Journal of Time Series Econometrics

Volume 2, *Issue* 1 2010 *Article* 4

Signal Extraction Revision Variances as a Goodness-of-Fit Measure

Tucker McElroy[*] Marc Wildi[†]

[*] U.S. Census Bureau, tucker.s.mcelroy@census.gov

[†] Institute of Data Analysis and Process Design, wlmr@zhaw.ch

Signal Extraction Revision Variances as a Goodness-of-Fit Measure[*]

Tucker McElroy and Marc Wildi

Abstract

Typically, model misspecification is addressed by statistics relying on model-residuals, i.e., on one-step ahead forecasting errors. In practice, however, users are often also interested in problems involving multi-step ahead forecasting performances, which are not explicitly addressed by traditional diagnostics. In this article, we consider the topic of misspecification from the perspective of signal extraction. More precisely, we emphasize the connection between models and real-time (concurrent) filter performances by analyzing revision errors instead of one-step ahead forecasting errors. In applications, real-time filters are important for computing trends, for performing seasonal adjustment or for inferring turning-points towards the current boundary of time series. Since revision errors of real-time filters generally rely on particular linear combinations of one- and multi-step ahead forecasts, we here address a generalization of traditional diagnostics. Formally, a hypothesis testing paradigm for the empirical revision measure is developed through theoretical calculations of the asymptotic distribution under the null hypothesis, and the method is assessed through real data studies as well as simulations. In particular, we analyze the effect of model misspecification with respect to unit roots, which are likely to determine multi-step ahead forecasting performances. We also show that this framework can be extended to general forecasting problems by defining suitable artificial signals.

KEYWORDS: model-diagnostics, nonstationary time series, real-time filtering, seasonality, signal extraction

[*] We are indebted to the careful comments of an anonymous referee.

1 Introduction

Generally speaking, time series models of economic data are misspecified, since in essence models are simplified portraits of underlying stochastic dynamics. The task of model diagnostics is then to identify mismatches pertinent to the goals of a particular analysis, so that faulty models can be improved accordingly. Not all such mismatches are desiderata, but rather only those that are relevant for a particular analysis; diagnostic tools should account for the purpose of a particular application by emphasizing model failures that are likely to adversely affect results. Traditional diagnostics in time series analysis focus on one-step ahead forecasting errors. Typical examples are (partial) autocorrelation functions of model residuals, as well as Ljung-Box (Ljung and Box, 1978) and Box-Pierce statistics. If the purpose of a particular application is short term one-step ahead forecasting, then these tools are appropriate. Yet sometimes there is interest in forecasting a time series over a longer horizon, and therefore the performance of a model over multiple forecast leads is more important than the modeling of short-term behavior. Model-based signal extraction, which implicitly utilizes multi-step ahead forecasting, is also more concerned with the long-term forecasting behavior of putative models.

Signal extraction is concerned with the definition and estimation of interesting components of a time series. In practice signal estimation for the concurrent time period, i.e., filtering or real-time estimation, are important because of the need for timely information (Findley, Bell, Monsell, Otto, and Chen, 1998). Unfortunately, symmetric filters cannot be used directly because future data hasn't been observed yet. Traditional methods overcome this difficulty by expanding series on both ends of the sample by backcasts and forecasts generated by a time series model – typically an AutoRegressive Integrated Moving Average (ARIMA) model – so that the symmetric filter can be used. If the coefficients of the symmetric filter decay slowly, then forecasts of longer horizons are emphasized. Therefore it is desirable – towards the end of producing acceptable concurrent signal extraction estimates – for a model to perform well with respect to all forecasting horizons simultaneously. The following example illustrates a challenging modeling problem for traditional diagnostics such as the Ljung-Box (LB) statistics[1].

Wildi (2008) compares the real-time – or concurrent – performance of several different signal extraction procedures, in a context where the general goal of analysis is to produce useful leading indicators. The so-called KOF Economic

[1] Letting $\widehat{\rho}_j$ denote the sample autocorrelation function of the prediction residuals obtained from a fitted model, the LB statistic is then defined via the formula $Q_h = n(n+2)\sum_{j=1}^{h} \widehat{\rho}_j^2/(n-j)$, where n is the sample size of the time series (Brockwell and Davis, 1991).

