# Data Transformation and Forecasting in Models with Unit Roots and Cointegration*

John C. Chao

*University of Maryland*

Valentina Corradi

*University of Exeter*

and

Norman R. Swanson

*Department of Economics, Texas A&M University*
*College Station, TX 77843-4228, U.S.A*
E-mail: nswanson@econ.tamu.edu

We perform a series of Monte Carlo experiments in order to evaluate the impact of data transformation on forecasting models, and find that vector error-corrections dominate differenced data vector autoregressions when the correct data transformation is used, but not when data are incorrectly tansformed, even if the true model contains cointegrating restrictions. We argue that one reason for this is the failure of standard unit root and cointegration tests under incorrect data transformation. © 2001 Peking University Press

*Key Words*: Integratedness; Cointegratedness; Nonlinear transformation.
*JEL Classification Numbers*: C22, C51.

* The authors wish to thank Frank Diebold and Clive W.J. Granger for useful suggestions which led to the writing of this paper. In addition, they are grateful to Claudia Olivetti for her efforts in carrying out many of the simulations reported here. Swanson thanks the Private Enterprise Research Center and the Bush Program in the Economics of Public Policy, both at Texas A&M University, for research assistance. Part of this paper was written while the third author was visiting the economics department of the University of California, San Diego, and he is grateful to members of the department for providing a stimulating research environment from which to carry out his research.

1529-7373/2001

## 1. INTRODUCTION

The purpose of this paper is to raise the issue of data transformation and its potential implications for the specification of forecasting models, and for forecasts from such models. This is done by first discussing unit root tests and cointegration, and then examining the impact of alternative assumptions placed on data generating processes before testing for unit roots and cointegration on subsequent predictions. Although we raise a number of issues, and in some cases suggest at least partial solutions, this paper is primarily meant to serve as a vehicle for underscoring the importance of the often ignored issue of data transformation on empirical model building. Thus, many issues are left unresolved.

In macroeconometrics, unit root tests are typically performed using logs. This is consistent with much of the real business cycle literature (see e.g. Long and Plosser (1983) and King, Plosser, Stock, and Watson (1991)) where it is suggested, for example, that GDP should be modeled in logs, given an assumption that output is generated according to a Cobb-Douglas production function. While this is sensible from a theoretical macroeconomic perspective, there is no clear empirical reason why logs should be used rather than levels, when performing unit root tests, particularly given that standard unit root tests assume linearity under both the null and the alternative, and violation of this linearity assumption can result in severe size and power distortion, both in finite and large samples (e.g. see Granger and Hallman (1991)). In addition, it is not always obvious by simply inspecting the data, for example, which transformation is 'appropriate', when modeling economic data (e.g. see Figure 1). Thus, it is reasonable to carefully address the problem of data transformation before running a unit root tests, for example. In a recent paper which is not discussed in detail here, Corradi and Swanson (2000) propose a framework for hypothesis testing in the presence of nonlinearity and nonstationarity. As a detailed illustration, they consider the problem of choosing between logs and levels before carrying out unit root and/or cointegration tests. An important feature of their test is that it is not subject to the difficulties discussed below when choosing between logs and levels using (possibly) integrated series.

The current convention is to define an integrated process of order $d$ (I$(d)$) as one which has the property that the partial sum of the $d$th difference, scaled by $T^{-1/2}$, satisfies a functional central limit theorem (FCLT). In this case, integratedness in logs does not imply integratedness in levels, and $vice-versa$. Thus, any *a priori* assumption concerning whether to model data in levels or logs has important implications for the outcome of unit root and related tests. For example, Granger and Hallman (1991) show that the percentiles of the empirical distribution of the Dickey-Fuller (1979) statistic constructed using $exp(X_t)$ are much higher, in absolute value, than the corresponding percentiles constructed using the original time series $X_t$, when $X_t$ is a random walk process. Thus, inference based on the Dickey-Fuller statistic using the exponential transformation leads to an overrejection of the unit root null hypothesis, when standard critical values are used. More recently it has been shown in Corradi (1995) that if $X_t$ is a random walk, then any convex transformation (such as exponentiation) is a submartingale, and any concave transformation (such as taking logs) is a supermartingale. However, while submartingales and supermartingales have a unit root component, their first differences do not generally satisfy typical FCLTs. Thus, Dickey-Fuller type tests no longer have well defined limiting distributions. Given all of the above considerations, it is of some interest to use a statistical procedure for selecting between linear and loglinear specifications, rather than simply assuming from the outset that a series is best modeled as linear or loglinear. Further, while Cox-type tests are available for the I(0) case, few results are available for the I(1) case.

The arguments used above carry over to the case of cointegration tests, and indeed to any statistical tests based on the use of partial sums of functionals of residuals, for example. As the use of cointegration tests is prevalent, however, we focus our discussion on them in this paper.

One of the areas where unit root and cointegration tests are crucial is in the construction of vector error correction (VEC) forecasting models. In order to illustrate this point, we simulate a real-time forecasting environment, where data are generated using cointegrated variables, and where models are estimated using data which are correctly or incorrectly transformed. Our primary focus is on the choice between log and level data, and we find that incorrect data transformation leads to poor forecasts from cointegrated models, relative to simpler models based on differenced data, even when the true data generating process exhibits cointegration. This may be due to imprecise estimation of cointegrating spaces when the correct data transformation is uncertain, for example, and may help to explain the mixed evidence concerning the usefulness of cointegration restrictions in forecasting (see e.g. the special issue of the Journal of Applied Econometrics (1996) on forecasting). The finding is based on an evaluation of VEC models and vector autoregressive (VAR) models using differenced and undifferenced data. Three additional findings based on our analytsis are that: (1) VEC models forecast-dominate differenced data VAR models when the correct data transformation is used. (2) The worst models based on correctly transformed data clearly dominate the best models based on incorrectly transformed data. (3) When the incorrect data transformation is used to construct forecasting models, differenced data VAR models outperform not only their VEC counterparts, but also VAR models based on undifferenced data. In order to shed further light on the issue of data transformation in VEC models, we examine the finite sample performance of cointegration tests under incorrect data transformation.

The rest of the paper is organized as follows. Section 2 discusses unit root and cointegration testing under data transformation, and Section 3 contains a discussion of forecasting models under data transformation as well as the results of our forecasting experiments. Concluding remarks are given in Section 4.

