Simulation and Prediction Evidence On the Usefulness of Seasonal Models

Norman R. Swanson[1,2] and Richard Urbach[2]
[1] Rutgers University
[2] DFA Capital Management, Inc.

this version: February 2007

Abstract

In this paper, we provide new evidence on the empirical usefulness of various simple seasonal models, and we underscore the importance of carefully designing on a case by case basis the criteria with which one judges alternative models.. This is done using a series of simulation and prediction experiments, as well as via discussion of the stochastic properties of seasonal unit root models. Our prediction experiments are based on analysis of a group of 14 variables have been chosen to closely mimic the set of indicators used by the Federal Reserve to help in setting U.S. monetary policy, and our simulation experiments are based on a comparison of simulated and historical distributions using a testing approach due to Corradi and Swanson (2007a). Our findings suggest that a simple version of the seasonal unit root (SUR) model performs very well for predicting 8 of 14 variables, when the forecast horizon is 1-step ahead. This result suggests that seasonal integration test results need to be interpreted with caution. Their poor finite sample properties may mislead investigators into believing that seasonal unit root models are not useful. However, for horizons of greater than one-step ahead, the SUR models perform poorly when used for prediction, suggesting that parameter estimation error is crucial to understanding the empirical performance of such models. This result is confirmed via a series of Monte Carlo experiments. Simple periodic autoregressions do not have this property, however, and indeed are found to perform very well in both prediction and simulation experiments, at all horizons cup to 60 months ahead.

JEL classification: C13, C22, C52, C53.
Keywords: seasonal unit root, periodic autoregression, difference stationary, prediction, simulation.

* Norman R. Swanson (nswanson@econ.rutgers.edu): Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA and DFA Capital Management, Inc. 100 Manhattanville Road, Suite 11, Purchase, NY 10577. Richard Urbach (ru@dfa.com): DFA Capital Management, Inc. 100 Manhattanville Road, Suite 11, Purchase, NY 10577. We owe a great many thanks to the editor, Hashem Pesaran, whose many comments and suggestions have led to a great many useful revisions to this paper. We also wish to thank two anonymous referees for providing useful comments. Finally, we are grateful to John Chao, Valentina Corradi and Eric Ghysels for useful comments and discussions. The authors have also benefited from the excellent research assistance provided by Geetesh Bhardwaj.

1 Introduction

One of the most important characteristics of econometric models is that they are generally viewed as *approximations* to some unknown underlying data generating process. Furthermore, the extent to which these *approximations* are adequate is often assessed by evaluation of their prediction and simulation properties. For example, insurance companies and banks are often interested in simulating future economic scenarios in order to evaluate measures of capital adequacy, possibly under a variety of alternative starting conditions, product pricing rules, and regulatory environments. Governments, on the other hand, often utilize macroeconometric forecasting models in order to predict economic variables of interest, such as interest rates and inflation, both of which play a crucial role in policy analysis. Our objective in this paper is to assess the adequacy of a number of simple seasonal models, all of which should be viewed as *approximations*, in terms of prediction and simulation.

Seasonal models in economics have received considerable renewed attention since the seminal paper of Hylleberg, Engle, Granger and Yoo (HEGY: 1990) on seasonal unit roots. From an empirical perspective, many papers have documented the prevalence of seasonal unit roots in the data, and evaluated the predictive performance of various seasonal models, including seasonal unit root models, deterministic seasonal models, and periodic autoregressions. From a theoretical perspective, any papers have developed new testing methods and suggested various nonlinear seasonal models. A very few important contributions in these area include: Hylleberg (1992,1994), Beaulieu and Miron (1993), Franses (1996), Miron (1996), Miron and Beaulieu (1996), Franses, Hoek and Paap (1997), Franses and Vogelsang (1998), Osborn, Heravi and Birchenhall (1999), Ghysels and Osborn (2001), Franses and Paap (2002), Osborn (2002), Osborn and Rodrigues (2002), Franses and van Dijk (2005), and the references cited therein.

The approach that we take in this paper is to consider four very simple econometric models and to examine fourteen different economic variables measuring seasonally unadjusted economic activity. The variables considered are chosen based on the economic and financial indicators used by the Federal Reserve to aid in the formulation of the nation's monetary policy (see below for further discussion). These variables examine various measures of industrial production, money, housing starts, employment, inventories, and sales; all of which are characterized by distinctive seasonal fluctuations. The models that we consider are: the seasonal unit root (SUR) model, a deterministic

seasonality (DS) model, a periodic autoregression (PAR) model, and a strawman random walk (RW) model, possibly with drift. Our analysis is based on the results of a series of Monte Carlo experiments, prediction experiments, and simulation experiments. Our Monte Carlo experiments evaluate SUR model and SUR tests, with emphasis on the effects of parameter estimation as forecast horizon increases. Our prediction experiments construct sequences of recursive ex -ante predictions using the alternative models, and assess mean square forecast error (MSFE) using point estimates as well as predictive accuracy tests. Finally, in a series of simulation experiments, we evaluate the distributional characteristics of data simulated using the four different model types. This is done using a recent testing methodology developed by Corradi and Swanson (2007a), where empirical distributions of the historical series and that of simulated series are compared, using a Kolmogorov type distributional test (see e.g. Andrews (1997)) based on distributional mean square error loss.

Our findings can be summarized as follows. We provide evidence that a simple version of the seasonal unit root (SUR) model performs very well for predicting various macroeconomic variables when the forecast horizon is 1-step ahead. This result suggests that seasonal integration test results need to be interpreted with caution. Their poor finite sample properties may mislead investigators into believing that seasonal unit root models are not useful. However, for horizons of greater than one-step ahead, the SUR models perform poorly when used for prediction, suggesting that parameter estimation error is crucial to understanding the empirical performance of such models. This result is confirmed via a series of Monte Carlo experiments. Interestingly, simple periodic autoregressions do not have this property, and indeed perform very well in both prediction and simulation experiments, at all horizons considered in this paper. Deterministic seasonality models also perform reasonably well at all forecast horizons, and indeed dominate PAR models for a small subset of our variables, including price and industrial production sub-component variables. Finally, by comparing simulation and prediction based evidence, we underscore the importance of carefully designing on a case by case basis the criteria with which one judges alternative models.

The rest of the paper is organized as follows. In Section 2, we outline our empirical methodology and elucidate some relevant analytical properties of SUR models. In Section 3, we describe the data used. Section 4 contains the results of our Monte Carlo experiments, and Section 5 summarizes our empirical findings. Concluding remarks and recommendations are given in Section 4.

2 Empirical Methodology

2.1 Models, Estimation and Diagnostic Tests

There are a plethora of models from whence the "best" model for a particular time series can be chosen. These include both linear and nonlinear models, parametric and nonparametric models, and stochastic and deterministic models, for example. However, there is substantial empirical evidence that parsimonious (linear) models often yield more accurate predictions than more complex (nonlinear) models. Given this fact, and in order to simplify our analysis, we consider a small group of widely used univariate models, and check whether any of these outperform SUR models when used for simulation and prediction. If we find evidence that predictions or simulations based on SUR models are less accurate than other models, say, then we have direct evidence against the usefulness of SUR models in practical applications. Whether the evidence that we gather can be generalized to the case of more complex linear and nonlinear models is left to future research.

Let $y_t = (1 - L) \ln X_t = \Delta \ln X_t$ be a scalar growth rate of some macroeconomic variable of interest. In the sequel we shall consider models of the form:

I. $y_t = \theta_0 + \varepsilon_t$ – *Random Walk (RW) Model*

II. $\phi(L) y_t = \Sigma_{s=1}^{S} \theta_s d_{s,t} + \varepsilon_t$ – *Difference Stationary (DS) Model (with deterministic seasonal components)*

III. $\phi(L) \triangle_S y_t = (1 + \theta L^s) \varepsilon_t$ – *Seasonal Unit Root (SUR) Model*

IV. $y_t = \theta_s + \Sigma_{i=1}^{p} \theta_{i,s} y_{t-i} + \varepsilon_t$ – *Periodic Autoregression (PAR) Model*

In these models, $\triangle_S$ denotes the S^{th} difference operator, where S is the number of seasons presumed to be in the data (i.e. $\triangle_S$ is the seasonal difference operator); $\phi(L) = (1 - \phi_1 L - ... - \phi_p L)$ is a standard lag polynomial of order p, expressed using the lag operator, L; and the θ_s denote seasonal intercepts, with associated and conformably defined dummy variables, $d_{s,t}$. For an exhaustive discussion of the seasonality models considered in this paper, see the books by Hylleberg (1992), Franses (1996), Miron (1996), and Ghysels and Osborn (2001); and for an interesting discussion on the usefulness of seasonal adjustment, see Franses (2001).