Barometer[2] is a monthly time series that combines several economic indicators related to banking, production, and housing, and is used as a leading indicator of Swiss GDP. Many of its constituent indicators are time series that are bounded by construction in the interval $[-100, 100]$. For purposes of illustration, we examine the series "Industry total: expected production" from the KOF data-base; this series measures projected industrial production, and is sometimes used as a proxy for Swiss Gross Domestic Product. Referring to this as Series 31 – displayed in Figure 1 with a solid line, TRAMO[3] selects the following Airline model

$$(1-B)(1-B^{12})X_t = (1-0.662B)(1-0.824B^{12})\varepsilon_t, \tag{1}$$

after adjustments for outliers and calendar effects. Here we utilize the notation B for the backshift operator, with X_t representing the time series under consideration, while ε_t is a white noise sequence. The Airline model is defined in Box and Jenkins (1976), and calendar effects and other pre-adjustment aspects of time series analysis are discussed at length in Findley, Bell, Monsell, Otto, and Chen (1998).

TRAMO represents a fairly conventional, state-of-the-art approach to time series modeling, including the estimation of fixed regression effects (e.g., additive outliers, level shifts, holiday regressors, trading day effects, etc.) as well as unit-root testing and Seasonal ARIMA (SARIMA) model identification procedures. We are mainly interested in this latter aspect of TRAMO, which is an automated model-selection procedure (although the user can intervene) that is discussed further in Maravall and Caporello (2004). Typical output of TRAMO includes model diagnostics such as the LB statistics; the results of such diagnostics for Series 31 can be seen in Figure 2.

Standard model assumptions are met; neither the autocorrelation nor the partial autocorrelation function nor the LB statistics suggest significant departures from the null hypothesis[4] that the Airline model is correctly specified. However, a simulation of the process defined by (1), plotted as the dotted line in Figure 1, highlights the fairly simple observation that Series 31 does not appear to be nonstationary. Indeed, the real series lacks the strong trend component that is typical of twice-integrated processes such as the Airline process.

Because Series 31 is *bounded* – as are many important economic time series (e.g., unemployment rates) – we expect a stationary model to be selected for Series

[2] Konjunkturforschungsstelle der ETH, or Institute of Business Cycle Research; see *www.kof.ethz.ch*.

[3] The TRAMO-SEATS for Windows (TSW) package is a widely-used software program for the seasonal adjustment of economic time series, and can be downloaded from the Bank of Spain (http://www.bde.es/servicio /software /econome). Maravall and Caparello (2004) is the most current documentation of both the program TSW and the seasonal adjustment method utilized therein.

[4] TRAMO provides additional diagnostic tools – such as heteroscedasticity and model stability tests – that did not either lead to a rejection of the specified model.

