections of the (true null) unit root hypothesis may occur as well in certain cases where the data is characterized also by the presence of breaks.

[4]Mentioning "over-rejections" is imprecise and "size distortions" is an even more inadequate expression to characterize the size properties of the test. Indeed, it must be stressed that the possibility of breaks is not allowed in the null hypothesis, this consisting only of the unit root. In this stricter sense, the size properties of the new test are excellent.

the corresponding no-break case. Hence, weighing size and power properties, the proposed test appears to be the most balanced of the three tests under scrutiny.

Notice also that to achieve an improved reliability, as in the test proposed by Kim and Perron (2009), one would have to know that one (or more) break(s) did really occur, the number of those breaks, and their exact date location and specific form, AO or IO (meaning additive outlier and innovative outlier, respectively). I believe that all this information is seldom available to practitioners and that an "all-terrain" test may be often useful.

Anyway, armed with the knowledge about the performance of the new test, I will analyse thoroughly the cases of rejection of the unit null, particularly those that appear to be surprising. Resorting to additional quantitative and qualitative information, solid conclusions appear to be reached almost everywhere. The exceptions are two countries for which there is mixed evidence, allowing only a tentative, somewhat shaky decision.

Recent empirical assessments of income convergence, using both a panel data approach and a time series/unit root testing one but allowing for breaks in the series tended to restore some of the hypothesis' initial aura, and went much further than the favourable initial results including in the converging group some countries not usually viewed as advanced. In this paper I find that such an optimistic view is not well sustained empirically. That is, even allowing for multiple breaks in the series of relative incomes and making an extra effort to maximize the power of the tests, i. e., to reject non-convergence, the favourable evidence is relatively weak, thereby tending to agree with the general results of a second generation of tests. That is, most of the results do not lend support to the hypothesis.

The remainder of this paper contains the following material. In the next section the arguments for level stationarity rather than trend stationarity as the alternative hypothesis in unit root tests for income convergence are briefly reviewed. Section 3 reviews the current versions of Fourier-type Dickey-Fuller tests for unit roots. In section 4, after adapting and criticizing those versions I propose the new test statistic. Its size and power performance in finite samples is analysed in the following section. Section 5 contains also an examination of a possible testing sequence, the union of rejections of the new test with the standard DF test. Section 6 presents the empirical results and the final section contains a comparison with recent empirical evidence and a further detailed discussion of the results.

## 2 Unit root tests: LSP rather than TSP

Let me represent the logarithms of *per capita* output for countries $i$ and $j$ with $y_{i,t}$ and $y_{j,t}$, respectively, the latter associated with the technological leader. As argued in Lopes (2016), the most adequate definition of income convergence, provided in Bernard and Durlauf (1996), requires that the income discrepancy $y_{i,t} - y_{j,t}$ is a level stationary process (LSP) rather than a trend stationary (TSP) one.

Strictly speaking, as the requirement is that the long run (MSE) optimal forecasts for the logs of both countries should not diverge, i. e.,

$$\lim_{k\to\infty} \mathrm{E}(y_{i,t+k} - y_{j,t+k} | \, \mathcal{F}_t) = 0, \tag{1}$$

with $\mathcal{F}_t$ denoting the set all information available at time $t$, stationarity around zero could be imposed. However, as the two economies may differ in important structural characteristics — rates of population growth, for instance —, a non-zero mean for the output gap is admissible. In Pesaran (2007) and in some further literature, the zero mean condition is considered as overly stringent. His simple growth model requires that several deep structural parameters must be identical for both economies; for instance, the savings rate and the steady-state growth rate of employment must be the same. Therefore, it is very frequently disregarded.

Transposed to the (DF) unit root testing framework, this condition implies that while a constant term is admissible in the regression equation, a (linear) trend term is not. Its inclusion would allow detecting the existence of *catching-up*, which is a weaker notion of convergence, but it is ruled out by equation (1). Rather clearly, the presence of a trend in the income discrepancy would mean that its long run forecasts would diverge instead of converging.

Moreover, as the power of unit root tests decreases as deterministic regressors are added to the test equation, the simple omission of the usual trend term in convergence tests, alone, is liable to improve their performance. Finally, as shown in Lopes (2016), this omission guarantees that convergence tests are consistent when the income discrepancy is a TSP, i. e., when the divergence is dominated by the presence of a (linear) deterministic trend (and deviations around that trend are stationary). Formally, it assures that

$$\lim_{T\to\infty} \Pr\left[\text{rejecting convergence} |\ (y_{i,t} - y_{j,t}) \sim TSP\right] = 1.$$

# 3 Fourier-type Dickey-Fuller tests: an overview

Although there are now several unit root tests incorporating the flexible Fourier approximation, due to its simplicity the most promising version appears to be the Dickey-Fuller one, proposed in Enders and Lee (2012a) and improved in Omay (2015) [5].

To simplify the notation, let me represent the income discrepancy or gap simply with $y_t$ $(= y_{i,t} - y_{j,t})$. The most general version of these tests departs from the basic equation

$$y_t = d(t) + \rho y_{t-1} + \gamma t + \epsilon_t,$$

where $d(t)$ denotes a general deterministic function of the time index, $t = 1, 2, \ldots, T$, and $\epsilon_t$ is assumed as $iid(0, \sigma^2)$. As usual, interest lies in testing the null hypothesis of a unit root ($\rho = 1$). But the function $d(t)$ is indeed very general because it encompasses any type of non-linearity and particularly one or several breaks of any kind (particularly when they are gradual or smooth[6]). This is precisely the strongest and most appealing feature of these tests: the nature, the number and the date(s) of the break(s) can be left unspecified; it is not necessary to know them *a priori* and it is not even necessary to know whether they really exist. Moreover, in case they do exist, it is also not even necessary to estimate them, avoiding any negative contamination arising from the estimation error. However, sometimes it is acknowledged that sudden and sharp breaks may require the traditional approach with dummy variables because the approximation works better with gradual

[5] The first version of these Fourier-type tests was proposed by Enders and Lee (2004) in the framework of Lagrange Multiplier tests but it lacks interest for the case that we study here because it applies only to trending time series. A KPSS-type, stationary test, was proposed in Becker, Enders and Lee (2006), and a DF-GLS-type, unit root test, is analysed in Rodrigues and Taylor (2012).

[6] When the breaks are abrupt the approximation that will be mentioned below can be poor. However, there is almost no research about this topic.

changes than with sharp ones.

This flexibility and robustness to breaks is achieved approximating the function $d(t)$ with a Fourier series expansion

$$d(t) = \alpha_0 + \sum_{k=1}^{n} \alpha_k \sin\left(\frac{2\pi kt}{T}\right) + \sum_{k=1}^{n} \beta_k \cos\left(\frac{2\pi kt}{T}\right),\ n \leq \frac{T}{2},$$

where $n$ denotes the number of approximating frequencies, $k$ is used to index the frequencies and $T$ represents the sample size.

In practice, using many frequencies will likely provoke an over-fitting problem and can lead to a substantial power loss. Therefore, at most two frequencies should be considered but the most frequent recommendation is to use only one.

$$\Delta y_t = c_1 + c_2 t + \phi y_{t-1} + c_3 \sin\left(\frac{2\pi kt}{T}\right) + c_4 \cos\left(\frac{2\pi kt}{T}\right) + u_t, \tag{2}$$

where $\phi = \rho - 1$, $k$ now denotes the single selected frequency and $u_t$ represents a zero mean stationary error term[7].

The asymptotic distribution of the unit root test statistics is invariant to the values of $c_3$ and $c_4$ but depends on the value of $k$. Since breaks push the spectral density function of the series towards zero, the usual recommendation is that the value of $k$ should be low. Sometimes the value $k = 1$ is mentioned as particularly adequate, as a reasonable approximation to many cases but, in practice, a selection/estimation problem now emerges. The initial break specification problem, consisting of determining the shape of the break(s), its date(s) and its number, is transformed into one of the selection of the particular frequency.

The most frequent solution consists of choosing $k$ from a small set of (small) integer values. Also, since the usual DF tests emerges as a special case when there is no non-linear deterministic component, a joint procedure to estimate $k$ and to decide which test to employ is recommended in Enders and Lee (2012a):

1. estimate $k$ using a grid search procedure over all integer values in the interval

[7] In this particular case with a single frequency, $\alpha_1$ and $\beta_1$ simply represent the amplitude and the displacement of the sinusoidal component.

$[1, k_{MAX}]$. Usually $k_{MAX} = 5$ and

$$\widehat{k} = \arg\min_k SSR(k),$$

$SSR$ denoting the sum of squared residuals of equation (2). As is usual in DF test regressions, this equation may require augmentation with lags of $\Delta y_t$ to whiten the residuals.

2. Perform a pre-test for non-linearity testing $H_0 : c_3 = c_4 = 0$ *vs.* $H_1 : c_3 \neq 0 \vee c_4 \neq 0$ using the usual $F$-statistic. Small sample conservative critical values, i. e., that are valid when the unit root null is imposed, are available in Enders and Lee (2012a)[8].

3. Decide which test to use on the basis of the previous test. In case the previous null hypothesis is rejected, tables of small sample critical values for the "with trend" and "no-trend" cases are also available in Enders and Lee (2012a) for $k \in \{1, 2, 3, 4, 5\}$. Otherwise, the usual ("linear") DF test should be performed.

Although the possibility of fractional frequencies was initially entertained (see, e.g., Enders and Lee (2004)), it was subsequently abandoned until it was recently recuperated by Christopoulos and Leon-Ledesma (2011) and by Omay (2015). Actually, a fractional frequency can provide a better fit to the data, i.e., a better approximation to the non-linear deterministic component, and hence it may improve substantially the power of the tests. Indeed, the simulation study by Nordström (2018) indicates that the tests allowing only integer frequencies can be completely powerless (i.e., have zero power) when the actual frequency in the DGP is fractional. This corroborates the idea that an incorrectly specified break can be as harmful to the properties of unit root tests as simply neglecting its presence.

On the other hand, fractional frequencies are considered to be better than integer ones to capture breaks occurring near the extremes of the sample. Therefore, hereafter I will consider the case where fractional frequencies are allowed. Omay

[8]Since it is a function of $\widehat{k}$ and since minimizing the sum of squared residuals is equivalent to maximizing this $F$-statistic, it is also sometimes denoted with $\max F(\widehat{k})$.

(2015) imposed $k_{MAX} = 2$ and tabulated the small sample distributions only for $k \in \{1.1, 1.2, 1.3, \ldots, 1.9\}$.

# 4 A simple proposal

The first task is to adapt the previous tests to the income convergence problem. The first and most obvious modification consists of dropping the linear trend term from equation (2). But the specification of the deterministic component is a particularly involved issue in these tests.

Since the parameter $k$ determines the frequency (and so the number) of cycles allowed over the span of the sample, for the typical samples of convergence tests, sized around $T = 60$, $k_{max} = 5$ appears clearly excessive, implying that relatively short-lived cycles, 12 years long only, would represent the "long-run". Values of 3 or 4, at most, appear more reasonable. And since an empirical analysis may provide useful insights, examining the estimated deterministic component of a few series may be helpful to establish a reasonable upper limit on the range of $k$.

