Federal Reserve Bank of Minneapolis
Research Department Staff Report 244

January 1998

Is There a Trend Break in U.S. GNP? A Macroeconomic Perspective

Lutz Kilian*

University of Michigan

Lee E. Ohanian*

Federal Reserve Bank of Minneapolis
and University of Minnesota

ABSTRACT

Unit root tests against trend break alternatives are based on the premise that the dating of the trend breaks coincides with major economic events with permanent effects on economic activity, such as wars and depressions. Standard economic theory, however, suggests that these events have large transitory, rather than permanent, effects on economic activity. Conventional unit root tests against trend break alternatives based on linear ARIMA models do not capture these transitory effects and can result in severely distorted inference. We quantify the size distortions for a simple model in which the effects of wars and depressions can reasonably be interpreted as transitory. Monte Carlo simulations show that in moderate samples, the widely used Zivot-Andrews (1992) test mistakes transitory dynamics for trend breaks with high probability. We conclude that these tests should be used only if there are no plausible economic explanations for apparent trend breaks in the data.

*We would like to thank seminar participants and, in particular, V. Corradi, F.X. Diebold, A. Inoue, C. Plosser, R.A. Stine, and a referee for helpful comments and suggestions. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

Since the work of Perron (1989) and Rappoport and Reichlin (1989), tests have been developed for the null of a unit root against the alternative of a trend stationary process with permanent changes in the trend, or *trend breaks.* Allowing for a break in the trend function often alters the outcome of tests for unit roots. Unit root tests against trend break alternatives now play an important role in time series analysis of macroeconomic data. Exhibit 1 shows that a variety of time series have been studied by many researchers using these procedures. These tests also have been used to pretest for unit roots in VAR models, as in Evans' (1989) study of unemployment and output dynamics.[1]

There has been a general tendency in the applied time series literature to associate any major economic event with potential trend breaks. In fact, unit root tests against trend break alternatives are based on the premise that the dating of the trend breaks coincides with major economic events with permanent effects on the level of economic activity. This basic view relating major events with underlying trend breaks dates back to Perron (1992):

"The estimated dates of break ... yield interesting conclusions about the identification of events that had a permanent effect on the levels of economic activity" (p. 144). In particular, the dating of the breaks is "associated with *major events* [emphasis added] that had a permanent effect on the behavior of the series creating a major change in intercept . . ., a change in the slope function . . .; or generally a combination of these events occurring around the same date" (p. 147).

In this paper, we argue that major economic events will generate important transitory movements in endogenous variables. In sharp contrast to the conventional view in the literature, this type of transitory movement is not adequately captured by the standard linear

ARIMA model. We show that researchers who rely on the linear ARIMA model will not be able to correctly identify these transitory dynamics. The inability to correctly identify these dynamics has serious implications for statistical inference. In particular, we find that these transitory dynamics can lead to rejection rates of the unit root null that far exceed the nominal size of the test.

For many aggregate time series, such as U.S. per capita gross national product (GNP), the estimated break dates coincide with wars or the Great Depression. Standard macroeconomic theory suggests that these events will have large transitory effects on aggregate variables such as GNP. For example, it is generally accepted that the substantial increase in output that occurred during World War II was the result of a huge temporary increase in government purchases brought about by the war effort. Between 1940 and 1945, government purchases rose over 400 percent, and real output nearly doubled over the same period. Immediately following the end of the war, in 1946, government spending fell by 75 percent, and real output declined sharply. This response of output to large temporary government spending shocks is consistent with both Keynesian models and neoclassical models. (See Barro (1981), Wynne (1989), and Ohanian (1997).) Similarly, the considerable decline in output associated with the Great Depression has been interpreted by many economists as a temporary response to a significant reduction in the money stock. (See, for example, Friedman and Schwartz (1963).) Figure 1 plots the log of real per capita GNP (y_t) in levels and in first differences. The substantial fluctuations in output during these two periods are clearly seen in this figure.

The view that some of the variation in GNP during the 1909–1970 period (the Nelson-Plosser (1982) sample) was transitory and that these transitory fluctuations could be substantial is not explicitly modeled under the null of unit root tests against trend break alternatives.

In this paper, we propose a simple statistical model that captures the idea that the sharp fluctuations in output that occurred during the Great Depression and World War II can reasonably be interpreted as transitory in nature. Our model generates time series as the sum of a latent random walk with drift and occasional large transitory movements driven by a regime-switching model. The regime switches are fully endogenous. Given our interest in large transitory movements in output during the Great Depression and World War II, the regime switches follow two independent Markov chains modeled after these two episodes in U.S. data.

We demonstrate that a process that is the sum of an integrated component with drift and an occasional large transitory component will generate data that in finite samples will be very difficult to distinguish from a trend stationary process with a trend break. In particular, the large temporary fluctuations in output that occur during wars and depressions may give the appearance of a change in intercept or trend slope or both in a trend stationary process. This demonstration suggests that in practice, unit root tests against trend break alternatives will tend to overreject the null.

To quantify the importance of these distortions for applied work, we conduct a Monte Carlo experiment. We study the rejection rates of the widely used Zivot-Andrews (1992) unit root tests for data generated from various unit root processes without trend breaks, but with occasional large transitory movements. Our main finding is that the Zivot-Andrews asymptotic test is strongly biased in small samples, with rejection rates as high as 60 percent at the 10 percent level for the data generating process (DGP) with transitory components and rejection rates as high as 35 percent for a pure random walk with drift DGP. Small-sample test statistics, which are rarely used in practice due to computational requirements, reduce this bias, but they do not eliminate the problem for the DGP with transitory components.

Other authors have argued that tests for unit roots may be sensitive to the nature of the DGP. Perhaps most prominently, Schwert (1987, 1989) shows that standard ADF tests overreject the unit root null if the true model is an ARIMA (0,1,1) model with an MA coefficient close to -1. Since the growth rates of many macroeconomic time series do not have MA coefficients near -1, however, many researchers have dismissed the notion that transitory dynamics in their data can interfere with statistical inference. Our study demonstrates that this practice can be a serious mistake.

There are fundamental differences, however, between our study and Schwert's point: We focus on trend break tests, not on tests against trend stationary alternatives. Our DGP is substantially different from Schwert's linear ARIMA model and can result in severe inference problems not foreseen by Schwert. The problem we study cannot be identified with the diagnostics that work for Schwert's problem. And procedures that result in correct inference in his problem do not in ours. Indeed, our results indicate that fundamental changes are needed in the application and interpretation of unit root tests against trend break alternatives.

Section 2 reviews the methodology of unit root tests against trend break alternatives. Section 3 describes the statistical model. Section 4 describes the simulation design. Section 5 presents the findings of the Monte Carlo analysis and compares our work to Schwert's in more detail. Section 6 concludes.

2. Unit Root Tests Against Trend Break Alternatives

Many methodologies exist for unit root tests against trend break alternatives, including, for example, those in Perron (1989), Rappoport and Reichlin (1989), Banerjee, Dolado, and Galbraith (1990), Perron (1990), Balke and Fomby (1991), Perron (1991), Perron and Vogelsang (1992a,b, 1993a,b), Park and Sung (1994), Stock (1994), Bradley and Jansen (1995), Perron and Vogelsang (1995), Nunes, Newbold, and Kuan (1996), and Montanes (1997). The early literature (for example, Perron (1989)) often determines the break points by inspecting the data. Today it is widely accepted that break points must be estimated endogenously. In this paper, we focus on the sequential break point selection tests developed by Zivot and Andrews (1992). These tests are very similar to the methodology used by Banerjee, Lumsdaine, and Stock (1992) and Perron (1997), and these tests are commonly used in applied work.[2] They also have recently been extended by Lumsdaine and Papell (1997) to allow for multiple trend breaks.

We consider three versions of the Zivot-Andrews tests that consider the same null hypothesis of a unit root, but differ in the type of structural break considered under the alternative. All tests assume that there is a one-time break under the alternative at date *TB*. Since the break point is assumed to be unknown, the Dickey-Fuller statistic is defined as the infimum of the sequence of the Dickey-Fuller statistics over all possible break points, not including the end points of the sample. Let $DU_t = 1(t > TB)$ and $DT_t = (t-TB)1(t > TB)$, where $t = 1, ..., T$ and $1(.)$ is the indicator function.

