Unit-Root Tests Are Useful for Selecting Forecasting Models

Francis X. Diebold
Department of Finance, Stern School of Business, New York University, New York, NY 10012, Department of Economics, University of Pennsylvania, Philadelphia, PA 19104, and National Bureau of Economic Research

Lutz Kilian
Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220, and Centre for Economic Policy Research (*lkilian@umich.edu*)

We study the usefulness of unit-root tests as diagnostic tools for selecting forecasting models. Difference-stationary and trend-stationary models of economic and financial time series often imply very different predictions, so deciding which model to use is tremendously important for applied forecasters. We consider three strategies: Always difference the data, never difference, or use a unit-root pretest. We characterize the predictive loss of these strategies for the canonical AR(1) process with trend, focusing on the effects of sample size, forecast horizon, and degree of persistence. We show that pretesting routinely improves forecast accuracy relative to forecasts from models in differences, and we give conditions under which pretesting is likely to improve forecast accuracy relative to forecasts from models in levels.

KEY WORDS: Model selection; Prediction; Pretest.

Difference-stationary and trend-stationary models of the same time series may imply very different predictions (e.g., Diebold and Senhadji 1996). Deciding which model to use is therefore tremendously important for applied forecasters. Rather than employing one or the other model by default, one may use a unit-root test as a diagnostic tool to guide the decision. In fact, one of the early motivations for unit-root tests was precisely to help determine whether to use forecasting models in differences or levels in particular applications (e.g., Dickey, Bell, and Miller 1986).

Much of the recent econometric unit-root literature has focused on the inability of unit-root tests to distinguish in finite samples the unit-root null from nearby stationary alternatives (e.g., Christiano and Eichenbaum 1990; Rudebusch 1993). But low power against nearby alternatives, which are typically the relevant alternatives in econometrics, is not necessarily a concern for forecasting. It has long been asserted, for example, that the accuracy of forecasts may be improved by employing a model in differences rather than a model in levels, if the root of the process is close to but less than unity (e.g., Box and Jenkins 1976, p. 192). Ultimately, the question of interest for forecasting is *not* whether unit-root pretests select the "true" model but whether they select models that produce superior forecasts. Surprisingly little is known about the efficacy of unit-root tests for this purpose.

The comparative merits of strategies such as "always difference," "never difference," or "sometimes difference, according to the results of a unit-root pretest" will in general depend on the degree of persistence of the true process, the forecast horizon of interest, the sample size, and the properties of the pretest. Hence, the purpose of this article is to explore systematically the extent to which pretesting for unit roots affects forecast accuracy for a variety of degrees of persistence, forecast horizons, and sample sizes.

We focus on the univariate trending autoregressive case with high persistence, which is of particular interest in economics and finance, and we proceed by Monte Carlo simulation as described in Section 1. The results are sharp and intuitive, and we summarize them with compact response surfaces in Section 2. In Section 3, we meld the results into practical prescriptions for applied work. In Section 4, we provide additional discussion: Looking backward, we interpret our results in the context of earlier literature on which they build, and looking forward, we sketch preliminary results of extensions that incorporate alternative estimators. Finally, in Section 5 we offer concluding remarks and directions for future research.

1. EXPERIMENTAL DESIGN

Here, as always, there is inescapable tension in experimental design. On the one hand, we want to examine a wide enough range of data-generating processes (DGP's) that the results will shed light on the behavior of alternative methods and on a range of empirically relevant situations. Clearly, we will want to examine a range of forecast horizons, degrees of persistence, and sample sizes. On the other hand, it is crucial that the DGP's examined be simple and their range small enough to promote manageable and interpretable Monte Carlo analysis.

Use of a first-order autoregressive [AR(1)] DGP, with differing degrees of persistence corresponding to different autoregressive parameter values, represents an appealing compromise. If, however, the analysis is to provide meaningful recommendations for applied work, we view the inclusion of a time trend as crucial. Trending behavior is routinely

Journal of Business & Economic Statistics
July 2000, Vol. 18, No. 3

present in economic and financial series and increases the amount of bias in the least squares estimator of autoregressive parameters, which has important implications for the performance of the alternative strategies of "always differencing," "never differencing," or "pretesting."

Hence, we examine a trending AR(1) process of the form

$$(y_t - a - bt) = \rho(y_{t-1} - a - b(t-1)) + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2),$$

$t = 1, 2, \ldots, T$. We can rewrite the process as $y_t = k_1 + k_2 t + \rho y_{t-1} + \varepsilon_t$, where $k_1 = a(1-\rho) + \rho b$ and $k_2 = b(1-\rho)$. Perhaps more intuitively, we can express the process in components form as the sum of a linear trend and an AR(1) process, $y_t = T_t + x_t$, where $T_t = a + bt$ and $x_t = \rho x_{t-1} + \varepsilon_t$. When $\rho = 1$, the process is a random walk with drift b, and when $\rho < 1$, the process is a linear trend with slope b buffeted by covariance stationary AR(1) shocks.

We parameterize the process to be consistent with U.S. postwar quarterly real gross national product (GNP) data by setting $a = 7.3707$, $b = .0065$, and $\sigma = .0099$. This parameterization is likely to be representative for many other trending macroeconomic time series as well. We examine $\rho \in \{.5, .9, .97, .99, 1\}$ and $T \in \{25, 30, 40, 50, 60, 70, 80, 100, 120, 140, 160, 180, 200, 240, 280, 320, 360, 400, 440, 480, 520, 560, 600, 640, 680, 720, 760, 800, 840, 880, 920, 960, 1{,}000\}$, which includes relevant degrees of persistence and sample sizes for annual, quarterly, monthly, weekly, and daily data.