Actually, the graphical representation of $\widehat{d(t)}$ shows that simply excluding the linear trend is not sufficient. Maintaining $k_{max} = 5$, in some cases the estimated frequency is precisely 5, which appears as clearly excessive, producing a deterministic component seemingly over-fitting the actual series, hardly justifiable, meaningless and possibly distorting inference.

This is the case, for instance, for Poland, whose gap to the technological leader, the US, is represented in figure 1. Using the estimation method of the previous section (with $k_{MAX} = 5$) produces precisely $\widehat{k} = 5$. Therefore, in figure 1, besides the gap for Poland the graph exhibits the estimated function

$$\widehat{c_1} + \widehat{\alpha_1} \sin\left(\frac{2\pi 5t}{T}\right) + \widehat{\beta_1} \cos\left(\frac{2\pi 5t}{T}\right),$$

which is represented with `det_compo`. One immediately wonders what sort of economic mechanism could have generated such a regular cyclical process to approximate the gap from the technology leader.

So many changes in level appear to be a rather implausible approximation to the gap. Since relatively short cycles are allowed to represent the long-run, the

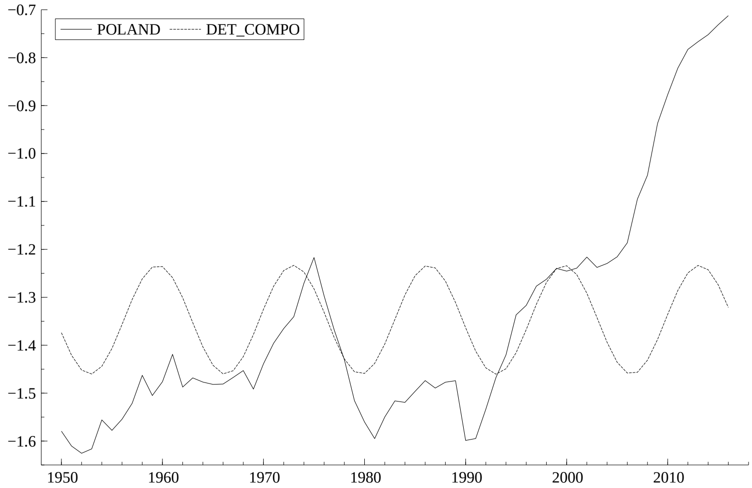


Figure 1: The output gap and the estimated deterministic component for Poland ($\widehat{k} = 5$)

estimated function contains 10 turning points, i.e., 10 breaks in level (that can even be seen as changes in the sign of a trend), which is clearly excessive for a series with only 67 observations. In case the deviations around this function behave as a stationary process, then the gap series could be called a "snake stationary process", a process that is devoid of any economic meaning. Moreover, such a process is not only unexplainable but it is also slippery to forecast, the in-sample fitting transmitting a misleading level of confidence that is completely erroneous; and indeed we can even observe already in the figure the last observations behaving rather differently, escaping completely from the snake like pattern and rapidly approaching a value much closer to zero.

Moreover, the case of Poland is not the only one in our data. Three more cases — Finland, Israel and Sweden — could be presented as illustrations of the same problem, totaling 4 out of the 24 countries, i.e., 16.6(6)% of the cases, and hence far from a negligible fraction. Such a large estimate of $k$ entails removing from the data information that appears to be unrelated with the long run or zero frequency but that, in practice, may be relevant to determine correctly their long run properties; in particular, the effect of shocks that are only temporary, whose persistence is low, appears to be incorrectly removed.

As previously mentioned, it is well known that structural breaks distort the spectral density function of time series at low frequencies. This justifies the usual recommendation that $k = 1$ or $k = 2$ (for the integer case) should be sufficient to handle the large majority of breaks. Since, on the other hand, one cannot find any argument supporting the use of $k_{MAX} = 5$, this upper limit appears as somewhat arbitrary, overly cautious, wasting power in irrelevant cases and, as previously illustrated, sometimes producing incredible attractor lines. As Enders and Lee (2012, p. 576) emphasize, "*there is little point in claiming that a series reverts to an arbitrarily evolving mean*". To this, one might add that there is also little point in claiming that a series *does not* revert to an arbitrarily evolving mean. Therefore, still based on some size concerned orientation, the first modification I propose is to restrict the upper limit of $k$ to 3, i.e., to consider the set of possible values for $k$ as $K = \{0, 0.1, 0.2, 0.3, \ldots, 3\}$. For the available sample size, with $T = 67$, this means that the shortest cycles representing the long-run are approximately 22 years long, which appears to be a reasonable assumption.

Notice also that, besides including fractional values between 0 and 1, the lower limit is set equal to zero – that is, in practice no trigonometric terms —, so that the standard linear case is also clearly included.

The main origin of the proposed test, however, relies on the feebleness of previously available critical values: besides referring to rather different sets for $k$, these assume that the selected frequency is known a priori, i. e., that it is given exogenously to the data, and that it is also previously known that there is a non-linear, preferably smooth component. In other words, both the data-dependent nature of the selection process for $k$ and the pre-test procedure for the test are completely neglected in published critical values.

More precisely, as the distribution of the test statistic depends on the frequency parameter $k$, published critical values are obtained as if it is previously known: its value is fixed in advance in the test regressions of the replications employed to produce the critical values and these regressions are run for each value belonging to the assumed set (which I denote with $K$). Therefore, although similar to the usual criticism made to Perron's (1989) initial work, this one here is much stronger: in complete contrast with the assumption made to tabulate critical values, $k$ needs to be estimated from the data, making the data-exogeneity assumption clearly inadequate.

Moreover, available critical values do not accommodate the pre-test sequence, that is, they implicitly assume that the presence of a break (or of any type of non-linear component) is certain, thereby *a priori* excluding an option that may be followed in practice: the implementation of a linear (standard) Dickey-Fuller test. Since both the usual DF critical values and those of the Fourier-type variety neglect this testing sequence, it is hard to believe that they produce tests that are free from size distortion problems. In this regard, notice that my proposal for the inclusion of the value 0 (zero) in the admissible set $K$ $(0 \in K)$ permits, from the outset, the case where there is no-break (or, more generally, no non-linear component). This possibility is actually more than allowed under the simple null hypothesis, which tests only the unit root while imposing that $c_3 = c_4 = 0$, i.e., breaks are allowed only under the alternative.

To circumvent the previous problems (and besides the previous enlargement of $K$) my proposal is to consider instead the minimum of the Fourier-type DF test statistics over the set of admissible values for $k$, i.e.,

$$\tau_{min}^{FDF} = \min_{k \in K} t_{\phi},$$

where $t_{\phi}$ denotes the $t$-ratio for $\phi$ in equation (2) without the linear trend term, as explained, and $K = \{0, 0.1, 0.2, \ldots, 3\}$. Besides relatively simple to compute, this test statistic neither assumes that $k$ is known *a priori* nor requires its previous estimation; instead, in the simplest case, this estimation is simultaneous with the computation of the test statistic. Moreover, in line with the flexibility of the Fourier approach, it also dispenses with the pre-test $F$ statistic. However, as is usual, the test regression may need to be augmented with lags of the dependent variable. Concerning this issue, I assume that this augmentation occurs only after obtaining the minimizer, i. e., after the estimation of $k$.

Quite naturally, this estimation occurs simultaneously with the calculation of the test statistic, that is, test regressions similar to (2) but without the linear trend term are run for all values of $k \in K$, and the estimated frequency is the one that produces the lowest value for $t_\phi$, i. e.,

$$\widehat{k} = \arg\min_{k \in K} t_\phi.$$

This idea is inspired in Zivot and Andrews' (1992) test, who overcome the well known endogeneity problem in the selection of the break date in Perron's (1989) work, proposing a minimization algorithm over all possible break fractions. As in their work, the procedure proposed here possesses two clear advantages:

a) it overcomes the Davies problem, estimating a parameter ($k$) that is unidentified under $H_0$ (as it really not even exists in that case), and

b) it adequately takes account of that estimation, i.e., it explicitly addresses the effect of that endogeneity, of that straightforward data dependency, on the distribution of the test statistic.

Thus, while previous tests are conditional upon a previously selected frequency, the $\tau_{min}^{FDF}$ test is a unconditional unit root test. Furthermore, it also dispenses with the test for linearity, suppressing another source for size distortion. Further still, again as in Zivot and Andrews (1992), the (min) form adopted for the statistic provides the least favourable result for the null hypothesis obtained with $t_\phi$, and hence it is expected to increase its rejection frequency when it is false, i. e., to produce a powerful test.

Table 1 contains some simulated critical values for this test. In all simulations the data generating mechanism is a simple random walk process, $y_t = y_{t-1} + \epsilon_t$, $\epsilon_t \sim iid\mathcal{N}(0,1)$ and the percentiles were obtained from simulations with $50,000$ replications. Sample sizes with $T = 50, 100, 200$ and 1000 observations were considered, this last case serving to approximate the asymptotic distribution. In spite of the focus in the case where, beyond the trigonometric terms, the test regression contains only the intercept term, the critical values for the other two usual cases are also presented. The three statistics are denoted with $\tau_{nc,min}^{FDF}$, $\tau_{c,min}^{FDF}$ and $\tau_{ct,min}^{FDF}$ for the no constant, no trend, and with trend cases, respectively.

Table 1. Critical values for the $\tau_{min}^{FDF}$ test

| $T$ | 1% | 5% | 10% |
|---|---|---|---|
| | no constant case, $\tau_{nc,min}^{FDF}$ | | |
| 50 | $-4.57$ | $-3.93$ | $-3.61$ |
| 100 | $-4.43$ | $-3.87$ | $-3.58$ |
| 200 | $-4.40$ | $-3.83$ | $-3.55$ |
| 1000 | $-4.35$ | $-3.82$ | $-3.55$ |
| | no trend case, $\tau_{c,min}^{FDF}$ | | |
| 50 | $-5.22$ | $-4.51$ | $-4.19$ |
| 100 | $-4.98$ | $-4.40$ | $-4.11$ |
| 200 | $-4.90$ | $-4.35$ | $-4.07$ |
| 1000 | $-4.84$ | $-4.30$ | $-4.03$ |
| | with trend case, $\tau_{ct,min}^{FDF}$ | | |
| 50 | $-5.70$ | $-4.99$ | $-4.65$ |
| 100 | $-5.40$ | $-4.82$ | $-4.53$ |
| 200 | $-5.30$ | $-4.76$ | $-4.48$ |
| 1000 | $-5.22$ | $-4.69$ | $-4.43$ |

# 5 Finite sample performance

To assess the finite sample performance of the proposed test two benchmarks were used: the standard DF test and the FDF test, a logical and natural competitor, considering only integer frequencies, albeit from 1 to 5, as in Enders and Lee (2004, 2012a, 2012b) and Rodrigues and Taylor (2012); i. e., for the FDF test, $K = \{1, 2, 3, 4, 5\}$. To circumvent the dependence of this test on the frequency parameter $k$, its estimate was neglected in the tabulation of the test statistic. That is, a Gaussian driftless random walk performed the role of the DGP and the critical values were collected regardless of the estimate $\widehat{k}$; actually, in each regression, these estimates were not even retained. These critical values are presented in table A of the Appendix.