- Model A allows for a one-time break in the mean of the deterministic trend (intercept break),

- Model B allows for a one-time break in the growth rate of the deterministic trend (slope break), and

- Model C allows for a simultaneous break in intercept and slope:

$$\text{Model A} \qquad \Delta y_t = \mu + \beta t + \theta DU_t + \alpha y_{t-1} + \sum_{i=1}^{k} c_i \Delta y_{t-i} + \varepsilon_t$$

$$\text{Model B} \qquad \Delta y_t = \mu + \beta t + \gamma DT_t + \alpha y_{t-1} + \sum_{i=1}^{k} c_i \Delta y_{t-i} + \varepsilon_t$$

$$\text{Model C} \qquad \Delta y_t = \mu + \beta t + \theta DU_t + \gamma DT_t + \alpha y_{t-1} + \sum_{i=1}^{k} c_i \Delta y_{t-i} + \varepsilon_t.$$

For all three models, we test $H_0 : \alpha = 0$ against the one-sided alternative. We follow Zivot and Andrews (1992) in determining the number of augmented lags k by Perron's (1989) sequential t-value procedure, starting with an upper bound of eight lags. Ng and Perron (1995) establish that this procedure has better small-sample properties than information-based lag order selection criteria.

The test rejects the null of a unit root if the minimum of the ADF statistic t_α over all possible break points TB falls below its critical value at a given significance level. The asymptotic critical values for t_α are taken from Zivot and Andrews (1992). Zivot and Andrews (1992) also propose a small-sample extension of their tests that involves resampling the ADF statistic under the null by fitting an ARIMA $(p, 1, q)$ model to the data. The finite-sample critical values of that test can be read off from the empirical distribution of t_α.

3. An Unobserved-Components Model with Occasional Transitory Shocks

This section describes the statistical model underlying the DGP for the simulation study. We base our analysis on the annual U.S. per capita output series (y_t) of Nelson and Plosser (1982) for 1909–1970. Large transitory responses in output to temporary government spending shocks or to monetary shocks arise naturally in dynamic equilibrium models. In principle, it is possible to generate data from a dynamic optimizing macroeconomic model with regime switching in the decision rule coefficients. However, for tractability reasons and to simplify the analysis, we focus on a univariate reduced form time series model that captures the same features.

We treat World War II and the Great Depression as random events that occur with positive probability along the sample path. In the model, wars and depressions follow two independent Markov chains modeled after World War II and the Great Depression. The Markov chains will be specified in detail later in this section. Wars and depressions are assumed to have only transitory effects. Let s_{1t} denote the state underlying the war variable w and s_{2t} the state underlying the depression variable d. The variables w_{s1t} and d_{s2t} measure the transitory effects of World War II and the Great Depression on output. They take on nonzero values if the respective underlying state is activated and zero values otherwise.

Output (y_t) is assumed to follow a process that is the sum of a latent random walk with drift (z_t) and the two state-dependent transitory components w_{s1t} and d_{s2t}. The specification of our model differs in important ways from standard outlier models in that the effects of wars and depressions are not permanent by construction and in that the random walk is not directly observable:

$$\text{(1)} \qquad \begin{aligned} y_t &= z_t + w_{s1t} + d_{s2t} \\ z_t &= \mu + z_{t-1} + \varepsilon_t \\ s_{1t} &= \Pi^w \ s_{1t-1} + v_t \\ s_{2t} &= \Pi^d \ s_{2t-1} + u_t \end{aligned}$$

where $\varepsilon_t \overset{iid}{\sim} N(0, \ \sigma_\varepsilon^2)$, v_t, and u_t are martingale difference sequences and Π^w and Π^d are the transition probability matrices associated with s_{it}, $i = 1, \ 2$. (See Hamilton (1994).) Lowercase letters denote natural logs.

The states s_{1t} underlying the transitory component w_{s1t} are defined so that wars in the model will have the same duration as World War II in the U.S. data. For the purpose of this model, the World War II period is assumed to last six years, corresponding to the years 1941–1946 in the U.S. data. We include 1946 in the war period, since we are interested in the transitory effects of the war on the level of economic activity. This is the year immediately after the war, in which government expenditures and output fell dramatically. For expositional purposes, we discount the possibility of longer-lasting effects. Define the *primitive states* $s^*_{1t} \in \{W,\ P\}$ corresponding to whether a particular year t is a war year or a peace year. By dividing the 62 observations of the Nelson-Plosser output series into overlapping blocks of six consecutive annual observations, we can completely characterize the Nelson-Plosser data by 12 states s_{1t}, each consisting of six consecutive annual observations. For example, the six-year period 1938–1943 is given by {P, P, P, W, W, W}, the period 1939–1944 is given by {P, P, W, W, W, W}, and so on.

s_{1t}	s^*_{1t}	s^*_{1t-1}	s^*_{1t-2}	s^*_{1t-3}	s^*_{1t-4}	s^*_{1t-5}
1	W	W	W	W	W	W
2	P	W	W	W	W	W
3	P	P	W	W	W	W
4	P	P	P	W	W	W
5	P	P	P	P	W	W
6	P	P	P	P	P	W
7	P	P	P	P	P	P
8	W	P	P	P	P	P
9	W	W	P	P	P	P
10	W	W	W	P	P	P
11	W	W	W	W	P	P
12	W	W	W	W	W	P

Based on the Nelson-Plosser data, it is straightforward to determine the empirical transition probabilities from s_{1t} to s_{1t+1}, conditional on the last six states s^*_{1t}. We obtain the following transition probability matrix Π^w in terms of the states s_{1t} and s_{1t+1}.

$$\begin{array}{cccccccccccccc}
 & & & & & & & s_{1t} & & & & & & \\
 & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
s_{1t+1} & 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 7 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{45}{46} & 0 & 0 & 0 & 0 & 0 \\
 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{46} & 0 & 0 & 0 & 0 & 0 \\
 & 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 & 11 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}$$

The states s_{2t} underlying d_{s2t} are constructed in a similar way, so that depressions in the model will have the same duration as the Great Depression in the U.S. data. For the purpose of the model, the Great Depression is assumed to last five years, corresponding to the years 1929–1933 in the U.S. data. Define $s^*_{2t} \in \{D, N\}$ corresponding to whether a particular year t is a depression year or a normal year. Then define 10 states s_{2t}, each consisting of five consecutive annual observations, such that these states completely characterize the observed U.S. data.

s_{2t}	s^*_{2t}	s^*_{2t-1}	s^*_{2t-2}	s^*_{2t-3}	s^*_{2t-4}
1	D	D	D	D	D
2	N	D	D	D	D
3	N	N	D	D	D
4	N	N	N	D	D
5	N	N	N	N	D
6	N	N	N	N	N
7	D	N	N	N	N
8	D	D	N	N	N
9	D	D	D	N	N
10	D	D	D	D	N

The corresponding empirical transition probabilities from s_{2t} to s_{2t+1} are summarized in the matrix Π^d.

$$\begin{array}{cccccccccccc}
 & & & & & & s_{2t} & & & & & \\
 & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
s_{2t+1} & 5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 6 & 0 & 0 & 0 & 0 & 1 & \frac{48}{49} & 0 & 0 & 0 & 0 \\
 & 7 & 0 & 0 & 0 & 0 & 0 & \frac{1}{49} & 0 & 0 & 0 & 0 \\
 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 & 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}$$

Nothing in the construction of the model depends on the duration of the wars and depressions or on the particular values adopted for the transition probabilities. For example, we could easily have adopted a standard two-state Markov chain with one-period memory instead. The reason for choosing this particular Markov chain model is that we want to rule out the possibility of wars and depressions of longer or shorter duration than those in the U.S. data. Our model is the simplest DGP which generates wars of the same duration as World War II and depressions of the same duration as the Great Depression.

4. Simulation Design

This section discusses the design of a Monte Carlo experiment to study the small-sample behavior of the asymptotic and bootstrapped Zivot-Andrews tests. Our simulation is based on the annual per capita GNP series of Nelson and Plosser for 1909–1970. We focus on this series because it has been analyzed by several authors and is central to a number of macroeconomic issues. We consider two basic DGPs for the Monte Carlo analysis.