Notice that the PAR model generalizes the DS model, as it allows the intercept *and slope* coefficients to vary according to the season. Notice also that both of these models generalize the RW model, which is used as a "strawman" model in our analysis. In all cases, two versions of the random walk are fitted, one with $\theta_0 = 0$ and one with $\theta_0 \neq 0$, and the model that performed better

is reported on. Finally, notice that the SUR model is the same as that specified by Bell (1987), and Bell shows that our DS and SUR models are equivalent when $\theta = 1$ (he actually sets $\phi(L) = 1$, but further shows that the result holds when the error follows a general ARMA process). This variety of SUR model has been seen to perform relatively well in empirical contexts (see e.g. Bell (1987), Ghysels and Osborn (2001), and the references cited therein).

All models are estimated using least squares, and the lag polynomial order is estimated using the Schwarz Information Criterion (SIC). The case where $p = 0$ was allowed for, and indeed picked in some instances. Consider the PAR(1) model, where $p = 1$. This model can be estimated via least squares using the equation: $y_t = \Sigma_{s=1}^{S} \theta_s d_{s,t} + \Sigma_{s=1}^{S} \theta_{1,s} d_{s,t} y_{t-1} + \varepsilon_t$. Here, under error normality and given fixed starting values, the least squares estimator is the maximum likelihood estimator, and standard asymptotics pertain (see Franses and Paap (2002) and Franses and van Dijk (2005) for further details). Notice that the PAR(1) model has a unit root when $\theta_{1,1}\theta_{1,2}\theta_{1,3} \cdots \theta_{1,s} = 1$. Clearly, the PAR(1) model nests the simple random walk model where $\triangle_1 y_t = (1 - \theta L) y_t = \theta_0 + \varepsilon_t$, with $\theta = 1$. In this case, the characteristic equation is $(1 - \theta^S z) = 0$, so that when $\theta = 1$, y_t has a single nonseasonal unit root, corresponding to the simple random walk model. Also, when $\theta = -1$, y_t has a seasonal unit root. Thus, as mentioned above, both seasonal and non-seasonal unit root processes such as those given in our RW, DS, and SUR models, are nested within PAR models (see HEGY (1990) for further details). Nonlinear variants of the above models are discussed in Franses and van Dijk (2005), Franses, de Bruin and van Dijk (2000) and Rodrigues (2001), and the papers cited therein. Boswijk and Franses (BF: 1996) outline tests for $\theta_{1,s}\theta_{2,s}\theta_{3,s}\theta_{4,s} = 1$ and for $\theta = 1$ and $\theta = -1$ (see Franses and Paap (2002) for a summary; and Granger and Teräsvirta (1993) for a summary of the types of nonlinear models used by these authors).

Diagnostic tests used in the sequel include Jarque-Bera, augmented Dickey-Fuller (ADF) (with lags selected via use of the SIC), BF periodic integration , and HEGY (called HEGY1) seasonal unit root tests. In addition, we use the HEGY2 test, which refers to a HEGY1 test with pre whitening, and involves estimating $\widehat{\triangle_S y_t} = \sum_{i=1}^{p} \widehat{\phi}_i \triangle_S y_{t-i}$. HEGY1 is then applied to pre-whitened data $y_t^* = \left(1 - \sum_{i=1}^{p} \widehat{\phi}_i L^i\right) y_t$ as suggested by Psaradakis (1997). Note that in this paper, we report test results for the null of no seasonal unit root based on examination of π_2 (see Beaulieu and Miron (1993) for complete details, and Swanson and Urbach (2007) for complete HEGY test results at all frequencies, which are omitted here for the sake of brevity).

One of the most recognized features of SUR models, the notion that *winter can become summer*,

is perhaps worth recalling before we turn to a discussion of the empirical methods used in the sequel. Consider the very simple model where $\triangle_S y_t = \theta_0 + \varepsilon_t$, where y_t is driven by a nonstationary process in each season. Now, these seasons might share a common drift, θ_0, but their evolution is clearly independent from one season to the next, given that the evolution of each season can be written as $y_t^s = \theta_0 + \varepsilon_{s,t}$, where $\varepsilon_{s,t} \sim iid\ N(0, \sigma^2)$, say, for $s = 1, ..., S$. Thus, $\triangle_S y_t$ can be viewed as being comprised of S independent random walks (see e.g. Osborn and Ghysels (2001) and Franses and van Dijk (2005)). This suggests that when used for simulation or prediction, one should expect values of y_t in different seasons to drift infinitely far apart, given enough time. Clearly this is a feature of SUR models which is not in accord with expected behavior of economic series, at least in the long run. However, all models should be viewed as approximations, and hence the above feature may be acceptable, as long as a particular SUR model still yields adequate predictions and simulated observations for a given time series, over horizons that the researcher is interested in. To illustrate the importance of this point, consider Figure 1, where inflation, measured as $y_t = \Delta \ln X_t$, where $X_t = CPI$, is simulated under the assumption that either (i) CPI follows a random walk with drift in logs, or (ii) $\triangle_S y_t = \theta_0 + \varepsilon_t$, and where models are estimated using monthly U.S. data for the period 1959-2005. Notice that y_t data generated via the SUR simulation model are increasing over time, with ever increasing volatility, a feature which is clearly not in accord with the historical record, and which suggests that various versions of the SUR model may yield poor results when used over long horizons.[1] While the fact that this feature arises in SUR models used in the current context is rather obvious, it is nevertheless a fact that is worth stressing, as it plays an important role in our subsequent findings.

2.2 Simulation and Predictive Accuracy Testing Methodology

We begin by summarizing the simulation based distributional accuracy test discussed in Corradi and Swanson (CS: 2007a) for comparing simulated data with historical data. Assume that our objective is to compare the joint distribution of the historical data with the joint distribution of a simulated series. Following CS, and for the sake of simplicity, we limit our attention in the section to the evaluation of the joint empirical distribution of (actual and model-based) current and previous

[1] Note that the two models fit the data equally well, based on examination of in-sample correlation and residual serial autocorrelation. Note also that further deatils of this property are given in the earlier working draft version of this paper (i.e. see Swanson and Urbach (2005)).

period values, say $y_t^* = (\Delta \ln X_t, \Delta \ln X_{t-1})$, of our variable of interest, $y_t = \Delta \ln X_t$. Consider m alternative econometric models, and set model 1 as the benchmark model. Let $\Delta \ln X_{j,n}(\widehat{\theta}_{j,T})$, $j = 1, ..., m$ and $n = 1, ..., N$, denote the series of interest, simulated under model j, where N denotes the length of the simulated sample, and $\widehat{\theta}_{j,T}$ denotes estimated parameters, using the T available historical observations. Thus, $y_{j,n}^*(\widehat{\theta}_{j,T}) = (\Delta \ln X_{j,n}(\widehat{\theta}_{j,T}), \Delta \ln X_{j,n-1}(\widehat{\theta}_{j,T}))$. (In general, one can set y_t^* to be any stationary vector of economic variables.) Also, let $F_0(u; \theta_0)$ denote the distribution of y_t^* evaluated at u and $F_j(u; \theta_j^\dagger)$ denote the distribution of $y_{j,n}^*(\theta_j^\dagger)$, where $\theta_j^\dagger$ is the probability limit of $\widehat{\theta}_{j,T}$, taken as $T \to \infty$, and where $u \in U \subset \Re^2$, possibly unbounded. Accuracy is measured in terms of square error. The squared (approximation) error associated with model i, $i = 1, ..., m$, is measured in terms of the (weighted) average over U of $E\left(\left(F_i(u; \theta_i^\dagger) - F_0(u; \theta_0)\right)^2\right)$. Thus, the rule is to choose Model 1 over Model 2 if:

$$\int_U E\left(\left(F_1(u; \theta_1^\dagger) - F_0(u; \theta_0)\right)^2\right)\phi(u)du < \int_U E\left(\left(F_2(u; \theta_2^\dagger) - F_0(u; \theta_0)\right)^2\right)\phi(u)du,$$

where $\int_U \phi(u)du = 1$ and $\phi(u) \geq 0$, for all u. The hypotheses of interest are:

$$H_0 : \max_{j=2,...,m} \int_U E\left(\left(F_0(u; \theta_0) - F_1(u; \theta_1^\dagger)\right)^2 - \left(F_0(u) - F_j(u; \theta_j^\dagger)\right)^2\right)\phi(u)du \leq 0$$

$$H_A : \max_{j=2,...,m} \int_U E\left(\left(F_0(u; \theta_0) - F_1(u; \theta_1^\dagger)\right)^2 - \left(F_0(u) - F_j(u; \theta_j^\dagger)\right)^2\right)\phi(u)du > 0.$$