## 2. UNIT ROOT AND COINTEGRATION TESTS

Given a series of observations on an underlying strictly positive process $X_t$, $t = 1, 2, \ldots$, our objective is to decide whether: (1) $X_t$ is an I(0) process (possibly around a linear deterministic trend), (2) $\log X_t$ is an I(0) process around a nonzero linear deterministic trend, (3) $X_t$ is an I(1) process (around a positive linear deterministic trend), and (4) $\log X_t$ is an I(1) process, (possibly around a linear deterministic trend). A natural approach to this problem is to construct a test that has a well defined limiting distribution under a particular DGP, and diverges to infinity under all of the other above DGPs.

While it is easy to define a test having a well defined distribution under one of (1)-(4), it not clear how to ensure that the test has power against all of the remaining DGPs. To illustrate the problem, consider the sequence $\hat{\epsilon}_t$, given as the residuals from a regression of $X_t$ on a constant and a time trend. Now, construct the test statistic proposed by Kwiatkowski, Phillips, Schmidt, and Shin (1992, hereafter KPSS):

$$S_T = \frac{1}{\hat{\sigma}_T^2} T^{-2} \sum_{t=1}^{T} \left( \sum_{j=1}^{t} \hat{\epsilon}_t^2 \right)^2 ,$$

where $\hat{\sigma}_T^2$ is a heteroskedasticity and autocorrelation (HAC) robust estimator of $var\left(T^{-1/2} \sum_{j=1}^{t} \epsilon_t\right)$. It is known from KPSS that if $X_t$ is I(0) (possibly around a linear deterministic trend), then $S_T$ has a well defined limiting distribution under the null hypothesis, while $S_T$ diverges at rate $T/l_T$ under the alternative that $X_t$ is an integrated process, where $l_T$ is the lag truncation parameter used in the estimation of the variance term in $S_T$. However, if the underlying DGP is $\log X_t = \alpha_1 + \delta_1 t + \sum_{j=1}^{t} \epsilon_j$, $\delta_1 > 0$ (i.e. $log X_t$ is a unit root process) then both $\hat{\sigma}_T^2$ and $T^{-2} \sum_{t=1}^{T} \left( \sum_{j=1}^{t} \hat{\epsilon}_j \right)^2$ will tend to diverge at a geometric rate, given that $X_t = \exp(\alpha_1 + \delta_1 t + \sum_{j=1}^{t} \epsilon_j)$. In this case it is not clear whether the numerator or the denominator is exploding at a faster rate. This problem is typical of all tests which are based on functionals of partial sums and variance estimators, and arises because certain *nonlinear* alternatives are not treatable using standard FCLTs.

So far we have analyzed the case in which we perform a test with $I(0)$ as the null hypothesis and $I(1)$ as the alternative. In this case, the statistic is typically constructed in terms of functionals of partial sums scaled by a variance estimator. Another common procedure is to test for the null of $I(1)$ versus the alternative of $I(0)$ using Dickey-Fuller type tests. To illustrate the problems associated with this approach, consider the following simple example. Assume that $\log X_t = \log X_{t-1} + \epsilon_t$, $\epsilon_t \sim iid(0, \sigma_\epsilon^2)$. However, we perform a Dickey-Fuller test using levels. For example, we compute $T(\hat{\alpha}_T - 1)$, where

$$\hat{\alpha}_T = \frac{\sum_{t=2}^T X_t X_{t-1}}{\sum_{t=2}^T X_{t-1}^2}.$$

Now, $X_t = \exp(\log X_{t-1} + \epsilon_t) = X_{t-1}\exp(\epsilon_t)$, so that we can write:

$$T(\hat{\alpha}_T - 1) = \frac{T\sum_{t=2}^T X_{t-1}^2 (e^{\epsilon_t} - 1)}{\sum_{t=2}^T X_{t-1}^2}.$$

Note that as $X_t = X_0 \exp(\sum_{j=1}^t \epsilon_j)$, standard unit root asymptotics no longer apply. However, by confining our attention to the case where $\epsilon \sim N(0, \sigma_\epsilon^2)$, we can examine the properties of $T(\hat{\alpha}_T - 1)$, thus gaining insight into the performance of a Dickey-Fuller test using an incorrect transformation of the data. Notice that $Ee^{\epsilon_t} = e^{\frac{1}{2}\sigma_\epsilon^2} > 1$. Thus, we might expect that $T(\hat{\alpha}_T - 1)$ tends to diverge to $+\infty$. However, Granger and Hallman (1991) find that this statistic tends to overreject the null of a unit root. One possible explanation for the difference between their finding and our intuition is that the distribution of $e^{\epsilon_t} - 1$ is highly skewed to the left, and has a lower bound of negative one. Thus, even though the mean of $e^{\epsilon_t} - 1$ is positive, this is due to the very long right-tail of the distribution. When $\epsilon_t$ is drawn from a standard normal distribution, however, most observations are rather close to zero (e.g. 95% are between 2 and -2). These data, when transformed using $e^{\epsilon_t} - 1$, are mainly between -0.86 and 6.4. Further, the median of the distribution of $e^{\epsilon_t} - 1$ is zero. Now, in the context of finite samples, this suggests that if we truncate the distribution of $e^{\epsilon_t} - 1$ to be, say, between -0.8 and 1, then the mean of this truncated distribution will actually be negative (as we draw relatively fewer observations close to the upper bound than negative observations close to the lower bound). In the context of generating data in finite samples, as Granger and Hallman did, this situation indeed seems to have occurred, resulting in mostly large negative values being calculated for the expression $T(\hat{\alpha}_T - 1)$. Put another way, the negative elements of $T\sum_{t=1}^T X_{t-1}^2 (e^{\epsilon_t} - 1)$ are usually quite large in magnitude, relative to most of the positive elements of the same sum. Of course, in large samples, and with large $\sigma_\epsilon^2$ we should expect that this result will not hold, as the effect of large positive draws from the distribution of $e^{\epsilon_t} - 1$ begins to dominate the overall sum $T\sum_{t=1}^{T} X_{t-1}^2(e^{\epsilon_t} - 1)$. This intuition suggests that Granger and Hallman's results, while holding for the usual sample sizes and the usual error variances observed in economic time series, should not hold generally. It further suggests that indeed using levels data when the true process is I(1) in logs will produce either overrejection of the unit root null (as Hallman and Granger show), or underrejection of the null. Interestingly, these arguments also suggest that for very special cases (i.e. appropriately chosen $\sigma_\epsilon^2$ and sample size), the empirical size of the Dickey-Fuller test may actually match the nominal size, even when the wrong data transformation is used! In summary, there appears to be a need to carefully consider which transformation is used when constructing unit root tests, as the wrong transformation may yield entirely misleading results.