We compare the performance of three forecasting models—AR(1) in levels with linear deterministic trend (L, for "levels"), random walk with drift (D, for "differences"), and the model suggested by Dickey–Fuller unit-root pretests using 5% finite-sample critical values (P, for "pretest"). For all models, the estimation method is ordinary least squares (OLS). The common objective is to forecast the level of the series at horizons h ranging from 1 to 100 periods ahead. Using common random numbers across models, we evaluate the performance of each model by its unconditional prediction mean squared error (PMSE) in 20,000 Monte Carlo trials. For each value of ρ, we calculate the ratios PMSE(D)/PMSE(L), PMSE(D)/PMSE(P), and PMSE(P)/PMSE(L) for all combinations of h and T.

2. RESULTS

In Figures 1–3 we shall show, for various values of ρ, response surfaces for PMSE(D)/PMSE(L), PMSE(D)/PMSE(P), and PMSE(P)/PMSE(L), for all combinations of forecast horizon and sample size. In particular, for each value of ρ, we show the relative PMSE as a function of h and T. We present unsmoothed response surfaces because they are quite smooth already and readily interpretable without additional smoothing.

2.1 D Versus L

Figure 1 makes clear that neither D nor L dominates always; the relative forecast accuracy in general depends on ρ, h, and T. Not surprisingly, for $\rho = 1$ the D model is uniformly more accurate than the L model because in that case D is the true model. The ratio PMSE(D)/PMSE(L) drops toward 0 as the forecast horizon h grows (for fixed sample size, T). This happens because the distortions resulting from the Dickey–Fuller small-sample bias, which plague the L model, are magnified as h grows. This effect, of course, is most pronounced for smaller sample sizes, for which the Dickey–Fuller bias is greatest. As a result,

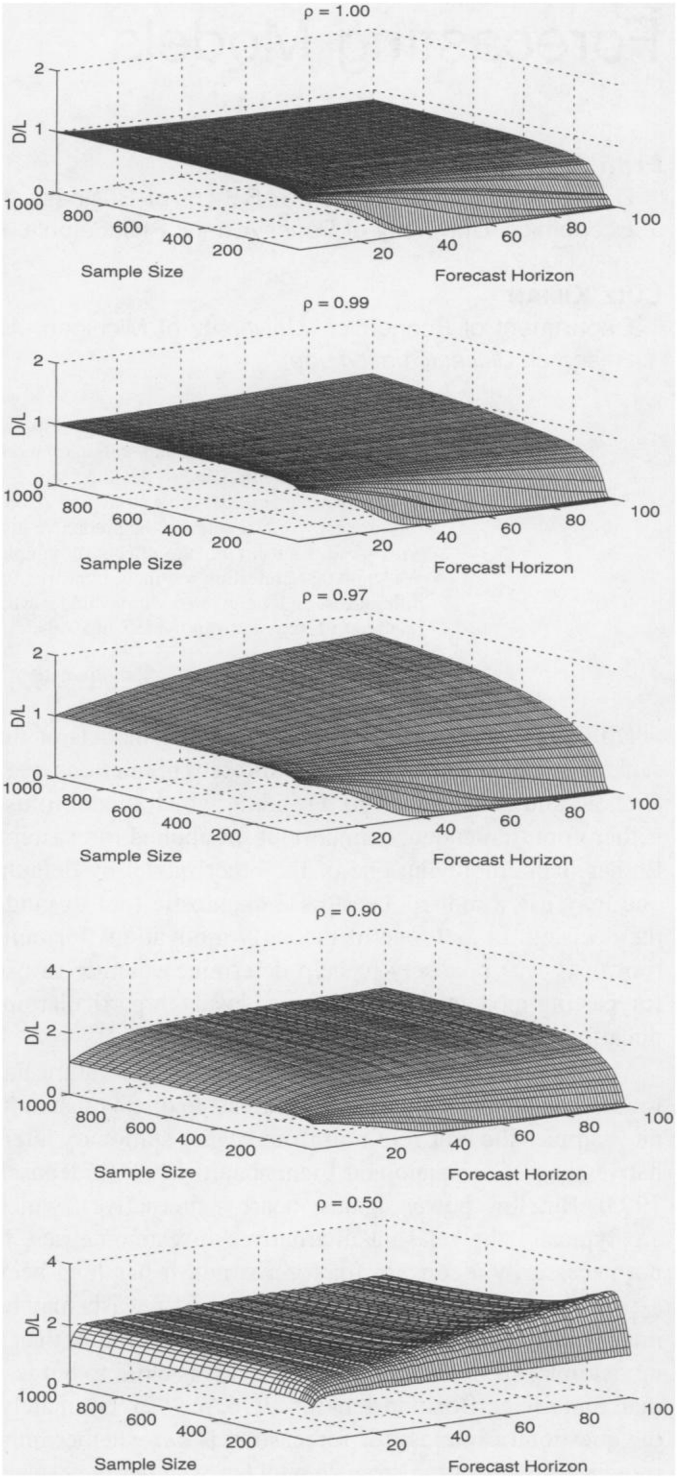

Figure 1. PMSE(D)/PMSE(L). Source: 20,000 Monte Carlo trials based on DGP $y_t = (a - a\rho + b\rho) + (b - b\rho)t + \rho y_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim N(0, .01^2)$, a = 7.3707, b = .0065. D denotes the model in first differences; L denotes the model in levels.

for fixed forecast horizon h, the ratio PMSE(D)/PMSE(L) drops toward 0 as T declines.

In contrast, for roots smaller than unity, D is false and would not be expected to dominate L always. That expectation is confirmed. For $\rho = .99$, for example, for sample sizes in excess of 600, forecasts from L are marginally more accurate than those from D. The D model forecast is least accurate for large h. This is to be expected because the error resulting from the false imposition of a unit root is compounded with rising h. In contrast, small-sample bias is of little concern for such large samples, and L is quite accurate. Nevertheless, for smaller sample sizes, forecasts from D continue to be more accurate than forecasts on the basis of the biased estimator associated with L, especially for long forecast horizons.