Under H_0, no model can provide a better approximation (in square error sense) to the distribution of Y_t than the approximation provided by model 1. If interest focuses on confidence intervals, so that the objective is to "approximate" $\Pr(\underline{u} \leq y_t \leq \overline{u})$, then the null and alternative hypotheses can be written in a form similar to that above, and appropriate statistics can be constructed. In order to test H_0 versus H_A, the relevant test statistic is $\sqrt{T}Z_{T,N}$, where $Z_{T,N} = \max_{j=2,...,m} \int_U Z_{j,T,N}(u)\phi(u)du$, and:

$$Z_{j,T,N}(u) = \frac{1}{T}\sum_{t=1}^{T}\left(1\{y_t^* \leq u\} - \frac{1}{N}\sum_{n=1}^{N} 1\{y_{1,n}(\widehat{\theta}_{1,T}) \leq u\}\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\left(1\{y_t^* \leq u\} - \frac{1}{N}\sum_{n=1}^{N} 1\{y_{j,n}^*(\widehat{\theta}_{j,T}) \leq u\}\right)^2,$$

where $\widehat{\theta}_{j,T}$ is an estimator of $\theta_j^\dagger$. (The first term in the above expression is called the *CS distributional loss* for Model "1", while the second term is the loss for Model "j".) Asymptotically valid critical values for this test can easily be constructed using the block bootstrap (see Corradi and Swanson (2007a) for further details). One crucial feature of the $Z_{T,N}$ test is that *all* models

under both hypotheses may be misspecified, as opposed to the usual practice of assuming correct specification under the null.

We now turn to a discussion of pointwise predictive accuracy testing. For a discussion of predictive accuracy and predictive model selection tests available in the current literature, the reader is referred to Corradi and Swanson (2006a) and the references cited therein. In the current paper, we use the Diebold and Mariano (DM: 1995) and West (1996) test, where the equal predictive accuracy null is tested using:

$$DM = \sqrt{P}\frac{\frac{1}{P}\sum_{t=R}^{T}\widehat{d}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}}\sum_{t=R+j}^{T-h}K\left(\frac{j}{M}\right)\left(\widehat{d}_{t+h}-\overline{d}\right)\left(\widehat{d}_{t+h-j}-\overline{d}\right)},$$

where $\widehat{d}_{t+h} = \widehat{u}^2_{1,t+h} - \widehat{u}^2_{2,t+h}$ under mean square error loss; $\overline{d} = \frac{1}{P-h+1}\sum_{t=R}^{T-\tau}\widehat{d}_{t+h}$; $\widehat{u}_{j,t+h}$ is an ex-ante forecast error constructed using a recursively estimated prediction model, for $j = 1, 2$; and P is the number of observations in the ex ante forecast period. Of note is that loss functions other than mean square error loss can be used, although in this paper we focus exclusively on mean square forecast error loss. The limiting distributions of the DM statistic is given in Theorems 3.1 and 3.2 in Clark and McCracken (2005), and for $h > 1$ contains nuisance parameters so that critical values cannot be directly tabulated, and hence Clark and McCracken (2005) use the Kilian (1999) parametric bootstrap to obtain critical values (see e.g. McCracken and Saap (2004) for discussion). Note, though, that for $h = 1$, when models are nonnested (see Corradi and Swanson (2002) for a discussion of predictive accuracy and predictive Granger causality tests with nested models), and assuming that parameter uncertainty vanishes asymptotically, the standard normal distribution applies in the case of the DM test, even when the heteroskedasticity and autocorrelation consistent standard error given in the numerator of DM is replaced with $\frac{1}{P-h+1}\sum_{t=R+j}^{T-h}\left(\widehat{d}_{t+h}-\overline{d}\right)^2$. Furthermore, nonstandard critical values which obtain in other cases are generally larger, so that rejection of the null hypothesis using percentiles of the normal distribution often implies rejection using nonstandard critical values. Finally, nonstandard critical values are usually quite close, in absolute magnitude, to standard normal values. For these reasons, we use standard normal rejection regions when we report significance of DM test statistics. These should, of course, only be taken as a rough guide. For a thorough discussion of predictive accuracy tests, based both on point MSFEs and densities, the reader is referred to Corradi and Swanson (2006a,b). For a discussion of predictive accuracy testing under recursive estimation schemes such as that used here, the reader is referred to Corradi and Swanson (2007b).

3 Data

In this paper, we examine 14 monthly U.S. series for the period 1959:1-2005:12, except where noted. The series include: two money stock variables (M1 and M2) and M3; two CPI series including CPI for all urban consumers, all items (*CPI1*), and energy (*CPI2*); housing starts (*H_Start*); industrial production (*IP*); total nonfarm employment (*NonF_Emp*); industrial production - automotive products (*IP_Auto*); industrial production - durable consumer goods (*IP_Dur*); industrial production - durable consumer goods (*IP_NDur*); durable goods shipments of new orders and unfilled orders (*D_Ship*); total inventories, manufacturing (*Invent*); retail sales (*Ret_sales*) - for the period 1967:1-2001:4; and motor vehicle unit retail sales (*Veh_sales*) - for the period 1967:1-2005:12. These variables have been extracted from the FRED (Federal Reserve Bank of St. Louis) database, the U.S. Census Bureau, and *http://wweconomagic.com*. Plots of the series are presented in Figure 1, where a variety of the series can be seen to exhibit cyclical variation consistent with seasonality.

Note that on the Federal Reserve Bank of New York website *http://www.newyorkfed.org/education/bythe.htm* it is stated that:

"In formulating the nation's monetary policy, the Federal Reserve considers a number of factors, including the economic and financial indicators which follow, as well as the anecdotal reports compiled in the Beige Book. Real Gross Domestic Product (GDP); Consumer Price Index (CPI); Nonfarm Payroll Employment Housing Starts; Industrial Production/Capacity Utilization; Retail Sales; Business Sales and Inventories; Advance Durable Goods Shipments, New Orders and Unfilled Orders; Lightweight Vehicle Sales; Yield on 10-year Treasury Bond; S&P 500 Stock Index M2"

Our group of 14 variables have been chosen to closely mimic this set of indicators, with the exception of yields and the S&P500 index, both of which were found to exhibit little seasonality.

4 Monte Carlo Results

In this section, the results of a small Monte Carlo experiment carried out in order to assess the finite sample impact of parameter estimation error on SUR model predictive performance, and HEGY test size are reported. Data are generated according to the following SUR model:

$$\phi(L)\triangle_S y_t = (1+\theta L^s)\varepsilon_t,$$

where $\phi(L) = 1 - \phi L$, and $\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$. Parameter values are chosen to mimic estimates obtained from the 14 monthly series used in our empirical analysis. In particular, we set $\theta = \{-0.6, -0.7, -0.8, -0.9\}$, $\phi = \{-0.5, 0.0, 0.5\}$, and $\sigma_\varepsilon = \{0.005, 0.01, 0.1\}$. Moreover, in order to mimic the data transformation used in our empirical analysis, we set $y_t = (1 - L) \ln X_t$, and $s = 12$.[2] Results based upon a (i) series of predictions experiments, and (ii) in-sample HEGY tests are reported in the table. With regard to the prediction experiments, in each simulation, a sample of 500 observations was drawn, and prediction models were recursively estimated, beginning with an in-sample period of 120-h+1 observations, and yielding a sequence of 380 ex-ante h-step ahead forecasts. Results reported in the first 12 columns of numerical entries in the Table 2 are based upon examination of these forecasts. Namely, the first 4 columns report the proportion of simulations for which the SUR model predictions were "MSFE-better" than those based on a Random Walk model (RW). The next 8 columns report the proportion of simulations for which the two models yielded "MSFE-better" predictions based upon application of the Diebold-Mariano (1995) predictive accuracy test at a 5% nominal significance level, and based on MSFE loss. Entries in the last three columns of the table indicate the proportion of times that the seasonal unit root (HEGY) and BF tests reject the null (see Table 1 and Section 2 for further details). All results are based on 500 Monte Carlo simulations.