Even if we decide to keep integratedness as a maintained assumption, and choose between I(1) in levels and I(1) in logs, or *vice versa*, we do not in general obtain a test which has unit asymptotic power. For example consider constructing a KPSS-type test using the first differences of the levels data (i.e. $\Delta X_t$). Under the null of I(1) in levels the statistic has the usual well defined limiting distribution. However, under the alternative of I(1) in logs it does not necessarily diverge to infinity. Again the reason for this result is that both the numerator and the denominator tend to diverge to infinity if they have a positive linear deterministic trend, and in general we cannot determine whether the numerator or the denominator is diverging at a faster rate.

Given the above issues, it may be useful to examine the finite sample performance of Johansen CI tests under incorrect data transformation. We turn next to this issue. Our approach is to examine the finite sample behavior of the Johansen (1988,1991) cointegration test using data generated according to the above parameterizations.

Table 1 reports the finite sample size and power of the Johansen trace test when applied to incorrectly transformed data. Data are generated according to the following VEC model:

$$\Delta Q_{1,t} = a + b(L)\Delta Q_{1,t-1} + cZ_{t-1} + \epsilon_t, \tag{1}$$

where $Q_{1,t} = (X_t, W_t')'$ is a vector if I(1) variables, $W_t$ a $n \times 1$ vector for some $n \geq 1$, $Z_{t-1} = dQ_{1,t-1}$ is a $r \times 1$ vector of I(0) variables, r is the rank of the cointegrating space, d is an $r \times (n+1)$ matrix of cointegrating vectors, $a$ is an $(n+1) \times 1$ vector, $b(L)$ is a matrix polynomial in the lag operator $L$, with $p$ terms, each of which is an $(n+1) \times (n+1)$ matrix, $p$ is the order of the VEC model, $c$ is an $(n+1) \times r$ matrix, and $\epsilon_t$ is a vector error term. For DGPs generated as linear in levels, we report rejection frequencies for $a = (a_1, a_2)\prime$, $a_1 = a_2 = \{0.0, 0.1, 0.2\}$, $b = 0$, $c = (c_1, c_2)\prime$, $c_1 = -0.2, c_2 = \{0.0, 0.2, 0.4, 0.6\}$, and $\sigma^2_{\epsilon_i} = 1.0$, $i = 1, 2$. For loglinear DGPs, $b$ and $c$ are as above, $a_1 = a_2 = \{0.0, 0.01, 0.02\}$, and $\sigma^2_{\epsilon_i} = 0.09$, $i = 1, 2$. Results for other parameterizations examined are qualitatively similar, and are available upon request from the authors. The results of the experiment are quite straightforward. First, the empirical size of the trace test statistic is severely upward biased, with bias increasing as $T$ increases. Further, and as expected, the finite sample power (all cases where $c_2 \neq 0$) increases rapidly to unity as $T$ increases. Thus, the null of no cointegration is over-rejected. Also, we know that estimators of cointegrating vectors are inconsistent under the wrong data transformation, even if the *true* cointegrating rank is known. Thus, it is perhaps not surprising that VEC models more clearly dominate VAR models (in differences) when the appropriate data transformation is used.

## 3. FORECASTING USING VECTOR ERROR CORRECTION MODELS

### 3.1. Discussion

A number of well known issues arise in the context of the specification and estimation of forecasting models which have obvious implications for the application of the above procedures. In particular: (1) As noted above, the gains to forecasting associated with the use of VEC models rather than simpler VAR models based on differenced data is not clear. (2) It is not clear whether models based on undifferenced data are dominated by VAR and VEC models based on differenced data, even when variables are I(1). (3) The choice of loss function, $f$, is not always obvious, and certainly this choice depends on the particular objective of the forecaster (see e.g. Christoffersen and Diebold (1996,1998), Pesaran and Timmerman (1994), Swanson and White (1997), and the references contained therein). (4) In a generic forecasting scenario it is not always obvious whether data should be logged or not, and it is not obvious how to compare forecasts of a variable arising from log and level versions of some generic model[1] (see e.g. Ermini and Hendry (1995)). For example, assume that one is interested in forecasting $Y_t$. In this context, there are a number of choices. First, we

[1] Note that here and below we assume that only linear forecasting models are being examined. If this were not the case, then this last issue would not be relevant, as any linear model estimated using levels data could obviously be transformed into some nonlinear model using logged data, and as long as heteroskedasticity etc. were appropriately modelled, there might be little to choose between the models, at least within the context of forecasting (see e.g. Granger and Swanson (1996).

**TABLE 1.**

Johansen Cointegration Test Performance under the Wrong Data Transformation [(*)]