DOI: 10.2202/1941-1928.1012

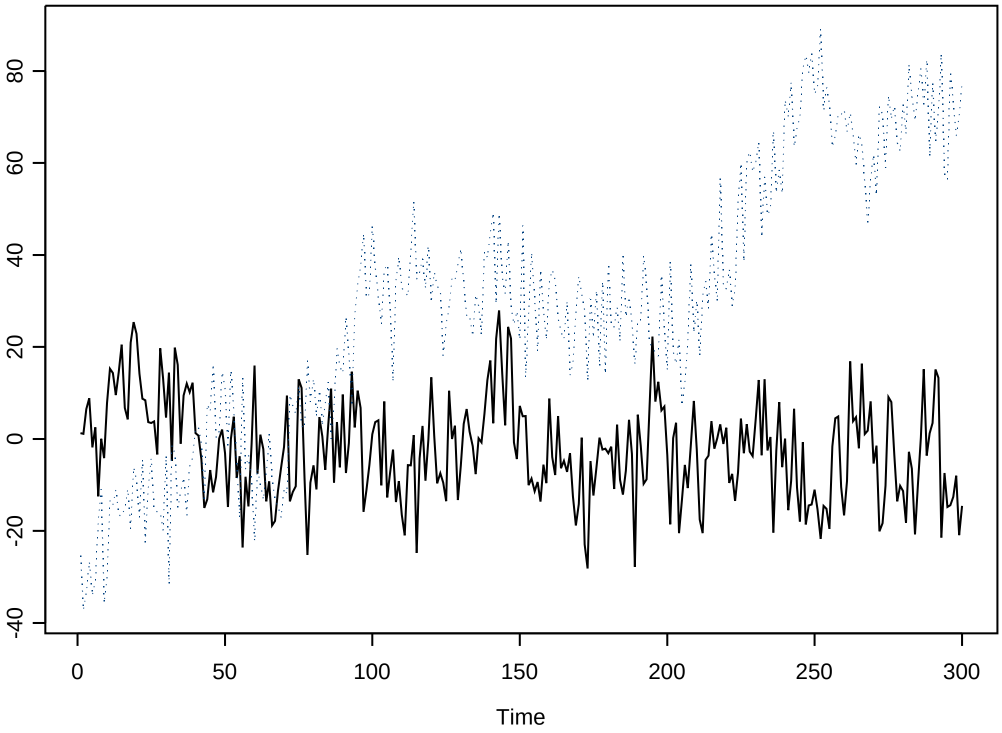

Figure 1: Series 31, a leading indicator that is incorporated in the KOF Economic Barometer, is displayed as a solid line against Time in months. The dotted lines represent a simulation from the model that was fitted to Series 31 by using TSW.

31 by TRAMO. This expectation is further bolstered upon surveying its sample Autocorrelation Function (ACF) plot (Figure 3), which shows no indications of either trend or seasonal nonstationarity behavior. Therefore the model adequacy indicated by TRAMO is quite disappointing[5].

The failure of conventional model diagnostics to reject the Airline model for Series 31 (with similar results for 35 other indicators that make up the KOF Economic Barometer) forms the central motivation of this paper. Now from a one-step ahead forecasting perspective the above model (1) performs well, thus confirm-

[5] This is not intended as a criticism of TRAMO, which utilizes the most recent advances in unit-root testing; similar results are also obtained with the automatic model selection procedure of X-12 ARIMA, the seasonal adjustment program of the U.S. Census Bureau (see Findley et al. (1998)).

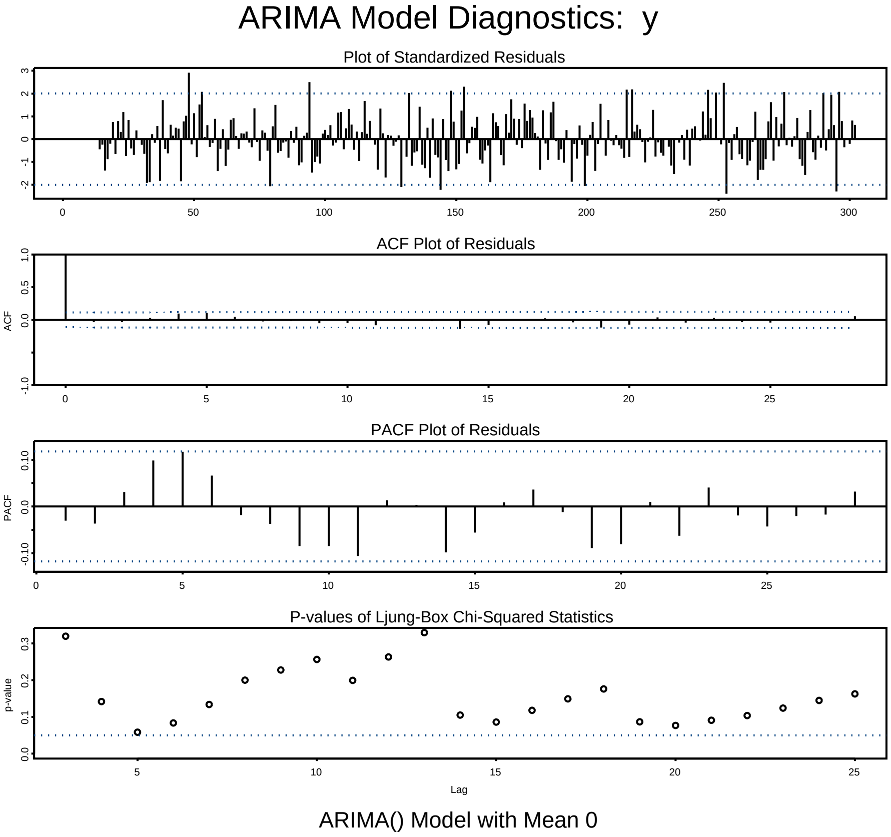

Figure 2: Model Diagnostic plots for Series 31, a leading indicator that is incorporated in the KOF Economic Barometer, generated by using TSW for the fitted Airline model. The top panel gives the time plot of the standardized model residuals, while the second panel gives the autocorrelation plot of such. The third panel is the partial autocorrelation plot of the model residuals, and finally the bottom panel has the p-values for the LB statistic (utilizing the asymptotic χ^2 distribution) at various lags h. The first three plots also have 95% confidence bands displayed as dotted lines, computed under the null hypothesis that the displayed series is white noise (i.e., the model is correctly specified), while the last plot has a dotted line at the 5% level – values above this indicate failure to reject the null hypothesis of uncorrelated model residuals.