The DGP of the Monte Carlo experiments is

$$y_t = \rho y_{t-1} + \alpha_1 \sin\left(\frac{2\pi kt}{T}\right) + \beta_1 \cos\left(\frac{2\pi kt}{T}\right) + \varepsilon_t, \quad 0 \leq \rho \leq 1$$

with $\varepsilon_t \sim iid\mathcal{N}(0,1)$, $k = 0.4, 0.8, 1.2, 1.6$ and 2 and each experiment consisted of $10,000$ Monte Carlo replications. Sample sizes with $T = 50, 100$ and 200 observations were considered. In all the cases, the reported rejection frequencies are based in 5% nominal critical values. While for the proposed test and for the FDF test these critical values were derived from the specific sample sizes, the usual and popular ($-2.86$) asymptotic critical value was used for the DF test (including a constant as the only deterministic term).

The experimental design follows those of Enders and Lee (2004, 2012a, 2012b) and Su and Nguyen (2013). Besides the no-break, linear case, with $(\alpha_1, \beta_1) = (0, 0)$, the pairs $(0, 3), (3, 0), (0, 5)$ and $(3, 5)$ were also used to generate data but notice that these exceed the simple unit root null hypothesis.

## 5.1 Rejection behaviour when the unit root is present

Percentage rejections for the true null hypothesis (with $\rho = 1$) for the three tests are presented in table 2 [9]. For the simple, exclusively linear (no-break) case of the unit root null, the size performance of all the tests is very good, with empirical size barely deviating from the nominal 5%.

However, when a non-linear component is added to the unit root, in general a situation of under-rejection emerges. This is most frequent and severe for the standard DF test but it also occurs very often for the two Fourier-type tests. Overall, the proposed test is the least severely affected by this problem.

The most serious problem for the integrated plus break case is, however, one of rejections in excess of the nominal size, the series appearing to be level stationary. This occurs mostly when $k$ is very low (0.4 and 0.8) but also when $k = 1.2$. While the problem is extremely severe in some cases, with "size" estimates rapidly attaining 100% of rejections when $T$ is only 50, it vanishes completely for $k > 1.2$. This is the well known "converse Perron phenomenon", firstly reported by Leybourne, Mills and Newbold (1998, LMN) for standard DF tests: a break in an I(1) series confounds the unit root test and it is considered as I(0).

What appears to be new or previously unreported (as far as I know) concerning this "phenomenon" is that:

[9] Notice that this table contains also the results for a different test procedure, the "UR" test, that will be presented only later.

a) it can also affect the size performance of tests designed to be robust to breaks (in terms of power), i.e., the Fourier-type tests;

b) it affects also standard DF tests when breaks are smooth (recall that the case reported in LMN is for abrupt breaks);

c) it is even more severe for the two Fourier-type tests than for the standard DF test, both in terms of the possible cases and of the magnitude of the size distortion.

For the proposed test, the origin of the problem is inherited from the Zivot and Andrews (1992) approach: the form adopted for the statistic to escape the endogeneity trap prohibits the presence of non-linearities under the null hypothesis as in the Zivot and Andrews test. Nevertheless, the reason why this problem inflates "size" only in some cases and not in others remains to be explained [10]. Therefore, although not surprising, this evidence is challenging: robustness to breaks appears to be hard to obtain in terms of size properties. As will become clear below, the flexibility of the Fourier approach is very useful in terms of power but methods to detect breaks and to handle them still cannot be dismissed when size control is the major concern. Moreover, these results imply that a particularly judicious scrutiny of empirical rejections is needed.

[10] Recall that since the proposed test allows the presence of breaks precisely only under the alternative hypothesis, I hold the view that one can hardly refer to this problem as one of "size distortions" or of "spurious rejections"

Table 2. Rejection frequencies (in %) for the unit root case (5% nominal size)

| | | | $T=50$ | | | | $T=100$ | | | | $T=200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $\alpha_1$ | $\beta_1$ | DF | FDF | $\tau_{c,min}^{FDF}$ | UR | DF | FDF | $\tau_{c,min}^{FDF}$ | UR | DF | FDF | $\tau_{c,min}^{FDF}$ | UR |
| | 0 | 0 | 5.56 | 4.74 | 4.85 | 6.40 | 5.85 | 4.84 | 4.97 | 5.25 | 5.45 | 4.64 | 5.09 | 5.19 |
| 0.4 | 0 | 3 | 98.18 | 99.61 | 99.72 | 99.68 | 99.94 | 100.0 | 100.0 | 100.0 | 99.97 | 100.0 | 100.0 | 100.0 |
| | 3 | 0 | 0.01 | 0.00 | 2.14 | 1.71 | 0.00 | 0.00 | 1.75 | 0.93 | 0.01 | 0.00 | 1.55 | 0.67 |
| | 0 | 5 | 99.98 | 100.0 | 100.0 | 100.0 | 99.97 | 100.0 | 100.0 | 100.0 | 99.98 | 100.0 | 100.0 | 100.0 |
| | 3 | 5 | 99.97 | 100.0 | 100.0 | 100.0 | 99.95 | 100.0 | 100.0 | 100.0 | 99.97 | 100.0 | 100.0 | 100.0 |
| 0.8 | 0 | 3 | 0.01 | 0.00 | 1.32 | 1.01 | 0.00 | 0.00 | 1.29 | 0.74 | 0.01 | 0.00 | 1.41 | 0.57 |
| | 3 | 0 | 3.24 | 50.35 | 97.98 | 97.26 | 25.16 | 92.81 | 100.0 | 100.0 | 86.37 | 99.98 | 100.0 | 100.0 |
| | 0 | 5 | 0.01 | 0.00 | 1.24 | 0.96 | 0.00 | 0.00 | 1.29 | 0.74 | 0.01 | 0.00 | 1.41 | 0.57 |
| | 3 | 5 | 0.01 | 0.00 | 1.24 | 0.96 | 0.00 | 0.00 | 1.29 | 0.74 | 0.01 | 0.00 | 1.41 | 0.57 |
| 1.2 | 0 | 3 | 0.01 | 12.29 | 22.61 | 18.27 | 0.10 | 42.89 | 93.60 | 86.29 | 0.11 | 92.78 | 100.0 | 100.0 |
| | 3 | 0 | 0.02 | 0.09 | 3.97 | 3.35 | 0.01 | 0.00 | 3.29 | 2.24 | 0.05 | 0.00 | 2.18 | 1.36 |
| | 0 | 5 | 0.11 | 51.74 | 72.25 | 64.71 | 0.12 | 97.24 | 100.0 | 100.0 | 0.13 | 100.0 | 100.0 | 100.0 |
| | 3 | 5 | 0.00 | 0.00 | 0.63 | 0.47 | 0.00 | 0.01 | 0.72 | 0.43 | 0.00 | 0.02 | 0.73 | 0.31 |
| 1.6 | 0 | 3 | 0.01 | 0.00 | 3.47 | 2.91 | 0.08 | 0.00 | 3.86 | 2.69 | 0.00 | 0.00 | 4.32 | 2.80 |
| | 3 | 0 | 0.00 | 0.01 | 1.07 | 0.81 | 0.00 | 0.00 | 0.91 | 0.55 | 0.02 | 0.01 | 0.62 | 0.31 |
| | 0 | 5 | 0.00 | 0.00 | 2.41 | 1.98 | 0.01 | 0.00 | 2.99 | 1.88 | 0.00 | 0.00 | 4.18 | 2.62 |
| | 3 | 5 | 0.00 | 0.00 | 0.46 | 0.40 | 0.00 | 0.00 | 0.63 | 0.32 | 0.00 | 0.00 | 0.60 | 0.30 |
| 2.0 | 0 | 3 | 0.00 | 1.27 | 0.77 | 0.53 | 0.00 | 1.17 | 0.55 | 0.26 | 0.00 | 1.16 | 0.60 | 0.25 |
| | 3 | 0 | 0.02 | 1.27 | 0.74 | 0.52 | 0.00 | 1.17 | 0.66 | 0.37 | 0.00 | 1.16 | 0.60 | 0.43 |
| | 0 | 5 | 0.00 | 1.27 | 0.49 | 0.38 | 0.00 | 1.17 | 0.37 | 0.15 | 0.00 | 1.16 | 0.43 | 0.17 |
| | 3 | 5 | 0.00 | 1.27 | 0.32 | 0.22 | 0.00 | 1.17 | 0.23 | 0.09 | 0.01 | 1.16 | 0.34 | 0.14 |

Note: the DGP is $y_t = y_{t-1} + \alpha_1 \sin\left(\frac{2\pi kt}{T}\right) + \beta_1 \cos\left(\frac{2\pi kt}{T}\right) + \varepsilon_t$ $(\rho = 1)$, with $\varepsilon_t \sim iid\mathcal{N}(0,1)$.

## 5.2 Power

Table 3, containing the finite sample power results for $\rho = 0.9$, clearly justifies the employment of the proposed test and a preference over the FDF test. Indeed, with a few exceptions, the proposed test generally dominates the two competitors in terms of estimated power performance, sometimes even overwhelmingly, particularly when $T = 200$ but also when $T = 100$.

One of the exceptions is the standard, purely linear (non-break) unit root case, where the standard DF test unsurprisingly dominates and the new test is the worst of the three. This means that the inclusion of the value zero in the set of admissible parameter values for $k$ is insufficient to warrant a reasonable performance when there is no non-linear component. This also suggests that a union of rejections strategy between the new test and the DF test may be beneficial, at least in that case.

The other exceptions are some of the cases when $T = 50$ only, where the power of the proposed ($min-$)test is low, and in few cases even lower than the power of the FDF test. Anyway, on one hand, provided that a non-linear component is present in the data, the new test is always more powerful than the DF test, sometimes much more so. On the other hand, when $T$ grows from 50 to 100 the growth of power of the proposed test is usually much faster than those of the two rival tests. Therefore, although not uniformly, the dominance of the new test is very clear.

Moreover, notice that in many break cases the estimated power of the new test is even larger than that of the DF test for the same sample size with no break case. Thus, the presence of breaks is often a very powerful boost to the power of the *min*-test. Furthermore, the results also suggest that the proposed test is always consistent, but the same cannot be said in some cases for the FDF test (for instance, when $k = 0.4$ and $(\alpha_1, \beta_1) = (3, 5)$) and, much less surprisingly, in most break cases for the DF test.