The trade-offs between the use of D and L become more pronounced as the persistence of the process declines. For $\rho = .97$, the ratio PMSE(D)/PMSE(L) exceeds 1 over much of the parameter space and is highest when both T and h are large. For small T and large h, however, the ratio still tends to approach 0. The poor relative performance of the L model for small T and large h is not only due to small-sample bias. In addition, the PMSE of L is inflated by occasional explosive estimates, resulting in absurd forecasts, especially at long forecast horizons. In contrast, the constraint implicit in D renders its forecasts more consistently reasonable, even when D is incorrect.

The problem with the L forecast is that for processes with large roots there is a nonnegligible probability in small samples of drawing an explosive estimate. As a result, using L, we occasionally encounter predictions based on "outlier" explosive models, which have extremely large prediction errors and dominate the PMSE. Typically in such cases the forecast dives toward minus infinity due to a slightly negative estimated trend coefficient and an estimated root in excess of unity. As a result, the PMSE does not improve at long horizons as the process reverts back to its mean, as one might have expected, because the effect of explosive forecasts on the PMSE obviously worsens for longer horizons. The problem does not arise in D because of the imposition of a unit root. Although the PMSE of D worsens for longer forecast horizons, as one would expect, the extent to which its PMSE deteriorates is dwarfed by the PMSE of L, which is inflated by the occasional explosive outliers. The net result is a ratio of PMSE(D)/PMSE(L) that approaches 0. In light of these phenomena, we also experimented with a mixed strategy (M), in which we used the L forecast unless the L forecast was explosive, in which case we replaced the L forecast with the D forecast. As expected, the small-sample forecast accuracy of M was much better than that of L, but, interestingly, the modification did not affect our qualitative results.

We also find that for small T, the ratio PMSE(D)/PMSE(L) *decreases* in h, whereas for large T it *increases* in h. This reversal makes sense. For small T, the loss in forecast accuracy from poor estimates of L is much greater than the loss from inappropriately using the model in differences, and the trade-off worsens as h increases. In contrast, for large T, the forecast from L is increasingly more accurate (because the least squares estimator is consistent), whereas using the model in differences (and thereby imposing a unit root) introduces a systematic distortion in forecasting, the effects of which are amplified with h.

The results for $\rho = .9$ are similar but even more pronounced. Differencing continues to improve forecast accuracy for small and moderate sample sizes, but as the persistence of the process declines, the gains are limited to increasingly smaller sample sizes. At the same time, for larger sample sizes, L becomes increasingly more accurate than the model in differences, especially as the forecast horizon increases.

It is interesting to note that in the case of $\rho = .9$, as well as several cases discussed later, for large T the ratio PMSE(D)/PMSE(L) (and later PMSE(D)/PMSE(P)) approaches 2 as h grows. This phenomenon occurs when T is large enough so that the parameters of the L model are estimated precisely (or, equivalently, for T large enough so that the unit-root null hypothesis tends to be rejected correctly, and the resulting trend-stationary model is estimated precisely). The explanation is simple: In population, when $\rho < 1$, the long-horizon forecast error from L is approximately the unconditional variance of the process, $\text{var}(y_t)$, whereas the long-horizon forecast error from the model in differences is approximately $\text{var}(y_{t+h} - y_t)$, which approximately equals twice the unconditional variance of the process. Appearance of these population results in the finite-sample Monte Carlo results requires a sample large enough to facilitate precise estimation and powerful unit-root testing.

Finally, for $\rho = .5$, the L model uniformly dominates the D model. Although this case is less interesting for applications and we shall therefore not dwell on it, it is interesting to note the emergence of a ridge in the response surface for small T, the height of which steadily increases in h.

Taken as a whole, the D versus L results appear driven by the fact that differencing provides insurance against problems due to small-sample bias and explosive root problems, at a cost. Those problems are most severe for small T and large h, so the insurance is more than worth its cost. Elsewhere in the parameter space, however, the situation is reversed. As a rule of thumb, the results suggest that one is better off differencing if the sample size is small or moderate and the process appears highly persistent, and conversely. Note in particular that the "we don't know and we don't care" view is explicitly refuted: Although the trend-stationary versus difference-stationary distinction is not important in some contexts, it most definitely makes a difference for forecasting. Moreover, the best forecasting model is not necessarily the true model; the ability of a unit-root pretest to select a good forecasting model is distinct from its ability to select the true model. The fact that neither D nor L dominates uniformly suggests that unit-root pretests may help to improve forecast accuracy. We now explore this possibility in detail.

2.2 D Versus P

Figure 2 makes clear that the pretesting strategy dominates that of routinely differencing the data for almost all sample sizes and forecast horizons. The reason is that our pretest takes the unit-root hypothesis as its null. For alternatives close to the unit-root null, the power of the pretest is low, so the pretest model reduces to the model in differences. Hence, P performs much like D did in Figure 1 when that model is a good approximation. On the other hand, for processes with roots far from the unit-root null, the Dickey–Fuller test is bound to find strong evidence against the null, in which case the pretest model reduces to L, which we know to be much more accurate than D when persistence is low.

