econometrics

Article

Unit Roots in Economic and Financial Time Series: A Re-Evaluation at the Decision-Based Significance Levels

Jae H. Kim [1,*] and **In Choi** [2]

[1] Department of Economics and Finance, La Trobe University, Melbourne VIC 3086, Australia
[2] Department of Economics, Sogang University, Seoul 04107, Korea; inchoi@gmail.com
[*] Correspondence: J.Kim@latrobe.edu.au; Tel.: +61-3-9479-6616

Academic Editors: Federico Bandi, Alex Maynard, Hyungsik Roger Moon and Benoit Perron
Received: 10 February 2017; Accepted: 29 August 2017; Published: 8 September 2017

Abstract: This paper re-evaluates key past results of unit root tests, emphasizing that the use of a conventional level of significance is not in general optimal due to the test having low power. The decision-based significance levels for popular unit root tests, chosen using the line of enlightened judgement under a symmetric loss function, are found to be much higher than conventional ones. We also propose simple calibration rules for the decision-based significance levels for a range of unit root tests. At the decision-based significance levels, many time series in Nelson and Plosser (1982) (extended) data set are judged to be trend-stationary, including real income variables, employment variables and money stock. We also find that nearly all real exchange rates covered in Elliott and Pesavento (2006) study are stationary; and that most of the real interest rates covered in Rapach and Weber (2004) study are stationary. In addition, using a specific loss function, the U.S. nominal interest rate is found to be stationary under economically sensible values of relative loss and prior belief for the null hypothesis.

Keywords: expected loss; line of enlightened judgement; power of the test; response surface

JEL Classification: C12; E30; F30

1. Introduction

Since the seminal study by Nelson and Plosser (1982), the presence of a unit root in economic and financial time series has been a highly controversial issue. It has compelling implications for a wide range of economic and financial theories. For example, a unit root in the real GNP contradicts the conventional view of the business cycle that a shock to the economy has a transitory effect (see Campbell and Mankiw 1987). However, a series of studies report mixed and inconclusive results regarding the presence of a unit root in the U.S. real GNP (see, for example, Rudebusch 1993; Diebold and Senhadji 1996; Murray and Nelson 2000; Papell and Prodan 2004; Darné 2009; Luo and Startz 2014). Other research areas where the presence of a unit root is contentious include empirical studies on the purchasing power parity (Lothian and Taylor 1996; Papell 1997); and the stationarity of real interest rate (Rose 1988; Rapach and Weber 2004). See Choi (2015) for further discussions on economic issues related to the presence of a unit root.

A major problem of unit root testing is that the power of the test is seriously low in small samples (see McCallum 1986, p. 406; Schwert 1989; DeJong et al. 1992). However, the low power is not fully taken into account in its practical implementations (Cochrane 1991, p. 283). Specifically, the test is almost exclusively conducted at the conventional level of significance (typically at 0.05), completely ignoring its power and other factors. To this end, a number of authors have raised serious concerns and criticisms about the fact that empirical researchers pay little attention to the power of tests (Arrow 1960; Hausman 1978; MacKinnon 2002, p. 633; Ziliak and McCloskey 2008; Startz 2014). It also has been pointed out that employing a conventional significance level is arbitrary and can lead to misleading results (Davidson and MacKinnon 1993, p. 79; Keuzenkamp and Magnus 1995; Lehmann and Romano 2005, p. 57). In the context of unit root testing, Maddala and Kim (1998, p. 128) question the appropriateness of using a conventional significance level.

Let α represent the level of significance which is the probability of rejecting the true null hypothesis (Type I error). The probability of Type II error (accepting the false null hypothesis) is denoted as β, with $(1 - \beta)$ being the power of the test.[1] It has been argued that when the power of the test is low, the value of α should be chosen at a much higher level than 0.05 (see, for example, Kish 1959[2]). More specifically, Winer (1962) states "when the power of the tests is likely to be low . . . , and when Type I and Type II errors are of approximately equal importance, the 0.3 and 0.2 levels of significance may be more appropriate than the 0.05 and 0.01 levels" (cited in Skipper et al. 1967)[3]. Arrow (1960, p. 73), in proposing the equal-probability test where the probabilities of Type I and II errors are set equal, demonstrates that, when the power of the test is low, the value of α should be set as high as 0.40 to balance the two error probabilities. Stressing that the value of α should be set as a decreasing function of sample size, Leamer (1978, Chapter 4) shows how a (decision-based) level of significance can be chosen by minimizing the expected loss, based on what he calls the line of enlightened judgement.[4] Fomby and Guilkey (1978) also show, through extensive Monte Carlo simulations, that the optimal significance level for the Durbin–Watson test should be around 0.5, much higher than the conventional one.

The purpose of this paper is to re-evaluate key past results of unit root testing at the decision-based significance levels chosen in explicit consideration of the power and expected loss, following Leamer (1978). It is found that the decision-based levels for popular unit root tests, such as the augmented Dickey–Fuller (Dickey and Fuller 1979; ADF) and DF–GLS tests (Elliott et al. 1996) are much higher than 0.05. In fact, they are in the 0.2 to 0.4 range for the sample sizes widely used in practice, under a symmetric loss function and equal chance for the null and alternative hypotheses. Through extensive simulations, we obtain simple calibration rules for the decision-based significance levels for the ADF, Phillips–Perron (Phillips and Perron 1988), DF–GLS and point optimal (ERS-P) tests of Elliott et al. (1996). When the ADF and DF–GLS tests are conducted at these levels, many time series in the Nelson–Plosser data set are found be trend-stationary including the real income and money stock. For the real exchange rates examined by Elliott and Pesavento (2006), the ADF and DF–GLS tests conducted at the decision-based level favor the stationarity for nearly all the rates, generating strong support for the purchasing power parity. Furthermore, most of the real interest rates series covered in Rapach and Weber (2004) study are found to be stationary at the decision-based levels. We demonstrate how the calibration rules for the decision-based significance levels for the Phillips–Perron and ERS–P tests are used to determine the presence of a unit root in the U.S. real GNP. We also consider a specific form of loss function proposed by Koop and Steel (1994) to test for the presence of a unit root in the U.S. nominal interest rate. We find that the conventional significance levels are justifiable only when the researcher assigns a heavy weight to the null hypothesis of a unit root (in the form of prior belief or loss), which may often be inconsistent with common economic reasoning.

This paper is organized as follows: Section 2 presents the line of enlightened judgement and the decision-based significance levels for the ADF and DF–GLS tests; Section 3 presents the calibration

[1] α and β in the context of unit root testing will be formally defined in Section 2.2.

[2] Reprinted in Morrison and Henkel (1970, p. 139).

[3] Reprinted in Morrison and Henkel (1970, p. 157).

[4] We note that Manderscheid (1965) and DeGroot (1975, p. 380) also propose the same method for choosing the decision-based significance level, without introducing the line of enlightened judgement.

rules based on asymptotic local power for a range of popular unit root tests; Section 4 re-evaluates past key results of unit root testing; and Section 5 concludes the paper.

2. Decision-Based Level of Significance for Unit Root Tests

There are renewed calls and growing evidence that hypothesis testing should be conducted at the level of significance chosen in consideration of the information specific to the problem under question. Kim and Ji (2015) demonstrate how Leamer (1978) method can be applied to empirical research in finance for more credible significance testing. Perez and Pericchi (2014) propose a calibration formula for the adaptive level of significance in the context of simple hypothesis testing, derived as a compromise between Bayesian and non-Bayesian methods. Pericchi and Pereira (2016) propose a general approach to determining the adaptive significance level by minimizing a weighted sum of Type I and II error probabilities, showing that their approach is closely related with the inference based on the Bayes ratio. In this section, we use the method based on Leamer (1978) line of enlightened judgement for the ADF and DF–GLS tests, which is a special case of the general approach of Pericchi and Pereira (2016). We present decision-based significance levels for these unit root tests under a range of sample sizes widely encountered in practice. We also examine the effects of other factors (prior probability, relative loss, starting values of the series) that can influence the choice of the decision-based level.

2.1. Decision-Theoretic Approach to Unit Root Testing

In this paper, we propose a decision-theoretic approach to unit root testing. Under this approach, we consider the losses from Type I and II errors of hypothesis testing, and choose the decision that minimizes the expected loss (see, for example, Arrow 1960; Leamer 1978; Das 1994; Poirier 1995). We observe that empirical researchers often make decisions in a dichotomous and mechanically way, based on an arbitrary threshold set by a conventional level of significance. We are concerned that this practice has rendered many empirical researchers make too rash decisions in favour of a unit root. As a result, there are a number of stylized facts that are not consistent with economic reasoning. For example, while most of economists believe that real exchange rates and interest rates are stationary, a large body of empirical studies report that these time series have a unit root. We argue that these stylized facts, inconsistent with economic reasoning or theories, are the results of conducting unit root tests at a conventional significance level. This point is in line with the concerns raised by the American Statistical Association (Wasserstein and Lazar 2016), with a statement that "Widespread use of 'statistical significance' (generally interpreted as '$p < 0.05$') as a license for making a claim that a scientific finding (or implied truth) leads to considerable distortion of the scientific process".

We note that failure to reject the null hypothesis of a unit root does not necessarily indicate the full support for a unit root. It tells us that we cannot rule out the presence of a unit root, while the alternative hypothesis is also likely. Under the decision-theoretic approach, we consider the expected loss under a range of the key parameters and choose the most likely action. With this approach, the researcher's prior belief for H_0 and H_1; and relative loss between Type I and II errors play important roles, which are totally ignored in conventional hypothesis testing. If the researcher believes that a unit root is unlikely based on economic reasoning, she should assign a low value to the prior probability for H_0 and/or a low value to the loss of Type I error (relative to that of Type II error). In so doing, the researcher is likely to reject the unit root hypothesis. As we shall see later in our applications, the use of conventional significance level for unit root testing is associated with the case where the researcher a priori believes that the null hypothesis of a unit root is highly likely and/or she assigns a high loss to Type I error (relative to that of Type II error). In many applications (e.g., real exchange rate or interest rates), such prior belief and relative loss values implied by a conventional significance level may not be consistent with economic reasoning. This decision-theoretic approach is also closely related with the Bayesian view of unit root test, where the decision is made based on the posterior odds ratio (see, for example, Sims 1988; Sims and Uhlig 1991)[5]. The latter compares the posterior probabilities under H_0 and H_1, which aid the decision to choose a more likely hypothesis (see Startz 2014).

2.2. Line of Enlightened Judgement and Decision-Based Significance Levels

It is well known that a trade-off between the two error probabilities (α and β) of hypothesis testing exists, with a higher (lower) value of α associated with a lower (higher) value of β. When α is set at 0.05, a low power means that the value of β is much higher than 0.05. For example, if the power is as low as 0.20, there is a serious imbalance between α and β, with the latter being 16 times higher than the former. As a result, the test is severely biased towards Type II error, with a consequence that a false null hypothesis fails to be rejected. By choosing a higher value of α in this case, say 0.3, one can achieve a balance between α and β, obtaining a higher power at the same time. The line of enlightened judgement (Leamer 1978) is formulated by plotting the combination of all possible α and β values, from which one can choose a desired combination in explicit consideration of the power and losses under Type I and II errors.