Table 3. Power estimates for 5% nominal tests (in %)

| | | | $T=50$ | | | | $T=100$ | | | | $T=200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $\alpha_1$ | $\beta_1$ | DF | FDF | $\tau^{FDF}_{c,min}$ | UR | DF | FDF | $\tau^{FDF}_{c,min}$ | UR | DF | FDF | $\tau^{FDF}_{c,min}$ | UR |
| | 0 | 0 | 13.34 | 10.98 | 7.94 | 12.50 | 32.94 | 21.95 | 13.72 | 25.33 | 86.52 | 64.38 | 43.93 | 76.94 |
| 0.4 | 0 | 3 | 0.06 | 0.00 | 11.17 | 9.12 | 0.06 | 0.00 | 63.50 | 51.16 | 0.07 | 0.00 | 99.81 | 99.28 |
| | 3 | 0 | 5.14 | 0.32 | 5.97 | 6.52 | 4.77 | 0.29 | 10.55 | 7.57 | 0.58 | 0.36 | 31.13 | 18.96 |
| | 0 | 5 | 0.06 | 0.00 | 33.12 | 28.99 | 0.07 | 0.00 | 99.57 | 98.88 | 0.07 | 0.00 | 100.0 | 100.0 |
| | 3 | 5 | 0.08 | 0.03 | 35.21 | 30.71 | 0.07 | 0.00 | 98.78 | 97.31 | 0.08 | 0.00 | 100.0 | 100.0 |
| 0.8 | 0 | 3 | 0.09 | 0.02 | 8.23 | 6.66 | 0.09 | 0.09 | 69.46 | 58.51 | 0.09 | 0.52 | 99.95 | 99.83 |
| | 3 | 0 | 0.03 | 0.00 | 11.13 | 9.19 | 0.02 | 0.00 | 13.08 | 8.03 | 0.02 | 0.00 | 29.89 | 17.23 |
| | 0 | 5 | 0.09 | 0.00 | 22.99 | 19.53 | 0.09 | 0.00 | 99.92 | 99.59 | 0.09 | 0.00 | 100.0 | 100.0 |
| | 3 | 5 | 0.01 | 0.00 | 2.97 | 2.36 | 0.08 | 0.00 | 81.18 | 73.08 | 0.09 | 0.00 | 100.0 | 100.0 |
| 1.2 | 0 | 3 | 0.08 | 84.28 | 57.99 | 52.00 | 0.05 | 99.11 | 98.45 | 96.19 | 0.05 | 94.59 | 100.0 | 100.0 |
| | 3 | 0 | 0.01 | 0.00 | 15.90 | 13.03 | 0.01 | 0.00 | 21.75 | 14.41 | 0.01 | 0.00 | 40.85 | 26.25 |
| | 0 | 5 | 0.12 | 99.99 | 99.33 | 98.70 | 0.06 | 100.0 | 100.0 | 100.0 | 0.05 | 100.0 | 100.0 | 100.0 |
| | 3 | 5 | 0.11 | 88.01 | 91.93 | 87.84 | 0.12 | 100.0 | 100.0 | 100.0 | 0.13 | 100.0 | 100.0 | 100.0 |
| 1.6 | 0 | 3 | 0.02 | 0.00 | 11.41 | 9.44 | 0.03 | 0.00 | 75.06 | 63.00 | 0.04 | 0.00 | 99.97 | 99.99 |
| | 3 | 0 | 0.02 | 0.00 | 4.84 | 3.90 | 0.02 | 0.00 | 22.61 | 15.16 | 0.01 | 0.00 | 49.26 | 33.20 |
| | 0 | 5 | 0.02 | 0.00 | 48.29 | 42.70 | 0.03 | 0.00 | 99.95 | 99.87 | 0.04 | 0.00 | 100.0 | 100.0 |
| | 3 | 5 | 0.00 | 0.00 | 1.18 | 0.92 | 0.01 | 0.00 | 25.62 | 15.72 | 0.03 | 0.00 | 100.0 | 100.0 |
| 2.0 | 0 | 3 | 0.00 | 3.84 | 3.59 | 2.86 | 0.00 | 38.91 | 36.87 | 26.69 | 0.04 | 99.91 | 99.52 | 98.72 |
| | 3 | 0 | 0.02 | 6.19 | 3.74 | 3.02 | 0.02 | 43.09 | 21.69 | 14.47 | 0.01 | 83.70 | 57.10 | 41.13 |
| | 0 | 5 | 0.00 | 5.97 | 8.53 | 7.06 | 0.02 | 93.84 | 89.95 | 83.85 | 0.04 | 100.0 | 100.0 | 100.0 |
| | 3 | 5 | 0.00 | 3.31 | 0.86 | 0.64 | 0.01 | 16.89 | 6.02 | 3.56 | 0.03 | 99.99 | 99.87 | 99.66 |

Note: the DGP is $y_t = 0.9y_{t-1} + \alpha_1 \sin\left(\frac{2\pi kt}{T}\right) + \beta_1 \cos\left(\frac{2\pi kt}{T}\right) + \varepsilon_t(\rho = 0.9)$, with $\varepsilon_t \sim iid\mathcal{N}(0,1)$.

## 5.3 The "union of rejections" (UR) testing sequence

As mentioned previously, in spite of the general power dominance of the new test, its weakness when breaks are absent may possibly be overcome through a union of rejections decision rule, combining the new test with a (previous) standard DF test[11]. This is indeed a very simple testing strategy, much simpler than the testing sequence proposed by Enders and Lee, and may reduce the cost associated with the robustness to breaks (or other non-linear components) of the proposed test.

The raw version of this strategy consists simply of the decision rule "reject the unit root null if either DF or $\tau_{min}$ rejects" and it is frequently adopted by many practitioners, albeit with tests different from the $\tau_{min}$. In such a crude form, it is quite obvious that it can easily attain a high power performance, but at the expense of some size distortion, possibly far exceeding the nominal level. Moreover, this is also a feature that it shares with the Enders and Lee testing sequence. A preferable size-adjusted variant is, however, easily implemented following Harvey et al. (2009): reject the unit root null hypothesis if either

$$\{DF < \gamma^\lambda \, cv_{DF}^\lambda\} \text{ or } \{\tau_{min} < \gamma^\lambda \, cv_{\tau_{min}}^\lambda\}$$

where $\lambda$ is the desired significance level, $\gamma^\lambda$ is a common scaling constant and $cv_{DF}^\lambda$ and $cv_{\tau_{min}}^\lambda$ are the corresponding $100\lambda\%$ asymptotic critical values of the DF and $\tau_{min}$ tests, respectively. Therefore, the constant $\gamma^\lambda$ ensures that the asymptotic size of the rule is $\gamma$, and it can be approximated through Monte Carlo simulation using a grid search procedure. In table 4 this constant is presented for the usual significance level $\gamma = 0.05$ for the three usual cases. It is this corrected version that is considered in the remainder of this paper.

Both size and power simulation results for this sequence are also available in tables 2 and 3, respectively. In terms of rejections when a unit root is really present, the UR strategy improves somewhat significantly the performance of the new test in only 3 of the most severe over-rejection cases, but it leaves us very far from eliminating them completely. On the other hand, in many other cases, it inherits

[11] The combination with a DF-GLS test was also considered but it was excluded *a priori* due to the problem of the initial condition, which is often present when testing for convergence with samples starting in 1950.

Table 4. 5% asymptotic critical values and size-
-correcting parameter (ascp) for the UR testing strategy

| | $cv_{DF}^{0.05}$ | $cv_{\tau_{min}}^{0.05}$ | 5% ascp |
|---|---|---|---|
| no const. | $-1.95$ | $-3.82$ | 1.094 |
| no trend | $-2.86$ | $-4.30$ | 1.072 |
| with trend | $-3.41$ | $-4.69$ | 1.054 |

Notes: the 5% asymptotic DF critical values are taken from Fuller (1996), table 10.A.2 (p. 642) and the 5% asymptotic critical values of the $\tau_{min}$ statistic are from table 1 of this paper with $T = 1000$. The size-correcting parameters are obtained with $10,000$ replications.

the conservative character of both component tests, which is not necessarily a desirable feature.

Concerning power, when breaks are absent the UR sequence really improves considerably the performance of the *min*-test, increasing the estimated rejection frequencies in more than 50%. But in many other cases the power of the UR combined test is significantly lower than the proposed test. For instance, when $T = 50$ this occurs with $k = 0.4$, $(\alpha_1, \beta_1) = (0, 5)$ and $(3, 5)$, when $k = 1.2$ with the pair $(0, 3)$ and with $k = 1.6$ with $(0, 5)$, and when $T = 100$ and $T = 200$ with many other cases.

All in all, the main benefit of the UR sequence lies in disciplining the informal procedure consisting of the sequential application of the two tests, searching for a rejection. As a formal procedure, it allows us controlling the overall size. However, its estimated size benefits are scarce and unclear and the power gains when breaks are absent are outweighed by significant power losses in many break cases. Hence, for the empirical case under analysis, the straightforward application of the *min*-test appears to be more adequate than the UR strategy: with such a long span of data and with time series depending on the economies of two countries it is likely that they include at least one break.

# 6 Empirical results: detailed analysis

The dataset is a subset of the recently updated and improved Maddison database. Since one of the purposes of this paper is to reassess the results obtained with other variants of Fourier-type tests and, particularly, by King and Ramlogan-Dobson

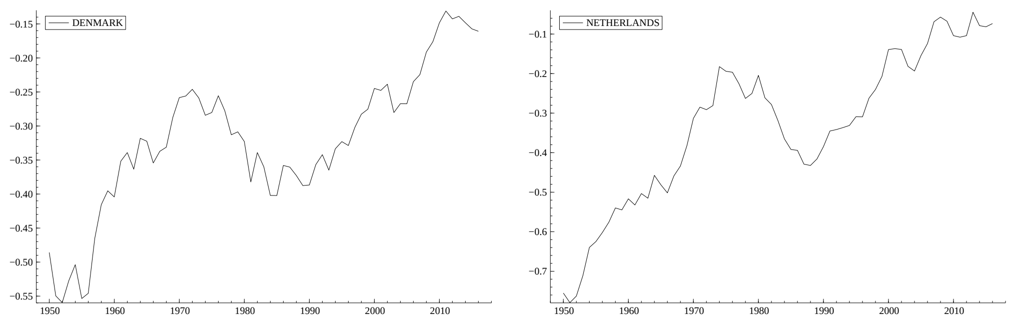


Figure 2: The cases of Denmark and the Netherlands

(2014) with LM tests, I have selected the same set of 25 countries as in their paper. These are, besides the US (the usual reference economy), Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Netherlands, New Zealand, Norway, Poland, Portugal, South Korea, Spain, Sweden, Switzerland, and the UK. The series that is selected to represent output *per capita* is called *CGDPpc* in Bolt et al. (2018), and it is considered the most reliable measure for assessing the degree of income convergence because it is based on multiple benchmark comparisons of prices and incomes across countries.

The empirical analysis will follow sequentially a "classical" perspective, based on standard unit root test statistics, the Enders and Lee testing strategy and the new test proposed in this paper, as well as the UR strategy.