ing the usefulness of traditional diagnostics for short-term projections. But for the application of real-time signal extraction, the multi-step ahead forecasting performance of a model is highly pertinent, and therefore one needs model diagnostics that can make identification discernments across a longer future horizon. We ar-

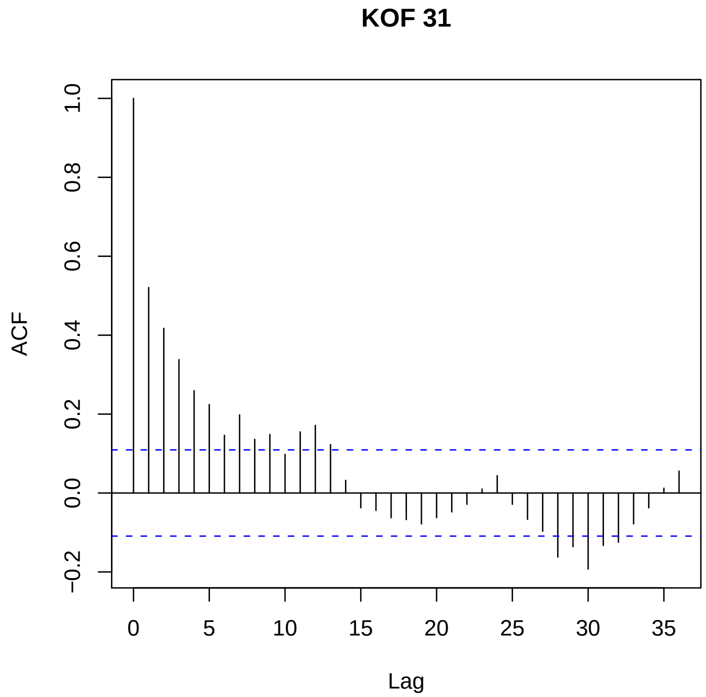

Figure 3: Sample ACF plot for Series 31, with lag displayed on the x-axis. The dotted lines represent 95% confidence bands under the assumption that the series is not serially correlated.

gue that specific diagnostics are needed that match the real-time signal extraction problem.

We propose to use signal extraction revision variances to assess the performance of a model over a long forecasting horizon, since signal extraction formulas implicitly utilize multi-step ahead forecasts. Signal extraction revisions and their variances have been studied for quite some time – see Pierce (1980), Maravall (1986), and Maravall and Caparello (2004) – but usually from the perspective that a semi-infinite sample of data extending into the infinitely remote past is available. The revision variance will tend to be unusually large when signal estimates are generated from faulty models, because abnormally large revisions will tend to occur in this case. Exact revision variances can be generated through model-based calculations, as described at length in McElroy and Gagnon (2008); these revision variances should coincide asymptotically with empirical revision sample variances if the model is correct. Therefore, a diagnostic test based on a comparison of revision variances should match the signal extraction problem. Note that such a test involves one- and multi-step ahead model forecast performance. The main contribution of the paper is the proposal of a new test statistic, denoted by RV, that matches the signal extraction problem and a derivation of its distribution under the hypothesis that the model fits the Data Generating Process (DGP).

Note that better models could of course be developed for bounded series such as the KOF – for one thing, transformations such as the logistic or tangent could be used to handle the boundedness (although some distortion would presumably be involved). While acknowledging this, our main thesis is that conventional models deemed adequate by diagnostics based on one-step ahead forecasting criteria should be wrong on *a priori* grounds, since their unit root structure fails to allow for long-term mean reversion (i.e., turning points). Surely better models exist (and should be used when practicable), but our point is that the classical diagnostics – unit root tests together with acceptable LB statistics – are telling us that an $I(2)$ specification for Series 31 is adequate. By emphasizing multi-step ahead forecasting in the diagnostic phase, we may be able to obtain statistically significant rejections of such over-specified models[6].

In Section 2 we discuss some of the background theory needed for a finite sample approach to signal extraction in a model-based context. We define the goodness-of-fit test statistic RV, and discuss its important finite sample and asymptotic properties under the null hypothesis that the given model is correct. Section 3 gives some of the details on implementing our testing procedure, with a discussion of the decomposition, structural, and direct approaches to defining a signal. In Section 4 we apply these concepts to several real series where there is a suspicion of model misspecification on *a priori* grounds; the series include sectoral leading indices used in the KOF Economic Barometer, as well as manufacturing data. Section 5 concludes, and mathematical proofs are in the Appendix.