| | | | T=100 | | T=250 | | T=500 | |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $c_1$ | $c_2$ | Trace1 | Trace2 | Trace1 | Trace2 | Trace1 | Trace2 |
| Panel A: DGP in logs | | | | | | | | |
| 0.0 | 0.0 | 0.0 | 0.339 | 0.426 | 0.596 | 0.733 | 0.94 | 0.95 |
| 0.01 | 0.0 | 0.0 | 0.298 | 0.383 | 0.534 | 0.671 | 0.914 | 0.924 |
| 0.02 | 0.0 | 0.0 | 0.310 | 0.376 | 0.560 | 0.644 | 0.914 | 0.906 |
| 0.00 | -0.2 | 0.2 | 0.978 | 0.982 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.00 | -0.2 | 0.4 | 0.998 | 0.996 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.00 | -0.2 | 0.6 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.01 | -0.2 | 0.2 | 0.973 | 0.965 | 1.000 | 0.999 | 1.000 | 1.000 |
| 0.01 | -0.2 | 0.4 | 0.997 | 0.994 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.01 | -0.2 | 0.6 | 1.000 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.02 | -0.2 | 0.2 | 0.967 | 0.956 | 0.999 | 0.999 | 1.000 | 1.000 |
| 0.02 | -0.2 | 0.4 | 0.997 | 0.990 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.02 | -0.2 | 0.6 | 1.000 | 0.997 | 1.000 | 1.000 | 1.000 | 1.000 |
| Panel B: DGP in levels | | | | | | | | |
| 0.1 | 0.0 | 0.0 | 0.152 | 0.163 | 0.354 | 0.235 | 0.657 | 0.395 |
| 0.2 | 0.0 | 0.0 | 0.498 | 0.208 | 0.922 | 0.434 | 0.999 | 0.763 |
| 0.1 | -0.2 | 0.2 | 0.989 | 0.965 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.1 | -0.2 | 0.4 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.1 | -0.2 | 0.6 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.2 | -0.2 | 0.2 | 1.000 | 0.965 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.2 | -0.2 | 0.4 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.2 | -0.2 | 0.6 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

[(*)] Notes: Entries are Johansen trace test statistic rejection frequencies (Trace1 uses intercept and Trace2 uses intercept and trend in test regressions), Panel A presents results when the DGP is in logs, and data are exponentiated; while in Panel B the DGP is in levels, and data are logged. For each panel, entries in the remaining rows report the empirical power (all based on 5% nominal level tests) when the wrong data transformation is used. 5000 Monte Carlo simulations were run (see above for further details).

must decide whether we want to forecast $Y_t$, $\Delta Y_t$, $\log Y_t$, or $\Delta \log Y_t$. Given this decision, we must decide how to compare the models. For example, if using differenced data, we may transform forecasts of $\Delta \log Y_t$ into forecasts of $\Delta Y_t$, or $vice - versa$, when comparing models.

### 3.2. Monte Carlo Results

In this section, we examine all of these issues by conducting a series of Monte Carlo experiments. We begin by assuming that we are interested in constructing forecasts using data which are generated according to the following VEC model:

$$\Delta Q_{1,t} = a + b(L)\Delta Q_{1,t-1} + cZ_{t-1} + \epsilon_t, \tag{2}$$

where $Q_{1,t} = (X_t, W_t')'$ is a vector if I(1) variables, $W_t$ a $n \times 1$ vector for some $n \geq 1$, $Z_{t-1} = dQ_{1,t-1}$ is a $r \times 1$ vector of I(0) variables, r is the rank of the cointegrating space, d is an $r \times (n+1)$ matrix of cointegrating vectors, $a$ is an $(n+1) \times 1$ vector, $b(L)$ is a matrix polynomial in the lag operator $L$, with $p$ terms, each of which is an $(n+1) \times (n+1)$ matrix, $p$ is the order of the VEC model, $c$ is an $(n+1) \times r$ matrix, and $\epsilon_t$ is a vector error term. In our experiments we let the integratedness of the series be unknown, the rank of the cointegrating space be unknown, $p$ be unknown, and we assume no prior knowledge concerning whether to log the data or not. Also, we set $n = 1$, and the order of the matrix lag polynomial equal to 0 or 1 (hence, $b(L) = b$, say, for simplicity). In all cases, we construct a sequence of $P$ 1-step ahead forecasts of $X_t$, and construct *average* mean square forecast error ($AMSE$), average mean absolute percentage forecast error ($AMAPE$), and average mean absolute deviation forecast error ($AMAD$) criteria, where the *average* is based on 1000 replications. Also, let $P = 50$ and $P = T/2$, where $T$ is the sample size. Lags are selected using the BIC criterion. All parameters (including the cointegrating rank) are re-estimated before each new forecast is formed, using an increasing window of observations, starting with $T - P$ observations, and ending with $T - 1$ observations, so that sequences of $P$ $ex - ante$ 1-step ahead forecasts are constructed for each replication. Data are generated according to the following parameterizations:

1. Data generated as loglinear: Samples are $T = 100, 250$, and 500 observations. $a = (a_1, a_2)\prime, a_1 = a_2 = \{0.0, 0.001, 0.002\}$; $b = (b_1, b_2), b_1 = (b_{11}, b_{21})\prime$, $b_2 = (b_{12}, b_{22})\prime$, and either $b_{12} = b_{21} = b_{11} = b_{22} = 0$, or $b_{12} = b_{21} = 0, b_{11} = -0.4, b_{22} = 0.2$; and $c = (c_1, c_2)\prime$, $c_1 = -0.2, c_2 = \{0.2, 0.4, 0.6\}$, $d = (1, -1)\prime$, $\epsilon_{i,t} \sim IN(0, \sigma^2_{\epsilon_i})$, $\sigma^2_{\epsilon_i} = 0.09$, $i = 1, 2$, and $E(\epsilon_{1,t}\epsilon_{2,t}) = 0$ for any $t$.
2. Data generated as linear in levels: The same parameterizations as above are used, except that $a_1 = a_2 = \{0.0, 0.1, 0.2\}$ and $\sigma^2_{\epsilon_i} = 1.0$, $i = 1, 2$.

In order to summarize our results, we group our Monte Carlo exercises into four experiments:

Experiment I: All simulations are based on data generated as loglinear (actual data are referred to as $\log X_t$). Model selection criteria (i.e. AMSE, AMAPE, and AMAD) are constructed using forecasts of $\log X_t$. We always estimate two types of models, one using logged data, and the other using levels data. For models estimated using logged data, we immediately have available the appropriate forecast, say $\widehat{\log X_t}$. However, for models estimated using levels data, we only have available a levels forecast, say $\hat{X}_t$, and we construct $\log \hat{X}_t$, in order to compare the model selection criteria across data transformations.

Experiment II: All simulations are based on data generated as loglinear (actual data are referred to as $\log X_t$). In this case, we construct model selection criteria using forecasts of $X_t$. For models estimated using logged data, we construct $\exp\left(\widehat{\log X_t}\right)$. Note here that we do not make the usual bias adjustment, as this bias adjustment is based on the presumption of normality, presumption which does not generally hold in economic data. For models estimated using levels data, we immediately have available the appropriate forecast, say $\hat{X}_t$.