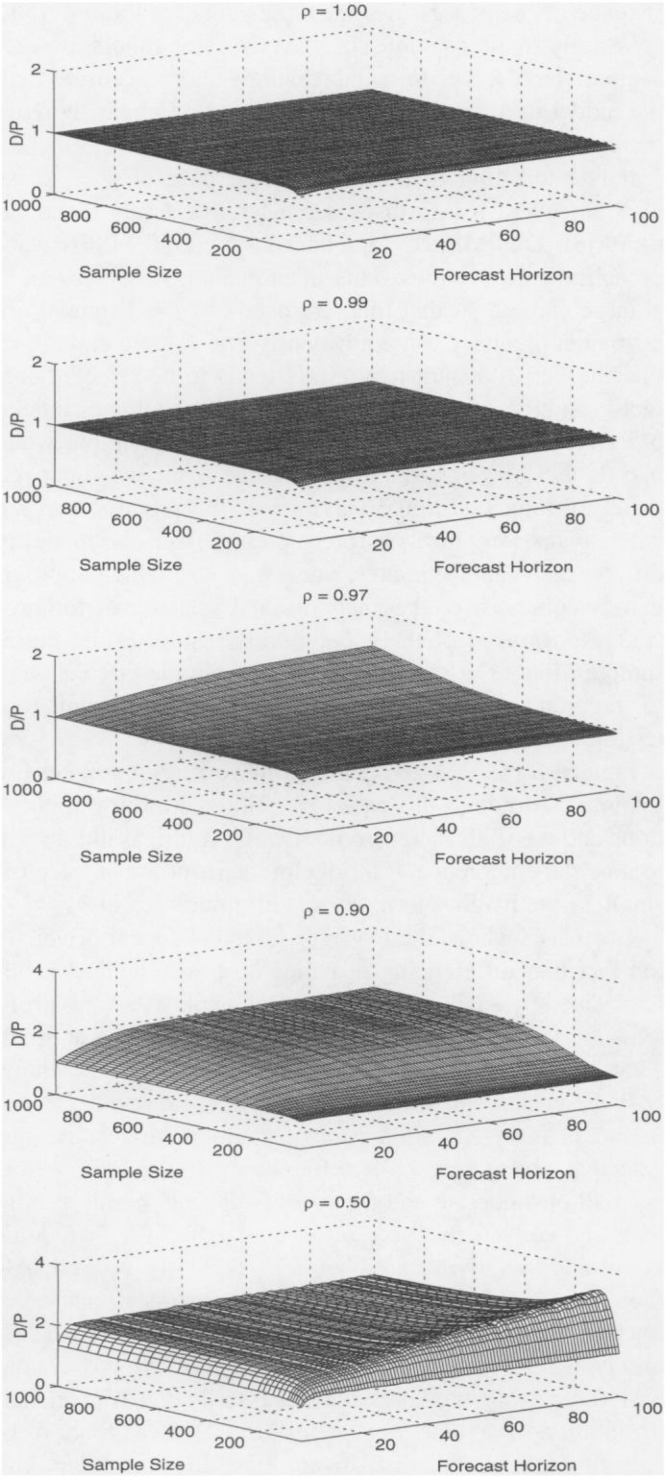

Figure 2. PMSE(D)/PMSE(P). Source: 20,000 Monte Carlo trials based on DGP $y_t = (a - a\rho + b\rho) + (b - b\rho)t + \rho y_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim N(0, .01^2)$, a = 7.3707, b = .0065. D denotes the model in first differences; P denotes the model after pretesting.

In particular, we find that for $\rho = 1$, when D is the true model, the pretest is unlikely to reject the model in differences, resulting in a PMSE(D)/PMSE(P) ratio very close to 1. Similar results hold for $\rho = .99$. For $\rho = .97$, P begins to exhibit important advantages over D. For small T, the test lacks power and rarely rejects, so P and D coincide, and PMSE(D)/PMSE(P) is effectively 1 regardless of h. As T grows, the test rejects the unit-root null more often, yet the ratio PMSE(D)/PMSE(P) remains close to 1. The reason is that, at least for small h, the PMSE for a highly persistent process in levels tends to be close to that of the equivalent model in differences. Because for large T the L model will be estimated rather precisely, the resulting forecast is about as accurate as that for the D model. In contrast, for both T and h large, the P forecast is considerably more accurate than the D forecast. This outcome is reflected in PMSE(D)/PMSE(P) ratios in excess of 1. The reason is that for long horizons the false imposition of a unit root (which is of little consequence for short horizons) becomes a liability.

This tendency becomes even more apparent for $\rho = .9$. Only for very small T, the relative accuracy of D and P remains similar. In general, the P model is much more accurate than the D model. Finally, consider the process with $\rho = .5$. In Figure 1 we showed that the loss in forecast accuracy from falsely adopting the model in differences is very high for $\rho = .5$. The Dickey–Fuller pretest has considerable power against this distant alternative, however, and almost always rejects the model in differences. Hence, P and L tend to coincide, and the PMSE(D)/PMSE(P) results in Figure 2 are almost identical to those in the corresponding panel of Figure 1 for PMSE(D)/PMSE(L).

2.3 P Versus L

In Figure 3, we directly compare P and L. For $\rho = 1$, pretesting gives similar results to differencing. Not surprisingly, pretesting uniformly dominates the levels model. Similar results hold, at least for small and moderate sample sizes, for $\rho = .99$. Figure 3 indicates that pretest-based forecasts are about as accurate as the level forecasts when $\rho = .5$. The most interesting results are for the intermediate region of $\rho = .97$ and $\rho = .9$. For small and moderate T and large h, Figure 3 shows evidence of a ridge on the response surface for $\rho = .9$. That ridge flattens and widens for $\rho = .97$, as the accuracy of the pretest model improves. Evidently those are cases for which we would like to have rejected the unit-root null hypothesis but did not. Although the root is far enough from the unit circle, and the sample size (albeit small) is large enough for the levels models to