A number of conclusions can be drawn from examination of the table. First, notice that in the point MSFE comparison, the *SUR* model is "MSFE-best" in every simulation for $h = 1$ and $h = 3$. However, for greater values of h, performance drops of, and indeed for $h = 12$, when $\theta = -0.9$ the RW model actually "wins" around 50% of the time; and around 75% of the time when $h = 60$. This suggests that parameter estimation error plays a substantive role when the forecast horizon is reasonably large, and when the MA coefficient is high, as may be expected to be the case in practice. Second, notice that The frequency of DM test "wins" agrees with the previous finding. Namely, for $h = 12$ and $h = 60$, the proportion of simulations for which the SUR model has a significantly lower MSFE than the *RW* model ranges from 25 to 20%, even for cases where $\theta = -0.6$. Given that the true DGP is a *SUR* model, and the the correct specification is nested within the class of estimated *SUR* models, it is quite clear that parameter estimation error truly plays a substantial

[2]Results are reported in Table 2 for all cases except $\theta = \{-0.7, -0.8\}$, which are omitted for the sake of brevity, and because results for these cases can be immediately inferred from examination of results for the cases where $\theta = \{-0.6, -0.9\}$.

role. Indeed, for $h = 60$ the random walk actually yields significantly lower MSFEs than the *SUR* model much more frequently than the number of times that SUR "wins", whenever $\theta = -0.9$. In some sense, this is not too surprising, given that seasonal difference operators tend to cancel out with "near" non-invertible MA operators, as pointed out by Bell (1987), and observed in finite sample experiments by Ghysels, Lee and Noh (1994) and Rodrigues and Osborn (1999). Finally, notice that BF, HEGY1, and HEGY2 tests are very poorly sized in our experiment (see last three columns in the table). These results are consistent with earlier findings of many authors, including those just mentioned, and suggests that when using monthly data, standard seasonal unit root tests may be an inadequate guide as to the usefulness of such models. For this reason, it may be preferable in empirical applications to simply fit the different models and assess which models perform "best", based on simulation and prediction experiments, for example.

5 Empirical Results

5.1 Basic Data Analysis

Given that BF and HEGY tests perform somewhat poorly using monthly data such as the 14 macroeconomic variables analyzed in this section, we begin by searching for evidence of seasonality using a variety of means. First, estimated spectra are plotted in Figure 2. To interpret these plots, recall that a frequency of $\omega_j = \frac{2\pi j}{T}$, corresponds to a period of $\frac{2\pi}{\omega_j} = \frac{T}{j}$ (see e.g. Hamilton (1994, chp. 6). In the plots j is reported on x-axis. For illustrative purposes, consider the time series *Durable goods shipments & unfulfilled orders*. The sample size is $T = 563$. The first peak occurs around $j = 93$,corresponds to a cycle of $563/93 \approx 6$ months. The second peak at $j = 140$, corresponds to a cycle of 4 months, the third cycle is 3 months, and fourth cycle is around 2 months. Similarly for Retail Sales, $T = 419$. Thus the spectrum peaks at $j \in (33, 68, 102, 136, 170)$ correspond to cycles at around 12, 6, 4, 3 and 2 months. Evidently, the spectra suggest that seasonality is prevalent in a number of the series. However, note in Table 1 that HEGY and BF tests find little evidence of seasonality. Given the poor empirical size of these tests, however, this evidence should be interpreted with care. Indeed, when deterministic seasonal dummy models were fit to the data, many associated coefficients were found to be significantly different from zero based upon the application of 5% nominal size tests. The last column of the Table 1 supports this finding, as the null of no deterministic seasonality is rejected for all 14 variables (see Urbach and Swanson (2007)

for complete tabulated HEGY test results and complete deterministic dummy variable estimation results). Overall, there is quite clear evidence of seasonality in our data. Furthermore, note that in Table 1, that all series fail to reject the null of a unit root, based on application of the ADF test, supporting our use of growth rates in our subsequent simulation and prediction experiments.

5.2 Simulation Experiments

In order to evaluate the performance of the different models from the perspective of simulation, our four estimated models, including *RW, DS, SUR, and PAR* were used to simulate N observations of y_t. The starting values for the simulations were fixed to be the last observation of the historical sample used in estimation. Note that the models used for simulation were the models outlined above (see Section 2.1), *including the error term*, where the error is assumed to be *iid* normal, and where the variance of the error term is estimated using the residuals of the fitted model. Thus, simulation models take forms such as $\triangle_{12} y_t = 1.66 + \varepsilon_t$, for example, where $\varepsilon_t \sim iidN(0, \sigma_\varepsilon^2)$, and σ_ε^2 is calibrated from the data.[3] Results were also tabulated for the case where errors were drawn randomly from the empirical distribution of the residuals. Findings in this case were qualitatively similar to those reported below, and hence have been omitted for the sake of brevity (complete results are available upon request). We consider simulation paths of length $N = \{5T, 10T\}$, where T is the historical sample size. All results are gathered in Table 3, and are based on the analysis of 100 simulation paths using the distribution comparison test discussed in Section 2.2 (further discussion of the experimental setup is contained in the footnote to Table 3; and a substantially more extensive simulation analysis is contained in Swanson and Urbach (2005).

Our results based on the simulation experiments can be summarized as follows. First, the null hypothesis that there is nothing to choose between the different models based on the application of the $Z_{T,S}$ test is not rejected for any variables , when the benchmark model is RW (compare the first column of entries in the table with the critical values reported in the 2^{nd} through 5^{th} columns of entries). However, this result may be in part be driven by poor finite sample test performance, as evidenced by the fact that there appear to be stark differences between the *CS distributional loss* measures reported in the last four columns of the table (see Section 2.2 for the definition of this MSE type distributional loss measure). In particular, notice that for *CPI1, PAR* and *RW CS loss*

[3]Note that all historical data were used in all estimations.

values are around one third as large as that of DS and one half as large as that of SUR. Overall, the PAR *CS loss* is lowest for 9 of 14 variables, and is either lowest or very close to lowest for 14 of 14 variables. In contrast, the SUR model is lowest or very close to lowest for 0 of 14 variables. Results for the other two models are more mixed, with both RW and DS performing well for 7 of 14 variables. However, it should be stressed that we have: (i) set $y_t^* = (\Delta \ln X_t, \Delta \ln X_{t-1})$; and (ii) set U to be the entire range of the historical data. For different choice of y_t^* and U, results may vary. For example, choosing U to correspond to the tail of the distribution, or choosing $\phi(u)$ to weight more important regions of the distribution may yield different results. Nevertheless, based upon our somewhat naive application of the CS test, there appears to be substantive evidence in favor of the more general PAR model, particularly with respect to the SUR model. Of course, if the objective were to simulate 12^{th} log differences, then the SUR model would be expected to perform much better (see Swanson and Urbach (2005) for evidence of this). Thus, it remains crucial for the empiricist to carefully design an appropriate loss function based on her/his particular requirements, prior to selecting amongst alternative models.

5.3 Prediction Experiments

The objective in the previous subsection was to assess the ability of the alternative models to simulate long paths of future observations. Clearly, each path can be thought of as containing predictions for many different horizons, some of which are very far in the future. While this approach to model assessment has obvious uses in a variety of contexts, it is quite different from an assessment of the models based on their ability to predict a variable for a single given horizon. Interestingly, relatively little evidence has been published comparing predictive performance of the 4 models analyzed in this paper (see Osborn, Heravi, and Birchenhall (1999) and Paap, Franses, and Hoek (1997)).

In this subsection we report the results of prediction experiments for fixed forecast horizons. Prediction models are constructed recursively, starting with $R = 120 - h + 1$ observations, and ending with $R = T - h$ observations, where T is the sample size, and h is the forecast horizon (set equal to 1, 3, 12, and 60 months). The results of these experiments are contained in Table 4, where MSFEs (Panel A) and DM predictive accuracy test statistics based on MSFE loss (Panel B) are reported. All DM tests are pairwise, and compare the benchmark RW model with the model denoted in the column header to the table. Negative values for DM statistics indicate that the

point MSFE associated with the benchmark model is lower than that for the other model. Starred DM test statistics indicate rejection of the predictive accuracy null using 5% nominal size critical values (see Section 2.2 for further details).

Our tabulated results are consistent with the following conclusions. First, The SUR model performs admirably well for predicting one-step ahead (see Panel A of the table). Indeed, the SUR model dominates the RW model for 8 of 14 series, based on application of the DM predictive accuracy test (see Panel B of the table), and for 10 of 14 series based on point MSFE comparison. This is a somewhat surprising finding, given the evidence of the HEGY tests that there is little seasonal integration; but is completely consistent with the finding of Osborn et al. (1999) that seasonal difference model perform well when used to predict 2-digit industrial production series at short horizons, even though they also find very little test-based evidence of seasonal integration. Thus, we again have evidence of the lack of reliability of in-sample integration tests.

Second, the results for $h = 3, 12$, and 60 are very different from those for $h = 1$. The SUR model is never dominant based on accuracy tests, and is only point MSFE dominant for one variable at one horizon. Thus, parameter estimation error plays a crucial role in empirical models of the variety examined here. Namely, even if one assumes that there is seasonal integration, in accord with our finding that SUR is dominant for $h = 1$, PEE ensures that SUR is soundly beaten at all higher horizons. At the very least, we must conclude that SUR is a poor approximation when used to predict horizons greater than $h = 1$.