Experiment III: All simulations are based on data generated as linear in levels (actual data are referred to as $X_t$). In this case, we construct model selection criteria using forecasts of $X_t$. All forecasts are constructed as in Experiment II.

Experiment IV: All simulations are based on data generated as linear in levels (actual data are referred to as $X_t$). In this case, we construct model selection criteria using forecasts of $\log X_t$. All forecasts are constructed as in Experiment I.

Based on the above framework, we compiled 48 tables of results, corresponding to Experiments I-IV, T= {100, 250, 500}, p = {1, 2} in the VAR(p) DGP, and P = {50, T/2}. Because the results are qualitatively similar, and for the sake of brevity, we present only four tables, corresponding to Experiments I-IV, T=100, p=1, and P=50. Complete results are available from the authors. Tables 2-5 summarize our findings, and our conclusions are grouped into answers to (1)-(4) above.

(1) Although the numerical differences are not great, the VEC model in differences always has a lower AMSE than the VAR model in differences when the correct data transformation is used to estimate the models and to compare the forecasts. This can be seen by comparing the first and third columns of entries in Tables 1 and 3. Furthermore, note that when the DGP is loglinear, and models are estimated using the correct data transformation, but forecasts are then transformed so that levels forecasts are compared when data are generated and estimated in logs (and vice-

**TABLE 2.**

Experiment I - DGP in logs, compare log forecasts of $Y_t$ $^{(*)}$

| | | | VEC in differences | | VAR in differences | | VAR in levels | |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $c_2$ | criterion | $(\widehat{\log X_t})$ | $\log \hat{X}_t$ | $(\widehat{\log X_t})$ | $\log \hat{X}_t$ | $(\widehat{\log X_t})$ | $\log \hat{X}_t$ |
| 0 | 0.2 | AMSE | .1008 | .2171 | .1034 | .2001 | .0953 | .1674 |
| | | AMAPE | 259.0 | 215.3 | 300.9 | 250.9 | 365.5 | 302.2 |
| | | AMAD | .2529 | .3166 | .2560 | .3089 | .2459 | .3002 |
| 0 | 0.4 | AMSE | .0978 | .2453 | .1002 | .2301 | .0950 | .1654 |
| | | AMAPE | 313.1 | 532.1 | 331.0 | 273.2 | 298.9 | 1176 |
| | | AMAD | .2492 | .3250 | .2521 | .3180 | .2455 | .3014 |
| 0 | 0.6 | AMSE | .0971 | .2722 | .0987 | .2515 | .0949 | .1716 |
| | | AMAPE | 241.2 | 237.7 | 214.9 | 311.4 | 303.9 | 277.2 |
| | | AMAD | .2482 | .3366 | .2501 | .3290 | .2454 | .3098 |
| 0.001 | 0.2 | AMSE | .1008 | .2103 | .1034 | .1944 | .1001 | .2823 |
| | | AMAPE | 313.3 | 235.4 | 273.3 | 247.8 | 261.2 | 278.8 |
| | | AMAD | .2529 | .3140 | .2560 | .3064 | .2521 | .3483 |
| 0.001 | 0.4 | AMSE | .0978 | .2384 | .1002 | .2254 | .0995 | .3177 |
| | | AMAPE | 253.0 | 290.8 | 263.5 | 257.4 | 306.0 | 260.6 |
| | | AMAD | .2492 | .3212 | .2521 | .3151 | .2513 | .3609 |
| 0.001 | 0.6 | AMSE | .0971 | .2653 | .0987 | .2424 | .0992 | .3609 |
| | | AMAPE | 216.3 | 210.4 | 212.5 | 238.7 | 242.7 | 282.1 |
| | | AMAD | .2482 | .3333 | .2501 | .3251 | .2510 | .3773 |
| 0.002 | 0.2 | AMSE | .1008 | .2058 | .1034 | .1865 | .1001 | .2751 |
| | | AMAPE | 607.2 | 472.2 | 618.9 | 455.8 | 341.8 | 352.3 |
| | | AMAD | .2530 | .3111 | .2560 | .3034 | .2521 | .3444 |
| 0.002 | 0.4 | AMSE | .0978 | .2249 | .1002 | .2166 | .0995 | .3067 |
| | | AMAPE | 236.0 | 315.3 | 239.6 | 258.0 | 396.0 | 274.3 |
| | | AMAD | .2491 | .3179 | .2521 | .3123 | .2513 | .3567 |
| 0.002 | 0.6 | AMSE | .0971 | .2552 | .0987 | .2320 | .0992 | .3431 |
| | | AMAPE | 232.2 | 300.1 | 247.7 | 314.4 | 224.9 | 1170 |
| | | AMAD | .2481 | .3294 | .2501 | .3211 | .2510 | .3716 |

$^{(*)}$Notes: Entries are averages of functions of forecast errors. The first, third, and fifth columns of entries correspond to models which are estimated using the "correct" data transformation, while for the second, fourth, and sixth columns, the "incorrect" data transformation is used. 1000 Monte Carlo replications were run (see above for further discussion).