In what follows, the line of enlightened judgement for unit root tests is presented. Following DeJong et al. (1992), we consider the time series:

$$Y_t = \gamma_0 + \gamma_1 t + X_t;\ X_t = \tau_1 X_{t-1} + \ldots + \tau_p X_{t-p} + u_t,\ (t = 1, \ldots, n), \tag{1}$$

where u_t is an independent error term with zero mean and fixed variance σ^2. The standardized initial value of (1) is denoted as $X_0^* \equiv X_0/\sigma = (Y_0 - \gamma_0)/\sigma$. The model (1) can be re-written in the ADF form as

$$\Delta Y_t = \delta_0 + \delta_1 t + \lambda Y_{t-1} + \sum_{j=1}^{p-1} \rho_j \Delta Y_{t-j} + u_t. \tag{2}$$

Taking the ADF test as an example, the test statistic for H_0: $\lambda = 0$; H_1: $\lambda < 0$ is ADF $= \hat{\lambda}/\text{se}(\hat{\lambda})$, where $\lambda \equiv (\tau - 1)$ and $\tau = \sum_{i=1}^{p} \tau_i$. Let $\hat{\lambda}$ be the least-squares (LS) estimator for λ and $\text{se}(\hat{\lambda})$ denote its standard error estimator. Note that $\alpha = P(\text{ADF} < CR(\alpha) \mid \lambda = 0)$, where $CR(\alpha)$ is the α-level critical value and $\beta = P(\text{ADF} > CR(\alpha) \mid \lambda < 0)$. The line of enlightened judgement is obtained by plotting all possible combinations of α and β.

A Monte Carlo experiment with the number of trials 10,000 is conducted to calculate the (α, β) values, using MacKinnon (1996) critical values. Following DeJong et al. (1992), the data is generated from model (2) with $p = 2$, $\rho_1 = 0.5$, $\delta_0 = \delta_1 = 0$, setting $\lambda = \lambda_1$, where λ_1 is a value of λ under H_1. In evaluating the power of the test, the choice of the value of λ_1 is important: in this study, we are guided by a past seminal study.[6] According to DeJong et al. (1992), $\lambda_1 \in [-0.15, 0]$ is a plausible range of the parameter values under H_1. In particular, they recommend $\lambda_1 = -0.15$ for annual time series and $\lambda_1 = -0.05$ and -0.01 for quarterly and monthly data, respectively. From a grid of α values between 0.01 and 0.99 with an increment of 0.02, the proportion of Type II error is obtained as an estimate of β. The standardized initial value $X_0{}^*$ is set initially at 1.5, which is the most plausible value according to DeJong et al. (1992).[7]

We also present the line of enlightened judgement for the DF–GLS test of Elliott et al. (1996), which is well known to have a higher power when the initial value is small. The test involves a simple

[5] For a comprehensive review of Bayesian unit root tests, see Choi (2015, Chapter 4).

[6] We note that this choice may be specific to the nature of the application at hand. This is also related with what Ziliak and McCloskey (2008) called the minimum oomph, which is the smallest value where the null hypothesis is violated economically.

[7] This choice will be generalized in Section 3 where the calibration rules for the decision-based significance levels are constructed under a wide range of starting values.

modification to the ADF test by de-trending the deterministic component using the GLS method. For the ADF and DF–GLS tests (for the model with a constant only), the lines of enlightened judgement are constructed using appropriate MacKinnon (1996) critical values. For the model with a constant and a linear trend, the DF–GLS test statistic follows a limiting distribution different from that of the ADF. For this case, we use the critical values from the asymptotic distribution of the test statistic obtained by simulation following Elliott et al. (1996).[8]

According to Leamer (1978), the expected loss from hypothesis testing is $p\alpha L_1 + (1-p)\beta L_2$, where $p \equiv P(\mathrm{H}_0)$, L_1 represents the loss of Type I error and L_2 that from the Type II error. Given the combinations of (α, β) values on the line of enlightened judgement, the level of significance α^* can be chosen so that the expected loss is minimized. The value of Type II error probability corresponding to α^* is denoted as β^*. The specific values of p, L_1 and L_2 depend on contexts and the researcher's prior belief. For the purpose of simplicity, we initially assume that $p = 0.5$ and $L_1 = L_2$, by which the minimization of the expected loss is simplified to that of $\alpha + \beta$. These assumptions mean that the researcher gives an equal weight to the two states of nature (H_0 and H_1) with a prior belief that: firstly, they are equally likely to be true, and, secondly, the losses from Type I and II errors are identical. In the analysis that follows, we will allow for general values of p and L's. In this paper, α^* is referred to as the decision-based significance level.

Figure 1 presents the lines of enlightened judgement for the ADF and DF–GLS tests for the model with a constant and a linear trend, when $\lambda_1 = -0.15$, under the sample sizes ranging from 60 to 130. These settings are suitable for annual time series. The line shifts towards the origin as the sample size increases, corresponding to lower values of β (or higher power) for a given value of α. The blue square dots represent the points of (α^*, β^*), where the expected loss ($\alpha + \beta$) is minimized. The decision-based significance levels of the DF–GLS test are much lower than those of the ADF test due to its higher power. For all sample sizes, they are in the neighborhood of 0.3, except when $n = 60$ for the ADF test, which is consistent with Winer (1962) assertion. They also decrease with the sample size as Leamer (1978) suggests. From Figure 1, when $n = 100$ and $\alpha = 0.05$, $\beta = 0.74$ and the power of the ADF test is only 0.26: i.e., as mentioned above, a case of low power with an obvious imbalance between α and β. However, if α is chosen to minimize the expected loss ($\alpha + \beta$), (α^*, β^*) = (0.31, 0.22) with a substantially higher power of 0.78. The expected loss is much higher when $\alpha = 0.05$, as expected. The critical value at the decision-based significance level is -2.54, which is much larger than the 5% critical value of -3.46. Similar results are evident for the DF–GLS test. Table 1 presents complete listing of the values indicated in Figure 1.

Figure 2 presents the lines of enlightened judgement associated with the ADF and DF–GLS tests for the model with a constant only, when $\lambda_1 = -0.05$ for the sample sizes ranging from 80 to 240. These settings are suitable for quarterly time series. Again, higher power associated with the DF–GLS test is evident with the lines for the DF–GLS test much closer to the origin than those of the ADF. When the sample size is 120 and $\alpha = 0.05$, the DF–GLS test is again severely biased towards the Type II error, with its β value more than 11 times higher than that of α. The power is 0.44, which is much higher than that of the ADF which is 0.13, as expected. However, at the decision-based significance level $\alpha^* = 0.23$, the DF–GLS test enjoys a substantially higher power of 0.92 with a balance between the two error probabilities. Overall, the decision-based significance levels for the DF–GLS test are in the neighborhood of 0.20 for a typical quarterly time series. Table 2 presents complete listing of the values indicated in Figure 2.

From Figures 1 and 2, we observe a tendency where the decision-based level (α^*) is more or less twice the size of the corresponding Type II error probability (β^*). This means that the test at the decision-based significance level tends to be conservative about the Type II error, in contrast with

[8] Cheung and Lai (1995) provide response surface estimates for the critical values of the DF–GLS test for the model with a linear trend. However, they are only applicable for 5% and 10% levels of significance.

the case of $\alpha = 0.05$ where the test is severely biased towards Type II error. It should be noted that the conventional levels of significance (such as 0.05) represent a poor benchmark level for these tests because they cannot be optimal under any sample sizes frequently encountered in practice. The level of significance in the 0.2 to 0.4 range may be "outrageously high" in comparison with the 0.05 level as Arrow (1960, p. 73) puts it, but the test will enjoy a considerably higher power with a balance between Type I and II error probabilities.

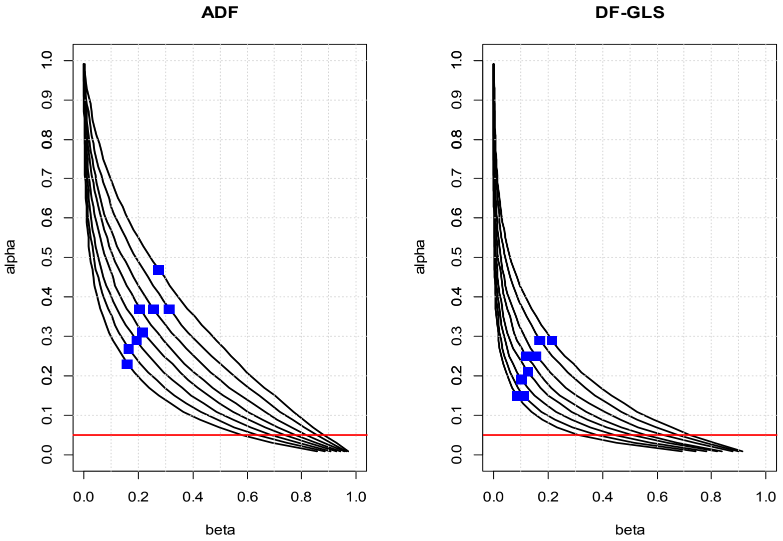

Figure 1. Lines of enlightened judgement ($\lambda_1 = -0.15$; model with a constant and a linear trend). The lines of enlightened judgement are plotted in black, corresponding to the sample sizes n = (60, 70, 80, 90, 100, 110, 120, 130) from the far right to the left. The red horizontal lines correspond to $\alpha = 0.05$. The square dots indicate the points where $\alpha + \beta$ is minimized.

Table 1. The values of α, β, and Power from Figure 1.

ADF Test								
n	α	β	Power	CR	α^*	β^*	Power	CR^*
	α fixed at 0.05				minimize $\alpha + \beta$			
60	0.05	0.88	0.12	−3.49	0.47	0.27	0.73	−2.22
70	0.05	0.86	0.14	−3.48	0.37	0.31	0.69	−2.41
80	0.05	0.82	0.18	−3.47	0.37	0.25	0.75	−2.41
90	0.05	0.78	0.22	−3.46	0.37	0.21	0.79	−2.41
100	0.05	0.74	0.26	−3.46	0.31	0.22	0.78	−2.54
110	0.05	0.69	0.31	−3.46	0.29	0.19	0.81	−2.58
120	0.05	0.64	0.36	−3.45	0.27	0.16	0.84	−2.63
130	0.05	0.58	0.42	−3.44	0.23	0.16	0.84	−2.72
DF–GLS Test								
n	α	β	Power	CR	α^*	β^*	Power	CR^*
	α fixed at 0.05				minimize $\alpha + \beta$			
60	0.05	0.72	0.28	−2.89	0.29	0.21	0.79	−2.05
70	0.05	0.68	0.32	−2.89	0.29	0.17	0.83	−2.05
80	0.05	0.62	0.38	−2.89	0.25	0.15	0.85	−2.13
90	0.05	0.56	0.44	−2.89	0.25	0.12	0.88	−2.13
100	0.05	0.51	0.49	−2.89	0.21	0.12	0.88	−2.24
110	0.05	0.44	0.56	−2.89	0.19	0.10	0.90	−2.29
120	0.05	0.38	0.62	−2.89	0.15	0.11	0.89	−2.40
130	0.05	0.31	0.69	−2.89	0.15	0.08	0.92	−2.40

CR: Critical value at 5% level; *CR**: Critical value associated with α^*.