The first DGP is based on the model (1) in Section 3. To generate data from (1), we need to specify values for the drift (μ) and innovation standard error (σ_ε) as well as the values that w_{s1t} and d_{s2t} take on if the war regime or if the depression regime is activated. These parameters are based on the Nelson-Plosser per capita GNP data. Under the null of the Zivot-Andrews test, the data are difference-stationary, so we proceed by regressing the first-differenced logged Nelson-Plosser data on a constant. The estimated drift ($\hat{\mu}$) over the sample period is 1.64 percent, with standard error 0.0084. The estimated residual standard error ($\hat{\sigma}_\varepsilon$) is 0.0653. However, the presence of measurement error, in particular for the pre-1929 period, and the possible presence of large transitory fluctuations during World War II and the Great Depression suggest that this estimate of the innovation variance of the permanent component is considerably inflated.[3] We therefore recalculate $\hat{\sigma}_\varepsilon$ based on the residual standard error of the same regression for annual postwar data. The revised estimate is $\hat{\sigma}_\varepsilon = 0.024$.[4]

The values of the transitory components d_{s2t} and w_{s1t} are specified as the transitory variation in output over the Great Depression (1929–1933) and World War II (1941–1946) in the Nelson-Plosser per capita GNP series. We approximate the transitory movements during these episodes as the deviations from drift in the logged data over the periods 1929–1933 (for the Great Depression) and 1941–1946 (for World War II):

$$\begin{aligned} d_{s2t} &= 1(s^*_{2t} = D)(\Delta y_t - \hat{\mu}) \\ w_{s1t} &= 1(s^*_{1t} = W)(\Delta y_t - \hat{\mu}). \end{aligned} \tag{2}$$

Figure 2 plots the observed sequences for d_{s2t} and w_{s1t}. Since the focus of this exercise is on the potential effects of occasional large transitory movements in the data on the performance of the Zivot-Andrews test, we abstract from other sources of transitory fluctuations in GNP, including the possibility of transitory fluctuations during World War I and the Korean War. While our definition of the transitory component is not the only one possible, it is consistent with common interpretations of the data, and it conveys a reasonable sense of the magnitude of the transitory movements in the data during these episodes.

Our main concern is that in small samples, large transitory movements, (w_{s1t}) and (d_{s2t}), may resemble trend breaks, leading conventional trend break tests to overreject the null of a unit root. To verify this conjecture, we will generate data from the model in (1):

$$\begin{aligned} y_t &= z_t + w_{s1t} + d_{s2t} \\ z_t &= \mu + z_{t-1} + \varepsilon_t \\ s_{1t} &= \Pi^w \ s_{1t-1} + v_t \\ s_{2t} &= \Pi^d \ s_{2t-1} + u_t. \end{aligned}$$

We will use our preferred parameter specifications

$$\mu = 0.0164, \ \sigma_\varepsilon = 0.024 \tag{3}$$

$$d_{s2t} = \{0.0370, -0.1311, -0.1048, -0.1833, -0.0410\} \text{ for } s_{2t}^* = D \tag{4}$$

$$w_{s1t} = \{0.1228, 0.0941, 0.0937, 0.0411, -0.0448, -0.1544\} \text{ for } s_{1t}^* = W \tag{5}$$

with Π^w and Π^d as given in Section 3. We will refer to this DGP as DGP 1.

For comparative purposes, we also include results for a random walk with drift with no transitory components. We will refer to this alternative DGP as DGP 2:

$$y_t = \mu + y_{t-1} + \varepsilon_t. \tag{6}$$

In this simple model, all shocks (ε_t) have permanent effects. Since the data for DGP 2 are generated under the null hypothesis of the Zivot-Andrews test, the relative frequency of rejections can be interpreted as the size of the test. We use identical values of μ and σ_ε for both DGPs (equations (1) and (6)) to isolate the effect of adding transitory components to the random walk process.

In addition to the two unrestricted stochastic processes DGP 1 and DGP 2, we conduct a sensitivity analysis in which we consider three restricted DGPs based on DGP 1, which we denote DGP 3, 4, and 5. These restricted processes are designed to isolate the marginal contribution of various factors to the performance of the Zivot-Andrews tests. These restricted processes analyze the following effects: the timing of wars and depressions, the sequencing of wars and depressions, and the number of wars and depressions.

In DGP 3, exactly one war and one depression occur in the sampling period. We allow for independent draws of wars and depressions at random dates over the 1910–1969 interval as in DGP 1. However, to make the data in the Monte Carlo study more comparable to the actual data, we only retain Monte Carlo draws with exactly one war and exactly one depression, and we require that both events be included in the sampling interval in their entirety. This means that the depression and the war can occur anywhere in the sample (with the exception of the end points), but not necessarily in their historical order, and possibly even at the same time.

In DGP 4, exactly one war and one depression occur in the sampling period, and the depression precedes the war. This DGP builds on DGP 3 in that we are still restricting ourselves to Monte Carlo trials with exactly one war and one depression, but in addition, we are trying to preserve the order in which the depression and the war occurred in the U.S. data. For that reason, we treat the entire sequence of events from 1929 through 1946 as one event (with zeros imposed for the transitory component for 1934–1940 to make the results compatible with DGP 5), and we allow the sequence to start at random points in the interval between 1910 and 1969. This means that the Great Depression and World War II can occur anywhere in the sample (with the exception of the starting point and the end point of the series), but that the depression will always precede the war by eight years as in the U.S. data.

In DGP 5, exactly one war and one depression occur in the sampling period, and the depression and the war occur on the same dates as in the U.S. data. This DGP builds on DGP 4, but in addition, we fix the dating of the depression and the war across all Monte Carlo trials, so that it is exactly identical to the dating of the Great Depression and World War II in the U.S. data. Thus, in each Monte Carlo trial, the Great Depression lasts from 1929 to 1933 and World War II from 1941 to 1946. However, we continue to draw permanent innovations ε_t to the DGP throughout the sample period, including the depression and the war. Thus, output realizations during these events are not fixed across Monte Carlo trials, and the resulting time series $\{y_t\}$ may look quite different from the historical per capita GNP series.

The five DGPs in our study are summarized in Exhibit 2. The number of Monte Carlo trials is 1,000 for each DGP. We average the number of rejections of the Zivot-Andrews test for Models A, B, and C across Monte Carlo trials and compare this number to the nominal size of the test at the 1, 2.5, 5, and 10 percent significance levels. The Monte Carlo standard errors for the rejection rates are 0.3, 0.5, 0.7, and 0.95 percentage points for the 1, 2.5, 5, and 10 percent tests, respectively. We consider two sample sizes in our experiments. We initially use a sample size of 62, which is the number of observations in the annual Nelson-Plosser per capita GNP data. This sample size is fairly common in applied work. To gain an understanding of how sample size affects the asymptotic tests, we subsequently conduct our experiments for a larger sample.

5. Simulation Results

A. Basic Findings

In this section, we report from our Monte Carlo simulation the behavior of the Zivot-Andrews tests in small samples. Table 1 presents the rejection rates of the null hypothesis of a unit root for Model A (intercept break), Model B (slope break), and Model C (simultaneous slope and intercept break) under DGP 1 (1) and DGP 2 (6). For completeness, Table 1 also shows the corresponding rejection rates for the unit root tests against trend stationary alternatives without breaks. Our simulation results for these processes for a sample of size 62 suggest that the asymptotic tests, which have been applied by nearly every researcher who has used the Zivot-Andrews tests, routinely mistake the transitory effects of wars and depressions for trend breaks. That result holds across all models and nominal sizes, but in relative terms, the accuracy of the test deteriorates with higher significance levels. For example, at the nominal 10 percent level, the test rejects the null of a unit root in favor of a trend break in 54 percent to 60 percent of all trials. At the nominal 1 percent level, the test rejects the null in 23 to 28 percent of all trials. Moreover, the rejection rates are not sensitive to the type of break considered. In addition, we find that there is considerable bias in the number of rejections even in the absence of transitory dynamics. In the pure random walk model (DGP 2), the size distortions of the test against trend break alternatives are an additional 6 percent to 17.5 percent higher, compared with rejection rates for the no-break alternatives. This finding establishes that trend break tests have a tendency to overreject the unit root null in small samples quite independently of the problems due to transitory dynamics we discussed earlier. The latter effect can be assessed by comparing the rejection rates for the no-break model under DGP 1 and DGP 2. Table 1 shows that the transitory dynamics in DGP 1 individually are responsible for an increase in rejection rates of up to 22.7 percent. More generally, of course, rejections may be due to both overfitting and the presence of transitory dynamics. For DGP 1, Table 1 allows us to decompose the total increase in the rejection rate into these two effects. The increases in rejection rates due to overfitting are between 50 percent and 104 percent of the size distortions due to transitory components. Thus, both problems are tremendously important for applied work based on asymptotic critical values.