**TABLE 3.**

Experiment II - DGP in logs, compare levels forecasts of $Y_t$ $^{(*)}$

| | | | VEC in differences | | VAR in differences | | VAR in levels | |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $c_2$ | criterion | $\exp(\widehat{\log X_t})$ | $\hat{X}_t$ | $\exp(\widehat{\log X_t})$ | $\hat{X}_t$ | $\exp(\widehat{\log X_t})$ | $\hat{X}_t$ |
| | | AMSE | .5940 | .8948 | .6224 | .8804 | .5693 | .9308 |
| 0 | 0.2 | AMAPE | 26.26 | 70.67 | 26.66 | 55.76 | 25.49 | 34.18 |
| | | AMAD | .3878 | .4357 | .3952 | .4339 | .3748 | .4336 |
| | | AMSE | .6962 | 1.5066 | .7306 | 1.5044 | .6872 | 1.5978 |
| 0 | 0.4 | AMAPE | 25.91 | 62.59 | 26.20 | 62.31 | 25.46 | 36.21 |
| | | AMAD | .3941 | .4529 | .3998 | .4509 | .3855 | .4565 |
| | | AMSE | .9296 | 2.7487 | .9550 | 2.7122 | .9200 | 3.9232 |
| 0 | 0.6 | AMAPE | 25.78 | 142.55 | 25.96 | 103.2 | 25.45 | 38.86 |
| | | AMAD | .4134 | .4872 | .4167 | .4830 | .4055 | .4980 |
| | | AMSE | .6355 | .9674 | .6662 | .9625 | .6254 | 1.0533 |
| 0.001 | 0.2 | AMAPE | 26.26 | 61.91 | 26.66 | 54.45 | 26.15 | 892.1 |
| | | AMAD | .3985 | .4481 | .4062 | .4467 | .3940 | .4653 |
| | | AMSE | .7470 | 1.6457 | .7842 | 1.6385 | .7529 | 2.2366 |
| 0.001 | 0.4 | AMAPE | 25.91 | 149.3 | 26.20 | 138.5 | 26.03 | 92.62 |
| | | AMAD | .4051 | .4664 | .4110 | .4642 | .4055 | .4914 |
| | | AMSE | 1.0005 | 3.0007 | 1.0279 | 2.9610 | 1.0011 | 5.5447 |
| 0.001 | 0.6 | AMAPE | 25.78 | 91.87 | 25.96 | 56.88 | 25.98 | 307.8 |
| | | AMAD | .4252 | .5016 | .4286 | .4973 | .4265 | .5363 |
| | | AMSE | .6801 | 1.0473 | .7136 | 1.0423 | .6698 | 1.1418 |
| 0.002 | 0.2 | AMAPE | 26.26 | 95.12 | 26.66 | 80.32 | 26.15 | 88.61 |
| | | AMAD | .4095 | .4612 | .4176 | .4599 | .4050 | .4780 |
| | | AMSE | .8022 | 1.7950 | .8423 | 1.7859 | .8086 | 2.4608 |
| 0.002 | 0.4 | AMAPE | 25.90 | 54.83 | 26.20 | 55.74 | 26.04 | 97.76 |
| | | AMAD | .4164 | .4802 | .4227 | .4779 | .4169 | .5050 |
| | | AMSE | 1.0774 | 3.2748 | 1.1070 | 3.2318 | 1.0780 | 6.1323 |
| 0.002 | 0.6 | AMAPE | 25.77 | 56.78 | 25.96 | 85.50 | 25.98 | 109.0 |
| | | AMAD | .4373 | .5164 | .4409 | .5121 | .4387 | .5520 |

$^{(*)}$Notes: See notes to Table 2.

**TABLE 4.**

Experiment III - DGP in levels, compare levels forecasts of $Y_t$ $^{(*)}$

| | | | VEC in differences | | VAR in differences | | VAR in levels | |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $c_2$ | criterion | $\hat{X}_t$ | $\exp(\widehat{\log X_t})$ | $\hat{X}_t$ | $\exp(\widehat{\log X_t})$ | $\widehat{X_t}$ | $\exp(\widehat{\log X_t})$ |
| | | AMSE | 1.1256 | 1.1257 | 1.1515 | 1.1514 | 1.0670 | 1.0670 |
| 0.0 | 0.2 | AMAPE | 0.8490 | 0.8490 | 0.8599 | 0.8599 | 0.8452 | 0.8280 |
| | | AMAD | 0.8450 | 0.8449 | 0.8550 | 0.8550 | 0.8240 | 0.8240 |
| | | AMSE | 1.0925 | 1.0931 | 1.1170 | 1.1175 | 1.0638 | 1.0638 |
| 0.0 | 0.4 | AMAPE | 0.8380 | 0.8389 | 0.8477 | 0.8478 | 0.8280 | 0.8284 |
| | | AMAD | 0.8336 | 0.8338 | 0.8425 | 0.8426 | 0.8234 | 0.8234 |
| | | AMSE | 1.0866 | 1.0868 | 1.1010 | 1.1016 | 1.0630 | 1.0630 |
| 0.0 | 0.6 | AMAPE | 0.8375 | 0.8376 | 0.8427 | 0.8428 | 0.8290 | 0.8290 |
| | | AMAD | 0.8317 | 0.8318 | 0.8367 | 0.8369 | 0.8233 | 0.8233 |
| | | AMSE | 1.1230 | 1.1250 | 1.1510 | 1.1520 | 1.1190 | 1.1240 |
| 0.1 | 0.2 | AMAPE | 0.7887 | 0.7890 | 0.7990 | 0.7997 | 0.7887 | 0.7900 |
| | | AMAD | 0.8440 | 0.8450 | 0.8555 | 0.8560 | 0.8440 | 0.8450 |
| | | AMSE | 1.0917 | 1.0947 | 1.1170 | 1.1190 | 1.1110 | 1.1150 |
| 0.1 | 0.4 | AMAPE | 0.7795 | 0.7800 | 0.7870 | 0.7880 | 0.7860 | 0.7880 |
| | | AMAD | 0.8334 | 0.8346 | 0.8425 | 0.8431 | 0.8400 | 0.8430 |
| | | AMSE | 1.0860 | 1.0880 | 1.1010 | 1.1020 | 1.1080 | 1.1130 |
| 0.1 | 0.6 | AMAPE | 0.7780 | 0.7790 | 0.7830 | 0.7840 | 0.7860 | 0.7870 |
| | | AMAD | 0.8310 | 0.8320 | 0.8370 | 0.8374 | 0.8395 | 0.8410 |
| | | AMSE | 1.125 | 1.129 | 1.151 | 1.155 | 1.119 | 1.129 |
| 0.2 | 0.2 | AMAPE | 0.7380 | 0.7398 | 0.7469 | 0.7478 | 0.7370 | 0.7400 |
| | | AMAD | 0.8450 | 0.8470 | 0.8550 | 0.8560 | 0.8440 | 0.8477 |
| | | AMSE | 1.0890 | 1.0940 | 1.1170 | 1.1200 | 1.1100 | 1.1210 |
| 0.2 | 0.4 | AMAPE | 0.7270 | 0.7287 | 0.7360 | 0.7370 | 0.7348 | 0.7380 |
| | | AMAD | 0.8320 | 0.8340 | 0.8420 | 0.8440 | 0.8400 | 0.8445 |
| | | AMSE | 1.0840 | 1.0890 | 1.1010 | 1.1040 | 1.1080 | 1.1190 |
| 0.2 | 0.6 | AMAPE | 0.7260 | 0.7270 | 0.7310 | 0.7325 | 0.7343 | 0.7375 |
| | | AMAD | 0.8300 | 0.8330 | 0.8370 | 0.8380 | 0.8395 | 0.8440 |