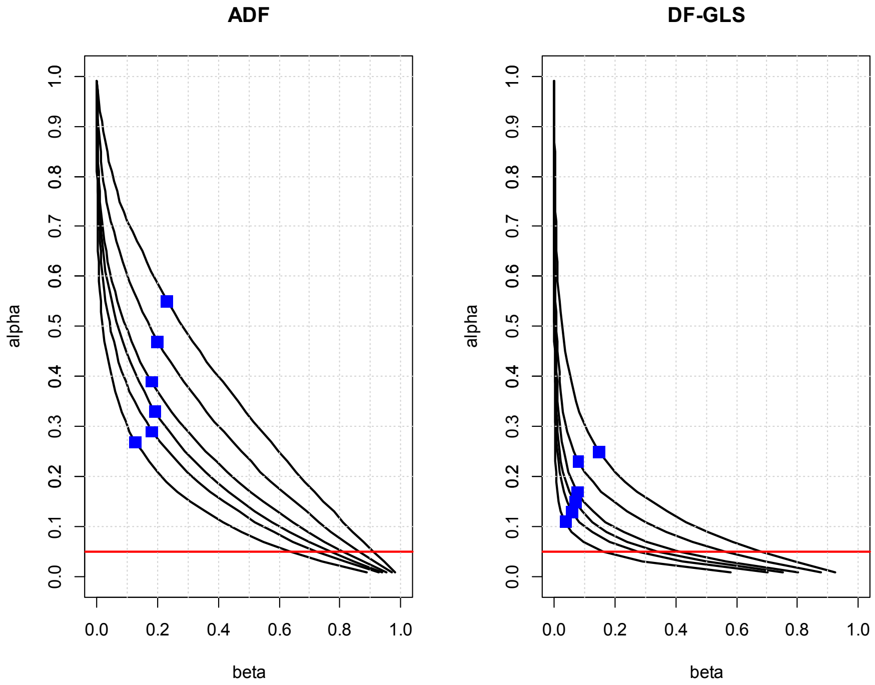

Figure 2. Lines of enlightened judgement ($\lambda_1 = -0.05$; model with a constant). The lines of enlightened judgement are plotted in black, corresponding to the sample sizes n = (80, 120, 160, 180, 200, 240) from the far right to the left. The red horizontal lines correspond to $\alpha = 0.05$. The square dots indicate the points where $\alpha + \beta$ is minimized.

Table 2. The values of α, β, and Power from Figure 2.

			ADF Test					
n	α	β	Power	CR	α^*	β^*	Power	CR^*
	α fixed at 0.05				minimize $\alpha + \beta$			
80	0.05	0.91	0.09	−2.90	0.55	0.23	0.77	−1.46
120	0.05	0.87	0.13	−2.89	0.47	0.20	0.80	−1.62
160	0.05	0.81	0.19	−2.88	0.39	0.18	0.82	−1.78
180	0.05	0.78	0.22	−2.88	0.33	0.19	0.81	−1.90
200	0.05	0.73	0.27	−2.88	0.29	0.18	0.82	−1.99
240	0.05	0.63	0.37	−2.87	0.27	0.13	0.87	−2.04
			DF–GLS Test					
n	α	β	Power	CR	α^*	β^*	Power	CR^*
	α fixed at 0.05				minimize $\alpha + \beta$			
80	0.05	0.68	0.32	−1.94	0.25	0.15	0.85	−1.08
120	0.05	0.56	0.44	−1.94	0.23	0.08	0.92	−1.14
160	0.05	0.41	0.59	−1.94	0.17	0.08	0.92	−1.33
180	0.05	0.34	0.66	−1.94	0.15	0.07	0.93	−1.40
200	0.05	0.27	0.73	−1.94	0.13	0.06	0.94	−1.48
240	0.05	0.16	0.84	−1.94	0.11	0.04	0.96	−1.57

CR: Critical value at 5% level; *CR**: Critical value associated with α^*.

2.3. Factors Affecting the Decision-Based Significance Level

Koop and Steel (1994, p. 99) consider the lack of formal development of loss function as a serious weakness of both Bayesian and classical unit root studies. They argue that the classical analysis has an implicitly defined loss function in choosing the level of significance, in which losses are asymmetric. That is, the use of a conventional level of significance (such as 0.05) implies an arbitrarily asymmetric loss function. While our analysis so far assumes a symmetric loss function ($L_1 = L_2$), it is possible that the value of the decision-based level changes in response to different values of relative loss from Type I and II errors. In addition, there are other factors that possibly affect the decision-based level; i.e., the probability for the null hypothesis (p) that is so far assumed to be 0.5; and the starting value of the series that may affect the power of a unit root test.

To examine the effects of the prior probability for H_0 and the relative loss, Figure 3 plots the decision-based significance levels for the ADF and DF–GLS tests (model with a constant only) as a function of p and relative loss. Letting $k = L_2/L_1$ (relative loss) and setting $L_1 = 1$ without loss of generality, the expected loss is expressed as $p\alpha + (1 - p)\beta k$. These values are calculated from the lines of enlightened judgement given in Figure 2 when $n = 120$, by minimizing the expected loss $p\alpha + (1 - p)\beta k$ under different values of p and k. It appears that they change sensitively to the value of p and k, and that the conventional levels of significance (such as 0.05 and 0.01) are justifiable only when p is high and k is low. That is, either when the researcher has a strong prior belief that H_0 is true (presence of a unit root) or when the loss from Type I error considerably outweighs the loss associated with Type II error. In the opposite case, the decision-based level can often be far higher than 0.50 for both tests. Under moderate values of p and k, the decision-based levels are in the 0.2 to 0.4 range.

It is well known that the power of unit root test changes sensitively to the initial value and the degree of autocorrelation of the error term (see DeJong et al. 1992; Müller and Elliott 2003). To examine the sensitivity, the decision-based levels of significance for the ADF and DF–GLS tests (model with a constant only) are reported in Table 3 when $n = 120$, under a range of X_0* and ρ_1. For the ADF test, under a reasonable value of X_0* (0 to 5), it appears that the decision-based level of significance is not sensitive to ρ_1. For the DF–GLS test, it changes sensitively, especially when the value of ρ_1 is negative, and tends to increase with the starting value. The decision-based level's sensitivity to the values of p and k and the starting value of the series are taken into account in the calibration rules that will be discussed in the following section.

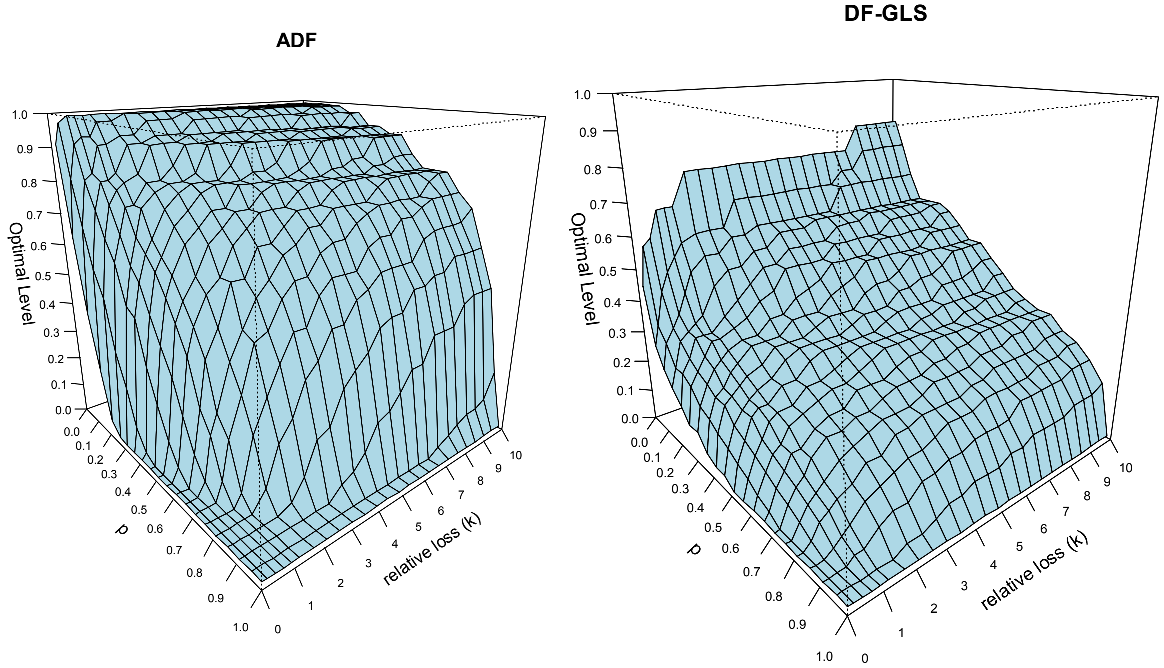

Figure 3. Decision-based level of significance, prior probability, and relative loss. Each figure plots the decision-based level of significance that minimizes the expected loss against $p = P(H_0)$ and $k = L_2/L_1$ (relative loss), for the ADF and DF–GLS tests (models with a constant only) when $n = 120$. A grid of p-values between 0 and 1 is used along with a grid of k-values between 0 and 10. All other settings for calculation are the same as those in Figure 2.

Table 3. Decision-based significance level, autocorrelation coefficient, and initial value.

	ADF				DF–GLS			
ρ_1	−0.5	0	0.5	0.9	−0.5	0	0.5	0.9
X_0*								
0	0.45	0.47	0.51	0.51	0.19	0.21	0.21	0.27
1	0.47	0.45	0.49	0.47	0.23	0.21	0.21	0.27
2	0.45	0.49	0.51	0.53	0.29	0.25	0.21	0.27
3	0.45	0.47	0.51	0.51	0.35	0.29	0.23	0.27
4	0.45	0.49	0.49	0.47	0.45	0.35	0.25	0.27
5	0.43	0.47	0.47	0.49	0.55	0.41	0.27	0.29
6	0.37	0.43	0.47	0.49	0.63	0.49	0.31	0.29
8	0.33	0.37	0.45	0.51	0.75	0.59	0.35	0.27
10	0.27	0.35	0.43	0.47	0.81	0.67	0.43	0.27

The entries are the decision-based significance level when $P(\mathrm{H}_0)$ = 0.5, L_1 = L_2, n = 120. ρ_1: the coefficient of the augmentation term given in (2); X_0*: standardized starting value of (1).

3. Calibration Rules Based on Asymptotic Local Power

In the previous section, we demonstrate how the decision-based significance level can be chosen in small samples. However, the choice depends on a range of factors such as sample size, value of λ_1, degree of autocorrelation, and data frequency. We also observe that the prior probability of the null hypothesis (p), relative loss from Type I and II errors (k), and starting values of the series (X_0*) play their roles. To simplify the choice in practice, it is useful to consider the asymptotic local power of a unit root test, which depends largely on the local-to-unity coefficient. There are advantages of using the asymptotic local alternatives: first, we do not have to fix the sample size as in Leamer (1978) method. Second, the value of the coefficient under H_1 does not need to be specified, but can be estimated from the data. Third, the values of nuisance parameters that are asymptotically negligible, such as the degree of autocorrelation, do not have to be specified. Building on this idea, we develop simple calibration rules for the decision-based significance levels of unit root tests, which use the value of local-to-unity coefficient as a key input.