2 Theory

We begin with a background discussion on model-based signal extraction in a finite-sample context; then we discuss signal extraction revisions for such estimates, and their autocovariance structure is provided in Proposition 1. We then define our goodness-of-fit test statistic RV and determine its statistical properties. Section 2.1 below consists of background material taken wholly from McElroy (2008a);

[6] Over-specification of nonstationarity has a precise definition: if the differencing polynomial $\delta(B)$ is sufficient to reduce a series to stationarity, then specifying $\delta(B)\tau(B)$ as the differencing polynomial – where $\tau(B)$ is a polynomial of degree greater than one with all roots located on the unit circle – is a case of over-specification.

Sections 2.2 and 2.3 present new material, defining the proposed revision variance statistic RV, as well as giving its basic properties.

2.1 Background on Signal Extraction

We consider the additive decomposition of our data vector $Y = (Y_1, Y_2, \cdots, Y_n)'$ into signal S and noise N, via $Y = S + N$. The signal might be the trend component, while the noise includes the seasonal and irregular components. Following Bell (1984), we let Y_t be an integrated process such that $W_t = \delta(B)Y_t$ is stationary, where B is the backshift operator and $\delta(z)$ is a polynomial with all roots located on the unit circle of the complex plane. (Also, $\delta(0) = 1$ by convention.) This $\delta(z)$ is referred to as the differencing operator of the series, and we assume it can be factored into relatively prime polynomials $\delta^S(z)$ and $\delta^N(z)$ (i.e., polynomials with no common zeroes), such that the series

$$U_t = \delta^S(B)S_t \qquad V_t = \delta^N(B)N_t \tag{2}$$

are mean zero stationary time series that are uncorrelated with one another. Note that $\delta^S = 1$ and/or $\delta^N = 1$ are included as special cases (in these cases either the signal or the noise or both are stationary). We let d be the order of δ, and d_S and d_N are the orders of δ^S and δ^N; since the latter operators are relatively prime, $\delta = \delta^S \cdot \delta^N$ and $d = d_S + d_N$.

As in Bell and Hillmer (1988), we assume Assumption A of Bell (1984) holds for the component decomposition, and we treat the case of a finite sample with $t = 1, 2, \cdots, n$ with $n > d$. Assumption A states that the initial d values of Y_t, i.e., the variables $Y_* = (Y_1, Y_2, \cdots, Y_d)$, are independent of $\{U_t\}$ and $\{V_t\}$. For a discussion of the implications of this assumption, see Bell (1984) and Bell and Hillmer (1988).

Now we can write (2) in a matrix form, as follows. Let Δ be a $(n-d) \times n$ matrix with entries given by $\Delta_{ij} = \delta_{i-j+d}$ (the convention being that $\delta_k = 0$ if $k < 0$ or $k > d$).

$$\Delta = \begin{bmatrix} \delta_d & \cdots & \delta_1 & 1 & 0 & 0 & \cdots \\ 0 & \delta_d & \cdots & \delta_1 & 1 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \delta_d & \cdots & \delta_1 & 1 \end{bmatrix}$$

The matrices Δ_S and Δ_N have entries given by the coefficients of $\delta^S(z)$ and $\delta^N(z)$, but are $(n-d_S) \times n$ and $(n-d_N) \times n$ dimensional respectively. This means that each row of these matrices consists of the coefficients of the corresponding differencing polynomial, horizontally shifted in an appropriate fashion. Hence

$$W = \Delta Y \qquad U = \Delta_S S \qquad V = \Delta_N N$$

where W, U, V, S, and N are column vectors of appropriate dimension. Then it is possible to write the mean square linear optimal estimate $\widehat{S}$ as a linear matrix operating on Y, i.e., $\widehat{S} = FY$. The error covariance matrix, i.e., the covariance matrix of $\widehat{S} - S$, is denoted by M. The formulas for F and M are given by:

$$F = \left(\Delta_S' \Sigma_U^{-1} \Delta_S + \Delta_N' \Sigma_V^{-1} \Delta_N\right)^{-1} \Delta_N' \Sigma_V^{-1} \Delta_N \tag{3}$$

$$M = \left(\Delta_S' \Sigma_U^{-1} \Delta_S + \Delta_N' \Sigma_V^{-1} \Delta_N\right)^{-1} \tag{4}$$

where Σ_X denote the covariance matrix for any random vector X.