## 6.1 Standard tests

Our standard tests are the well known and the simple but adequate $\tau_c^{DF}$ and the inadequate $\tau_{ct}^{DF}$ test statistics, and the $\tau_c^{DF-GLS}$ statistic of Elliot et al. (1996), with results presented in table 5.

The evidence for non-divergence for Austria, France, Germany, Italy and Japan produced by the $\tau_c^{DF}$ which was previously found in Lopes (2016) is confirmed, in some cases with a change in the significance level. But compared with those results, now the same type of evidence for non-divergence for Denmark and the Netherlands has evaporated. While for Denmark there appears indeed to have

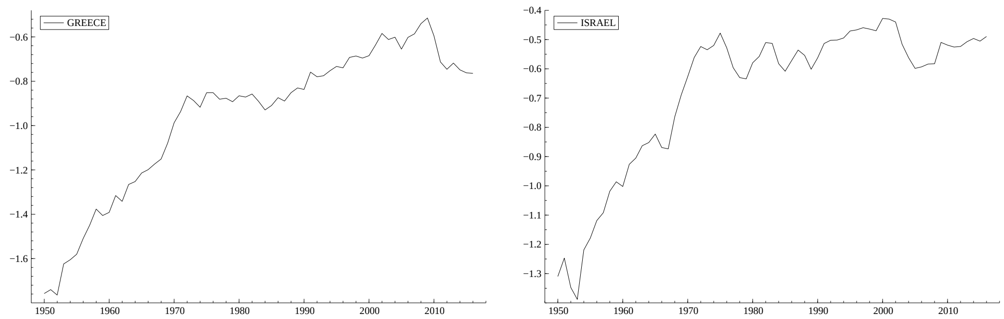


Figure 3: The cases of Greece and Israel

occurred a widening of the gap with the US in the last years of the sample, the result for the Netherlands is much more surprising because one cannot find a similar graphical evidence (see figure 2).

What appears to be disconcerting is the favourable outcome for Greece, in spite of the rather visible reversal of the converging process that has occurred in the last 8 years of the sample. This result appears to be due mainly to some stability of the income gap in a very substantial and final fraction of the sample, that is, albeit at a much lower level than the US, it appears that Greece has attained its steady state. A similar phenomenon appears to be responsible for the result for Israel: the graphical analysis suggests that Israel has attained its steady state in the 1970s and since then the income gap has been rather stable, with a rather stationary visual shape. In other words, although neither Greece nor Israel had attained the same level of *per capita* output as the US, the difference has remained limited and rather stable in the last 40-45 years of the sample.

Finally notice that the DF-GLS test allows rejecting the unit root null only for Canada, an outcome seemingly contradicting the superior power properties of this test over the simpler DF-OLS version. I believe that this evidence must not be taken into consideration because it is known that DF-GLS tests are very sensitive to the initial conditions and these are large for most countries because in 1950 their economies were still greatly affected by the Second World War. Therefore, this clearly appears to be a case where the recommendation by Muller and Elliot (2003) should be adopted, i.e., the simpler and usually less powerful DF-OLS test must supersede its DF-GLS relative, except for the case of Canada. Moreover, notice that the rejection for this country only concerns precisely an economy whose

Table 5 – Standard unit root test statistics

| | $\tau_c^{DF}$(nlag) | $\tau_{ct}^{DF}$ | $\tau_c^{DF-GLS}$ |
|---|---|---|---|
| Australia | $-1.32$ (3) | $-1.37$ (3) | $-1.04$(3) |
| Austria | $-3.53$ (0) ** | $-2.91$ (0) | 0.09(2) |
| Belgium | $-1.92$ (5) | $-1.74$ (0) | 0.07(0) |
| Canada | $-2.42$ (0) | $-2.80$ (0) | $-2.07$(1)** |
| Denmark | $-1.89$ (3) | $-2.49$ (4) | $-0.52$(3) |
| Finland | $-2.07$ (4) | $-1.86$ (4) | 0.04(1) |
| France | $-2.87$ (0) ** | $-2.03$ (0) | $-0.27$(1) |
| Germany | $-4.04$ (1) *** | $-3.11$ (6) | $-0.02$(3) |
| Greece | $-3.38$ (0) ** | $-1.01$ (0) | $-0.28$(5) |
| Hungary | $-2.22$ (6) | $-3.33$(6)* | 0.09(6) |
| Ireland | 0.03 (6) | $-1.96$ (6) | 0.43(6) |
| Israel | $-3.89$ (2) *** | $-2.62$ (2) | 0.02(0) |
| Italy | $-4.51$ (5) *** | $-2.47$ (5) | $-0.24$(6) |
| Japan | $-3.26$ (1) ** | $-1.34$ (6) | $-0.19$(2) |
| Netherlands | $-1.56$ (6) | $-2.85$ (6) | $-0.22$(6) |
| New Zealand | $-1.27$ (3) | $-0.58$ (3) | $-0.64$(3) |
| Norway | $-0.85$ (5) | $-2.66$ (5) | $-0.34$(5) |
| Poland | 1.08 (5) | $-0.16$ (5) | 1.24(5) |
| Portugal | $-1.49$ (1) | $-2.61$ (3) | 0.27(3) |
| South Korea | $-0.66$ (6) | $-2.17$ (6) | 0.40(6) |
| Spain | $-1.32$ (0) | $-2.54$ (2) | 0.50(0) |
| Sweden | $-1.47$ (0) | $-2.23$ (0) | 0.05(0) |
| Switzerland | $-2.21$ (1) | $-2.70$ (1) | 0.77(0) |
| U.K. | $-1.22$ (0) | $-2.64$ (3) | $-0.69$(0) |

Notes: 1) "***", "**", and "*" represent rejections of the (unit root) null hypothesis at the 1%, 5% and 10%, respectively; 2) the general-to-specific $t$-sig ($GTS$) procedure was employed to select the lag truncation parameter, with asymptotic 10% level tests and initiating the testing sequence with ($k_{MAX}=$) 6 lagged terms (the AIC method was also employed, producing identical or very similar results).

territory was not involved in WW2, i.e., one whose initial condition is closer to the remaining sample points.

On the other hand, this evidence further illustrates the improved power performance of the level stationarity analysis compared to the trend stationarity one: the previous evidence for the rejection of the unit root null disappears for all the previous seven countries when the $\tau_{ct}^{DF}$ is used and somewhat surprisingly it appears only for Hungary. Recalling the discussion of section 2, this only means that the Hungarian income gap can be considered as trend stationary, not that Hungary has converged; instead, it appears to be in a catching-up process, the gap steadily decreasing as time passes, following a process dominated by a linear deterministic trend, with deviations or fluctuations around that trend that behave in a stationary fashion.

## 6.2 Enders and Lee testing strategy results

In table 6 I present the results of the Enders and Lee (2012a) procedure. The conservative $\max F(\widehat{k})$ allows rejecting the linear null hypothesis at the 5% level for 3 countries only: Australia, New Zealand and Poland. Of these, the $\tau_c^{FDF}$ test statistic is able to reject the unit root null at 5% level only for Australia. For the remaining 21 countries the sequence produces exactly the same results as in the previous subsection because it neglects its pre-testing character.

These results appear unsatisfactory and unreliable:

a) given the number of countries and the length and nature of the sample period, the small number of rejections resulting from the $\max F(\widehat{k})$ test seems to testify that its power is low, affecting also the overall power properties of the procedure;

b) the rejections of the second test in the sequence may reflect a serious problem of size distortion, affecting both the DF and the $\tau^{FDF}$ tests.

Concerning this, recall that both statistics result from a sequential procedure whose nature is ignored in the derivation of their null distributions; moreover, recall also that available critical values for $\tau_c^{FDF}$ incorrectly assume the exogeneity of $\widehat{k}$.

Table 6 – Enders and Lee test statistics ($k$ integer)

| | $\max F(\widehat{k})$ | $\widehat{k}$ | $\tau_c^{FDF}$(nlag) | $\tau_c^{DF}$ |
|---|---|---|---|---|
| Australia | 7.65** | 1 | −4.07 (2) ** | — |
| Austria | 5.68 | 4 | — | −3.53 (0) ** |
| Belgium | 2.81 | 2 | — | −1.92 (5) |
| Canada | 5.43 | 2 | — | −2.42 (0) |
| Denmark | 2.31 | 2 | — | −1.89 (3) |
| Finland | 1.69 | 5 | — | −2.07 (4) |
| France | 6.12 | 2 | — | −2.87 (0) ** |
| Germany | 5.28 | 4 | — | −4.04 (1) *** |
| Greece | 2.85 | 2 | — | −3.38 (0) ** |
| Hungary | 1.09 | 4 | — | −2.22 (6) |
| Ireland | 3.21 | 1 | — | 0.03 (6) |
| Israel | 2.84 | 5 | — | −3.89 (2) *** |
| Italy | 2.88 | 4 | — | −4.51 (5) *** |
| Japan | 2.26 | 3 | — | −3.26 (1) ** |
| Netherlands | 3.89 | 2 | — | −1.56 (6) |
| New Zealand | 8.02** | 1 | −3.14 (3) | — |
| Norway | 1.79 | 2 | — | −0.85 (5) |
| Poland | 8.07** | 5 | 1.73 (0) | — |
| Portugal | 3.47 | 3 | — | −1.49 (1) |
| South Korea | 5.86 | 4 | — | −0.66 (6) |
| Spain | 2.37 | 4 | — | −1.32 (0) |
| Sweden | 2.43 | 5 | — | −1.47 (0) |
| Switzerland | 1.79 | 2 | — | −2.21 (1) |
| U.K. | 6.28 | 3 | — | −1.22 (0) |

Notes: 1) a "**" in the $\max F(\widehat{k})$ statistic represents a rejection at the 5% level using the critical value for $T = 100$ in Enders and Lee (2012b, EL) (7.58); for 10% it is 6.35; 2) again the general-to-specific $t$-sig ($GTS$) procedure was employed to select the lag truncation parameter, with asymptotic 10% level tests and initiating the testing sequence with ($k_{MAX} =$) 6 lagged terms; 3) the "**" in the $\tau_c^{FDF}$ test statistic denotes a rejection at the 5% level (the 5% critical value for $T = 100$ and $k = 1$ from EL is −3.81).

## 6.3 The FDF and $\tau_{min}^{FDF}$ test results

In table 7 I present the results for the $FDF_c$ and $\tau_{c,min}^{FDF}$ and $\tau_{nc,min}^{FDF}$ test statistics. Recall that the first test is not the one by Enders and Lee (2012a) and that it is not strictly valid because it neglects the dependence of the distribution on the estimated value of $k$. It is a useful benchmark against which the properties of the new test could be assessed but its use cannot be straightforwardly recommended.

Nonetheless, the results of the first two columns are disappointing: the small sample properties of the FDF and the $\tau_{c,min}^{FDF}$ statistics and, in particular, the power properties of this one promised a number of rejections of the unit root null superior to the one produced by the DF statistic but rather the opposite occurs. Actually, the new test is able to reject the unit root at 5% or lower only for Israel, Italy and South Korea. At the 10% level one gets rejections for Australia and France as well, i.e., in total only 5 countries, less than the corresponding 7 rejections obtained with the much simpler and non-robust, prone to power deficiencies due to breaks $\tau_c^{DF}$. Although the statistic is (negative and) large for all these cases, now one does not get a rejection neither for Austria nor for Germany, Greece and Japan.