To achieve this, we follow Elliott et al. (1996) to generate the asymptotic local power as a function of local-to-unity coefficient $c \equiv n(1-\tau)$, estimated as $\hat{c} = n(1-\hat{\tau})$, where $\hat{\tau}$ is the LS estimator for τ. Since $\hat{c} = c + O_p(1)$, we note that the choice of $\hat{c}$ should be made carefully in practice. We propose that, to improve the estimation of $\hat{\tau}$ in small samples, bias-correction proposed by Stine and Shaman (1989) and Kim (2004) be employed (see Sections 4.4 and 4.5). In addition, we propose that empirical researchers try a range of values around $\hat{c}$ for c when they use the calibration rule. The researcher may choose this value subjectively based on economic reasoning or descriptive measures such as time plot or autocorrelation function. If those chosen values provide largely consistent inferential outcomes, we may be able to arrive at an optimal statistical decision.

Figure 4 presents the lines of enlightened judgement for the ADF and DF–GLS tests (the model with a constant only) under a selected values of c when n = 500. All other computational details are the same as before. As might be expected, the decision-based significance level α^* is a decreasing function of c. This is because the tests gain a higher power as the model moves away from the unit root. To obtain the calibration rules, we calculate the asymptotic local power (and the values of β) for a grid of c values ranging from 0.1 to 30 with an increment of 0.6, under different values of p- and k-values used in Figure 3; and X_0* values ranging from 0 to 5. For all combinations, the decision-based level is chosen so that the expected loss $p\alpha$ + $(1-p)\beta k$ is minimized. In addition to the ADF and DF–GLS tests, we obtain the calibration rules for the Phillips–Perron and ERS–P tests, which are also widely used in practice. For each test, we calculate 157,700 decision-based levels from the combinations of p, k, c and X_0* values (21 × 25 × 50 × 6).

Our initial estimates of response surfaces are summarized as follows:

- Model with a constant only

  ADF: $\hat{\alpha}^* = 0.825 - 0.505p + 0.028k - 0.025c - 0.002X_0^*$

  Phillips–Perron: $\hat{\alpha}^* = 0.827 - 0.516p + 0.029k - 0.025c - 0.002X_0^*$

  DF-GLS: $\hat{\alpha}^* = 0.525 - 0.339p + 0.018k - 0.019c + 0.025X_0^*$

  ERS-P: $\hat{\alpha}^* = 0.509 - 0.309p + 0.017k - 0.018c + 0.026X_0^*$
- Model with a constant and a linear trend

  ADF: $\hat{\alpha}^* = 0.933 - 0.655p + 0.037k - 0.023c - 0.001X_0^*$

  Phillips–Perron: $\hat{\alpha}^* = 0.932 - 0.665p + 0.037k - 0.023c - 0.001X_0^*$

  DF-GLS: $\hat{\alpha}^* = 0.802 - 0.546p + 0.031k - 0.024c + 0.017X_0^*$

  ERS-P: $\hat{\alpha}^* = 0.793 - 0.541p + 0.031k - 0.023c + 0.018X_0^*$

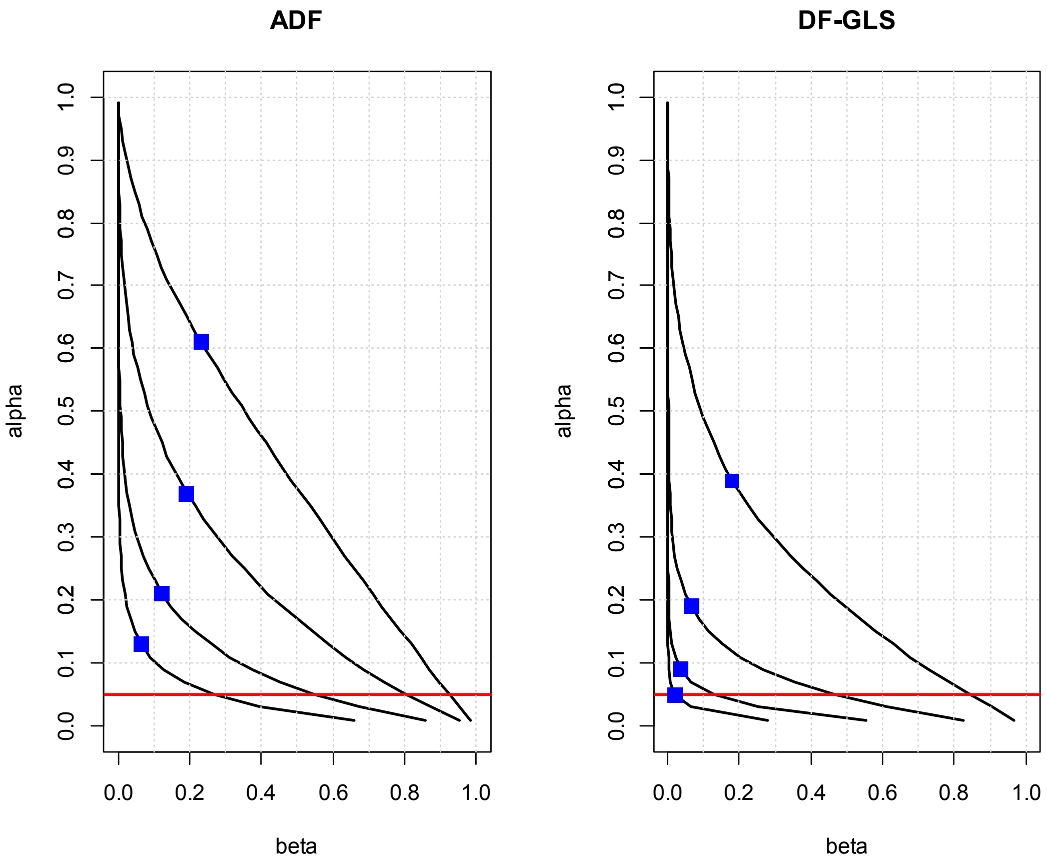

Figure 4. Lines of enlightened judgement based on asymptotic local power (model with a constant). The lines of enlightened judgement are plotted in black, corresponding to the local-to-unity coefficients c = (2.5, 7.5, 12.5, 17.5) from the far right to the left (n = 500). The red horizontal lines correspond to α = 0.05. The square dots indicate the points where $\alpha + \beta$ is minimized.

These response surfaces indicate that the decision-based significance level is related negatively to the local-to-unity coefficient (c) and $p \equiv P(\mathrm{H}_0)$, while it is positively related to the relative loss k. To the starting values, the decision-based levels of the ADF and Phillips–Perron tests show negative but small responses, while those of the DF–GLS and ERS–P tests show positive and large responses.

Although simple and easy-to-understand, the above response surfaces do not show satisfactory predictive abilities, with their R^2 values ranging from 0.64 to 0.75. To improve their predictive power, we fit the response surfaces for the decision-based significance level as a polynomial in c, given the values of p, k and X_0*. We specify α* as a sole function of c, conditionally on the values of (p, k, X_0*) as

$$\alpha^*(c \mid p, k, X_0^*) = a_0 + a_1c + a_2c^2 + a_3c^3 + a_4c^4 + a_5c^5. \tag{3}$$

The above response surfaces are estimated from 50 pairs of (α^*, c) values. These response surfaces provide high R^2 values often close to one, with satisfactory predictive ability. An estimate of the decision-based level is obtained by plugging the estimated values of c and a's into (3). That is,

$$\hat{\alpha}^*(c \mid p, k, X_0^*) = \hat{a}_0 + \hat{a}_1c + \hat{a}_2c^2 + \hat{a}_3c^3 + \hat{a}_4c^4 + \hat{a}_5c^5,$$

where $\hat{a}_i$'s are the LS estimators for a_i's. The estimated value of α* is obtained as $\hat{\alpha}^*(\hat{c} \mid p, k, X_0^*)$. As we shall see in Section 4, $\hat{c}$ can be estimated using a bias-corrected estimator for $\hat{\tau}$, while the initial value is estimated as $\hat{X}_0^* = (Y_0 - \hat{\gamma}_0)/\hat{\sigma}$, where $\hat{\gamma}_0$ and $\hat{\sigma}$ are estimated values for γ_0 and σ in (1). The values of p and k are chosen by the researcher, while p = 0.5 and k = 1 represent the case where a researcher is neutral or impartial between Type I and II errors.

Table 4 reports the estimated values of the decision-based significance level calculated in this way against the true values, for a range of p- and k-values when X_0* = 3 and c = 10.3. It can be seen that the estimated values are fairly close to the true values indicating satisfactory performance of the calibrated rules given in (3). Overall, the root mean squared forecast error is around 0.01 on average, with the standard deviation of around 0.007. The conventional levels are close to the decision-based counterparts, only when the value of p is high and the value of k is low. This means that they are optimal only when the researcher places a heavier weight on the null hypothesis. In Section 4.4, we demonstrate with an example of how these calibration rules can be used in practice. The *R* function calculating the estimated decision-based significance levels from these calibrations rules is available in the Supplementary Materials.

Table 4. Decision-based level of significance: exact vs. estimated values.

(*p*, *k*)	**ADF**		**DF–GLS**	
	True	**Estimated**	**True**	**Estimated**
(0.25, 0.25)	0.27	0.30	0.25	0.26
(0.25, 1)	0.63	0.65	0.49	0.49
(0.25, 4)	0.83	0.84	0.67	0.67
(0.5, 0.25)	0.03	0.03	0.07	0.05
(0.5, 1)	0.39	0.37	0.29	0.30
(0.5, 4)	0.71	0.70	0.53	0.53
(0.75, 0.25)	0.01	0.01	0.01	0.01
(0.75, 1)	0.07	0.06	0.11	0.10
(0.75, 4)	0.47	0.46	0.37	0.36
	ADF		**DF–GLS**	
	True	**Estimated**	**True**	**Estimated**
(0.25, 0.25)	0.27	0.29	0.25	0.26
(0.25, 1)	0.63	0.64	0.49	0.49
(0.25, 4)	0.83	0.84	0.69	0.67
(0.5, 0.25)	0.03	0.03	0.07	0.05
(0.5, 1)	0.35	0.36	0.31	0.31
(0.5, 4)	0.67	0.69	0.55	0.54
(0.75, 0.25)	0.01	0.01	0.01	0.01
(0.75, 1)	0.07	0.06	0.11	0.10
(0.75, 4)	0.43	0.43	0.37	0.35

The model with intercept and time trend is used, setting X_0^* = 3 and c = 10.3. The estimated values are obtained from the calibration rules given in (3). ($p \equiv P(\text{H}_0)$, $k = L_2/L_1$).