B. Sensitivity Analyses

Relative Size of Transitory Components and Permanent Shocks

Table 2 analyzes how sensitive these results are to changes in the magnitude of the transitory component (measured in multiples of our preferred values of d and w). Clearly, the number of rejections in general depends on the extent to which transitory effects dominate the time path of the output series. That, in turn, depends on the size of the permanent innovations (ε_t) (and, as we will show later, on the sample size). To test the robustness of our findings, we vary the size of the transitory effects in multiples of 0, 1/2, 1, and 2 while holding fixed the noise level as measured by the innovation standard deviation of the population process (σ_ε). Depending on the size of the transitory effect, the number of rejections varies from 33 percent to 76 percent. In all cases, the number of rejections increases with the ratio of the size of the transitory effect to the noise level. These results show how important the existence of transitory effects in the data can be for the performance of unit root tests against trend break alternatives. Interestingly, even for the pure unit root case without any transitory dynamics

(DGP 2), the rejection rates of the test are much too high. At the nominal 10 percent level, the test rejects in 23 to 35 percent of all trials.

Pretesting for Unit Roots

Tables 1 and 2 demonstrate that the Zivot-Andrews test rejects with high probability the null of a unit root test in favor of a trend break, even if the data do not contain such a break by construction. However, these findings are based on the premise that no pretesting has preceded the analysis. It seems reasonable to assume that many researchers would not consider structural break alternatives if standard ADF tests already rejected the unit root in favor of trend-stationarity. Such a view is consistent with the sequence of published work on the Nelson-Plosser data series. This suggests that we conduct additional simulations in which we mimic this practice by discarding all Monte Carlo trials for which standard ADF tests reject the null of a unit root at the nominal 10 percent level (using interpolated finite-sample critical values) and replacing them with new draws. As Nunes, Newbold, and Kuan (1996) point out, the critical values compiled by Zivot and Andrews (1992) are no longer valid once we recognize the existence of pretesting. Nunes, Newbold, and Kuan (1996) propose a modification of the Zivot-Andrews test to account for data mining. However, their modification has not been adopted in applied work, and here we are interested in the question of how misleading the results of the Zivot-Andrew tests are in actual practice. It therefore is useful to study the performance of the Zivot-Andrews tests the way they are currently applied in the literature. Moreover, additional simulation results suggest that rejection rates after pretesting are systematically lower than the actual rates by up to 18 percentage points and hence can be regarded as a conservative lower bound to the true rejection rates.

The results after pretesting are displayed in Table 3. The discussion focuses on the nominal 10 percent level. We begin with DGP 1. That process differs from DGP 1 in Table 1 only to the extent that we discard apparently trend stationary trials and replace them with new draws. As before, the number of wars and depressions in the sample is unrestricted. Events can occur at any point in time and in any order. The result is a drop in the number of spurious rejections from 54–60 percent to 34–46 percent after pretesting. Similarly, for the pure random walk model, after pretesting, the rejection rates drop as low as 15–31 percent. However, the effective size for DGP 2 still far exceeds the nominal size.

Restricted DGPs

While the Monte Carlo simulation results just discussed are perfectly valid as an indicator of the unconditional or average performance of the Zivot-Andrews test, they do not necessarily tell us how likely it is that the evidence for a trend break in the actual U.S. per capita GNP series is spurious. The reason is that our model can generate trials with possibly many world wars and great depressions or, for that matter, none at all, whereas the U.S. data are characterized by exactly one such event. We therefore proceed with a number of simulations conditional on key features of the U.S. data. For example, to isolate the marginal effects of the number of big events, in DGP 3, we consider Monte Carlo trials that include exactly one war and one depression in the sampling interval. We also make certain that these wars and depressions are included in their entirety in the sampling interval (excluding the end points), just as in the U.S. data, and we pretest all trials. As a result, the number of rejections in Table 3 rises slightly from 34 to 37 percent for Model A and from 45 to 46 percent for Model B, but jumps from 46 to 62 percent for Model C, close to the original result in Table 1.

To determine how sensitive these results are to the timing and sequencing of the Great Depression and World War II, we conduct two more experiments. To investigate the importance of the order in which the Great Depression and World War II occurred, in DGP 4, we define a new Markov chain model in which we treat the entire 1929–1946 episode as one sequence (with zeros imposed for the years 1934–1940 to ensure compatibility of the results with DGP 5). After pretesting and discarding multiple events, we find that, on average, the sequencing of the events has little impact on the rejection rates. When we compare DGP 4 with DGP 3, the rejection rates rise slightly by about 6 percent for Model A and 3 percent for Model B and fall by 5 percent for Model C. Finally, we consider the case in which the dating of the Great Depression and World War II is fixed at the historical dates throughout all trials. After pretesting, in DGP 5, the rejection rates rise even further to 54 percent for Model B, drop slightly to 39 percent for Model A, and remain constant at 57 percent for Model C.

To summarize, the results in Table 3 are consistent with several of the initial findings, even after accounting for pretesting and possible restrictions on the DGP. First, asymptotic tests always reject too often in small samples, and the distortions can be substantial. These findings hold across all DGPs, although the behavior of the test is worse under DGPs that include occasional large transitory components. For example, under DGP 2, the test for Model B rejects the null of a unit root up to 30 percent of the time, and under the other DGPs, the test for Model C favors the trend break hypothesis in six out of ten cases, even in the absence of a break. Second, rejection rates increase relative to the nominal size while the significance level is raised. For example, while the nominal 10 percent test typically rejects four to six times too often, the nominal 5 percent test rejects five to nine times too often, and the nominal 1 percent test up to 27 times too often! Restricting attention to series containing exactly one war and one depression can make some difference depending on the model used, but does not alter the basic result. Moreover, the rejection probabilities are fairly robust to the sequence and dating of the wars and depressions once attention is restricted to draws with exactly one war and one depression in the sampling interval.

Large-Sample Results

So far, all results are based on the widely used Nelson-Plosser per capita GNP series with a sample size of 62 annual observations. However, some studies use the asymptotic Zivot-Andrews test for samples as small as 44 annual observations (Raj and Slottje (1994)) or 27 annual observations (Alba and Papell (1995)). In that case, the performance of the test is likely to be even worse.[5] At the other extreme, one study by Sadorsky (1994) uses as many as 169 annual observations on average tariff rates. It is reasonable to expect the performance of the asymptotic structural break test to improve as the sample size becomes large. To address this issue, we double the sample size to 124 observations and repeat the experiments summarized in Table 3. This sample size roughly corresponds to the number of annual observations currently available for per capita GNP. More importantly, it is larger than the sample sizes typically used in the literature. Among all the studies we survey, only three use more annual observations.[6]

The results of the experiments with 124 observations are presented in Table 4. The performance of the asymptotic test improves, but the size distortions remain substantial. Although the asymptotic test now seems to perform well for Model A under DGP 2, it continues to overreject for Models B and C under the same DGP, and for all models under the other DGPs. For example, the nominal 10 percent test rejects up to five out of ten times in the presence of transitory components (down from six out of ten times) and up to two out of ten times for the pure integrated process (down from three out of ten times). At the 1 percent level, the asymptotic tests rejects up to 26 times too often in the presence of transitory components (down from 27 times) and up to two times too often in the pure random walk case (down from six times).

Performance of the Bootstrapped Test Statistic

The evidence presented in Tables 1 through 4 suggests strongly that the asymptotic Zivot-Andrews test is not reliable in sample sizes typically used in applied work. We therefore proceed with the analysis of the bootstrapped version of the same test for the sample of size 62. These tests are rarely used in practice, because they can be computationally expensive. In fact, of all the studies in our survey that use endogenous break point selection tests, only Zivot and Andrews (1992), Sadorsky (1994), and Lumsdaine and Papell (1997) have bootstrapped the test statistic, and the latter paper uses an unreasonably small number of replications to reduce the computational cost. The computational burden of compiling small-sample test statistics is even greater in the context of Monte Carlo simulation. We therefore restrict our attention to DGP 4 (treating the entire 1929–1946 episode as one random event) and DGP 2 (random walk model) and focus on the nominal 10 percent test for the sample size of 62. For each of the Monte Carlo trials, we fit an ARIMA model under the null and calculate the bootstrapped critical values based on 1,000 bootstrapped replications of the Zivot-Andrews ADF statistic.[7] For 500 Monte Carlo trials, the Monte Carlo standard error of the size estimate is 1.3 percent. More precise tests would require considerably more replications. For example, to obtain stable tails of the bootstrapped distribution of the ADF statistic at the 5 percent level would require 2,000 bootstrapped replications. At the 1 percent level, at least 5,000 replications would be needed.