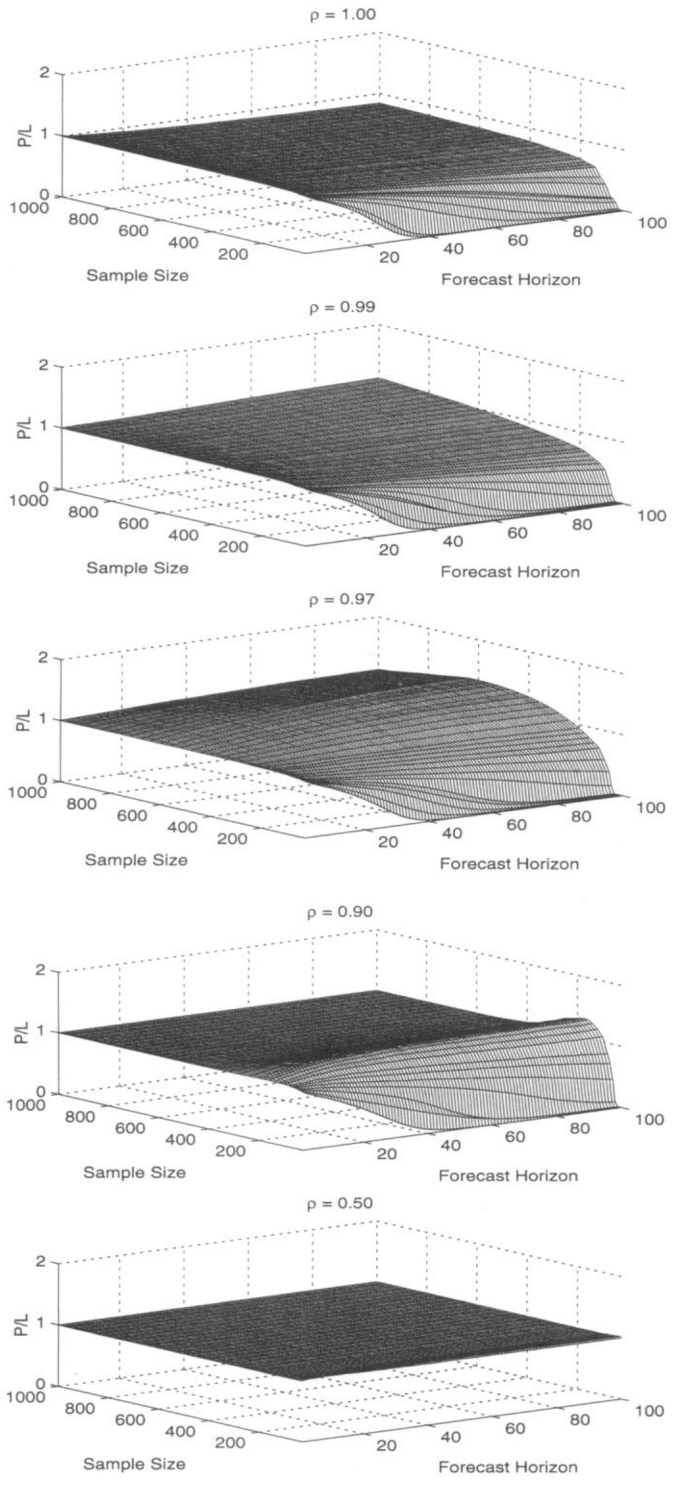

Figure 3. PMSE(P)/PMSE(L). Source: 20,000 Monte Carlo trials based on DGP $y_t = (a - a\rho + b\rho) + (b - b\rho)t + \rho y_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim N(0, .01^2)$, $a = 7.3707$, $b = .0065$. P denotes the model after pretesting; L denotes the model in levels.

be reasonably accurate, the Dickey–Fuller test is not powerful enough to detect the absence of a unit root. This observation suggests that more powerful unit-root tests such as the DF-GLS test of Elliott, Rothenberg, and Stock (1996) could be used to flatten the ridge and to improve forecast accuracy. It is not obvious, however, that more powerful tests would be beneficial in all regions of the parameter space. As we showed earlier, in some cases incorrectly using the model in differences rather than the correct model in levels will actually improve forecast accuracy, so more powerful unit-root pretests may actually worsen forecast accuracy in those regions. More research is needed to quantify these trade-offs.

2.4 A Summary Assessment

Taken as a whole, the results cast the pretesting strategy in a favorable light. P dominates D uniformly, which makes clear that the Box–Jenkins strategy of routinely differencing to achieve stationarity is not to be recommended for constructing forecasting models. P does not dominate L uniformly, but it nevertheless dominates over much of the design space, which similarly casts doubt on a strategy relying on asymptotics by routinely specifying forecasting models in levels.

3. SOME PRACTICAL ADVICE

Given the wide range of sample sizes and forecast horizons, it is difficult to translate the results in Figures 1–3 into concrete practical advice. Moreover, the DGP based on quarterly real GNP may not be representative for other frequencies. We therefore repeated the simulation exercise for selected sample sizes and forecast horizons for DGP's specifically chosen to be representative for each frequency. Because the pretesting strategy clearly dominates differencing, we focus on the choice between pretesting and routinely forecasting on the basis of the level model.

Table 1 summarizes the simulation design for each frequency. The quarterly DGP based on U.S. real GNP is identical to the DGP defined in Section 1. The annual DGP is based on 125 observations for U.S. per capita real GNP as defined by Diebold and Senhadji (1996). The daily DGP is based on the Dow Jones stock-price index for 1/1/74–4/2/98, and the monthly DGP is based on the U.S. industrial production index (DRI code: IP) for the postwar period.

For annual data (say, T = 40–160 and h = 1–100), we find that pretesting unambiguously improves forecast accuracy for all forecast horizons and sample sizes if the root of the DGP is .97 or higher. For $\rho = .9$, pretesting still improves forecast accuracy for sample sizes as high as 70 but does not perform as well as the L model in larger samples. For $\rho = .5$, the two models are tied. In practice, this result suggests using pretests for datasets of up to 70 annual observations and for all larger sample sizes, provided the process is likely

Table 1. Data-Generating Processes

	Parameter		
	a	*b*	σ
Annual	−6.0674	.0173	.0500
Quarterly	7.3707	.0065	.0099
Monthly	3.3654	.0024	.0105
Daily	5.1126	.0004	.0095

NOTE: The DGP is $(y_t - a - bt) = \rho(y_{t-1} - a - b(t - 1)) + \varepsilon_t$, where ε_t is normally distributed with standard deviation σ. We choose annual and quarterly parameters to be representative of those for U.S. real GNP, monthly parameters for U.S. industrial production, and daily parameters for the U.S. Dow-Jones stock-price index. See the text for details.

to be highly persistent. In the remaining cases, the L model is preferred.

For quarterly data (say, $T = 80$–200 and $h = 1$–16), the P model is more accurate for all forecast horizons and sample sizes, provided the root of the process is .97 or higher. For $\rho = .9$, the level model is uniformly more accurate, and for $\rho = .5$ the models are tied. We conclude that pretesting should be used for all processes with roots of .97 or higher and the L model for processes with smaller roots.