4. Re-Evaluation of Past Empirical Results

In this section, using the decision-based significance level and the calibration rules, we examine the extended Nelson–Plosser data set for U.S. macroeconomic time series; the real exchange rates covered by Elliott and Pesavento (2006); the real interest rates studied by Rapach and Weber (2004); and the nominal interest rates used by Neely and Rapach (2008).

4.1. Extended Nelson–Plosser Data

Table 5 reports the results for the extended Nelson–Plosser data. With the ADF test at the 5% level, every time series, excepting the real GNP, real per capita GNP, and unemployment rate, is found to have a unit root. The real GNP and real per capita GNP have their *p*-values close to 0.05, which leads to accepting the null hypothesis at the 1% level of significance. However, if the decision-based significance level obtained in Figure 1 (α^* = 0.37) is utilized, the presence of unit roots in the real GNP and real per capita GNP is clearly rejected. At these levels, the employment and money stock series are also found to be trend-stationary, in contrast to the outcomes at the conventional level. Similar results are evident when the DF–GLS test is used. That is, the unit root hypotheses for the real GNP and real per capita GNP are rejected at the decision-based significance level (α^* = 0.25), and so are those for the employment and money stock series (α^* = 0.21). For all the other time series, the inferential outcomes of the ADF and DF–GLS tests are consistent at the conventional and decision-based levels of significance.

It is interesting to observe that the results at the decision-based level are largely in agreement with the Bayesian results of Schotman and Dijk (1991). At the decision-based levels, nine time series are found to be difference-stationary, namely the nominal GNP, GNP deflator, consumer prices, wages, real wages, velocity, interest rate, industrial production, and common stock prices. Schotman and Dijk (1991) find eight of these (excluding industrial production) to have a unit root based on the Bayesian method (with posterior probability higher than 0.75). Overall, similarly to Schotman and Dijk (1991), we find that the real variables are found to be trend-stationary while the nominal ones are difference-stationary, at the decision-based significance levels. As mentioned earlier, it has been shown that the method of choosing the decision-based significance level is closely related with the Bayesian inference (see DeGroot 1975, p. 381; Pericchi and Pereira 2016).

Table 5. Extended Nelson–Plosser data: annual U.S. macroeconomic time series to 1988.

	n	**ADF**				**DF–GLS**			
		***p*-Value**	**Decision (α = 0.01/0.05)**	**α^***	**Decision***	***p*-Value**	**Decision (α = 0.01/0.05)**	**α^***	**Decision***
Real GNP	80	0.05	Accept	0.37	Reject	0.05	Accept	0.25	Reject
Nominal GNP	80	0.58	Accept	0.37	Accept	0.49	Accept	0.25	Accept
Real per capital GNP	80	0.04	Accept/Reject	0.37	Reject	0.06	Accept	0.25	Reject
Industrial Production	129	0.26	Accept	0.23	Accept	0.27	Accept	0.15	Accept
Employment	99	0.18	Accept	0.31	Reject	0.04	Accept/Reject	0.21	Reject
Unemployment Rate	99	0.01	Reject	0.31	Reject	0.01	Reject	0.21	Reject
GNP deflator	100	0.70	Accept	0.31	Accept	0.73	Accept	0.21	Accept
Consumer Prices	129	0.91	Accept	0.22	Accept	0.61	Accept	0.15	Accept
Wages	89	0.53	Accept	0.37	Accept	0.37	Accept	0.25	Accept
Real Wages	89	0.75	Accept	0.37	Accept	0.51	Accept	0.25	Accept
Money Stock	100	0.18	Accept	0.31	Reject	0.10	Accept	0.21	Reject
Velocity	120	0.78	Accept	0.27	Accept	0.87	Accept	0.15	Accept
Interest Rate	89	0.98	Accept	0.37	Accept	0.33	Accept	0.25	Accept
Common Stock Prices	118	0.64	Accept	0.27	Accept	0.63	Accept	0.15	Accept

A model with a constant and a linear trend is used for all series; the *p*-values of the ADF test are obtained from the statistics reported in Schotman and Dijk (1991) using MacKinnon (1969) method, while that of the DF–GLS is derived from the simulated asymptotic distribution; α^*: the decision-based significance level from Figure 1. Decision*: Decision for H_0 at α^* given in Figure 1. The lag orders used are the same as those used by Nelson and Plosser (1982).

4.2. Elliott–Pesavento Data

Elliott and Pesavento (2006) examine stationarity of fifteen currencies' real exchange rates using quarterly data from 1973 to 2003. To improve the power of unit root tests, they consider the co-variation of the real exchange rates with other relevant economic variables. In so doing, they report the results of the ADF and DF–GLS tests. Table 6 presents the ADF and DF–GLS statistics, along with the inferential outcomes based on the decision-based significance level obtained in Figure 2. For the ADF test, at the 5% level, all the exchange rates are found to have a unit root with all statistics greater than the critical value of −2.89. For the ERS–GRS test, at the 1% level, all the rates are found to have a unit root; while at the 5% level, all the rates except for those of Belgium, Denmark, France, Germany, Italy, Norway, and Sweden are found to have a unit root. Hence, the purchasing power parity is not strongly supported by both of the tests at the conventional levels of significance. However, at the decision-based levels obtained in Figure 2 (α^* = 0.47 for the ADF and α^* = 0.23 for the DF–GLS), both the ADF and DF–GLS tests reject the unit root hypothesis for all the real exchange rates, except for the Canadian rate. These results represent strong empirical support for the validity of purchasing power parity. We note that the two tests provide unanimous results when the decision-based levels of significance are employed.

Table 6. Elliott–Pesavento data: quarterly real exchange rates from 1973 to 2003 (n = 120).

	ADF				**DF–GLS**			
	Statistic	***p*-Value**	**Decision (α = 0.01/0.05)**	**Decision* (α^* = 0.47)**	**Statistic**	***p*-Value**	**Decision (α = 0.01/0.05)**	**Decision* (α^* = 0.23)**
Austria	−1.729	0.414	Accept	Reject	−1.155	0.225	Accept	Reject
Belgium	−2.319	0.168	Accept	Reject	−2.133	0.032	Accept/Reject	Reject
Canada	−1.297	0.629	Accept	Accept	−0.487	0.503	Accept	Accept
Denmark	−2.507	0.116	Accept	Reject	−2.318	0.020	Accept/Reject	Reject
Finland	−2.377	0.150	Accept	Reject	−1.847	0.061	Accept	Reject
France	−1.955	0.306	Accept	Reject	−1.965	0.047	Accept/Reject	Reject
Germany	−1.996	0.288	Accept	Reject	−2.006	0.043	Accept/Reject	Reject
Italy	−1.966	0.301	Accept	Reject	−1.975	0.047	Accept/Reject	Reject
Japan	−2.265	0.185	Accept	Reject	−1.208	0.207	Accept	Reject
Netherlands	−1.755	0.401	Accept	Reject	−1.714	0.082	Accept	Reject
Norway	−2.178	0.215	Accept	Reject	−2.135	0.032	Accept/Reject	Reject
Spain	−1.928	0.319	Accept	Reject	−1.478	0.130	Accept	Reject
Sweden	−2.219	0.201	Accept	Reject	−1.997	0.044	Accept/Reject	Reject
Switzerland	−2.499	0.118	Accept	Reject	−1.521	0.120	Accept	Reject
UK	−2.363	0.154	Accept	Reject	−1.703	0.084	Accept	Reject

The ADF and DF–GLS statistics (model with a constant only) are re-produced from Elliott and Pesavento (2006). The p-values are calculated using MacKinnon (1996) method. The asterisk indicates the rejection of the null hypothesis of a unit root at 5% level of significance. Decision*: Decision for H_0 at α^*. The critical value for the ADF test corresponding to α^* = 0.47 is −1.62, while that of the DF–GLS test corresponding to α^* = 0.23 is −1.14 for n = 120.

4.3. Rapach–Weber Data

Rapach and Weber (2004) employ a range of unit root tests to examine stationarity of real interest rates of a number of international capital markets. Using quarterly data from 1957 to 2000, they report results of the ADF and DF–GLS tests for 10 capital markets. The results are reported in Table 7. If the ADF test is used, the presence of unit root cannot be rejected at the 1% level for all the rates; while, at the 5% level, all the rates except for those of Denmark and the UK are found to have a unit root. Hence, the results are strongly in favor of the presence of unit root in real interest rates. At the decision-based significance level of 0.33, the results of the ADF test are largely reversed. That is, the unit root hypothesis is rejected for all the rates except for those of the Netherlands and New Zealand, providing the evidence that eight out of ten rates are stationary. With the DF–GLS test, the unit root hypothesis is rejected for one (four) of ten real interest rates, at the 1% (5%) level of significance. At the decision-based level of 0.15, the DF–GLS test rejects the null hypothesis for six of ten rates. Hence, at the decision-based significance level, both the unit root tests are in favor of the stationarity of real interest rates.

Table 7. Rapach–Weber data: quarterly real interest rates from 1957 to 2000 (n = 173).

	ADF				DF–GLS			
	Statistic	**p-Value**	**Decision (α = 0.01/0.05)**	**Decision* (α^* = 0.33)**	**Statistic**	**p-Value**	**Decision (α = 0.01/0.05)**	**Decision* (α^* = 0.15)**
Belgium	−2.22	0.200	Accept	Reject	−1.99	0.045	Accept/Reject	Reject
Canada	−2.12	0.237	Accept	Reject	−1.78	0.071	Accept	Reject
Denmark	−2.93	0.044	Accept/Reject	Reject	−1.33	0.169	Accept	Accept
France	−2.08	0.253	Accept	Reject	−2.23	0.025	Accept/Reject	Reject
Ireland	−2.35	0.158	Accept	Reject	−1.10	0.245	Accept	Accept
Italy	−2.42	0.138	Accept	Reject	−1.44	0.139	Accept	Reject
Japan	−2.49	0.120	Accept	Reject	−2.06	0.038	Accept/Reject	Reject
Netherlands	−1.44	0.562	Accept	Accept	−0.99	0.287	Accept	Accept
New Zealand	−1.35	0.606	Accept	Accept	−1.11	0.241	Accept	Accept
UK	−2.98	0.039	Accept/Reject	Reject	−2.64	0.009	Reject	Reject

The ADF and DF–GLS statistics (model with a constant only) are re-produced from Rapach and Weber (2004). The p-values are calculated using MacKinnon (1996) method. The asterisk indicates the rejection of the null hypothesis of a unit root at 5% level of significance. Decision*: Decision for H_0 at α^*. The critical value for the ADF test corresponding to α^* = 0.33 is −1.90, while that of the DF–GLS test corresponding to α^* = 0.15 is −1.40 when n = 180.

4.4. *Application of the Calibration Rules*

In this sub-section, we demonstrate how the calibration rules can be used in practice, allowing for general values of p and k. We take the real GNP from the extended Nelson–Plosser data set as an example, and employ the calibration rules for the Phillips–Perron and ERS–P tests. For the real GNP in natural log (denoted Y), LS estimation of an AR(2) model with a constant and a linear time trend provides the following results:

$$Y_t = 0.81 + 0.006t + 1.23Y_{t-1} - 0.41Y_{t-2}.$$

The Phillips–Perron statistic is −2.83 with the p-value of 0.19, indicating the acceptance of the unit root hypothesis at the 5% level of significance. The ERS–P test statistic is 5.64, leading to the acceptance of the null hypothesis of a unit root at the 1% level with its p-value slightly less than 0.05. Hence, at the conventional levels of significance, these two tests provide evidence that favors the presence of a unit root in the real GNP.