Table 5 presents the results for the bootstrapped test statistic after pretesting. The bootstrapped test statistic clearly performs much better than the asymptotic test, but it does not eliminate the size distortions. Moreover, the performance of the bootstrapped test statistic is erratic. In the pure random walk case with no transitory components (DGP 2), it does not reject the null often enough for Models A and B. The rejection rates are significantly different from the nominal size at the 95 percent significance levels. With transitory components in the data (DGP 4), the bootstrapped test tends to reject the null too often in Models A and C. For Model C, the rejection rate is 23 percent, more than twice the nominal size. For Model A, the distortions are less severe, but we can still reject the hypothesis that the test is accurate at the 95 percent significance level.

C. A Comparison of Our Analysis with Schwert's

Our study should not be confused with the very different work of Schwert (1987, 1989). Schwert shows that tests for unit roots against a simple trend stationary alternative are sensitive to the assumption that the data are generated by a pure AR process. When the process contains an MA component, the finite-sample distribution of the DF statistic may be far from the tabulated distribution. Schwert shows that the ADF test tends to significantly overreject the unit root null if the true model is an ARIMA (0,1,1) model with an MA coefficient near unity. He also presents evidence suggesting that this distortion can be corrected by recalculating the critical values of the ADF test under the null of an ARIMA (0,1,1) model.

While Schwert's work is well known, applied researchers have abstracted from it. This is largely because the growth rates of many economic time series, such as GNP, do not have moving average roots near unity. For example, estimating an ARIMA (0,1,1) model for real per capita GNP yields

$$\Delta y_t = 0.0163 + \varepsilon_t + 0.3066\varepsilon_{t-1},\ \sigma_\varepsilon = 0.0618.$$

The estimated coefficient suggests that the moving average root is far from -1, unlike in Schwert's analysis. Based on this evidence, it is understandable why applied researchers have abstracted from the effect of moving average components. In fact, we conduct a Monte Carlo analysis based on 1,000 trials from this ARIMA (0,1,1) DGP and find that for a sample of size 62, the size distortion of the standard ADF test with conventional critical values is only 5.8 percent for the nominal 10 percent level test.

The DGP we study differs considerably from Schwert's basic linear ARIMA (0,1,1) model, as do its implications for applied work. First, the motivation for, and nature of, the transitory dynamics is very different. The motivation stressed by Schwert for linear ARIMA models is aggregation and measurement error. The motivation for dynamics in this study is that transitional fluctuations in endogenous variables arise naturally in equilibrium from occasional shocks associated with events such as World War II and the Great Depression.

Second, while both DGPs contain transitory components, these components manifest themselves in very different ways. In particular, procedures that can diagnose the problem of MA components analyzed by Schwert fail to identify the serious problems with inference in our DGP. Estimating an ARIMA (0,1,1) model for our DGP yields a moving average coefficient of about 0.3, which is roughly equal to that in the data. As a result, researchers familiar with Schwert's critique would fail to detect any warning signs of the problems of inference we document. In fact, given the value of the estimated MA coefficient, those familiar with Schwert's analysis and related studies seem to have the view that there are no important problems with inference arising from transitory dynamics. For example, Perron and Vogelsang (1992a) conduct a Monte Carlo test of trend break tests using an ARMA (1,1) process. For an AR coefficient of unity and an MA coefficient of 0.5, they find no evidence of size distortions. For an MA coefficient of -0.5, there is some size distortion, but they find that adding augmented lags corrects the problem. Even researchers who have not completely dismissed Schwert-type problems (for example, De Haan and Zelhorst (1993)) flatly state that using a large number of augmented lags protects them from the Schwert critique.

Third, although there is no evidence of a serious Schwert-type problem with this DGP, there are other serious problems due to transitory dynamics. For example, the same ADF test against the trend stationary alternative that has a size distortion of 5.8 percent for the ARIMA DGP yields a size distortion of 30.1 percent for our DGP. Moreover, even if we follow Schwert's suggestion to recalculate the critical values of the ADF test under the null of an ARIMA $(p, 1, q)$ model, the test will continue to overreject the null in the presence of occasional large transitory shocks. Bootstrapped versions of the ADF tests under the ARIMA null hypothesis yield a rejection rate of just 8.5 percent for the ARIMA (0,1,1) DGP (with a Monte Carlo standard error of 1.85 percent), but a rate of 17.7 percent for the DGP with occasional transitory movements. Thus, the bootstrap removes the size distortions for the ARIMA DGP, but it does not for our DGP.

All of these findings highlight how important it is for applied researchers to understand the effects of a few large transitory shocks on inference and the inability of standard statistical procedures to diagnose the problem.

6. Summary and Conclusion

Allowing for a break in the trend function can alter the outcome of tests for unit roots. Unit root tests against trend break alternatives are being used by many macroeconomists to answer a variety of applied questions. These tests are based on the premise that the dating of the trend breaks is associated with major economic events with permanent effects on economic activity. We have argued that this assumption is not plausible for many aggregate time series, such as U.S. per capita GNP, for which estimated break dates coincide with wars or the Great Depression. There is considerable agreement among macroeconomists that major events, such as the Great Depression and World War II, are likely to have large transitory effects on economic activity. Standard unit root tests against trend break alternatives do not explicitly model these occasional large transitory effects and, as a result, can give misleading answers in practice.

To quantify how important these distortions are for applied work, we conducted a Monte Carlo experiment. We proposed a simple reduced-form model that captures the idea that there are occasional large transitory variations in output that reflect major events, such as the Great Depression and World War II. Our model consists of the sum of a latent random walk and a transitory component driven by a regime-switching model. By construction, this process contains a unit root, but does not include any trend breaks. We studied the rejection rates of the widely used Zivot-Andrews (1992) unit root tests for data generated by this process.

Our results raise serious questions about the reliability of standard unit root tests against trend break alternatives in small samples. We found that at the nominal 10 percent level, the asymptotic Zivot-Andrews tests reject the null of a unit root in favor of a trend break in up to 60 percent of all Monte Carlo trials. Even for the pure random walk model, the test rejected the null hypothesis in as many as 35 percent of the trials at the nominal 10 percent level. The results are robust to several alternative specifications of the model. The performance of the test improves with larger sample sizes, but we showed that even for relatively large samples, the rejection rates of the asymptotic tests remain high. Computationally expensive bootstrapped tests, which have been used rarely in applied work, are somewhat more reliable, but still perform erratically.

The intuitive explanation for the poor performance of these tests is that occasional large transitory movements in the data, such as the Great Depression and the expansion during World War II, and genuine structural breaks in a trend stationary model are hard to distinguish in small samples. It is possible that our results may be sensitive to the precise measurement of the transitory component in U.S. GNP. Our aim in this paper has not been to precisely interpret a particular historical episode, but to furnish a plausible example of how the Zivot-Andrews tests perform in the presence of occasional large fluctuations that can be given sensible transitory interpretations. Regardless of the true size of the transitory component, we showed that a slight modification of the statistical model consistent with the standard economic interpretation of the data is sufficient to render the Zivot-Andrews tests inaccurate in small samples.

We conjecture that recent extensions of the Zivot-Andrews tests to allow for multiple breaks (Lumsdaine and Papell (1997)) are likely to suffer from the same type of problem. Our experience in this paper suggests that multiple-break tests would also tend to pick out large movements in the data, such as those that occurred during the Great Depression and World War II, as trend breaks. Another interesting extension of this research would be to consider the behavior of the tests in the presence of transitional dynamics following large shocks. In particular, both the Great Depression and World War II were periods of low capital accumulation, which suggests that capital was low relative to its steady state following these episodes. The neoclassical growth model of Solow (1956) predicts a temporary, but very persistent, period of relatively fast economic growth until capital returns to its steady-state growth path. This suggests that transitional dynamics following a large temporary shock may appear to be a trend break in the data.

We conclude from our Monte Carlo analysis that the many applied researchers working in this area should use considerable caution in applying unit root tests allowing for trend break alternatives in small and moderately sized samples. In particular, we have shown that the evidence in favor of a trend break in the Nelson-Plosser per capita GNP series is likely to have been spurious. Our findings suggest that many of the reported rejections of the unit root hypothesis in the literature based on such tests may have been spurious as well. While this paper has focused on the widely used Zivot-Andrews test, we conjecture that other trend break tests will have similar problems in the presence of occasional large transitory movements in the data. Our evidence against the trend break hypothesis does not imply that there actually are unit roots in the data. Rather, it suggests that existing tests may not be very informative in the sample sizes typically encountered in macroeconomics, unless the possibility of occasional large transitory movements in the data can be ruled out a priori.