For monthly data (say, $T = 240$–480 and $h = 1$–48), pretesting improves forecast accuracy for $\rho = 1$ and for $\rho = .99$ for all forecast horizons and sample sizes considered. For $\rho = .97$ and $\rho = .9$, however, the L model is at least as accurate as the P model, and for $\rho = .5$ the two methods are tied. This finding suggests that pretesting is useful only for processes with roots of .99 or higher, and in all other cases the L model will be more accurate.

For daily data (say, $T = 360$–720 and $h = 1$–90), pretesting only improves forecast accuracy uniformly for $\rho = 1$. For $\rho = .99$, the performance is mixed, with the P model being more accurate for sample sizes of fewer than 600 days at all horizons. For larger sample sizes, the L model is slightly more accurate at long forecast horizons and roughly as accurate as the P model at shorter horizons. For $\rho = .97$, the L model is uniformly more accurate. For $\rho = .9$, the L model is slightly more accurate for small T, especially for $T < 500$, except at very short horizons. For larger sample sizes the differences vanish. For $\rho = .5$ the two methods are tied. This finding suggests that pretesting is useful for forecasting daily data only if the data are very persistent with roots of .99 or higher. For other applications, the L model is likely to be more appropriate.

Our advice may appear to be circular in that it often depends on knowledge of the true root. In practice, however, OLS point estimates of the roots for quarterly macroeconomic data are typically in excess of .97, estimates for monthly data are in excess of .99, and estimates for daily data are well in excess of .99. These differences in dominant roots across sampling frequencies make sense when viewed in terms of the implied half-life of the response to an innovation (see Caner and Kilian 1999). Moreover, the presence of small-sample bias suggests that these OLS estimates, if anything, understate the true roots. We therefore conclude that pretesting is recommended for virtually all forecasting exercises involving trending macroeconomic data.

4. DISCUSSION

Here we provide additional discussion of our Monte Carlo results. Looking backward, we interpret them in the context of earlier theoretical and empirical literatures on which they build, and looking forward, we sketch preliminary results of extensions to incorporate alternative estimators.

4.1 Relationship to the Literature

Our work builds on, and complements, a small literature dating back almost a decade. Stock (1990) found in a particular application that model specification in levels versus differences matters little but pointed out that in general it will. Some preliminary evidence in favor of pretesting was presented by Campbell and Perron (1991), who studied 1-step-ahead and 20-step-ahead forecasts made using autoregressive models. They showed that one loses little by pretesting relative to using the true model, and sometimes one actually gains. In their study, autoregressive models in levels did best for series that are near white noise, whereas autoregressive models in differences did best for series that are near a random walk. Cochrane (1991), in a comment on Campbell and Perron, explored longer forecast horizons. He compared level and difference-stationary models but did not discuss pretesting. Neither did Franses and Kleibergen (1996), who studied the out-of-sample forecasting accuracy of trend-stationary and difference-stationary models for the Nelson–Plosser dataset.

Apart from the contemporaneous and independent contribution of Clements and Hendry (1999), whose results nicely complement ours, the extant work most closely related to ours is that of Stock (1996) and Stock and Watson (1998). Stock (1996) showed that pretesting may be useful from a local-to-unity asymptotic perspective and presented some Monte Carlo evidence for the AR(1) model without trend. In recent contemporaneous and independent work, Stock and Watson (1998) provided a comprehensive empirical study of the out-of-sample accuracy of macroeconomic forecasting models; one of their conclusions was that autoregressive models based on unit-root pretests tend to perform well.

Our Monte Carlo results complement and strengthen both the largely theoretical work of Stock (1996) and the purely empirical work of Stock and Watson (1998). Our analysis is closer in spirit to Stock's, but there are important differences. Stock focused narrowly on documenting problems in long-horizon forecasting from models with roots close to unity. Moreover, he did not consider models with trend, and he fixed the ratio h/T in his Monte Carlo analysis. Our analysis, in contrast, is wider in scope. It includes a grid of alternative values of $\rho, h,$ and T, corresponding to applications of autoregressive forecast models using daily, weekly, monthly, quarterly, and annual data. To the extent that results can be compared directly, ours and Stock's tend to agree; however, we find stronger evidence in favor of pretesting than did Stock (1996), reflecting the greater importance of small-sample bias in models with trends.

4.2 Extensions to Other Estimators

The OLS estimator is the most commonly used estimator in practice, but there are other estimators of interest in forecasting. For example, Canjels and Watson (1997) documented that, for processes with roots close to unity, the feasible generalized least squares estimator of Prais and Winsten (1954) provides the best estimates of the trend coefficient. Given the obvious importance of accurate trend estimates, especially at long-forecast horizons, a comparison of the forecast accuracy of the Prais–Winsten estimator to the OLS estimator used in this article seems useful. In addition, it will be worthwhile to study bias-corrected OLS forecasts, insofar as our simulation results are consistent with the view that much of the advantage of falsely imposing a unit root in borderline stationary processes is due to the elimination of OLS small-sample bias. A natural conjecture is that the mean squared error of forecasts from trend-stationary models may be improved by replacing the OLS autoregressive coefficient estimates by bias-corrected coefficient estimates. Such corrections have been used successfully in the closely related area of impulse response analysis. For example, Andrews and Chen (1994) reported that approximate median bias corrections for univariate autoregressive models may reduce the mean squared error of impulse response estimates for at least some parameter ranges and horizons. Alternative bias corrections based on the mean bias of the autoregressive coefficient estimates were explored by Rudebusch (1993) and Kilian (1998) based on work by Shaman and Stine (1988) and Pope (1990).