Third, The PAR model is reasonably accurate. In particular, PAR is point MSFE and DM test dominant for 8 of 14 variables, relative to the RW benchmark, for $h = 1$. Furthermore, PAR performs approximately as well as the RW model for 4 of the remaining 6 variables. Even more importantly, PAR continues to dominate RW, and indeed all other model, with respect to the same set of 8 variables across *all* forecast horizons. Overall, the set of variables for which PAR dominates include all variables other than our price and IP sub-component variables. This favorable evidence in support of the PAR model mimics the evidence presented in the previous subsection, where the PAR model was found to be the best performing model, based on CS *distributional loss.*

We have evidence that SUR

6 Concluding Remarks

We have provided evidence that a simple version of the seasonal unit root (SUR) model performs very well for predicting various macroeconomic variables when the forecast horizon is 1-step ahead. This result suggests that seasonal integration test results need to be interpreted with caution. Their poor finite sample properties may mislead investigators into believing that seasonal unit root models are not useful. However, for horizons of greater than one-step ahead, the SUR models perform poorly when used for prediction, suggesting that parameter estimation error is crucial to understanding the empirical performance of such models. This result is confirmed via a series of Monte Carlo experiments. Interestingly, simple periodic autoregressions do not have this property, and indeed perform very well in both prediction and simulation experiments, at all horizons considered in this paper. Deterministic seasonality models also perform reasonably well at all forecast horizons, and indeed dominate PAR models for a small subset of our variables, including price and IP sub-component variables. Finally, by comparing simulation and prediction based evidence, we underscore the importance of carefully designing on a case by case basis the criteria with which one judges alternative models.

It remains to further uncover both theoretical and empirical evidence concerning the role that parameter estimation error plays in the context of seasonal models. Additional areas for future research include the study of dynamically more complex models than those considered here, including various nonlinear models; and the development of new models that can be used to accurately model various different transformations of a variable when there is substantial seasonality in the data.

7 References

Andrews, D.W.K., (1997), A Conditional Kolmogorov Test, *Econometrica*, 65, 1097-1128.

Beaulieu, J.J. and J. Miron, (1993), Seasonal Unit Roots and Deterministic Seasonals in Aggregate U.S. Data, *Journal of Econometrics*, 55, 305-328.

Bell, W.R., (1987), A Note on Overdifferencing and the Equivalence of Seasonal Time Series Models with Monthly Mean and Models with $(0,1,1)_{12}$ Seasonal Parts When $\theta = 1$, *Journal of Business and Economic Statistics*, 5, 383-387.

Boswijk, H.P. and P.H. Franses, (1996), Unit Roots in Periodic Autoregressions, *Journal of Time Series Analysis*, 17, 221-245.

Clark, T.E., and M.W., McCracken (2005), Evaluating Direct Multi-Step Forecasts, *Econometric Reviews*, 24, 369-404.

Corradi, V. and N.R. Swanson, (2002), A Consistent Test for Out of Sample Nonlinear Predictive Ability, *Journal of Econometrics,* 110, 353-381.

Corradi, V. and N.R. Swanson, (2006a), Predictive Density Evaluation, in: *Handbook of Economic Forecasting*, eds. Clive W.J. Granger, Graham Elliot and Allan Timmermann, Elsevier, Amsterdam, pp. 197-284.

Corradi, V. and N.R. Swanson, (2006b), Predictive Density and Conditional Confidence Intervals Accuracy Tests, *Journal of Econometrics*, 135, 187-228.

Corradi, Valentina and Norman R. Swanson, (2007a), Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data, *Journal of Econometrics*, 136, 699-723.

Corradi, Valentina and Norman R. Swanson, (2007b), Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes, *International Economic Review*, 48, ??-??.

Diebold, F.X., and R.S. Mariano, (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Franses, P.H., (1996), *Periodicity and Stochastic Trends in Economic Time Series*, Oxford, Oxford University Press.

Franses, P.H., (1996), Recent Advances in Modelling Seasonality, *Journal of Economic Surveys*, 10, 299-345.

Franses, P.H., (2001), Some Comments on Seasonal Adjustment, *Revista de Economia del Rosaro*, 4, 9-16.

Franses, P. H., P. De Bruin, and D. van Dijk (2002), Seasonal Smooth Transition Autoregression, Econometric Institute Report 2000-06/A, Erasmus University.

Franses, P.H., H. Hoek, and R. Paap, (1997), Bayesian Analysis of Seasonal Unit Roots and Seasonal Mean Shifts, *Journal of Econometrics,* 78, 359-380.

Franses, P.H. and R. Paap, (2002), Forecasting with Periodic Autoregressive Time Series Models, in: *A Companion to Economic Forecasting,* eds. M.P. Clements and D.F. Hendry, Oxford, Blackwell Publishers, pp. 432-452.

Franses, P.H. and D. van Dijk, (2005), The Forecasting Performance of Various Models for Seasonality and Nonlinearity for Quarterly Industrial Production, *International Journal of Forecasting, 21, 87-102.*

Franses, P.H. and T. Vogelsang, (1998), On Seasonal Cycles, Unit Roots, and Mean Shifts, *Review of Economics and Statistics*, 80, 231-240.

Ghysels E. and D.R. Osborn, (2001), *The Econometric Analysis of Seasonal Time Series*, Cambridge, MA, Cambridge University Press.

Ghysels E., H.S. Lee and J. Noh, (1994), Testing for Unit Roots in Seasonal Time Series - Some Theoretical Extensions and A Monte Carlo Investigation, *Journal of Econometrics, 62, 415-442.*

Granger C.W.J. and T. Teräsvirta, (1993), *Modelling Nonlinear Economic Relationships*, Oxford, Oxford University Press.

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

Hylleberg, S., (1992), *Modelling Seasonality*, Oxford, Oxford University Press.

Hylleberg, S., (1994), Modelling Seasonal Variation, in: *Nonstationary Time Series Analysis and Cointegration*, eds. C.P. Hargreaves, Oxford, Oxford University Press.

Hylleberg, S., R.F. Engle, C.W.J. Granger and B.S. Yoo, (1990), Seasonal Integration and Cointegration, *Journal of Econometrics,* 44, 215-238.

Kilian, L., (1999), Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?, *Journal of Applied Econometrics*, 14, 491-510.

McCracken. M.W. and S. Sapp, (2004), Evaluating the Predictive Ability of Exchange Rates Using Long Horizon Regressions: Mind your p's and q's. *Journal of Money, Credit and Banking*, forthcoming.

Miron, J.A., (1996), *The Economics of Seasonal Cycles*, Cambridge, MA, MIT Press.

Miron, J.A. and J.J. Beaulieu, (1996), What Have Macroeconomists Learned About Business Cycles from the Study of Seasonal Cycles, *Review of Economics and Statistics*, 78, 54-66.

Osborn, D.R., (2002), Unit Roots versus Deterministic Representations of Seasonality for Forecasting, in: *A Companion to Economic Forecasting,* eds. M.P. Clements and D.F. Hendry, Oxford, Blackwell Publishers, pp. 409-431.

Osborn, D.R., S. Heravi, and C.R. Birchenhall, (1999), Seasonal Unit Roots and Forecasts of Two-Digit European Industrial Production, *International Journal of Forecasting*, 15, 27-47.

Osborn, D.R. and P.M.M. Rodrigues, (2002), Asymptotic Distributions of Seasonal Unit Root Tests: A Unifying Approach, *Econometric Reviews,* 21, 221-241.

Paap, R., P.H. Franses, andP H. Hoek, (1997), Mean Shifts, Unit Roots, and Forecasting Seasonal Time Series, *International Journal of Forecasting, 13*, 357-36

Psaradakis, Z., (1997), Testing for Unit Roots in Time Series with Nearly Deterministic Seasonal Variation, *Econometric Reviews*, 16, 421-439.

Rodrigues, P.M.M., (2001), Near Seasonal Integration, *Econometric Theory, 17*, 70-86.

Rodrigues, P.M.M. and D.R. Osborn, (1999), Performance of Seasonal Unit Root Tests for Monthly Data, *Journal of Applied Statistics, 26*, 985-1004.

Swanson, N.R. and R. Urbach, (2005), Simulation and Prediction Evidence on the Usefulness of Seasonal Unit Root Models, Working Paper, Rutgers University.

Swanson, N.R. and R. Urbach, (2007), Simulation and Prediction Evidence on the Usefulness of Seasonal Unit Root Models: Additional Results, Working Paper, Rutgers University.

West, K., (1996), Asymptotic Inference About Predictive Ability, *Econometrica*, 64, 1067-1084.