$^{(*)}$ Notes: See notes to Table 2. Data are generated according to the following process: $\Delta Q_{1,t} = a + b\Delta Q_{1,t-1} + cZ_{t-1} + \epsilon_t$, where $Q_{1,t} = (X_t, W_t)'$ is a 2x1 vector of I(1) variables, $\epsilon_t$ is a 2x1 vector whose components are distributed $IN(0,1)$, and $Z_t = X_t - W_t$. We set $a_1 = a_2 \in \{0, 0.1, 0.2\}$,the initial value $Q_{1,0} = 100$, $c_1 = -0.2$, $c_2 \in \{0.2, 0.4, 0.6\}$, and $b_{11} = b_{12} = b_{21} = b_{22} = 0$.

**TABLE 5.**

Experiment IV - DGP in levels, compare log forecasts of $Y_t$ $^{(*)}$

| | | | VEC in differences | | VAR in differences | | VAR in levels | |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $c_2$ | criterion | $\log(\widehat{X_t})$ | $\widehat{(\log X_t)}$ | $\log(\widehat{X_t})$ | $\widehat{(\log X_t)}$ | $\log(\widehat{X_t})$ | $\widehat{(\log X_t)}$ |
| | | AMSE | .000113 | .000113 | .000116 | .000116 | .000106 | .000106 |
| 0.0 | 0.2 | AMAPE | .183739 | .183751 | .186229 | .186203 | .178603 | .178630 |
| | | AMAD | .008455 | .008455 | .008569 | .008568 | .008219 | .008220 |
| | | AMSE | .000110 | .000110 | .000113 | .000113 | .000106 | .000106 |
| 0.0 | 0.4 | AMAPE | .181387 | .181454 | .183676 | .183691 | .178645 | .178652 |
| | | AMAD | .008345 | .008348 | .008450 | .008451 | .008219 | .008219 |
| | | AMSE | .000109 | .000109 | .000111 | .000111 | .000107 | .000107 |
| 0.0 | 0.6 | AMAPE | .181064 | .181089 | .182714 | .182698 | .178679 | .178680 |
| | | AMAD | .008328 | .008329 | .008404 | .008403 | .008219 | .008219 |
| | | AMSE | .000097 | .000098 | .000100 | .000100 | .000096 | .000097 |
| 0.1 | 0.2 | AMAPE | .168229 | .168442 | .170436 | .170534 | .167104 | .167475 |
| | | AMAD | .007864 | .007874 | .007967 | .007972 | .007811 | .007829 |
| | | AMSE | .000095 | .000095 | .000097 | .000097 | .000096 | .000097 |
| 0.1 | 0.4 | AMAPE | .165849 | .166013 | .168085 | .168203 | .166955 | .167370 |
| | | AMAD | .007751 | .007759 | .007856 | .007861 | .007803 | .007823 |
| | | AMSE | .000094 | .000094 | .000096 | .000096 | .000096 | .000097 |
| 0.1 | 0.6 | AMAPE | .165568 | .165752 | .167187 | .167314 | .167008 | .167463 |
| | | AMAD | .007737 | .007746 | .007812 | .007818 | .007804 | .007826 |
| | | AMSE | .000085 | .000085 | .000087 | .000088 | .000084 | .000085 |
| 0.2 | 0.2 | AMAPE | .154989 | .155192 | .157030 | .157234 | .153972 | .154645 |
| | | AMAD | .007349 | .007359 | .007446 | .007456 | .007301 | .007334 |
| | | AMSE | .000083 | .000083 | .000085 | .000085 | .000084 | .000085 |
| 0.2 | 0.4 | AMAPE | .152627 | .152992 | .154854 | .155043 | .153822 | .154581 |
| | | AMAD | .007236 | .007254 | .007342 | .007352 | .007293 | .007330 |
| | | AMSE | .000082 | .000083 | .000084 | .000084 | .000084 | .000085 |
| 0.2 | 0.6 | AMAPE | .152423 | .152833 | .154014 | .154255 | .153858 | .154681 |
| | | AMAD | .007226 | .007246 | .007301 | .007313 | .007294 | .007333 |

$^{(*)}$Notes: See notes to Table 4.

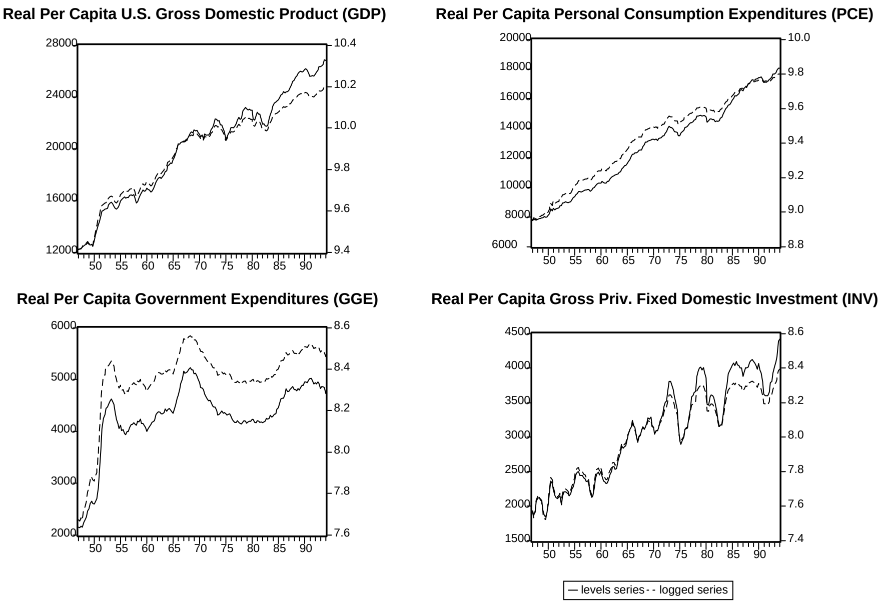
versa for levels-linear DGPs), then the VEC models again always have lower AMSE values than their difference VAR counterparts (compare the first and third columns of entries in Tables 2 and 4). However, the VEC model clearly does not AMSE-dominate the VAR model in differences when incorrectly transformed data are used in forecast construction (compare the second and fourth column of entries in Tables 1-4). Thus, VEC models do appear to uniformly dominate VAR model in differences, but only when the correct data transformation is used for model estimation, regardless of whether levels or log forecasts are ultimately compared. One reason for this finding may be that CI vectors and ranks are not precisely estimated when incorrectly transformed data are used.