We estimate the local-to-unity coefficient as $\hat{c} \equiv n(1 - \hat{\tau})$, where $\hat{\tau}$ is an estimator for AR(1) coefficient. For an AR(p) model with $p > 1$, the sum of AR coefficient estimators is used as an estimator for $\hat{\tau}$. One may use the LS estimator, but it is well known to be biased in small samples under-estimating the value of τ, which may result in over-estimation of c. Due to this problem, we propose the use of Kim (2004) bias-corrected estimators for AR(p) parameters unbiased to order n^{-1}, which is a generalized version of the bias-corrected estimator for the AR(1) model of Orcutt and Winokur (1969). Kim (2004) method makes use of the asymptotic bias formulae derived by Stine and Shaman (1989), and employs Kilian (1998) stationarity-correction in the event that bias-correction pushes the model to non-stationarity[9]. The bias-corrected estimation gives

$$Y_t = 0.62 + 0.004t + 1.27Y_{t-1} - 0.40Y_{t-2}, \tag{4}$$

and the resulting estimate for c is 10.62. This value is substantially smaller than the estimate of c based on the LS estimator, which is 14.11. The estimate of the standardized starting value X_0* is 2.77.

The values of c and X_0* estimated from (4) indicate that the estimated decision-based significance level reported in Table 4 are applicable to this case. For the Phillips–Perron test with the p-value of 0.19, the null hypothesis of a unit root is rejected at all combinations of p and k since the p-value is lower than the estimated decision-based levels, except when $(p, k) \in \{(0.5, 0.25), (0.75, 0.25), (0.75, 1)\}$. For the

[9] The R package BootPR (Kim 2015) provides computational resources for this bias-corrected estimation.

ERS–P test whose p-value is slightly less than 0.05, the null hypothesis is rejected for all the cases, except when (p, k) = (0.75, 0.25). That is, the unit root hypothesis is supported only under special cases where either the researcher strongly believes in the presence of a unit root or her loss from Type I error is substantially higher than that of Type II error, as might be expected. The evidence strongly suggests that, in general, the unit root hypothesis for the real GNP cannot be supported at the decision-based significance level. As a further check of robustness, in Table 8, we present the values of decision-based levels calculated from the calibration rules for a range of c and X_0* values, under the assumption of the neutral researcher with p = 0.5 and k = 1. These levels indicate again that the unit root hypothesis of the real GDP cannot be defended under a wide range of scenarios.

Table 8. Decision-based levels of significance under a range of c and X_0* values (p = 0.5, k = 1).

	Phillips–Perron		**ERS–P**	
X_0*	**c = 5**	**c = 15**	**c = 5**	**c = 15**
0	0.52	0.25	0.49	0.17
5	0.50	0.24	0.49	0.27

c: the value of the local-to-unity coefficient; X_0*: standardized initial value.

As we have seen in this section, the values of decision-based significance level change under different values of p and k, which has a consequential impact on the inferential outcome. Hence, the values of p and k should be chosen carefully depending on the contexts of investigation. For example, if one strongly believes in the economic theory that the time series under investigation is stationary, she may choose a value of p close to 0. In the event that a Type II error leads to a huge loss relative to that of a Type I error, one may choose a large value of k. However, when such contexts do not dictate (as in many academic research and practical applications), the most reasonable values of p and k are 0.5 and 1. The use of a conventional level of significance as a routine benchmark implies that the researcher is employing arbitrary and unknown values of p and k, which may often be inconsistent with economic reasoning. We note that Startz (2014) also makes a similar point.

4.5. Decision-Based Significance Level under a Specific Loss Function

The analysis so far has been conducted under generic values of relative loss without introducing a specific loss function. In practice, a researcher may wish to use a loss function suitable to the application at hand. In this section, we conduct unit root testing of the U.S. nominal interest rate, employing the loss function Koop and Steel (1994) propose. The data is three-month Treasury bill rate, quarterly from 1953:01 to 2007:02 (219 observations), taken from Neely and Rapach (2008). The nominal interest rate is well expected to be stationary since it represents a rate of return set by the central bank to stabilize the economy. However, many previous empirical studies could not reject the unit root hypothesis (see, for example, Neely and Rapach 2008) at a conventional level of significance. For the model with constant only and with the order of augmentation 7, the ADF test statistic is -2.47 (p-value = 0.1376), while the DF–GLS test statistic is -1.56 (p-value = 0.1126), both indicating that the null hypothesis of a unit root cannot be rejected at the 10% level of significance. From the estimated model, the sum of AR coefficients (bias-corrected) is 0.9571, which gives the estimated value of c of 9.3748.

The loss function of Koop and Steel (1994) is based on the predictive variance of the model under H_0 and H_1. We consider a researcher who is concerned with the prediction of a U.S. nominal interest rate, whose losses from Type I and II errors depend on the degree of under-estimation or over-estimation of predictive variance. The predictive variance is given by

$$g_h(\lambda) = \sum_{i=0}^{h-1} \lambda^{2i} + \frac{2}{n(n^2-1)} \sum_{i=1}^{h} \sum_{i=1}^{h} r(i,j)\lambda^{2h-i-j},$$

where λ is the AR(1) coefficient, n is the sample size, and h is the forecasting period, while $r(i,j)$ is a constant term whose form is given in Koop and Steel (1994). In this paper, we set $\lambda = 1$ under H_0 and use the bias-corrected estimate of the AR(1) coefficient (or the sum of AR coefficients) under H_1. Koop and Steel (1994) propose the loss functions of the form

$$L_1 = \max(1, g(H_1)/g(H_0)) + \delta \max(1, g(H_0)/g(H_1)) - (1 + \delta),$$

$$L_2 = \max(1, g(H_0)/g(H_1)) + \delta \max(1, g(H_1)/g(H_0)) - (1 + \delta).$$

Note that $\delta \geq 1$ is the parameter that controls the asymmetry of loss function, which gives a heavier penalty when the predictive variance is under-estimated. If $\delta = 1$, the loss function is symmetric and equal penalty is imposed, resulting in the relative loss value (k) of 1. A value of δ greater than 1 means that under-estimation of predictive variance is more costly and a larger loss is incurred for Type I error ($k \equiv L_2/L_1 = 1/\delta$).

Table 9 presents the decision-based level of significance calculated using the calibration rule for the ADF and DF–GLS tests. We present the decision-based levels of significance calculated with a range of parameter values. The inferential outcomes are not much sensitive to different values of c and X_0^*, but they change dramatically to the values of p and δ. If the researcher believes that H_0 is unlikely or equally likely ($p \leq 0.5$) and/or she is neutral about the losses of Type I and II errors (δ = 1), then H_0 is clearly rejected under nearly all the values of c and X_0* considered: the decision-based levels are well above the p-values of the tests. However, if the researcher believes that H_0 is highly likely ($p = 0.9$) and is greatly concerned about the occurrence of Type I error ($\delta = 5$), then H_0 cannot be rejected under all the values of c and X_0* considered: the decision-based levels are smaller than the p-values. The decision-based levels for the latter case are close to the conventional levels, indicating that unit root testing conducted at a conventional level is associated with the researcher's prior belief and losses that are not consistent with economic reasoning. Given the economic nature of the nominal interest rate, which is widely believed to be stationary, a researcher who gives an equal or less weight to H_0 ($p \leq 0.5$ and/or $\delta = 1$) is likely to make an economically sensible decision.

As a further note, we stress that the conventional level of significance may be well justified for the situation where a higher weight should be given to Type I error, in the form of large p or δ values. We note that there are applications where such a weighting is appropriate. For example, if unit root testing is conducted as a pre-test for choosing a forecasting model, one may justifiably use a conventional level, in view of the findings of Diebold and Kilian (2000) that unit root tests are useful for selecting forecasting models. In other words, the choice of p- and δ-values should be made carefully in practice with consideration of the nature of application at hand. As discussed above, the values of c and X_0* are also important inputs for the decision-based level, but we note that test outcomes are insensitive to these values. We have conducted an extensive sensitivity analyses for the extended Nelson–Plosser data, Elliott–Pesavento data, and Rapach–Weber data examined in Section 4.1 to Section 4.3. We have found that the inferential outcomes at the decision-based significance levels are not sensitive to a wider range of c and X_0* values. The detailed results are presented in the Appendix A.

Table 9. Decision-based levels of significance under the Koop and Steel (1994) loss function (using the calibration rule) for the U.S. nominal interest rate.

	ADF		**DF–GLS**	
	$\delta = 1$	**$\delta = 5$**	**$\delta = 1$**	**$\delta = 5$**
$c = 9.37$				
$X_0{}^* = 1$				
$p = \mathrm{P}(\mathrm{H}_0)$				
0.1	0.64	0.45	0.24	0.17
0.5	0.32	0.05	0.13	0.07
0.9	0.01	0.01	0.03	0.01
$X_0{}^* = 3$				
$p = \mathrm{P}(\mathrm{H}_0)$				
0.1	0.63	0.45	0.32	0.22
0.5	0.31	0.05	0.18	0.09
0.9	0.01	0.01	0.03	0.01
$c = 7$				
$X_0{}^* = 1$				
$p = \mathrm{P}(\mathrm{H}_0)$				
0.1	0.89	0.70	0.51	0.33
0.5	0.48	0.01	0.25	0.07
0.9	0.01	0.01	0.01	0.01
$X_0{}^* = 3$				
$p = \mathrm{P}(\mathrm{H}_0)$				
0.1	0.88	0.70	0.55	0.37
0.5	0.48	0.01	0.28	0.07
0.9	0.01	0.01	0.01	0.01
$c = 13$				
$X_0{}^* = 1$				
$p = \mathrm{P}(\mathrm{H}_0)$				
0.1	0.45	0.30	0.14	0.11
0.5	0.21	0.07	0.09	0.05
0.9	0.02	0.01	0.03	0.01
$X_0{}^* = 3$				
$p = \mathrm{P}(\mathrm{H}_0)$				
0.1	0.43	0.30	0.22	0.17
0.5	0.21	0.07	0.13	0.08
0.9	0.02	0.01	0.03	0.01

δ: the value which determines the degree of asymmetry in the loss function of Koop and Steel (1994), $k = 1/\delta$, where $k = L_2/L_1$ is the relative loss while L_i represent the loss of Type i error (i = I, II). The values in red font are those less than the p-values of the ADF and DF–GLS test (0.1376 and 0.1126, respectively) for H_0 of a unit root in the U.S. nominal interest rate, which leads to the acceptance of H_0.