Most importantly, we conclude that tests like the Zivot-Andrews test are likely to be useful for some macroeconomic time series, but not for all series. Our analysis highlights the dangers of automatically applying these tests to macroeconomic data and suggests that applied researchers need to give careful thought to the nature of the exogenous events that may have triggered permanent changes in economic growth and to alternative economic explanations for possible trend breaks based on endogenous transition dynamics. We conclude that a purely statistical analysis of the trend properties of economic time series is not sufficient. In addition to the statistical analysis, researchers should consider whether there are important shocks in the process, and economic theory should be used to understand how those shocks can affect the endogenous variables under study.

Notes

[1] In some cases, allowing for structural breaks under the alternative reverses the conclusions of earlier VAR studies. (See Gamber and Joutz (1993) and Fernandez (1994).)

[2] See, for example, Raj (1992), Raj and Slottje (1994), Sadorsky (1994), and Serletis (1995).

[3] Romer (1989) argues that after adjusting for data construction and collection methods, the variability of pre- and postwar fluctuations is fairly similar.

[4] We also verify that the random walk model with drift is an adequate representation of U.S. per capita real GNP in the postwar period. For the first five coefficients of the autocorrelation function of the first-differenced data, we cannot reject the null that the autocorrelations are zero at conventional significance levels.

[5] Alba and Papell (1995, p. 267) conclude that they were still able to find strong evidence against the unit root hypothesis, despite the short time span of their data. In contrast, our simulation evidence suggests that their results may have been obtained because of the short time span.

[6] While our analysis is based on annual data, some studies analyze postwar data with as many as 160 quarterly observations or monthly data with close to 400 observations. However, in practice, power considerations dictate the use of the data with the longest time span, and postwar quarterly or monthly data are currently available for a span shorter than 50 years, even fewer than our sample size of 62 annual observations. (For a similar argument, see Perron (1992).)

[7] Zivot and Andrews allow a maximum order of five for p and q in selecting the best-

fitting ARIMA $(p, 1, q)$ model. In contrast, we impose a maximum order of three for both p and q because of computational considerations.

References

Alba, J.D., and D.H. Papell (1995), "Trend Breaks and the Unit-Root Hypothesis for Newly Industrializing and Newly Exporting Countries," *Review of International Economics* 3(3), 264–274.

Balke, N.S., and T.B. Fomby (1991), "Shifting Trends, Segmented Trends, and Infrequent Permanent Shocks," *Journal of Monetary Economics* 28(1), 61–85.

Banerjee, A., J. Dolado, and J.W. Galbraith (1990), "Recursive Tests for Unit Roots and Structural Breaks in Long Annual GNP Series," unpublished manuscript, Department of Economics, University of Florida.

Banerjee, A., R.L. Lumsdaine, and J.H. Stock (1992), "Recursive and Sequential Tests of the Unit-Root and Trend-Break Hypotheses: Theory and International Evidence," *Journal of Business and Economic Statistics* 10(3), 271–287.

Barro, R.J. (1981), "Output Effects of Government Purchases," *Journal of Political Economy* 89(6), 1086–1121.

Ben-David, D., R.L. Lumsdaine, and D.H. Papell (1995), "The Unit Root Hypothesis in Long-Term Output: Evidence from two Structural Breaks for 16 Countries," unpublished manuscript, University of Houston.

Ben-David, D., and D.H. Papell (1995), "The Great Wars, the Great Crash, and Steady State Growth: Some New Evidence About an Old Stylized Fact," *Journal of Monetary Economics* 36(3), 453–475.

Bradley, M.D., and D.W. Jansen (1995), "Unit Roots and Infrequent Large Shocks: New International Evidence on Output Growth," *Journal of Money, Credit, and Banking*

27(3), 876–893.

Carlino, G.A., and L.O. Mills (1993), “Are U.S. Regional Incomes Converging? A Time Series Analysis,” *Journal of Monetary Economics* 32(2), 335–346.

Cheung, Y.-W., and M.D. Chinn (1996) “Deterministic, Stochastic, and Segmented Trends in Aggregate Output: A Cross-Country Analysis,” *Oxford Economic Papers* 48(1), 134–162.

Christiano, L.J. (1992), “Searching for a Break in GNP,” *Journal of Business and Economic Statistics* 10(3), 237–250.

Culver, S.E., and D.H. Papell (1995), “Real Exchange Rates under the Gold Standard: Can They Be Explained by the Trend Break Model?” *Journal of International Money and Finance* 14(4), 539–548.

Culver, S.E., and D.H. Papell (1997), “Is There a Unit Root in the Inflation Rate? Evidence from Sequential Break and Panel Data Methods,” *Journal of Applied Econometrics* 12, 435–444.

De Haan, J., and D. Zelhorst (1993), “Does Output Have a Unit Root? New International Evidence,” *Applied Economics* 25(7), 953–960.

Duck, N.W. (1992), “UK Evidence on Breaking Trend Functions,” *Oxford Economic Papers* 44(3), 426–439.

Edison, H.J., and E. O’N. Fisher (1991), “A Long-Run View of the European Monetary System,” *Journal of International Money and Finance* 10(1), 53–70.

Evans, G.W. (1989), “Output and Unemployment Dynamics in the United States: 1950–1985,” *Journal of Applied Econometrics* 4(3), 213–237.

Evans, M.D.D., and K.K. Lewis (1995), “Do Expected Shifts in Inflation Affect Estimates of the Long-Run Fisher Relation?" *Journal of Finance* 50(1), 225–253.

Fernandez, D.G. (1994), "Breaking Trends and the Money-Output Correlation," unpublished manuscript, SAIS, Johns Hopkins University.

Friedman, M., and A.J. Schwartz (1963), *A Monetary History of the United States, 1867–1960,* Princeton University Press, Princeton, NJ.

Gamber, E.N., and F.L. Joutz (1993), "The Dynamic Effects of Aggregate Demand and Supply Disturbances: Comment," *American Economic Review* 83(5), 1387–1393.

Hamilton, J.D. (1994), *Time Series Analysis,* Princeton University Press, Princeton, NJ.

Husted, S. (1992), "The Emerging U.S. Current Account Deficit in the 1980s: A Cointegration Analysis," *Review of Economics and Statistics* 74(1), 159–166.

Loewy, M.B., and D.H. Papell (1996), "Are U.S. Regional Incomes Converging? Some Further Evidence," *Journal of Monetary Economics* 38(3), 587–598.

Lumsdaine, R.L., and D.H. Papell (1997), "Multiple Trend Breaks and the Unit-Root Hypothesis," *Review of Economics and Statistics* 79(2), 212–218.

Nelson, C.R., and C.I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics* 10(2), 139–162.

Ng, S., and P. Perron (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(429), 268–281.

Nunes, L.C., P. Newbold, and C.-M. Kuan (1996), "Testing for Unit Roots with Breaks: Evidence on the Great Crash and the Unit Root Hypothesis Reconsidered," discussion paper, University of Illinois at Urbana-Champaign.

Montanes, A. (1997) "Level Shifts, Unit Roots and Misspecification of the Breaking Date," *Economics Letters* 54(1), 7–13.

Ohanian, L.E. (1997), "The Macroeconomic Effects of War Finance in the United States: World War II and the Korean War," *American Economic Review* 87(1), 23–40.

Park, J.Y., and J. Sung (1994), "Testing for Unit Roots in Models with Structural Change," *Econometric Theory* 10(5), 917–936.

Perron, P. (1989), "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica* 57(6), 1361–1401.

Perron, P. (1990), "Testing for a Unit Root in a Time Series with a Changing Mean," *Journal of Business and Economic Statistics* 8(2), 153–162.

Perron, P. (1991), "A Test for Changes in a Polynomial Trend Function for a Dynamic Time Series," unpublished manuscript, Department of Economics, Princeton University.

Perron, P. (1992), "Trend, Unit Root, and Structural Change: A Multi-Country Study with Historical Data," *American Statistical Association: Proceedings of the Business and Economics Statistics Section,* 144–149.

Perron, P. (1997), "Further Evidence on Breaking Trend Functions in Macroeconomic Variables," *Journal of Econometrics* 80(2), 355–386.

Perron, P., and T.J. Vogelsang (1992a), "Nonstationarity and Level Shifts with an Application to Purchasing Power Parity," *Journal of Business and Economic Statistics* 10(3), 301–320.

Perron, P., and T.J. Vogelsang (1992b), "Testing for a Unit Root in a Time Series with a Changing Mean: Corrections and Extensions," *Journal of Business and Economic Statistics* 10(4), 467–470.

Perron, P., and T.J. Vogelsang (1993a), "A Note on the Asymptotic Distribution of Unit Root Tests in the Additive Outlier Model with Breaks," *Revista de Econometria* 13(2), 181–207.