One possible explanation for this apparent contradiction lies in the likely interference or noise produced by the lag length selection method, which has simply not played any role in the simulation study. To this one may object, on one hand, that all the rival methods must be equally affected. On the other hand, trying to remedy the problem one may consider reversing the order of the procedures for the estimation (minimization) of $k$ and for the lag augmentation but this does not seem reasonable. Anyway, to further investigate this issue, the calculation of the test statistic was redone with a less size concerned t$-sig$ method, more power oriented, using 5% level tests for the simplification process. This has produced shorter lag lengths for four countries but only two different decisions: the test statistic for Germany changes to $-5.17(\widehat{k} = 0)$, allowing a rejection at 1%, and the one for Italy changes to $-5.12$ ($\widehat{k} = 0$ also), changing only the rejection level from 5% to 1%. For the remaining cases there is no relevant change. Anyway, notice also that the estimate originally obtained for Germany for $k$ (zero, with one lag) already implied the rejection of the unit root null via the DF test statistic obtained previously, i. e., in a reversed UR testing sequence.

Inspired by the same evidence, a rather different argument could sustain that

Table 7 – FDF and $\tau_{min}^{FDF}$ test statistics

| | $FDF_c(\widehat{k})$ [nlag] | $\tau_{c,min}^{FDF}(\widehat{k})$ [nlag] | $\tau_{nc,min}^{FDF}(\widehat{k})$ [nlag] |
|---|---|---|---|
| Australia | $-4.07(1)[2]^{**}$ | $-4.43(1.4)[0]^{*}$ | $-3.02(0.3)[0]$ |
| Austria | $-3.13(4)[0]$ | $-3.56(1.2)[0]$ | $-5.75(2.9)[0]^{***}$ |
| Belgium | $-1.90(2)[5]$ | $-2.51(2.7)[5]$ | $-2.64(2.5)[5]$ |
| Canada | $-3.72(2)[0]^{*}$ | $-3.72(2.0)[0]$ | $-2.65(0.1)[0]$ |
| Denmark | $-1.10(2)[2]$ | $-2.52(0.1)[4]$ | $-3.28(0.8)[2]$ |
| Finland | $-2.07(5)[4]$ | $-3.44(2.2)[5]$ | $-2.87(2.3)[4]$ |
| France | $-4.45(2)[5]^{**}$ | $-4.38(2.0)[5]^{*}$ | $-3.33(0.7)[5]$ |
| Germany | $-0.18(4)[6]$ | $-3.98(0.0)[1]^{+++}$ | $-7.24(0.6)[0]^{***}$ |
| Greece | $-3.08(2)[0]$ | $-3.84(2.4)[0]$ | $-4.30(2.4)[0]^{**}$ |
| Hungary | $-2.04(4)[6]$ | $-2.64(1.9)[0]$ | $-3.90(0.7)[0]^{*}$ |
| Ireland | $-1.82(1)[6]$ | $-2.53(0.7)[6]$ | $-2.93(1.0)[6]$ |
| Israel | $-3.81(5)[2]^{*}$ | $-4.72(2.8)[5]^{**+}$ | $-3.66(2.7)[2]^{**}$ |
| Italy | $-4.80(4)[5]^{***}$ | $-4.64(2.5)[4]^{**}$ | $-3.79(2.8)[1]^{*}$ |
| Japan | $-3.84(3)[1]^{*}$ | $-3.80(2.7)[6]$ | $-3.99(2.6)[1]^{**}$ |
| Netherlands | $-2.12(2)[0]$ | $-0.85(2.2)[3]$ | $-1.78(2.4)[6]$ |
| New Zealand | $-3.14(1)[3]$ | $-2.71(1.0)[5]$ | $-3.76(0.3)[3]^{*}$ |
| Norway | $-0.92(2)[0]$ | $-2.77(0.4)[5]$ | $-2.47(0.8)[5]$ |
| Poland | $1.73(5)[0]$ | $-1.98(0.1)[3]$ | $-3.62(0.7)[6]^{*}$ |
| Portugal | $-2.10(3)[4]$ | $-3.02(0.1)[5]$ | $-3.29(2.6)[4]$ |
| South Korea | $-0.52(4)[5]$ | $-5.65(0.5)[0]^{***}$ | $-2.60(1.1)[6]$ |
| Spain | $-1.01(4)[0]$ | $-2.40(0.8)[0]$ | $-3.10(2.4)[0]$ |
| Sweden | $-1.78(5)[2]$ | $-2.13(0.1)[1]$ | $-2.24(0.1)[0]$ |
| Switzerland | $-0.92(2)[0]$ | $-2.60(3.0)[3]$ | $-3.26(3.0)[1]$ |
| U.K. | $-0.73(3)[2]$ | $-3.76(0.6)[6]$ | $-2.90(0.1)[3]$ |

Notes: 1) "***", "**" and "*" denote rejections at the 1%, 5% and 10% levels, respectively, using the critical values presented in this paper for $T = 50$; 2) again the general-to-specific $t$-sig ($GTS$) procedure was employed to select the lag truncation parameter in every case, with asymptotic 10% level tests and initiating the testing sequence with ($k_{MAX} =$) 6 lagged terms; 3) "+++" and "**+" denote rejections at 1% but with the $GTS$ $t$-sig method employing 5% level tests.

the results do not support the existence of breaks in the series: after all, this arguing goes, since the test which is robust to breaks produces less, not more evidence for stationarity, breaks must be largely absent from the data. However, this argument is rather feeble: although the proposed test is indeed more powerful — sometimes much more powerful — than the DF test under the stationary alternative, in many cases its power is still very low, as a closer inspection of table 4 confirms. For instance, when $T$ is only 50 (here $T = 67$), $\rho = 0.9$, $k = 2$ and $(\alpha_1, \beta_1) = (3, 5)$, the false unit root null is expected to be rejected in only 0.86% of the cases. This is an extreme case but other cases exist, mostly when $(\alpha_1, \beta_1) = (3, 5)$, where the estimated power of the $min$-test is rather low, and particularly below the fixed nominal size.

Overall, these results inspired a further search for a more powerful test. While in statistical terms this search lead to a small step away, in terms of the concrete problem the adopted solution appears to be counter-intuitive because the requirement for non-divergence becomes more demanding, not less. In fact, additional power can be obtained adopting the strict interpretation of equation (1), i.e., dropping the constant term from the test regression. Requiring stationarity around a zero mean provides, simultaneously, a more stringent condition for convergence and, insofar as a deterministic regressor that becomes irrelevant is omitted, a more powerful unit root test.

The results for the *min* version of this test are also presented in table 7, in its last column. Although, as expected, the number of rejections of divergence increases, three distinct cases are worth considering:

a) for Germany, Israel and Italy the rejection may be viewed as simply confirming identical previous results;

b) for Austria, Greece, Japan and New Zealand the novelty of the rejection is far from surprising because previous results were already close to it. Therefore, these cases seem to serve as good illustrations of the gain in power.

c) However, the cases of Hungary and Poland appear as dubious, not only because the rejections necessitate a size of 10% but mostly because previous tests were rather unfavourable to the hypothesis and the graphical analysis (see figure 4) does not lend any support to such a decision. In both cases, a

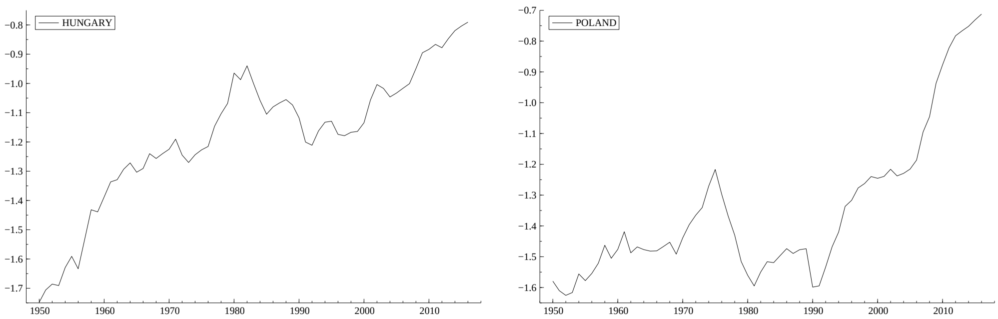


Figure 4: The cases of Hungary and Poland

catching-up process initiated around 1990 is clearly visible but it seems far from attaining stability, even at a (much) lower level than the leader (as in the cases of Greece and Israel).

A closer inspection of this last case does not allow drawing any firm conclusion. The simulation results of table 2 show that indeed the over-rejection problems tends to concentrate around the values of $k$ which lie close to the (common) estimate for Hungary and Poland ($k = 0.4$ and $0.8$ and $\widehat{k} = 0.7$, respectively). But, on the other hand, the estimates for $\alpha_1$ and $\beta_1$ are far from all the cases of the simulation study: again in both cases they are small and symmetric. Moreover, a unit root test allowing for a smooth break under the null, the one by Lanne et al. (2003), produces strong supporting evidence for the unit root, contradicting the result of the $\tau^{FDF}_{nc,min}$ test, thereby confirming the suspicion that this test is producing an erroneous inference. While acknowledging that the ground is not solid, I believe that this conclusion is the most plausible, and consider, at least provisionally, that both Hungary and Poland are still divergent cases.

# 7 Comparison and final discussion

Summing up the evidence produced by the proposed *min* test, only 10 decisions for non-divergence were obtained: 3 at the 10% level (Australia, France and New Zealand), 5 at the 5% level (Greece, Israel, Italy, Japan and South Korea), and only 2 at the 1% level (Austria and Germany). Moreover, recall that in some cases — most notably those of Greece and Israel, but also, for instance, Italy and Japan

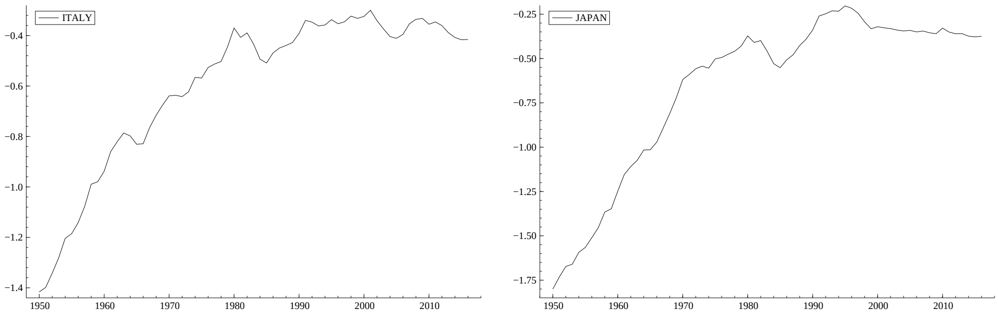


Figure 5: The cases of Italy and Japan

at a closer level to the leader — an outcome for non-divergence means that the relative income difference to the reference economy has remained limited and rather stable for some time, not that the level of *per capita* income of this latter country has been attained.