5. Conclusions

This paper re-evaluates the key past results of unit root testing at the decision-based significance level chosen based on the line of enlightened judgement (Leamer 1978). Previous studies exclusively adopt the conventional level, which are arbitrary and not optimal by any criterion. More importantly, the use of conventional level of significance completely ignores the low power associated with the unit root test. In this paper, we choose the level of significance by minimizing the expected loss from Type I and II errors, in explicit consideration of the power of the test, following Leamer (1978). When the level of significance is chosen under a symmetric loss function with an assumption that the null and alternative hypotheses are equally likely to be true, we find that these decision-based levels for the ADF and DF–GLS tests are in the range of 0.2 to 0.4 for the sample sizes frequently encountered in practice. These values are well above the conventional levels, consistent with Winer (1962) conjecture and Nelson and Plosser (1960) proposal when the power of the test is low. We also propose calibration rules based on asymptotic local power for several unit root tests, which are simple to use in practice with the value of local-to-unity coefficient as a key input.

At the decision-based significance levels, we find many time series in the extended Nelson–Plosser data set to be trend-stationary, including the real (per capita) GNP, unemployment rate, employment, and money stock. On the other hand, the price time series (consumer prices, wages, common stock prices, and GNP deflator) and nominal GDP are found to have a unit root. These findings are largely consistent with the Bayesian results of Schotman and Dijk (1991), who find that the real variables are trend-stationary while nominal ones are difference-stationary. For the real exchange rates studied by Elliott and Pesavento (2006), both the ADF and DF–GLS tests demonstrate strong support for purchasing power parity at the decision-based levels, in contrast with the results at a conventional level. In addition, most of the real interest rates studied by Rapach and Weber (2004) are found not to have a unit root, based on the ADF and DF–GLS tests at the decision-based levels of significance. We also apply the decision-based level of significance to testing for a unit root in the U.S. nominal interest rate under a specific loss function, and found that the null hypothesis of a unit root is highly likely to be rejected unless the researcher gives a heavier weight to the null hypothesis.

The results obtained in this study strongly suggest that the conventional levels of significance represent a rather poor benchmark for popular unit root tests. They may be justifiable only when the researcher has a strong prior belief that the unit root is present or when the loss of Type I error disproportionately outweighs that of Type II error, which often may be inconsistent with common economic reasoning in practice. We propose that empirical researchers take a decision-theoretic approach and choose the level carefully in consideration of a range of factors, including the power of the test, for more sound empirical analysis, especially in the context of unit root testing. Mindless and mechanical use of the *p*-value with the conventional levels should be avoided, as Engsted (2009) and Kim and Ji (2015) point out. This is especially so in light of the recent statement made by the American Statistical Association (Wasserstein and Lazar 2016) expressing grave concerns that improper use of the *p*-value criterion is distorting the scientific process and invalidating many scientific conclusions.

In response to the univariate unit root test's low power, a number of panel unit root tests with substantially higher power have been proposed (see, for an up-to-date review, Choi 2015). For these tests, it is highly likely that the decision-based significance level should be set at a much lower level than the conventional one. With a low probability of Type II error, a panel test at a conventional level (e.g., 0.05) may be severely biased towards the Type I error, with a consequence that a true null hypothesis is rejected too often. By lowering the level of significance, a balance between the two error probabilities can be attained (see, for a related discussion, Kim and Ji 2015). There are also a number of unit root tests that incorporate the effects of structural breaks (see, for a review, Choi 2015). The empirical results based on these tests may also be re-evaluated at the decision-based significance level. Its application to construction of confidence interval for the unit root coefficient (Andrews and Guggenberger 2014; Stock 1991) would also be an interesting exercise. We leave these lines of research as possible future research topics.

Supplementary Materials: The following are available online at http://www.mdpi.com/2225-1146/5/3/41/s1: *R* function that returns the decision-based significance levels based on the calibration rules.

Acknowledgments: All computations are conducted using *R* (R Core Team 2014) and its package urca (Pfaff 2008). The R codes used for this study are available on request. Constructive comments from Amelie Charles, Olivier Darne, Graham Elliott, Tom Engsted, Lutz Kilian, Tae Hwy Lee, Marc Paolella, Luis Pericchi, Abul Shamsuddin, Dick Startz, Tom Stanley, and Herman van Dijk are gratefully acknowledged. An earlier version of the paper was presented at the LEMNA, University of Nantes, Nantes, France; University of Toyama, Toyama, Japan; and at the 2016 Asian Meeting of the Econometric Society, Kyoto, Japan. We would also like to thank the referees and editors for their constructive comments.

Author Contributions: Both authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Further Sensitivity Analyses

The purpose of this appendix is to present further empirical analyses when the decision-based level of significance is applied to the extended Nelson–Plosser data, Elliott–Pesavento data (quarterly real exchange rates), and Rapach–Weber data (quarterly real interest rates), examined in Section 4.1 to Section 4.3 of the paper. In particular, we pay attention to the sensitivity of the results to the local-to-unity coefficients (c) and the standardized initial value (X_0*). We set p = 0.5; k = 1, in keeping with the empirical analyses in Section 4.1 to Section 4.3. Table A1 below presents the decision-based levels of significance obtained from the calibration rule given in Section 3 of the paper, for the ADF and DF–GLS tests. We consider a wider range of c and X_0* values, as in the paper.

Table A1. Decision-based levels of significance for a range c and X_0* values (p = 0.5, k = 1).

	Model with Constant Only				**Model with Constant and Time Trend**			
	c				***c***			
	2.5	7.5	12.5	17.5	2.5	7.5	12.5	17.5
X_0* = 1								
ADF	0.58	0.38	0.22	0.13	0.60	0.46	0.31	0.22
DF–GLS	0.40	0.17	0.09	0.05	0.56	0.41	0.20	0.14
X_0* = 3								
ADF	0.59	0.38	0.22	0.13	0.59	0.46	0.31	0.21
DF–GLS	0.41	0.21	0.14	0.09	0.56	0.41	0.23	0.17
X_0* = 5								
ADF	0.59	0.38	0.22	0.12	0.60	0.46	0.30	0.21
DF–GLS	0.44	0.28	0.21	0.17	0.55	0.44	0.28	0.22

The entries are the decision-based levels of significance obtained from the calibration rule given in Section 3 of the paper. c: the value of the local-to-unity coefficient; X_0*: standardized initial value; p = Prob(H_0); $k = L_2/L_1$.

Table A2 below presents inferential outcomes of the ADF and DF–GLS tests when the decision-based level of significance is applied to the extended Nelson–Plosser data set.

Table A2. Extended Nelson–Plosser data: annual U.S. macroeconomic time series to 1988 (Model with constant and linear trend).

	***p*-Value**		**Sensitivity Analysis**	
	ADF	**DF–GLS**	**ADF**	**DF–GLS**
Real GNP	0.05	0.05	H_0 is rejected for all c values	H_0 is rejected for all c values
Nominal GNP	0.58	0.49	H_0 is rejected when c = 2.5	H_0 is rejected when c = 2.5
Real per capital GNP	0.04	0.06	H_0 is rejected for all c values	H_0 is rejected for all c values
Industrial Production	0.26	0.27	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected only when c = 2.5, 7.5 (*)
Employment	0.18	0.04	H_0 is rejected for all c values	H_0 is rejected for all c values
Unemployment Rate	0.01	0.01	H_0 is rejected for all c values	H_0 is rejected for all c values
GNP deflator	0.70	0.73	H_0 is accepted for all c values	H_0 is accepted for all c values
Consumer Prices	0.91	0.61	H_0 is accepted for all c values	H_0 is accepted for all c values
Wages	0.53	0.37	H_0 is rejected when c = 2.5	H_0 is rejected when c = 2.5, 7.5
Real Wages	0.75	0.51	H_0 is accepted for all c values	H_0 is rejected when c = 2.5
Money Stock	0.18	0.10	H_0 is rejected for all c values	H_0 is rejected for all c values
Velocity	0.78	0.87	H_0 is accepted for all c values	H_0 is accepted for all c values
Interest Rate	0.98	0.33	H_0 is accepted for all c values	H_0 is rejected when c = 2.5, 7.5
Common Stock Prices	0.64	0.63	H_0 is accepted for all c values	H_0 is accepted for all c values

The asterisk (*) represents the case where the results are sensitive to the value of X_0*.

The results can be summarized as follows:

- The real GNP and real per capita GNP are found to be trend-stationary for all c and X_0* values, for both ADF and DF–GLS tests.
- The employment and money stock are found to be trend-stationary for all c and X_0* values, for both ADF and DF–GLS tests.
- The nominal GNP is found to be trend-stationary only when c = 2.5. For other c values, it is found to be difference-stationary.
- Other nominal and price variables are found to be difference-stationary for all values of c and X_0* values.

- The results are nearly the same as those reported in Section 4.1 of the paper, showing little sensitivity to the c and X_0^* values.

Table A3 below presents inferential outcomes of the ADF and DF–GLS tests when the decision-based level of significance is applied to the Elliott–Pesavento data set.

Table A3. Elliott–Pesavento data: quarterly real exchange rates from 1973 to 2003 (n = 120, Model with constant only).

	***p*-Value**		**Sensitivity Analysis**	
	ADF	**DF-GLS**	**ADF**	**DF-GLS**
Austria	0.414	0.225	H_0 is rejected when c = 2.5	H_0 is rejected when c = 2.5 (*)
Belgium	0.168	0.032	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected for all c values
Canada	0.629	0.503	H_0 is accepted for all c values	H_0 is accepted for all c values
Denmark	0.116	0.020	H_0 is rejected for all c values	H_0 is rejected for all c values
Finland	0.150	0.061	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected for all c values
France	0.306	0.047	H_0 is rejected when c = 2.5, 7.5	H_0 is rejected for all c values
Germany	0.288	0.043	H_0 is rejected when c = 2.5, 7.5	H_0 is rejected for all c values
Italy	0.301	0.047	H_0 is rejected when c = 2.5, 7.5	H_0 is rejected for all c values
Japan	0.185	0.207	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected when c = 2.5 (*)
Netherlands	0.401	0.082	H_0 is rejected when c = 2.5	H_0 is rejected for all c values
Norway	0.215	0.032	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected for all c values
Spain	0.319	0.130	H_0 is rejected when c = 2.5, 7.5	H_0 is rejected when c = 2.5, 7.5 (*)
Sweden	0.201	0.044	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected for all c values
Switzerland	0.118	0.120	H_0 is rejected for all c values	H_0 is rejected when c = 2.5, 7.5 (*)
UK	0.154	0.084	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected for all c values

The asterisk (*) represents the case where the results are sensitive to the value of X_0*.

The results can be summarized as follows:

- For the ADF test, most of real exchange rates are found to be stationary under a wide range of c values, showing little sensitivity to the X_0* value. For eight time series, the null hypothesis is rejected for all c values or $c \in \{2.5, 7.5, 12.5\}$.
- For the DF–GLS test, most of real exchange rates are found to be stationary under a wide range of c values. For ten time series, the null hypothesis is rejected for all c values considered.
- When the DF–GLS test is used, there are cases where the results are sensitive to the choice of X_0* value, which might be expected from the nature of the DF–GLS test.

Table A4 below presents inferential outcomes of the ADF and DF–GLS tests when the decision-based level of significance is applied to the Rapach–Weber data set.

Table A4. Rapach–Weber data: quarterly real interest rates from 1957 to 2000 (n = 173, Model with constant only).