(2) Undifferenced data VAR models AMSE-dominate difference VEC and VAR models around 50% of the time when correctly transformed data are used in estimation and forecast comparison (compare the first, third, and fifth columns of entries in Tables 1-4). Thus, in some cases, the simplicity of levels models appears to dominate more complex models, even when there is cointegration. However, it must be stressed that this finding only holds when *correctly* transformed data are used in forecast model construction (see below).

(3) The choice of loss function, $f$, does appear to make a difference in our experiments. In particular, the VEC model no longer beats the VAR in differences in every single instance, when the AMAPE and AMAD are used to compare models based on correctly transformed data. Also, the undifferenced data VAR model which is based on correctly transformed data is less frequently "better" than the VEC model when AMAPE and AMAD (rather than AMSE) are used to compare models. Thus, the choice of loss criterion appears to play an important role in model selection, even when criteria which are very similar, such as AMSE, AMAPE, and AMAD, are used.

(4) Choosing the data transformation in some cases appears to play a crucial role when comparing difference VEC and VAR models. For example, consider comparing models using AMSE. Note that the worst of the forecasting models based on correctly transformed data (i.e. choose the worst performer from columns 1, 3, and 5 of the entries in Tables 1-4) is almost always better than the best of the forecasting models based on incorrectly transformed data (i.e. choose the best performer from columns 2, 4, and 6 of Tables 1-4). Furthermore, this result is much more apparent for the loglinear DGPs reported on in Tables 1 and 2. When data is generated as level-linear (Tables 3 and 4), there appears to be surprisingly little to choose between data transformation, although the correct transformation is still usually “better” based on our criteria. Finally, there appears to be little to choose between transforming the forecasts from different forecasting models into levels or into logs to facilitate comparisons between

**FIG. 1.** Logs Versus Levels Linear Representations for Some U.S. Macroeconomic Series



Real Per Capita U.S. Gross Domestic Product (GDP)

Real Per Capita Personal Consumption Expenditures (PCE)

Real Per Capita Government Expenditures (GGE)

Real Per Capita Gross Priv. Fixed Domestic Investment (INV)

Note: All plots are of quarterly U.S. series for the period 1947:1-1994:1.

forecasts based on different models. This can be seen by noting that the ordinal ranking of the different forecasting models is the same when the corresponding entries in either Tables 1 and 2 or Tables 3 and 4 are compared.

In addition to the above findings, the following points are worth noting. First, although models based on data which have not been differenced are often AMSE-best when data are correctly transformed, they are almost always AMSE-worst when comparing forecast performance based on models estimated with incorrectly transformed data (compare column 6 entries with all other entries in Tables 1-4). This suggests that models with undifferenced data should perhaps only be used in the I(1) context when one is rather sure that the data are correctly transformed. As it is often difficult to ascertain the "correct" data transformation to use, forecasting models based on data which have not been differenced should thus be used with caution in practical applications.

Second, we have only indirect evidence on the usefulness of the BIC criterion when data are incorrectly transformed. In particular, one of the reasons why models with incorrectly transformed data perform so poorly relative to models estimated with correctly transformed data may be that there is no guarantee that the correct lag order will be chosen, even in the limit, when the wrong data transformation is used.

## 4. CONCLUSIONS

In a series of Monte Carlo experiments we show that data transformation matters when forecasting using vector error correction models, and when testing for unit roots and cointegration using standard approaches which assume linearity both under the null and under the alternative. We also show that incorrect data transformation leads to poor forecasts from cointegrated models, relative to simpler models based on differenced data, even when the true data generating process exhibits cointegration. This finding may be due to imprecise estimation of cointegrating spaces when the correct data transformation is uncertain, and may help to explain the mixed evidence concerning the usefulness of cointegration restrictions in forecasting.

## REFERENCES

Christoffersen, P. and F. X. Diebold, 1996, Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics* **11**, 651-572.

Christoffersen, P. and F. X. Diebold, 1998, Cointegration and long-horizon forecasting. *Journal of Business and Economic Statistics* **16**, 450-458.

Corradi, V., 1995, Nonlinear transformation of integrated time series. *Journal of Time Series Analysis* **16**, 539-550.

Corradi, V. and N. R. Swanson, 2000, Sample conditioned inference. Working Paper. Texas A&M University.

Dickey, D. A. and W. A. Fuller, 1979, Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**, 427-431.

Ermini L. and D. F. Hendry, 1995, Log income versus linear income: An application of the encompassing principle. Working Paper. University of Hawaii at Manoa.

Granger, C. W. J. and J. Hallman, 1991, Nonlinear transformations of integrated time series. *Journal of Time Series Analysis* **12**, 207-224.

Granger, C. W. J. and N. R. Swanson, 1996, Further developments in the study of cointegrated variables. *Oxford Bulletin of Economics and Statistics* **58**, 537-553.

Johansen, S., 1988, Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**, 231-254.

Johansen, S., 1991, Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* **59**, 1551-1580.

*Journal of Applied Econometrics*, (1996), Special issue on forecasting in economics.

King, R. G., C. I. Plosser, J. H. Stock, and M. M. Watson, 1991, Stochastic trends and economic fluctuations. *American Economic Review* **81**, 819-840.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin, 1992, Testing for the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* **54**, 159-178.

Long, J. B. and C. I. Plosser, 1983, Real business cycles. *Journal of Political Economy* **91**, 39-69.

Pesaran, M. H. and A. G. Timmerman, 1994, The use of recursive model selection strategies in forecasting stock returns. Working Paper. University of California, San Diego.

Sargent, T. J., 1998, The conquest of American inflation. *Marshall Lecture Series.* The University of Cambridge.

Swanson, N. R. and H. White, 1997, A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics* **79**, 540-550.