Here we briefly report some preliminary results for the accuracy of forecasts based on the iterated Prais–Winsten (IPW) estimator (as described by Park and Mitchell 1980) and based on the exactly median-unbiased estimator (MU) of Andrews (1993) for the AR(1) model with trend. The exactly median unbiased estimator is feasible only in the AR(1) model, but it provides a useful benchmark. Note that bias corrections of slope parameters need not lower the prediction mean squared error because the forecast variance will tend to increase. Hence, the usefulness of bias corrections in forecasting is an empirical question. We find that the MU forecasts are clearly dominated overall by the L forecasts. The reason is that the MU procedure occasionally produces very poor estimates of the trend behavior of the process, especially near the unit circle. We therefore will not pursue the MU forecast any further and focus on the IPW forecast instead.

Although the IPW forecasts from the levels model tend to dominate the L forecasts of the same model, they do not dominate the D forecast. This finding suggests the possibility that pretesting may further improve the accuracy of the IPW forecasts. It is not straightforward to compare the pretest strategy to that of estimating the levels model by the IPW estimator, however. The reason is that the standard Dickey–Fuller pretest is based on the OLS estimator. If we are interested in using other estimators, the usual Dickey–Fuller critical values do not apply. Thus, we would have to generate a new set of critical values for use with the IPW estimator to isolate the effects of pretesting, an exercise that is beyond the scope of this study.

Instead, we compare the strategy of always using the IPW forecast to the strategy of using the IPW forecast only if the standard OLS-based Dickey–Fuller test rejects the null of a unit root (and using the D forecast if the test does not reject). We find that, for the parameter regions of practical interest in macroeconomics, this pretest strategy tends to dominate the IPW forecast for the quarterly, monthly, and daily DGP's discussed in Section 3. For the annual DGP, the results are mixed. For $\rho = .97$ or higher, pretesting tends to result in more accurate forecasts than the IPW forecast. For $\rho = .9$, however, the results depend on the sample size. For sample sizes of, at most, 70, the P forecast tends to be more accurate, but for larger sample sizes the IPW forecasts dominate.

Overall, these results are qualitatively very similar to the results in Section 3 for P versus L. Our simulations clearly show that the usefulness of pretesting is not limited to the OLS estimator. We defer to future work a more careful investigation of the pretest strategy for the IPW estimator based on appropriately modified pretests.

5. CONCLUDING REMARKS AND DIRECTIONS FOR FUTURE RESEARCH

Difference-stationary and trend-stationary models of the same series may imply very different predictions. Deciding which model to use is thus tremendously important for applied forecasters, and unit-root pretests may provide a formal criterion for deciding whether to difference the data. Very little is known, however, about the usefulness of unit-root tests as diagnostic tools for selecting a forecasting model. In an effort to remedy this situation, we conducted a Monte Carlo study in which we explored systematically the extent to which pretesting for unit roots improves forecast accuracy in a canonical AR(1) model with trend, for a variety of sample sizes, forecast horizons, and degrees of persistence. We found strong evidence that pretesting improves forecast accuracy relative to routinely differencing the data. We also characterized in detail the conditions under which pretesting is likely to improve forecast accuracy relative to forecasts from models in levels and provided some practical advice.

There are many useful directions for future research. Given the narrow confines of our AR(1) DGP, the results by necessity are tentative, and there are many obvious but nevertheless important variations on the applications considered in this article. For example, we chose to focus on just one of many pretests for unit roots, and we ignored asymptotic refinements of unit-root tests based on bootstrap theory. Moreover, Stock (1996) showed that the asymptotically more powerful DF-GLS test of Elliott et al. (1996) may further improve forecast accuracy. Our analysis confirmed that there are important potential advantages to the use of more powerful unit-root tests in some regions of the parameter space, but it also showed that low power in some cases may improve forecast accuracy. This finding suggests that there are likely to be trade-offs between different unit-root pretests in terms of their power properties. Working with daily data, for example, may call for different unit-root tests than working with annual data. Future research will have to quantify these trade-offs. In addition, it would be of interest to explore tests that take L rather than D as the null hypothesis (see Kwiatkowski, Phillips, Schmidt, and Shin 1992; Leybourne and McCabe 1994). Pretest procedures based on such unit-root tests might be expected to dominate L for the same reason that the Dickey–Fuller pretest dominates D.

Another limitation of our Monte Carlo analysis is the greatly simplified lag structure of the DGP. Further research is required to verify the robustness of our findings in models with richer dynamics. We also deliberately ignored the issue of lag-order uncertainty at this stage of the analysis. Future work will have to address the fact that the population model is unknown in practice and may not even be of finite lag order. Appropriate data-based lag-order selection procedures for the class of (autoregressive moving average) ARMA(p, q) models were discussed, for example, by Ng and Perron (1995).

A third limitation of our analysis is our focus on univariate models. Univariate models are of central interest in many applications and often have proved superior to multivariate forecasting models, but they are not the only model in use. For example, Stock (1990, 1994) noted that results for univariate models do not bear directly on macroeconomic forecasting, which is typically multivariate. Future research undoubtedly will have to include vector-valued processes. Although the standard augmented Dickey–Fuller (ADF) test used in this article is widely used as a pretest for vector autoregressions, a similar analysis for multivariate cointegration tests would be useful. One would conjecture that imposing cointegration in small samples ought to improve forecast accuracy, whether or not cointegration holds exactly. There is reason to believe, however, that imposing cointegration may be less important than commonly thought. For example, Christoffersen and Diebold (1998) showed that, when forecasting cointegrated systems at long horizons, imposing the correct order of integration is crucial, but imposing cointegration is not.

A fourth extension would be to allow for endogenously selected deterministic trend breaks under the alternative. In particular, piecewise linear deterministic trend models may forecast more accurately than linear models. The ADF test considered in this article does not allow for trend breaks, but tests like those developed by Zivot and Andrews (1992) do. Moreover, alternative procedures for the simultaneous determination of the trend model and of the order of integration have been proposed, for example, by Phillips and Ploberger (1994).

A fifth extension would be to examine the robustness of the results to structural change, in light of recent work by Clements and Hendry (1998) indicating that certain specifications may be relatively more robust to structural change than others.