	***P*-Value**		**Sensitivity Analysis**	
	ADF	**DF–GLS**	**ADF**	**DF–GLS**
Belgium	0.200	0.045	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected for all c values
Canada	0.237	0.071	H_0 is rejected when c = 2.5, 7.5	H_0 is rejected when c = 2.5, 7.5, 12.5 (*)
Denmark	0.044	0.169	H_0 is rejected for all c values	H_0 is rejected only when c = 2.5, 7.5
France	0.253	0.025	H_0 is rejected when c = 2.5, 7.5	H_0 is rejected for all c values
Ireland	0.158	0.245	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected only when c = 2.5 (*)
Italy	0.138	0.139	H_0 is rejected when c = 2.5, 7.5, 12.5	H_0 is rejected only when c = 2.5, 7.5 (*)
Japan	0.120	0.038	H_0 is rejected for all c values	H_0 is rejected for all c values
Netherlands	0.562	0.287	H_0 is rejected only when c = 2.5	H_0 is rejected only when c = 2.5, 7.5
New Zealand	0.606	0.241	H_0 is accepted for all c values	H_0 is rejected when c = 2.5, 7.5 (*)
UK	0.039	0.009	H_0 is rejected for all c values	H_0 is rejected for all c values

The asterisk (*) represents the case where the results are sensitive to the value of X_0*.

The results can be summarized as follows:

- For the ADF test most of real interest rates are found to be stationary under a wide range of c values, showing little sensitivity to the X_0* value. For six time series, the null hypothesis is rejected for all c values or $c \in \{2.5, 7.5, 12.5\}$. Two additional time series are found to be stationary when $c \in \{2.5, 7.5\}$.
- For the DF–GLS test, most of the real exchange rates are found to be stationary under a wide range of c values. For five time series, the null hypothesis is rejected for all c values considered or $c \in \{2.5, 7.5, 12.5\}$. Three additional time series are found to be stationary when $c \in \{2.5, 7.5\}$.
- When the DF–GLS test is used, there are cases where the results are sensitive to the choice of X_0* value, which might be expected from the nature of the DF–GLS test.

References

Andrews, Donald W. K., and Patrik Guggenberger. 2014. A Conditional-Heteroskedasticity-Robust Confidence Interval for the Autoregressive Parameter. *Review of Economics and Statistics* 96: 376–81. [CrossRef]

Arrow, Kenneth. 1960. Decision theory and the choice of a level of significance for the *t*-test. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Edited by Ingram Olkin. Palo Alto: Stanford University Press, pp. 70–8.

Campbell, John Y., and N. Gregory Mankiw. 1987. Are Output Fluctuations Transitory? *Quarterly Journal of Economics* 102: 857–80. [CrossRef]

Cheung, Yin-Wong, and Kon S. Lai. 1995. Lag order and critical values of a modified Dickey–Fuller test. *Oxford Bulletin of Economics and Statistics* 57: 411–19. [CrossRef]

Choi, In. 2015. *Almost All about Unit Roots: Foundations, Developments, and Applications*. New York: Cambridge University Press.

Cochrane, John H. 1991. A critique of application of unit root tests. *Journal of Economic Dynamics and Control* 15: 275–84. [CrossRef]

Darné, Olivier. 2009. The uncertain unit root in real GNP: A re-examination. *Journal of Macroeconomics* 31: 153–66. [CrossRef]

Das, C. 1994. Decision making by classical test procedures using an optimal level of significance. *European Journal of Operational Research* 73: 76–84. [CrossRef]

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

DeJong, David N., John C. Nankervis, N. E. Savin, and Charles H. Whiteman. 1992. Integration versus trend stationary in time series. *Econometrica* 60: 423–33. [CrossRef]

DeGroot, Morris. 1975. *Probability and Statistics*, 2nd ed. Reading: Addison-Wesley.

Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–31.

Diebold, Francis X., and Lutz Kilian. 2000. Unit-Root Tests Are Useful for Selecting Forecasting Models. *Journal of Business and Economic Statistics* 18: 265–73.

Diebold, Francis X., and Abdelhak S. Senhadji. 1996. The uncertain root in real GNP: Comment. *American Economic Review* 86: 1291–98.

Elliott, Graham, and Elena Pesavento. 2006. On the Failure of Purchasing Power Parity for Bilateral Exchange Rates after 1973. *Journal of Money, Credit, and Banking* 38: 1405–29. [CrossRef]

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock. 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64: 813–36. [CrossRef]

Engsted, Tom. 2009. Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak. *Journal of Economic Methodology* 16: 393–408. [CrossRef]

Fomby, Thomas B., and David K. Guilkey. 1978. On Choosing the Optimal Level of Significance for the Durbin–Watson test and the Bayesian alternative. *Journal of Econometrics* 8: 203–13. [CrossRef]

Hausman, Jerry A. 1978. Specification Tests in Econometrics. *Econometrica* 46: 1251–71. [CrossRef]

Keuzenkamp, Hugo A., and Jan R. Magnus. 1995. On tests and significance in econometrics. *Journal of Econometrics* 67: 103–28. [CrossRef]

Kilian, Lutz. 1998. Small sample confidence intervals for impulse response functions. *The Review of Economics and Statistics* 80: 218–30. [CrossRef]

Kim, Jae H. 2004. Bootstrap Prediction Intervals for Autoregression using Asymptotically Mean-Unbiased Parameter Estimators. *International Journal of Forecasting* 20: 85–97. [CrossRef]

Kim, Jae H. 2015. BootPR: Bootstrap Prediction Intervals and Bias-Corrected Forecasting. R package version 0.60. Available online: http://CRAN.R-project.org/package=BootPR (accessed on 9 February 2016).

Kim, Jae H., and Philip Inyeob Ji. 2015. Significance Testing in Empirical Finance: A Critical Review and Assessment. *Journal of Empirical Finance* 34: 1–14. [CrossRef]

Kish, Leslie. 1959. Some statistical problems in research design. *American Sociological Review* 24: 328–38. [CrossRef]

Koop, Gary, and Mark F. J. Steel. 1994. A Decision-Theoretic Analysis of the Unit-Root Hypothesis Using Mixtures of Elliptical Models. *Journal of Business and Economic Statistics* 12: 95–107.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Lehmann, Erich L., and Joseph P. Romano. 2005. *Testing Statistical Hypothesis*, 3rd ed. New York: Springer.

Lothian, James R., and Mark P. Taylor. 1996. Real exchange rate behavior: The recent float from the perspective of the past two centuries. *Journal of Political Economy* 104: 488–510. [CrossRef]

Luo, Sui, and Richard Startz. 2014. Is it one break or ongoing permanent shocks that explains U.S. real GDP? *Journal of Monetary Economics* 66: 155–63. [CrossRef]

MacKinnon, James G. 1996. Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics* 11: 601–18. [CrossRef]

MacKinnon, James G. 2002. Bootstrap inference in Econometrics. *Canadian Journal of Economics* 35: 615–44. [CrossRef]

Maddala, G. S., and In-Moo Kim. 1998. *Unit Roots, Cointegration and Structural Changes*. Cambridge: Cambridge University Press.

Manderscheid, Lester V. 1965. Significance Levels-0.05, 0.01, or? *Journal of Farm Economics* 47: 1381–85. [CrossRef]

McCallum, Bennett T. 1986. On "Real" and "Sticky-Price" Theories of the Business Cycle. *Journal of Money, Credit and Banking* 18: 397–414. [CrossRef]

Morrison, Denton E., and Ramon E. Henkel, eds. 1970. *The Significance Test Controversy: A Reader*. New Brunswick: Aldine Transactions.

Müller, Ulrich K., and Graham Elliott. 2003. Testing for unit roots and the initial condition. *Econometrica* 71: 1269–86. [CrossRef]

Murray, Christian J., and Charles R. Nelson. 2000. The uncertain trend in U.S. GDP. *Journal of Monetary Economics* 46: 79–95. [CrossRef]

Neely, Christopher, and David E. Rapach. 2008. Real Interest Rate Persistence: Evidence and Implications. *Federal Reserve Bank of St. Louis Review* 90: 609–42.

Nelson, Charles R., and Charles R. Plosser. 1982. Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10: 139–62. [CrossRef]

Orcutt, Guy H., and Herbert S. Winokur. 1969. First order autoregression: Inference, estimation and prediction. *Econometrica* 37: 1–14. [CrossRef]

Papell, David H. 1997. Searching for stationarity: Purchasing power parity under the current float. *Journal of International Economics* 43: 313–32. [CrossRef]

Papell, David H., and Ruxandra Prodan. 2004. The uncertain unit root in US real GDP: Evidence with restricted and unrestricted structural change. *Journal of Money, Credit and Banking* 36: 423–27. [CrossRef]

Perez, María-Eglée, and Luis Raúl Pericchi. 2014. Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics and Probability Letters* 85: 20–24. [CrossRef] [PubMed]

Pericchi, Luis Raúl, and Carlos Pereira. 2016. Adaptive significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics* 30: 70–90. [CrossRef]

Pfaff, Bernhard. 2008. *Analysis of Integrated and Cointegrated Time Series with R*, 2nd ed. New York: Springer.

Phillips, Peter, and Pierre Perron. 1988. Testing for a Unit Root in Time Series Regression. *Biometrika* 75: 335–46. [CrossRef]

Poirier, Dale J. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge: MIT Press.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available online: http://www.R-project.org/ (accessed on 9 February 2016).

Rapach, David E., and Christian E. Weber. 2004. Are real interest rates really nonstationary? New evidence from tests with good size and power. *Journal of Macroeconomics* 26: 409–30. [CrossRef]

Rose, Andrew K. 1988. Is the real interest rate stable? *Journal of Finance* 43: 1095–112. [CrossRef]

Rudebusch, Glenn D. 1993. The uncertain unit root in real GNP. *American Economic Review* 83: 264–72.

Schotman, Peter C., and Herman K. van Dijk. 1991. On Bayesian Roots to Unit Roots. *Journal of Applied Econometrics* 6: 387–401. [CrossRef]

Schwert, G. William. 1989. Testing for unit roots: A Monte Carlo investigation. *Journal of Business and Economic Statistics* 7: 147–59.

Sims, Christopher. 1988. Bayesian Scepticism on Unit Root Econometrics. *Journal of Economics Dynamics and Control* 12: 463–74. [CrossRef]

Sims, Christopher, and Harald Uhlig. 1991. Understanding Unit Rooters: A Helicopter Tour. *Econometrica* 59: 1591–99. [CrossRef]

Skipper, James K., Anthony L. Guenther, and Gilbert Nass. 1967. The sacredness of 0.05: A note on concerning the use of statistical levels of significance in social science. *The American Sociologist* 2: 16–18.

Startz, Richard. 2014. Choosing the More Likely Hypothesis. *Foundations and Trends in Econometrics* 7: 119–89. [CrossRef]

Stine, Robert A., and Paul Shaman. 1989. A fixed point characterization for bias of autoregressive estimators. *The Annals of Statistics* 17: 1275–84. [CrossRef]

Stock, James H. 1991. Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series. *Journal of Monetary Economics* 28: 435–59. [CrossRef]

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]

Winer, Benjamin J. 1962. *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

Ziliak, Stephen Thomas, and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: The University of Michigan Press.