Figure 1: Simulated CPI Levels and 1st Log Differences
Simulation Models Calibrated Using Monthly U.S. Data for the Period 1991-2004

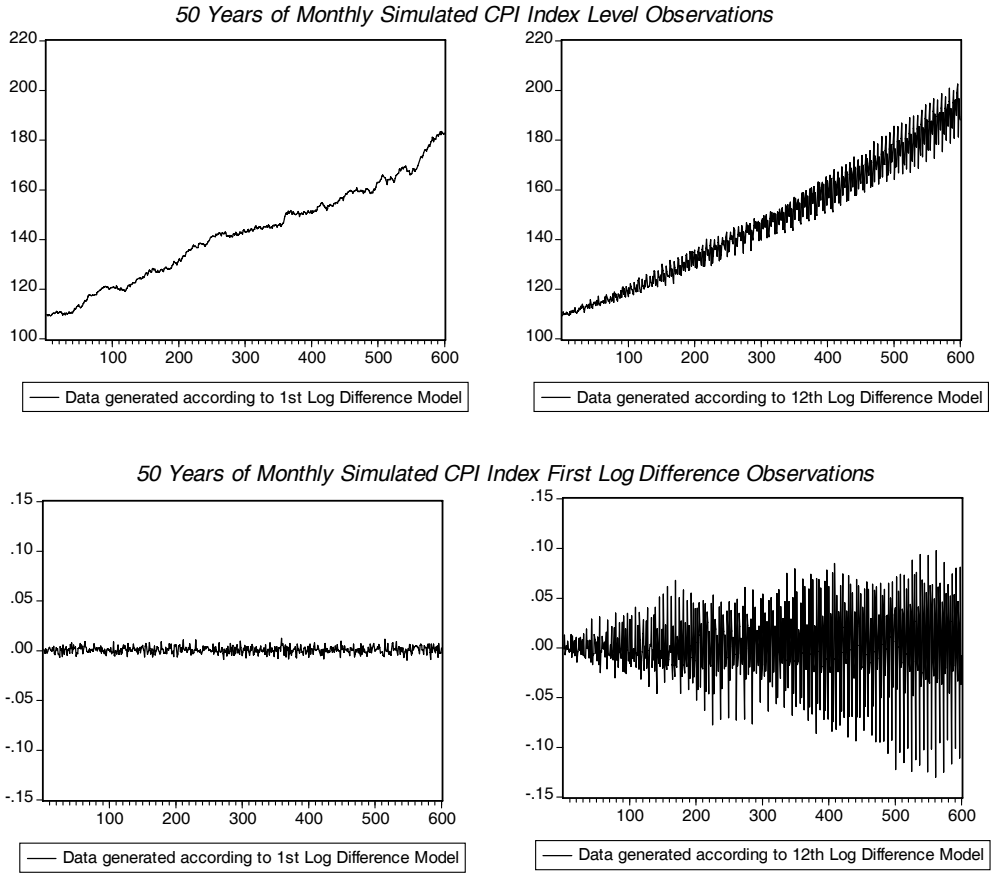

Figure 2: Macroeconomic Variables - Growth Rates for the Period 1959:1-2005:12

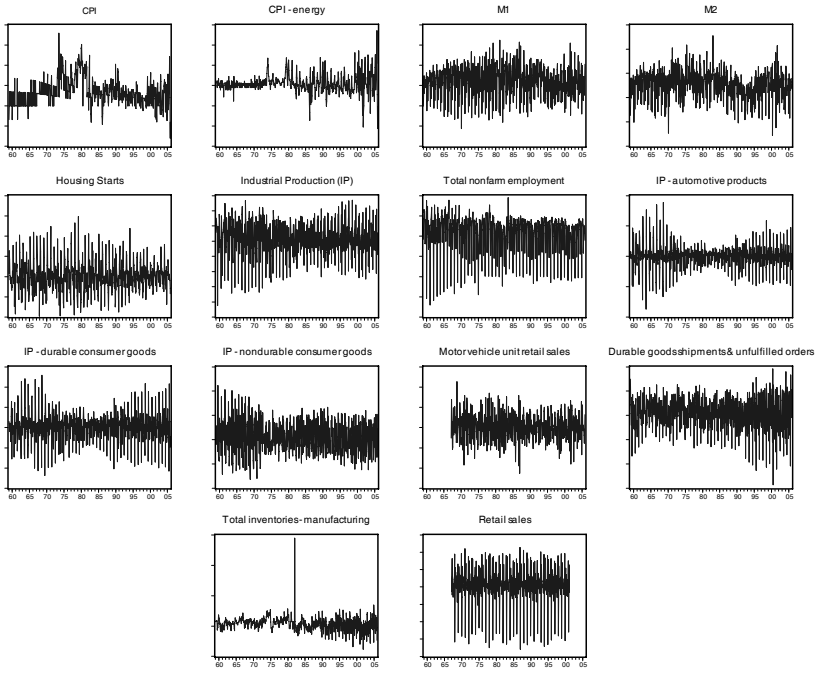

Figure 3: Macroeconomic Variables - Estimated Spectra

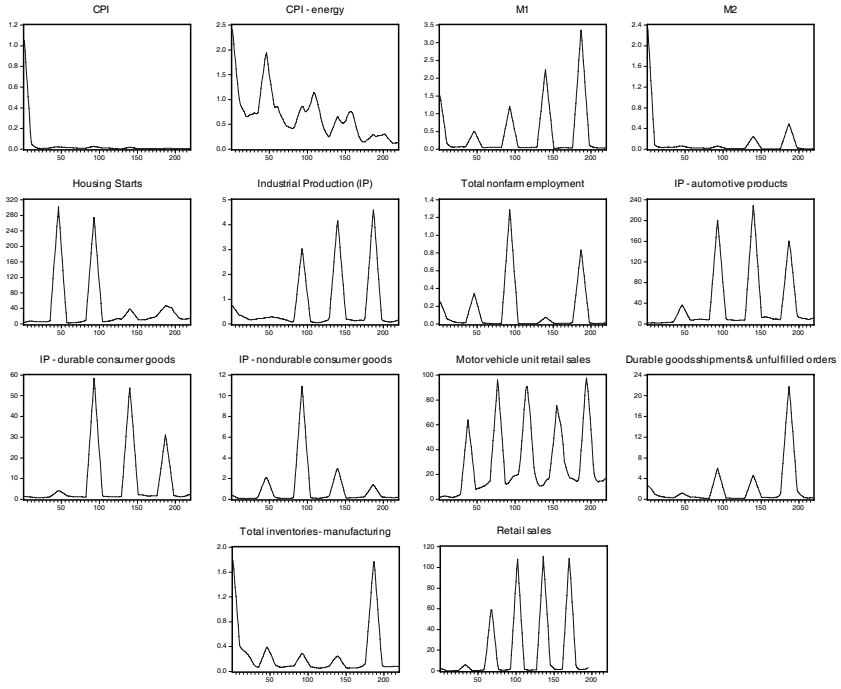

Table 1: Seasonality Evidence - Diagnostic Statistics for Macroeconomic Variables (*)

Series	*Mean*	*Stnd Dev*	*Skewness*	*Kurtosis*	*Jarque Bera*	*ADF*	*BF*	*HEGY*1	*HEGY*2	*DUMMY_LR*
CPI1	0.0041	0.0032	0.9168	4.2301	83.488	-1.31*	15.20	-7.45	-6.99	6.60*
CPI2	0.0042	0.0165	0.4440	5.5488	124.76	-0.22*	6.84*	-8.39	-7.41	6.39*
M1	0.0045	0.0158	-0.4169	2.8838	12.134	-1.16*	5.79*	-3.99	-4.21	142.0*
M2	0.0058	0.0061	-0.3555	3.5043	13.014	-2.08*	29.97	-4.53	-4.67	85.3*
H_Start	0.0023	0.1610	0.5117	3.9207	32.452	-4.29	1.66*	-7.49	-7.18	105.0*
IP	0.0023	0.0222	-0.5048	3.5371	22.393	-1.31*	51.84	-5.63	-5.59	138.7*
NonF_Emp	0.0017	0.0086	-1.5923	4.9595	239.43	-1.54*	158.69	-3.64	-3.77	383.0*
IP_Auto	0.0026	0.1366	-0.1845	6.6850	234.87	-0.76*	3.40*	-2.50*	-2.65*	36.4*
IP_Dur	0.0027	0.0694	-0.4980	4.7093	67.019	-0.86*	0.21*	-2.59*	-2.66*	82.8*
IP_NDur	0.0016	0.0251	0.2133	2.5401	6.7389	-3.18	187.14	-4.68	-4.35	61.0*
D_Ship	0.0045	0.0344	0.7707	4.0412	59.249	-1.94*	6.36*	-4.26	-5.37	112.8*
Invent	0.0044	0.0120	3.5928	48.027	35604	-1.91*	83.15	-5.27	-4.99	20.4*
Veh_Sales	-0.0002	0.1373	0.0748	3.0104	0.3470	-1.63*	3.27*	-5.74	-6.26	34.9*
Ret_Sales	0.0062	0.1126	-1.2410	5.3646	201.24	-2.37*	18.89	-4.27	-4.58	813.2*