These results are presented in table 8, together with a summary of some previous evidence reported in the literature[12]. The main inference that this comparison delivers is that Fourier-type unit root tests are not always as favourable to the hypothesis as previous studies implied. Both in Christopoulos and Leon-Ledesma (2011, CL-L) and in King and Ramlogan-Dobson (2014, KR-D) the evidence for non-divergence was overwhelming, with only the exception of one country (Japan) in the former study. That is, although not starkly contrasting with it, the evidence gathered here is much less supportive of the hypothesis.

However, in both cases a strict comparison is not feasible: in CL-L the sample size is much larger but it is cross-sectionally restricted to 13 countries that are usually considered as developed or high-income. The set of countries considered here coincides with the one of KR-D, their sample ending in 2008 but, most importantly, their Fourier-type tests allow a linear trend term in the deterministic component[13], which I consider inadmissible in my approach.

Nonetheless, the results of this paper clearly contrast with those previously produced with the Fourier approach to unit root tests, and more generally they

[12]Surprisingly, the number of empirical studies that are comparable to this one is very small. This is because very often the methods that are employed are rather disparate.

[13]The Lagrange-multiplier framework used in KR-D requires the presence of this component and hence it is not adaptable to the framework adopted here.

Table 8 — Comparison of test results

| | CHLL(2008) | CL-L(2011) | KR-D(2014) | L(2016) | this paper |
|---|---|---|---|---|---|
| database | PWT | Maddison | Maddison | Maddison | Maddison |
| sample | 1950-2000 | 1900-2008 | 1950-2008 | 1950-2008 | 1950-2016 |
| Australia | nd1 | nd1 | nd5 | | nd10 |
| Austria | nd5 | nd5 | nd5 | nd5 | nd1 |
| Belgium | | nd1 | nd5 | | |
| Canada | | nd5 | nd5 | nd5 | |
| Denmark | — | nd1 | nd5 | nd5 | |
| Finland | | nd5 | nd5 | | |
| France | | nd5 | nd5 | nd10 | nd10 |
| Germany | — | nd10 | nd5 | nd1 | nd1 |
| Greece | — | — | nd5 | — | nd5 |
| Hungary | — | — | nd5 | — | |
| Ireland | — | — | nd5 | — | |
| Israel | — | — | nd5 | — | nd5 |
| Italy | — | nd5 | nd5 | nd5 | nd5 |
| Japan | | | nd5 | nd1 | nd5 |
| Netherlands | nd1 | nd1 | nd5 | nd1 | |
| New Zealand | — | — | nd5 | — | nd10 |
| Norway | | nd1 | nd5 | | |
| Poland | — | — | nd10 | — | |
| Portugal | — | — | nd5 | — | |
| South Korea | — | — | nd5 | — | nd5 |
| Spain | — | — | nd5 | — | |
| Sweden | | nd5 | nd5 | | |
| Switzerland | | — | nd5 | | |
| U.K. | nd1 | nd5 | nd5 | | |

Notes: 1) "PWT" represents the Penn World Tables. 2) a blank entry represents a non rejection of the UR hypothesis, i.e., a result for divergence. 3) "ndx" represents a rejection of the UR hypothesis at the x% level, i.e., a result for non divergence. 4) CHLL (2008) represents Chong, Hinich, Liew and Lim (2008), who use a KSS (Kapetanios, Shin and Snell (2003)) unit root test only in those cases where, in a first stage, a linearity test finds evidence for non-linearity in the series; this excludes Denmark, Germany and Italy from further consideration. 5) CL-L denotes Christopoulos and Leon-Ledesma (2011), where the unit root test is a two-step FDF test where the non-linear deterministic component is removed in the first step with a Fourier expansion with no trend (as in this paper), and a unit root is tested in the residuals against both a linear and a logistic (stationary) smooth transition autoregressive (LSTAR) alternative using the $(inf - t)$ test of Park and Shintani (2016). 6) KR-D (2014) represents King and Ramlogan-Dobson (2014), where 2 different Fourier type tests are used over a LM unit root test variant but allowing a linear trend term in the deterministic component. In one of the tests the deviations from the trend are allowed to follow an exponential STAR (ESTAR). In both these papers the joint nature of the procedure is not considered in the evaluation of the $p$-value. 7) L(2016) denotes Lopes (2016).

contradict a recent wave of optimism concerning the convergence hypothesis, represented, for instance, by Desli and Gkoulgkoutsika (2021, DG): "*most of the studies concerned with developed countries find evidence of convergence*". Note, however, that mostly due to the procedures employed, these studies that DG refer are not comparable with this one. A parallel optimistic perspective has also recently appeared concerning low and medium income countries but it is soundly refuted in Johnson and Papageorgiou (2020).

Qualitatively, the evidence gathered here is much closer to the one in Chong, Hinich, Liew and Lim (2008, CHLL), where the framework is not formally one of allowing for breaks but non-linearities are allowed, both in the deviations around the trend and in this latter component as well (in dissonance with the approach adopted here). The general picture painted by this evidence is, therefore, much less favourable to the hypothesis than the one that usually transpires from recent studies and, particularly, from those that resort to the Fourier approach. A partial explanation for this lies in the diversity of the conditions embedded in the testing procedures. For instance, in Ceylan and Abiyev (2016), the optimistic view concerning the countries of the European Union is likely to derive (at least) partially from the adopted benchmark: not the technological leader but the country-average (see also Islam, 2003, on this subject)[14].

The evidence presented here appears to be particularly robust to the possibility of undetected convergence. Not only due to the improved robustness of DF tests in terms of power but also to the (*min*) form of the statistic and, at at later stage of the analysis, to the adoption of an even more powerful (*no-constant*) version of the test statistic. Furthermore, having identified the main weakness of the proposed procedure as one where, in a few cases, a finding for non-existing convergence arises from the coexistence of a unit rot with breaks, a test less prone to this deficiency will find less, not more, evidence for convergence.

The present evidence is also qualitatively very close to my previous one (Lopes, 2016), based on a previous version of the Maddison database, ending in 2008 and therefore with a significantly smaller sample size, but making no allowance for the possibility of breaks in the level of the series. The shifting composition of the

[14] For an opposite outcome concerning the EU but using rather different methods see Franks et al. (2018).

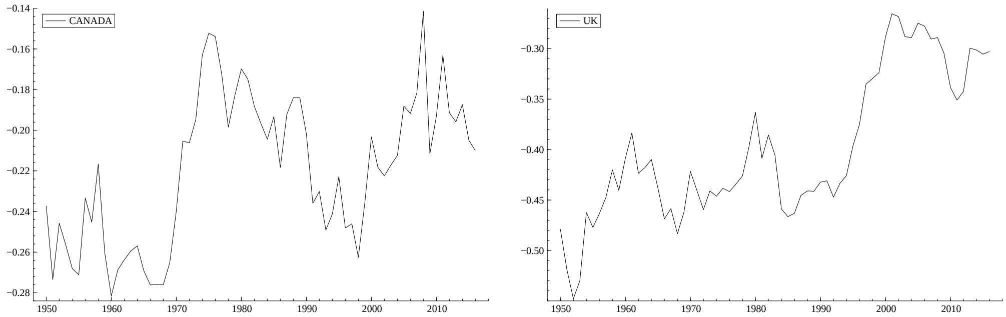


Figure 6: The cases of Canada and the U.K.

evidence therefore seems to result mainly from two conflicting forces:

a) both the augmentation of the sample size and the allowance for breaks tend to produce tests with improved power properties, lending support to convergence;

b) however, the developments that have followed the burst of the global financial crisis appear having derailed some economies from their steady path, at least in relative terms, making them move away from the convergence trajectory.

This seems to be the case particularly for Canada and Denmark, now negatively labelled as non-converging, in stark contrast with my previous results (see figure 2 for the case of Denmark and figure 6 below for Canada). Moreover, the graphical analysis indicates that several other economies appear having been particularly disturbed by the events initiated in 2007–2008: this is the case clearly for Australia, Finland, and Spain, and not so significantly for Belgium and Sweden; the case of the U.K. is similar but the relative decay began sooner, closer to the beginning of the century (see figure 6). The inauguration of the financial crisis therefore seems to mark the beginning of a new phase of transitional dynamics for many countries and it is yet unclear when and how (and whether) it will stabilize.

On the other hand, I reiterate that the view of convergence that I adopt here is distinct from the usual catching-up approach, which allows and even rests on the presence of a trend component in the logged output gap. If that was the case, the economies of Ireland and, most notoriously, those of Norway and Switzerland had not only converged with the leader but have even overtook it, becoming the new

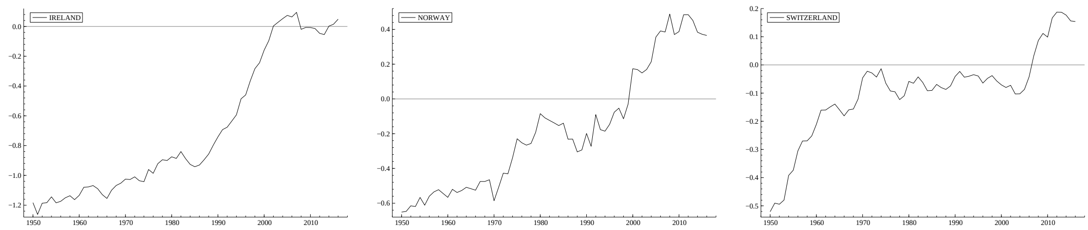


Figure 7: The cases of Ireland, Norway and Switzerland

leaders. In the case of these three economies, the interpretation of the outcome of unit root tests is that a stable path has not yet been achieved; they appear to be still in a process of transitory dynamics as well, approaching their steady state from a level above that of the (current) leader.

Although exclusively based in the past, the lack of support to convergence that was found has something to say about the future too. A shift in the global balance of economic and political power would require a rather strong convergence of poor to rich countries. Its absence in the past indicates that it is unlikely to occur in the future as well.

A final remark concerns the economic significance of the statistical evidence. Both Durlauf, Johnson and Temple (2005, DJT) and Johnson and Papageorgiou (2020, JP) question the relevance of unit root tests allowing for breaks to assess income convergence. In particular, JP argue that the "*interpretation of breaks is unclear"*. However, on one hand, as DJT emphasize, the time series approach to convergence is "*largely statistical in nature"* (p. 589). Now, the failure of standard unit root tests to reject divergence when it is false because the gap series displays some sort of discontinuity is a failure of these statistical methods, which need to be patched to produce a less fragile result. On the other hand, JP themselves provide the key to the interpretation problem when they make the question "*do the breaks represent large exogenous shocks?"* Indeed, at least returning to Perron's (1989) original argumentation, time series breaks may be viewed as a "*device*" to remove large shocks from the mean (trend) function without investing a lot of effort with its adequate modelling. It then follows that, in fact, these results are not free from the testing model. They are conditional upon it, as is usually the case.