Perron, P., and T.J. Vogelsang (1993b), "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis: Erratum,"*Econometrica* 61(1), 248–249.

Perron, P., and T.J. Vogelsang (1995), "Additional Tests for a Unit Root Allowing for a Break in the Trend Function at Unknown Time," unpublished manuscript, Department of Economics, Cornell University.

Pischke, J.-S. (1991), "Measuring Persistence in the Presence of Trend Breaks," *Economics Letters* 36(4), 379–384.

Raj, B. (1992), "International Evidence on Persistence in Output in the Presence of an Episodic Change," *Journal of Applied Econometrics* 7(3), 281–293.

Raj, B., and D.J. Slottje (1994), "The Trend Behavior of Alternative Income Inequality Measures in the United States from 1947–1990 and the Structural Break," *Journal of Business and Economic Statistics* 12(4), 479–487.

Rappoport, P., and L. Reichlin (1989), "Segmented Trends and Non-Stationary Time Series," *Economic Journal* 99, 168–177.

Romer, C.D. (1989), "The Prewar Business Cycle Reconsidered: New Estimates of Gross National Product, 1869–1908," *Journal of Political Economy* 97(1), 1–37.

Sadorsky, P. (1994), "The Behavior of U.S. Tariff Rates: Comment," *American Economic Review* 84(4), 1097–1103.

Schwert, G.W. (1987), "Effects of Model Specification on Tests for Unit Roots in Macroeconomic Data," *Journal of Monetary Economics* 20(1), 73–103.

Schwert, G.W. (1989), "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business and Economic Statistics* 7(2), 147–159.

Serletis, A. (1995), "Random Walks, Breaking Trend Functions, and the Chaotic Structure of the Velocity of Money," *Journal of Business and Economic Statistics* 13(4), 453–458.

Solow, R.M. (1956), "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics* 70(1), 65–94.

Stock, J.H. (1994), "Deciding Between I(1) and I(0)," *Journal of Econometrics* 63(1), 105–131.

Wynne, M. (1989), "The Aggregate Effects of Temporary Government Purchases," unpublished Ph.D. dissertation, University of Rochester.

Zelhorst, D., and J. De Haan (1995), "Testing for a Break in Output: New International Evidence," *Oxford Economic Papers* 47(2), 357–362.

Zivot, E., and D.W.K. Andrews (1992), "Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis,"*Journal of Business and Economic Statistics* 10(3), 251–270.

Exhibit 1: Applications of Unit Root Tests Against Trend Break Alternatives

Nelson-Plosser data	**Perron (1989), Rappoport and Reichlin (1989), Perron (1991), Zivot and Andrews (1992), Nunes, Newbold, and Kuan (1996), Lumsdaine and Papell (1997), Perron (1997)**
tariff rates	**Sadorsky (1994)**
real exchange rates	**Edison and Fisher (1991), Perron and Vogelsang (1992a), Culver and Papell (1995)**
net exports	**Husted (1992)**
aggregate income inequality	**Raj and Slottje (1994)**
regional dispersions of income	**Carlino and Mills (1993), Loewy and Papell (1996)**
commodity prices	**Perron (1990)**
interest rates	**Perron (1990), Perron and Vogelsang (1992b), Duck (1992), Evans and Lewis (1995)**
unemployment rates	**Perron (1990), Perron and Vogelsang (1992b)**
price level	**Balke and Fomby (1991), Duck (1992)**
inflation	**Evans and Lewis (1995), Culver and Papell (1997)**
money	**Duck (1992)**
velocity of money	**Serletis (1995)**
U.S. postwar output	**Perron (1989), Balke and Fomby (1991), Pischke (1991), Christiano (1992), Zivot and Andrews (1992)**
U.S. long-run output	**Banerjee, Dolado, and Galbraith (1990), Balke and Fomby (1991)**
international output	**Banerjee, Lumsdaine, and Stock (1992), Raj (1992), Perron (1992), De Haan and Zelhorst (1993), Alba and Papell (1995), Bradley and Jansen (1995), Ben-David and Papell (1995), Ben-David, Lumsdaine, and Papell (1995), Zelhorst and De Haan (1995), Cheung and Chinn (1996), Perron (1997)**

Exhibit 2: Summary of Data Generating Processes Used in Monte Carlo Study

A. Unrestricted Data Generating Processes 1 and 2:

DGP 1: **Sum of random walk with drift and two transitory components driven by independent Markov chains based on World War II and the Great Depression**

DGP 2: **Random walk with drift**

B. Restricted Data Generating Processes 3, 4, 5:

Like DGP 1, but with the following additional restrictions:

DGP 3: **Exactly one war and one depression in the sample period.**

DGP 4: **Exactly one war and one depression in the sample period, and the depression precedes the war.**

DGP 5: **Exactly one war and one depression in the sample period, and the depression and the war occur on the same dates as in the U.S. data.**

Figure 1: Nelson-Plosser Per Capita GNP Series (1909–1970)

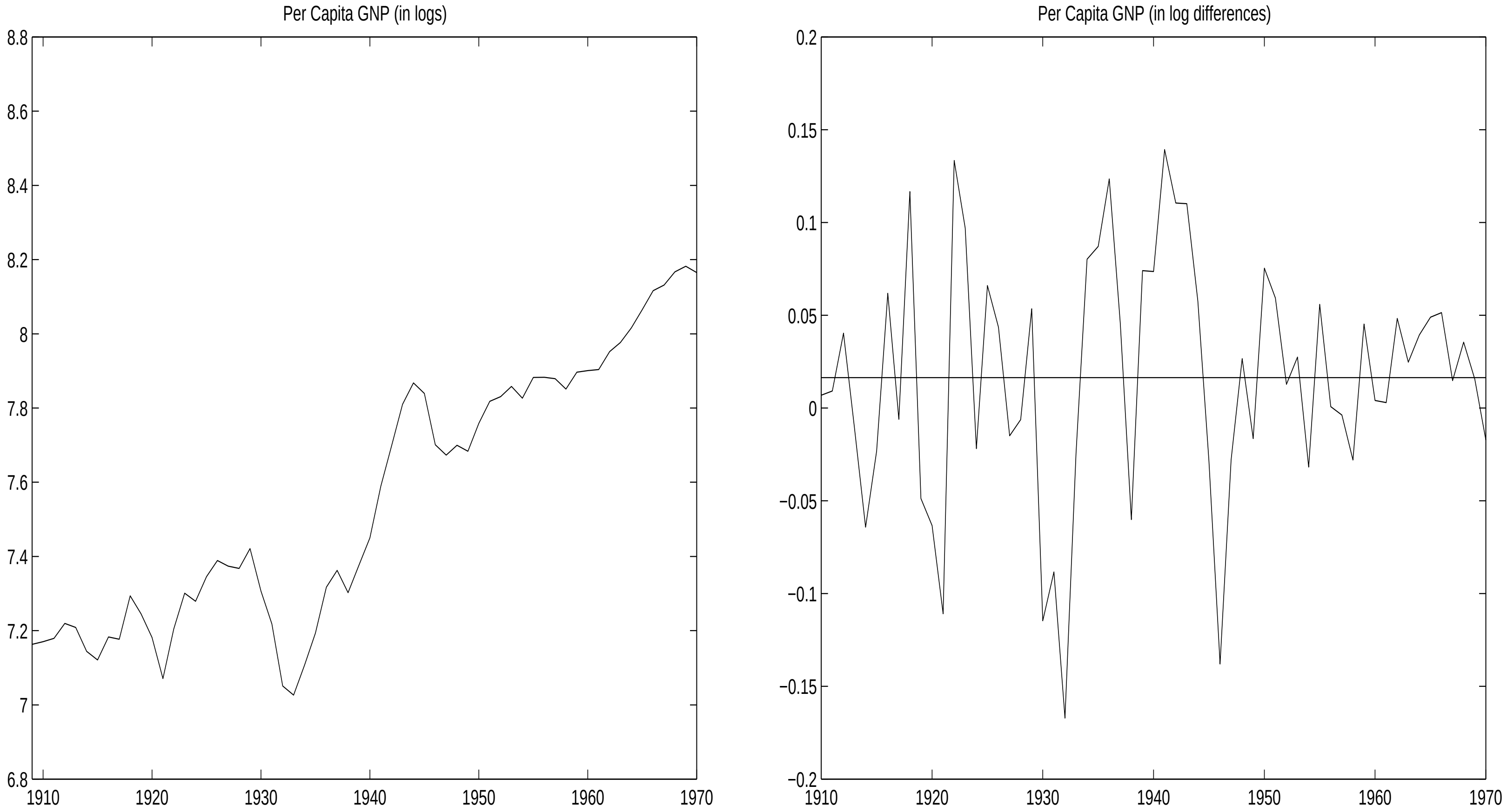

Figure 2: Transitory Components Measured as Percent Deviations from Drift in Nelson-Plosser Per Capita GNP Series (1909–1970)