(*) Various summary measures for the 14 series examined in the paper are contained in this table. The series consist of: two money stock variables (M1 and M2); two CPI series including CPI for all urban consumers, all items (CPI1), and energy (CPI2); housing starts (H_Start); industrial production (IP); total nonfarm employment (NonF_Emp); industrial production - automotive products (IP_Auto); industrial production - durable consumer goods (IP_Dur); industrial production - durable consumer goods (IP_NDur); durable goods shipments of new orders and unfilled orders (D_Ship); total inventories, manufacturing (Invent); retail sales (Ret_sales) - for the period 1967:1-2001:4; and motor vehicle unit retail sales (Veh_sales) - for the period 1967:1-2005:12. Data are monthly U.S. figures for the period 1959:1-2005:12, except where noted. Mean, Stnd Dev, Skewness, Kurtosis, Jarque-Bera, and Dummy_LM statistics are calculated using log differenced data, while unit root tests are carried out using logged data. For the unit root tests, starred entries represent cases where the null of unit root fails to reject at a 5% nominal level. HEGY test statistics are reported for seasonal frequency π_2 (see above for further discussion), and starred entries denote a failure to reject the seasonal unit root null at a 5% level. The last column reports likelihood ratio statistics for testing the null hypothesis that $\theta_s = \theta, s = 1 \cdots 12$ in the standard deterministic dummay variable model (i.e. see model DS above). Starred entries indicate rejection of the null at 5% level. Further details are contained in Section 2.

Table 2: Monte Carlo Results: MSFE Comparison, DM Predictive Accuracy Analysis, and Seasonality Test Performance (*)

DGP	*Point MSFE comparison*				*DM Test Win Frequency*								*Seasonality Test Size*		
$(\theta, \sigma_\varepsilon, \phi)$	$h=1$	$h=3$	$h=12$	$h=60$	$h=1$		$h=3$		$h=12$		$h=60$		BF	$HEGY1$	$HEGY2$
					SUR	RW	SUR	RW	SUR	RW	SUR	RW			
$(-0.6, 0.005, 0.0)$	1.00	1.00	0.87	0.67	1.00	0.00	1.00	0.00	0.30	0.00	0.07	0.05	0.64	0.46	0.45
$(-0.9, 0.005, 0.0)$	1.00	1.00	0.70	0.20	1.00	0.00	0.97	0.00	0.11	0.05	0.02	0.27	0.34	1.00	1.00
$(-0.6, 0.01, 0.0)$	1.00	1.00	0.90	0.80	1.00	0.00	1.00	0.00	0.32	0.00	0.10	0.05	0.61	0.38	0.36
$(-0.9, 0.01, 0.0)$	1.00	0.97	0.70	0.30	0.97	0.00	0.95	0.00	0.14	0.11	0.04	0.19	0.34	1.00	1.00
$(-0.6, 0.1, 0.0)$	1.00	1.00	0.90	0.87	1.00	0.00	1.00	0.00	0.35	0.01	0.17	0.04	0.54	0.41	0.39
$(-0.9, 0.1, 0.0)$	1.00	0.95	0.75	0.25	1.00	0.00	0.94	0.00	0.15	0.07	0.03	0.26	0.30	1.00	1.00
$(-0.6, 0.005, 0.5)$	1.00	1.00	0.86	0.76	1.00	0.00	1.00	0.00	0.26	0.01	0.17	0.07	0.58	0.34	0.33
$(-0.9, 0.005, 0.5)$	1.00	0.97	0.57	0.27	1.00	0.00	0.96	0.00	0.17	0.12	0.02	0.32	0.35	1.00	1.00
$(-0.6, 0.01, 0.5)$	1.00	1.00	0.82	0.71	1.00	0.00	1.00	0.00	0.31	0.02	0.11	0.10	0.62	0.38	0.36
$(-0.9, 0.01, 0.5)$	1.00	0.98	0.62	0.22	1.00	0.00	0.98	0.00	0.22	0.12	0.02	0.26	0.34	1.00	1.00
$(-0.6, 0.1, 0.5)$	1.00	1.00	0.90	0.84	1.00	0.00	1.00	0.00	0.28	0.00	0.14	0.02	0.58	0.42	0.43
$(-0.9, 0.1, 0.5)$	1.00	0.99	0.50	0.21	1.00	0.00	0.97	0.00	0.10	0.18	0.02	0.25	0.33	1.00	1.00
$(-0.6, 0.005, -0.5)$	1.00	1.00	0.86	0.76	0.95	0.00	1.00	0.00	0.24	0.01	0.16	0.06	0.62	0.32	0.35
$(-0.9, 0.005, -0.5)$	1.00	0.96	0.66	0.36	1.00	0.00	0.95	0.00	0.16	0.12	0.06	0.31	0.41	1.00	1.00
$(-0.6, 0.01, -0.5)$	1.00	1.00	0.94	0.74	1.00	0.00	1.00	0.00	0.34	0.01	0.14	0.03	0.60	0.37	0.39
$(-0.9, 0.01, -0.5)$	1.00	0.97	0.67	0.19	1.00	0.00	0.96	0.00	0.11	0.12	0.02	0.32	0.33	1.00	1.00
$(-0.6, 0.1, -0.5)$	1.00	1.00	0.84	0.68	1.00	0.00	1.00	0.00	0.24	0.01	0.18	0.04	0.57	0.34	0.37
$(-0.9, 0.1, -0.5)$	1.00	0.96	0.49	0.17	1.00	0.00	0.95	0.01	0.09	0.11	0.07	0.24	0.30	1.00	1.00

(*) Results based upon a (i) series of prediction experiments, and (ii) in-sample HEGY testings are reported in this table. With regard to the prediction experiments, in each simulation a sample of 500 observations was drawn, and prediction models were recursively estimated, begininng with an in-sample period of $R = 120 - h + 1$ observations, and yielding a sequence of 380 ex-ante h-step ahead forecasts. Results reported in the first 12 columns of numerical entries in the table are based upon examination of these forecasts. Namely, the first 4 columns report the proportion of simulations for which the SUR model predictions were "MSFE-better" than those based on the RW model. The next 8 columns report the proportion of simulations for which the two models yielded "MSFE-better" predictions based upon application of the Diebold-Mariano (1995) predictive accuracy test at a 5% nominal significance level, based on MSFE loss (see Section 4 for further details). Entires in the last three columns of the table indicate the proportion of times that the seasonal unit root (HEGY) and BF tests reject the null (see Table 1 and Section 2 for further details). All results are based on 500 Monte Carlo simulations.

Table 3: Distributional Accuracy Tests Based on the Comparison of Historical and Simulated Data - Benchmark Model is RW (*)