A final extension, and perhaps the most novel and interesting in our view, would be to consider unit-root test sizes other than 5% and to determine how the performance of P relative to D and L varies with test size. In particular, it should be possible to tune the test size to optimize the performance of P. The present fairly stringent size of 5% leads to domination of D by P because the pretest selects D except when there is strong evidence against D but fails to produce domination of L by P. It is possible that increasing the test size would leave largely intact the domination of D by P but could bring us closer to a similar domination of L by P. At any rate, there is certainly no reason to think that the arbitrary size of 5% is necessarily close to optimal. Hence, our results on the generally good performance of the pretesting strategy are conservative—some simple additional tuning could cast the pretest strategy in an even more favorable light.

ACKNOWLEDGMENTS

For helpful comments, we thank participants at the June 1998 North American meeting of the Econometric Society, Montreal, and the December 1998 EC^2 meeting, Stockholm. We especially thank (without implicating) David Hendry, Phil Howrey, Chris Murray, Paul Newbold, Allan Timmermann, and Tim Vogelsang. The National Science Foundation provided research support.

[Received January 1999. Revised September 1999.]

REFERENCES

Andrews, D. W. K. (1993), "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models," *Econometrica*, 61, 139–165.

Andrews, D. W. K., and Chen, H.-Y. (1994), "Approximately Median-Unbiased Estimation of Autoregressive Models," *Journal of Business & Economic Statistics*, 12, 187–204.

Box, G. E. P., and Jenkins, G. W. (1976), *Time Series Analysis, Forecasting and Control* (2nd ed.), Oakland, CA: Holden-Day.

Campbell, J. Y., and Perron, P. (1991), "Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots," in *NBER Macroeconomics Annual, 1991*, eds. O. Blanchard and S. Fischer, Cambridge, MA: MIT Press, pp. 141–200.

Caner, M., and Kilian, L. (1999), "Size Distortions of Tests of the Null Hypothesis of Stationarity: Evidence and Implications for the PPP Debate," unpublished manuscript, University of Michigan, Dept. of Economics.

Canjels, E., and Watson, M. (1997), "Estimating Deterministic Trends in the Presence of Serially Correlated Errors," *Review of Economics and Statistics*, 79, 184–200.

Christiano, L. J., and Eichenbaum, M. (1990), "Unit Roots in Real GNP: Do We Know and Do We Care?" *Carnegie-Rochester Conference Series on Public Policy*, 32, 7–82.

Christoffersen, P. F., and Diebold, F. X. (1998), "Cointegration and Long-Horizon Forecasting," *Journal of Business & Economic Statistics*, 16, 450–458.

Clements, M. P., and Hendry, D. F. (1998), "How to Win Forecasting Competitions in Economics," unpublished manuscript, University of Warwick and Nuffield College, Oxford, Dept. of Economics.

——— (1999), "Forecasting With Difference-Stationary and Trend-Stationary Models," unpublished manuscript, University of Warwick and Nuffield College, Oxford, Dept. of Economics.

Cochrane, J. H. (1991), "A Comment on Campbell and Perron," in *NBER Macroeconomics Annual, 1991*, eds. O. Blanchard and S. Fischer, Cambridge, MA: MIT Press, pp. 201–210.

Dickey, A. D., Bell, W. R., and Miller, R. B. (1986), "Unit Roots in Time Series Models: Tests and Implications," *The American Statistician*, 40, 12–26.

Diebold, F. X., and Senhadji, A. S. (1996), "The Uncertain Unit Root in Real GNP: Comment," *American Economic Review*, 86, 1291–1298.

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996), "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836.

Franses, P. H., and Kleibergen, F. (1996), "Unit Roots in the Nelson–Plosser Data: Do They Matter for Forecasting?" *International Journal of Forecasting*, 12, 283–288.

Kilian, L. (1998), "Small-Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics*, 80, 218–230.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992), "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root," *Journal of Econometrics*, 54, 159–178.

Leybourne, S. J., and McCabe, B. P. M. (1994), "A Consistent Test for a Unit Root," *Journal of Business & Economic Statistics*, 12, 157–166.

Ng, S., and Perron, P. (1995), "Unit Root Tests in ARMA Models With Data Dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90, 268–281.

Park, R. E., and Mitchell, B. M. (1980), "Estimating the Autocorrelated Error Model With Trended Data," *Journal of Econometrics*, 13, 185–201.

Phillips, P. C. B., and Ploberger, W. (1994), "Posterior Odds Testing for a Unit Root With Data-Based Model Selection," *Econometric Theory*, 10, 774–808.

Pope, A. L. (1990), "Biases of Estimators in Multivariate Non-Gaussian Autoregressions," *Journal of Time Series Analysis*, 11, 249–258.

Prais, S. J., and Winsten, C. B. (1954), "Trend Estimators and Serial Correlation," Discussion Paper 383, Cowles Foundation,

Rudebusch, G. D. (1993), "The Uncertain Unit Root in Real GNP," *American Economic Review*, 83, 264–272.

Shaman, P., and Stine, R. A. (1988), "The Bias of Autoregressive Coefficient Estimators," *Journal of the American Statistical Association*, 83, 842–848.

Stock, J. H. (1990), "Unit Roots in Economic Time Series: Do We Know and Do We Care? A Comment," *Carnegie-Rochester Conference Series on Public Policy*, 32, 63–82.

——— (1994), "Unit Roots, Structural Breaks, and Trends," in *Handbook of Econometrics* (Vol. 4), eds. R. Engle and D. McFadden, Amsterdam: North-Holland, pp. 2739–2841.

——— (1996), "VAR, Error Correction, and Pretest Forecasts at Long Horizons," *Oxford Bulletin of Economics and Statistics*, 58, 685–701.

Stock, J. H., and Watson, M. W. (1998), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," unpublished manuscript, Harvard University, Kennedy School, and Princeton University, Woodrow Wilson School.

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 & Economic Statistics*, 10, 251–270.