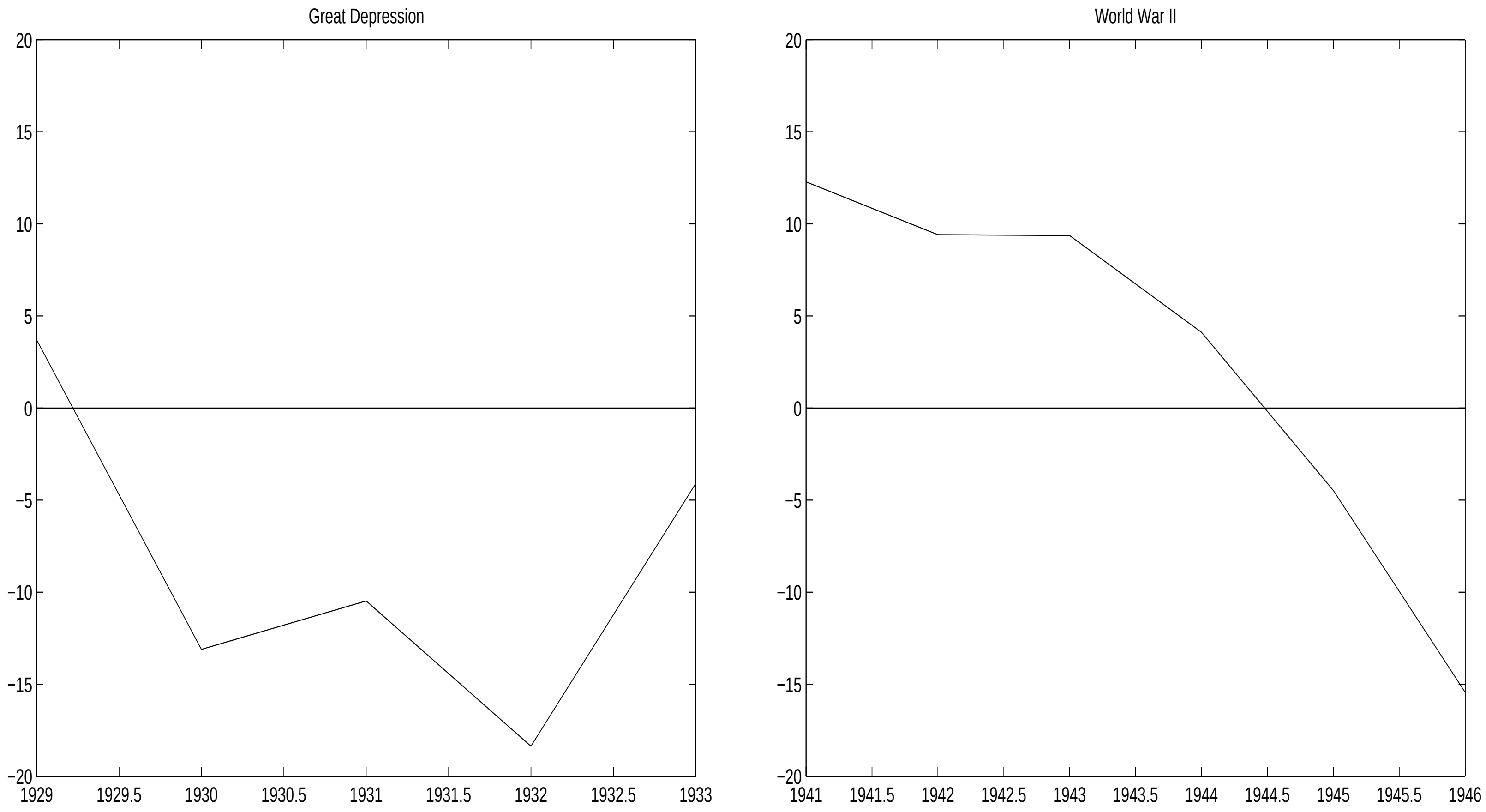

Table 1: Rejection Rates of Asymptotic Unit Root Tests for T = 62 (in percent)

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
No-Break Model	1	15.0	23.0	31.6	40.1
	2	3.0	5.2	9.1	17.4

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
Model A	1	23.0	34.9	42.8	53.7
	2	5.9	10.9	16.1	23.4

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
Model B	1	25.7	35.0	46.7	59.1
	2	8.1	14.3	22.2	34.9

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
Model C	1	27.5	38.6	48.1	60.0
	2	7.0	12.7	18.0	28.9

Source: Based on 1,000 Monte Carlo trials.

Table 2: Rejection Rates of Asymptotic Unit Root Tests for T = 62 (in percent)
Sensitivity Analysis

	Multiples of Transitory Effects	Nominal Size			
		1.0	2.5	5.0	10.0
Model A	2	41.8	53.7	61.9	69.0
	1*	23.0	34.9	42.8	53.7
	1/2	10.0	17.0	24.6	32.6
	0**	5.9	10.9	16.1	23.4
Model B	2	39.5	49.3	59.5	70.3
	1*	25.7	35.0	46.7	59.1
	½	14.7	22.9	31.4	45.8
	0**	8.1	14.3	22.2	34.9
Model C	2	48.1	57.8	66.2	76.0
	1*	27.5	38.6	48.1	60.0
	½	13.5	19.2	27.2	39.4
	0**	7.0	12.7	18.0	28.9

Source: Based on 1,000 Monte Carlo trials.

* DGP 1

** DGP 2

Table 3: Rejection Rates of Asymptotic Unit Root Tests for T = 62 (in percent) After Pretesting

		Nominal Size			
	DGP	**1.0**	**2.5**	**5.0**	**10.0**
Model A	**1**	**11.3**	**18.8**	**24.5**	**34.1**
	3	**11.5**	**20.1**	**26.8**	**37.1**
	4	**13.2**	**22.1**	**31.3**	**43.3**
	5	**16.2**	**24.0**	**30.5**	**39.4**
	2	**2.7**	**5.2**	**8.6**	**14.6**

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	**1.0**	**2.5**	**5.0**	**10.0**
Model B	**1**	**18.1**	**25.3**	**35.5**	**45.2**
	3	**14.3**	**21.6**	**31.5**	**46.1**
	4	**16.6**	**24.8**	**36.3**	**49.2**
	5	**23.9**	**32.6**	**41.5**	**54.4**
	2	**6.8**	**12.4**	**19.8**	**31.2**

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	**1.0**	**2.5**	**5.0**	**10.0**
Model C	**1**	**18.5**	**26.7**	**34.4**	**46.2**
	3	**24.9**	**36.6**	**46.6**	**61.8**
	4	**22.8**	**33.5**	**42.9**	**56.7**
	5	**27.8**	**37.3**	**46.3**	**56.8**
	2	**4.7**	**9.1**	**13.3**	**23.1**

Source: Based on 1,000 Monte Carlo trials.

Table 4: Rejection Rates of Asymptotic Unit Root Tests for T = 124 (in percent) After Pretesting

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
Model A	1	16.0	23.4	28.1	34.2
	3	9.0	16.0	20.3	31.5
	4	7.2	13.3	18.9	28.4
	5	9.6	15.1	20.0	26.4
	2	0.9	3.0	6.0	10.6

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
Model B	1	22.0	26.7	33.9	41.0
	3	12.1	18.3	25.8	35.9
	4	12.9	20.3	27.9	37.9
	5	16.0	22.5	27.7	37.0
	2	3.3	6.7	10.7	19.7

Source: Based on 1,000 Monte Carlo trials.

		Nominal Size			
	DGP	1.0	2.5	5.0	10.0
Model C	1	26.6	34.2	40.2	47.4
	3	12.5	19.3	26.7	38.1
	4	14.0	21.5	29.7	40.5
	5	17.4	25.1	31.4	40.0
	2	3.1	5.7	9.5	15.6

Source: Based on 1,000 Monte Carlo trials.

Table 5: Rejection Rates of Bootstrapped Unit Root Tests for T = 62 (in percent) After Pretesting

	10 Percent Nominal Size		
DGP	**Model A**	**Model B**	**Model C**
4	**13.0**	**10.4**	**22.6**
2	**5.4**	**6.8**	**9.0**

Source: Based on 500 Monte Carlo trials with 1,000 bootstrapped replications each.

`