# References

[1] Becker, R., Enders, W. and Lee, J. (2006). A stationarity test in the presence of an unknown number of smooth breaks, *Journal of Time Series Analysis*, 27 (3), 381-409.

[2] Bernard, A. B. and Durlauf, S. N. (1996). Interpreting tests of the convergence hypothesis, *Journal of Econometrics*, 71, 161-73.

[3] Bolt, J., Inklaar, R., de Jong, H. and van Zanden, J. L. (2018). Rebasing 'Maddison': new income comparisons and the shape of long run economic development, *CGDC Research Memorandum*, University of Groningen.

[4] Ceylan, R. and Abiyev, V. (2016). An examination of convergence hypothesis for EU-15 countries, *International Review of Economics and Finance*, 45, 96-105.

[5] Chong, T. T.-L., Hinich, M. J., Liew, V. K.-S. and Lim, K.-P. (2008). Time series tests of nonlinear convergence and transitional dynamics, *Economics Letters*, 100, 337-339.

[6] Christopoulos, D. K. and Leon-Ledesma, M. A. (2011). International output convergence, breaks and asymmetric adjustment, *Studies in Nonlinear Dynamics and Econometrics*, 15 (3), article 4.

[7] Desli, E. and Gkoulgkoutsika, A. (2021). Economic convergence among the world's top-income economies, *The Quarterly Review of Economics and Finance*, vol. 80, issue C, 841-53.

[8] Durlauf, S. N., Johnson, P. A. and Temple, J. R. W. (2005). Growth Econometrics, *in* Aghion, P. and Durlauf, S. N. (eds.), *Handbook of Economic Growth*, vol. 1A, Elsevier B. V., 555-677.

[9] Elliot, G., Rothenberg, T. J. and Stock, J. H. (1996). Efficient tests for an autoregressive unit root, *Econometrica*, 64 (4), 813-36.

[10] Enders, W. and Lee, J. (2004). Testing for a unit root with a nonlinear Fourier function, *working paper*.

[11] Enders, W. and Lee, J. (2012a). The flexible Fourier form and Dickey-Fuller type unit root tests, *Economics Letters*, 117, 196-199.

[12] Enders, W. and Lee, J. (2012b). A unit root test using a Fourier series to approximate smooth breaks, *Oxford Bulletin of Economics and Statistics*, 74 (4), 574-599.

[13] Franks, J., Barkbu, B., Blavy, R., Oman, W. and Schoelermann, H. (2018). Economic Convergence in the Euro Area: Coming Together or Drifting Apart? *IMF working paper* 18/10.

[14] Fuller, W. A. (1996). *Introduction to Statistical Time Series*, 2nd. ed., John Wiley & Sons.

[15] Greasley, D. and Oxley, L. (1997). Time-series based tests of the convergence hypothesis: some positive results, *Economics Letters*, 56, 143-47.

[16] Harvey, D. I., Leybourne, S. J. and Taylor, A. M. R. (2009). Unit root testing in practice: dealing with uncertainty over the trend and initial condition, *Econometric Theory*, 25, 587-636.

[17] Islam, N. (2003). What have we learnt from the convergence debate?, *Journal of Economic Surveys*, vol. 17 (3), 309-62.

[18] Kapetanios, G., Shin, Y. and Snell, A. (2003). Testing for a unit root in the nonlinear STAR framework, *Journal of Econometrics*, 112, 359-79.

[19] Kiliç, R. (2011). Testing for a unit root in a stationary ESTAR process, *Econometric Reviews*, 30, 274-302.

[20] Kim, D. and Perron, P. (2009). Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses, *Journal of Econometrics*, 138, 1-13.I

[21] King, A., and Ramlogan-Dobson (2014). Are income differences within the OECD diminishing? Evidence from Fourier unit root tests, *Studies in Nonlinear Dynamics and Econometrics*, 18 (2), 185-199.

[22] Johnson, P. and Papageorgiou, C. (2020). What remains of cross-country convergence? *The Journal of Economic Literature*, 58(1), 129-75.

[23] Lanne, M, Lütkepohl, H. and Saikkonen, P. (2003). Test procedures for unit roots in time series with level series at unknown time, *Oxford Bulletin of Economics and Statistics*, 65, 91-115.

[24] Lee, J. and Strazicich, M. C. (2001). Break point estimation and spurious rejections with endogenous unit root tests, *Oxford Bulletin of Economics and Statistics*, 63 (5), 535-58.

[25] Leybourne, S. J., Mills, T. C. and Newbold, P. (1998). Spurious rejections by Dickey Fuller tests in the presence of a break under the null, *Journal of Econometrics* 87, 191-203.

[26] Li, Q. and Papell, D. (1999). Convergence of international output: time series evidence for 16 OECD countries, *International Review of Economics and Finance*, 8, 267-280.

[27] Lopes, A. S. (2016). A simple proposal to improve the power of income convergence tests, *Economics Letters*, 138, 92-95.

[28] Müller, U. K. and Elliot, G. (2003). Tests for unit roots and the initial condition, *Econometrica*, 71, 1269-86.

[29] Nordström, M. (2018). On the use of integer and fractional flexible Fourier form Dickey-Fuller unit root tests, *working paper*, Lund University.

[30] Omay, T. (2015). Fractional frequency flexible Fourier form to approximate smooth breaks in unit root testing, *Economics Letters*, 134, 123-6.

[31] Park, J. Y. and Shintani, M. (2016). Testing for a unit root against transitional autoregressive models, *International Economic Review*, vol. 57 (2), 635-64.

[32] Perron, P. (1989). The great crash, the oil price shock and the unit root hypothesis, *Econometrica*, 57, 1361-1401.

[33] Pesaran, M. H. (2007). A pair-wise approach to testing for output and growth convergence, *Journal of Econometrics*, 138, 312-55.

[34] Rodrigues, P. M. M. and Taylor, A. M. R. (2012). The flexible Fourier form and local generalised least squares de-trended unit root tests, *Oxford Bulletin of Economics and Statistics*, 74 (5), 736-59.

[35] Su, J.-J. and Nguyen, J. K. (2013). Alternative unit root testing strategies using the Fourier approximation, *Economics Letters*, 121, 8-11.

[36] Zivot, E. and Andrews, D. W. K. (1992). Further evidence on the great crash, the oil-price shock, and the unit root hypothesis, *Journal of Business and*

*Economic Statistics*, 10, 251-270.

## Statements and Declarations

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

The author has no relevant financial or non-financial interests to disclose.

# 8 Appendix

Table A. Critical values for the FDF test

| $T$ | 1% | 5% | 10% |
|---|---|---|---|
| | no constant case, $FDF_{nc}$ | | |
| 50 | −3.38 | −2.48 | −1.99 |
| 100 | −3.40 | −2.60 | −2.11 |
| 200 | −3.41 | −2.65 | −2.17 |
| 1000 | −3.48 | −2.76 | −2.29 |
| | no trend case, $FDF_c$ | | |
| 50 | −4.65 | −3.97 | −3.62 |
| 100 | −4.52 | −3.91 | −3.60 |
| 200 | −4.46 | −3.88 | −3.57 |
| 1000 | −4.41 | −3.85 | −3.55 |
| | with trend case, $FDF_{ct}$ | | |
| 50 | −5.32 | −4.65 | −4.31 |
| 100 | −5.12 | −4.53 | −4.24 |
| 200 | −5.04 | −4.47 | −4.20 |
| 1000 | −4.96 | −4.43 | −4.16 |

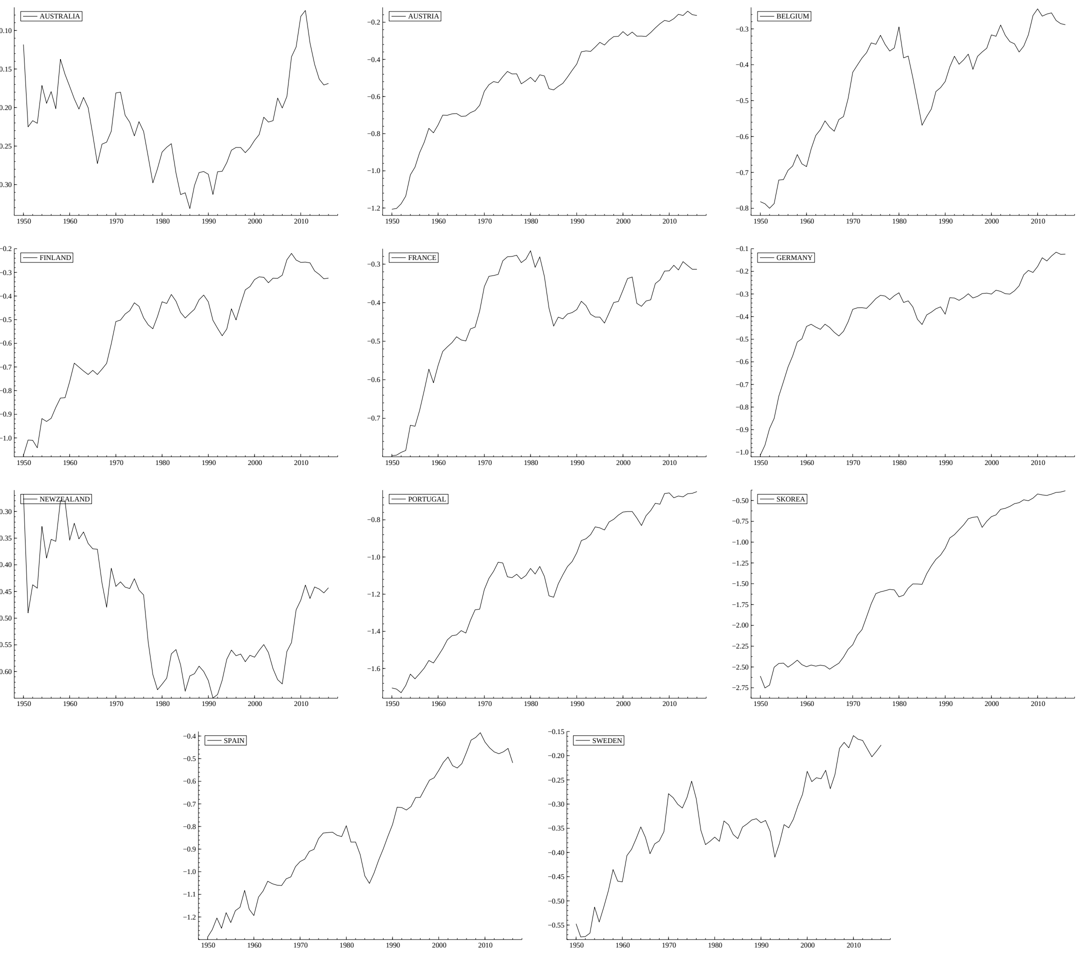


Figure 8: The remaining countries: Australia, Austria, Belgium, Finland, France, Germany, New Zealand, Portugal, South Korea, Spain and Sweden.