Series	S,l	Z	*Crit Val* (Z^*)		*Crit Val* (Z^{**})		*CS Distributional Loss*			
			10%	5%	10%	5%	RW	DS	SUR	PAR
CPI1	5T,4	-0.0031	2.893	2.9863	2.8997	2.9954	1.5634	4.7753	2.5993	1.5665
	10T,12	-0.0048	2.6184	2.8337	2.6068	2.8256	1.5631	4.9371	2.6047	1.5679
CPI2	5T,4	-0.0067	0.2552	0.2652	0.2518	0.2673	1.1283	1.1412	1.3952	1.1351
	10T,12	-0.0022	0.2602	0.2676	0.2613	0.2710	1.1285	1.1446	1.3959	1.1307
M1	5T,4	0.0174	0.6724	0.704	0.6696	0.6990	2.2174	2.2000	2.9743	2.2013
	10T,12	0.0167	0.6774	0.7037	0.6729	0.7040	2.2167	2.2000	2.9743	2.2005
M2	5T,4	0.0046	0.8092	0.9512	0.8069	0.9393	1.8239	2.4718	3.1176	1.8193
	10T,12	0.0042	0.9044	1.0065	0.9025	1.0039	1.8234	2.4716	3.157	1.8192
H_Start	5T,4	0.0077	0.0627	0.0796	0.0658	0.0814	1.9628	1.9551	2.0481	1.9570
	10T,12	0.0104	0.0784	0.0859	0.0789	0.0877	1.9650	1.9546	2.0514	1.9575
IP	5T,4	0.0151	0.3824	0.4107	0.3845	0.4119	2.0792	2.0677	2.5311	2.0641
	10T,12	0.0167	0.4067	0.4172	0.4059	0.4183	2.0808	2.0667	2.5327	2.0641
NonF_Emp	5T,4	0.1401	0.6085	0.6974	0.5943	0.6931	2.0522	1.9366	2.8813	1.9121
	10T,12	0.1237	0.6524	0.7127	0.6629	0.7195	2.0363	1.9397	2.8811	1.9126
IP_Auto	5T,4	0.0231	0.1317	0.1372	0.1317	0.1358	1.3325	1.3291	1.4475	1.3094
	10T,12	0.0198	0.11	0.1167	0.1098	0.1153	1.3288	1.3299	1.4159	1.309
IP_Dur	5T,4	0.0217	0.2924	0.3102	0.2938	0.3075	1.8412	1.8301	2.1449	1.8195
	10T,12	0.0219	0.3205	0.3381	0.3201	0.3388	1.8421	1.8307	2.1624	1.8202
IP_NDur	5T,4	0.0015	0.3974	0.4088	0.4014	0.407	2.2186	2.2929	2.6528	2.2171
	10T,12	0.0018	0.3969	0.4051	0.3967	0.4062	2.2195	2.2971	2.6557	2.2177
D_Ship	5T,4	0.0058	0.2615	0.2807	0.2605	0.2777	1.8453	1.8419	2.1582	1.8395
	10T,12	0.0082	0.2638	0.2783	0.2668	0.2782	1.8466	1.8408	2.1629	1.8384
Invent	5T,4	-0.0178	0.5438	0.5506	0.5426	0.5497	0.7184	0.8054	1.2663	0.7362
	10T,12	-0.0186	0.5291	0.5354	0.5292	0.5354	0.7183	0.7976	1.2663	0.7369
Ret_Sales	5T,4	0.1019	0.3401	0.3461	0.3411	0.353	1.9304	2.0016	2.2441	1.8285
	10T,12	0.1004	0.3368	0.3507	0.3366	0.3502	1.9291	2.0002	2.2407	1.8286
Veh_Sales	5T,4	0.0011	0.0556	0.0627	0.0538	0.0623	1.7596	1.7616	1.8359	1.7585
	10T,12	0.0006	0.0495	0.0553	0.0502	0.0553	1.7590	1.7606	1.8259	1.7584

(*) $Z_{T,S}$ test statistics (called Z in the table), and associated distributional loss measures (denoted *CS Distributional Loss*) are reported, where S denotes the length of the simulated data series used in test statistic construction, and T is the historical sample length, assumed to be the entire historical sample period (see footnote to Table 1). Note that this historical period is also the period used to estimate the models. As usual, the models are denoted by RW, DS, SUR, and PAR (see Section 2 for further details). The $Z_{T,S}$ test is designed to facillitate selection amongst alternative simulation models via comparison of simulated and historical distribtutions In the test, the benchmark model is RW, against which all other model are compared. The null hypothesis corresponds to the case where no alternative model outperforms the benchmark. Critical values that are constructed both assuming that T/S approaches $\gamma > 0$ (Z^{**}) and assuming that T/S approaches 0 (Z^*), as T and S increase are reported in the 3th through 6th columns of entries (see Section 2 and Corradi and Swanson (2007a) for complete details). Further, l is the bootstrap block length, and all statistics are based on a grid of 20x20 values for u, distributed uniformly across the historical data range. Bootstrap empirical distributions are constructed using 100 bootstrap replications.

Table 4: Predictive Accuracy Test Results for Various Macroeconomic Variables (*)

	$h=1$			$h=3$			$h=12$			$h=60$		
Series	*DS*	*SUR*	*PAR*	*DS*	*SUR*	*PAR*	*DS*	*SUR*	*PAR*	*DS*	*SUR*	*PAR*
Panel A: MSFE for DS, SUR and PAR, Relative to the RW Model												
CPI1	1.24	1.17	1.35	0.98	2.41	1.27	0.99	2.35	1.25	0.98	1.74	1.10
CPI2	0.88	1.68	2.37	1.07	3.28	1.04	1.03	3.23	1.03	1.04	2.23	1.02
M1	1.19	0.53	0.57	1.03	3.28	0.45	2.38	3.73	0.45	2.34	2.36	0.45
M2	1.97	0.88	0.85	1.57	5.69	0.75	9.05	5.61	0.78	9.55	2.53	0.76
H_Start	1.57	0.98	0.46	5.49	1.58	0.43	2.00	2.50	0.43	2.05	1.57	0.43
IP	1.26	0.53	0.79	1.27	3.90	0.67	1.91	3.32	0.65	2.00	2.97	0.65
NonF_Emp	9.69	0.14	0.55	1.09	3.77	0.36	1.37	3.53	0.36	1.47	2.07	0.37
IP_Auto	0.98	0.39	2.94	1.05	1.66	3.65	1.06	2.13	3.67	1.10	2.38	3.67
IP_Dur	0.94	0.38	1.55	1.01	2.24	1.42	1.45	2.50	1.45	1.41	2.24	1.45
IP_NDur	2.20	0.56	2.5	1.37	2.19	2.06	1.58	2.58	2.11	1.65	2.15	2.11
D_Ship	1.05	0.52	0.43	1.26	2.95	0.46	1.01	2.99	0.46	1.00	2.12	0.46
Invent	2.38	1.65	1.18	1.2	3.86	0.85	1.28	4.21	0.95	1.43	2.48	0.94
Ret_Sales	7.87	0.08	0.06	2.75	0.77	0.07	2.55	1.58	0.06	2.61	2.64	0.06
Veh_Sales	1.07	1.67	1.55	2.85	2.36	1.20	1.47	2.49	1.23	1.63	2.39	1.23
Panel B: DM Predictive Accuracy Test Statistics - Benchmark is the RW Model												
CPI1	-1.51	-1.36	-2.26*	0.77	-3.81*	-2.62*	0.33	-4.67*	-2.57*	0.85	-3.53*	-1.64
CPI2	2.01*	-2.77*	-3.79*	-1.82	-3.53*	-1.11	-1.55	-3.05*	-0.97	-1.48	-3.16*	-0.62
M1	-4.2*	3.07*	5.29*	-3.85*	-5.23*	4.11*	-8.33*	-4.39*	4.07*	-8.22*	-4.46*	4.08*
M2	-7.51*	0.73	1.64	-6.01*	-5.19*	3.07*	-2.16*	-3.95*	2.68*	-1.89	-4.32*	3.00*
H_Start	-4.15*	0.37	5.8*	-5.26*	-2.81*	6.04*	-5.68*	-5.78*	6.04*	-5.75*	-3.75*	6.04*
IP	-5.12*	3.87*	2.33*	-6.24*	-5.1*	3.42*	-7.95*	-4.82*	3.73*	-8.35*	-5.18*	3.73*
NonF_Emp	-3.96*	7.37*	6.65*	-6.36*	-7.96*	6.88*	-6.16*	-5.14*	6.85*	-6.08*	-7.18*	6.82*
IP_Auto	0.89	4.37*	-5.63*	-4.78*	-3.22*	-5.51*	-5.83*	-6.63*	-5.59*	-5.23*	-4.6*	-5.59*
IP_Dur	2.29*	4.68*	-3.75*	-3.04*	-5.16*	-2.89*	-6.86*	-7.87*	-3.10*	-6.65*	-5.34*	-3.11*
IP_NDur	-9.32*	6.09*	-4.81*	-4.99*	-4.02*	-4.63*	-4.45*	-4.79*	-4.78*	-6.03*	-5.34*	-4.79*
D_Ship	-2.98*	4.45*	6.17*	-5.19*	-5.25*	6.82*	-1.82	-8.5*	6.82*	0.00	-6.03*	6.82*
Invent	-3.48*	-0.89	-1.83	-2.3*	-2.46*	2.48*	-2.09*	-1.98*	1.41	-2.09*	-2.25*	1.55
Ret_Sales	-2.62*	4.52*	4.59*	-4.07*	1.95	4.6*	-8.13*	-5.29*	4.61*	-8.19*	-5.28*	4.61*
Veh_Sales	-2.85*	-2.89*	-2.89*	-4.09*	-2.61*	-2.16*	-4.92*	-2.57*	-2.36*	-5.87*	-3.36*	-2.36*

(*) MSFEs and DM predictive accuracy test statistics based on MSFE loss are reported in the two panels of this table. All DM tests are pairwise, and compare the benchmark *RW* model with the model denoted in the column header to the table. Negative values for DM statistics indicate that the point MSFE associated with the benchmark model is lower than that for the other model. Starred DM test statistics indicate rejection of the predictive accuracy null using 5% nominal size critical values (see Section 2 for further details). Prediction models are constructed recursively, starting with $R = 120\text{-}h\text{+}1$ observations, and ending with $R = T - h$ observations, where T is the sample size, and h is the forecast horizon (set equal to 1, 3, 12, and 60 months ahead). For variable definitions, refer to Table 1.