Now these basic notions are generalized slightly for the development needed below. We will be considering samples of varying dimension; denote the signal extraction matrix of dimension m by $F^{(m)}$, and the MSE matrix by $M^{(m)}$. Also e_m denotes the mth unit vector in $\mathbb{R}^l$, where the dimension $l \geq m$ will be apparent from the context. We introduce a general notation for signal extraction estimates: $\widehat{S}_{t|_s^m}$. This is an estimate of S_t, which is a linear function of the data $Y_s, Y_{s+1}, \cdots, Y_m$ such that the associated error $\widehat{S}_{t|_s^m} - S_t$ is uncorrelated with the data $Y_s, Y_{s+1}, \cdots, Y_m$ under Assumption A. Such a signal extraction estimate has minimum Mean Squared Error (MSE) among all estimates that are linear in the data. Note that Assumption A has to do with the initial values $Y_1, \cdots, Y_d$, which may not even be a part of the sample $Y_s, \cdots, Y_m$ (e.g., say $s > d$). The actual initial values in this sample are $Y_s, \cdots, Y_{s+d-1}$, but these can be expressed as a linear combination of the initial values Y_*. Therefore Assumption A does indeed guarantee the validity of all the signal extraction formulas for samples computed at subsequent time periods.

We make a final distinction. Any model-based signal extraction matrix will have the form F given by (3), though we allow that the model may be misspecified. That is, any of δ^S, δ^N, Σ_U, or Σ_V may be in error. If we wish to denote the "true" specifications of these quantities, we place a tilde over it, e.g., $\widetilde{\Sigma}_U$ is the true autocovariance matrix of U_t, whereas Σ_U denotes the matrix implied by our model. Misspecifying δ^S and δ^N is a worse error than the misspecification of Σ_U and Σ_V (see the discussion at the end of Section 2).

2.2 Revisions

The main concept in revision calculations is to consider a "window-sample" of size n; this is a sample $Y_{t+1}, Y_{t+2}, \cdots, Y_{t+n}$ for some $t = 0, 1, \cdots, N-1$, where N denotes the number of windows that we consider (not to be confused with the noise vector N). We focus on the concurrent signal extraction estimate, where we are interested in the signal at time $t+n$; simple extensions of our method can deal with the signal considered at other time points within the sample. Hence we consider signal extraction estimates $\widehat{S}_{t+n|_{t+1}^{t+n}}$, and are interested in the revision error that occurs if our sample was increased by a further $h > 0$ data points; the revised estimate would then be $\widehat{S}_{t+n|_{t+1}^{t+n+h}}$. Using the convention that the revision is "new minus old," the revision equals

$$\varepsilon_t = \widehat{S}_{t+n|_{t+1}^{t+n+h}} - \widehat{S}_{t+n|_{t+1}^{t+n}}.$$

Of course the revision ε_t depends on n and h as well as t, but these will be held fixed throughout our analysis, so they don't enter the notation for the revision. If the nonstationary operators δ^S and δ^N have been correctly specified, then ε_t will be a stationary sequence; this is because $\widehat{S}_{t+n|_{t+1}^{t+n+h}}$ and $\widehat{S}_{t+n|_{t+1}^{t+n}}$ will have no noise nonstationarity, and will both contain signal nonstationarity in such a manner that their difference is in fact stationary. The following proposition describes some of the important statistical properties of these revisions. Let $\underline{e}_n$ denote the nth unit vector in $\mathbb{R}^{n+h}$, whereas e_n denotes the nth unit vector in $\mathbb{R}^n$.

Proposition 1 *Assume that the signal extraction conditions of Section* 2 *hold, and in particular that δ^S and δ^N are correctly specified (though Σ_U and Σ_V need not be). Then the sequence of revisions ε_t is weakly stationary with mean zero and autocovariance sequence*

$$\begin{aligned}\gamma_\varepsilon(k) = & \left(\underline{e}_n' M^{(n+h)} \Delta_S' \Sigma_U^{-1} \left[1_{n+h-d_S}\, 0_k\right] - e_n' M^{(n)} \Delta_S' \Sigma_U^{-1} \left[1_{n-d_S}\, 0_{k+h}\right]\right) \widetilde{\Sigma}_U \\ & \left(\left[0_k\, 1_{n+h-d_S}\right]' \Sigma_U^{-1} \Delta_S M^{(n+h)} \underline{e}_n - \left[0_k\, 1_{n-d_S}\, 0_h\right]' \Sigma_U^{-1} \Delta_S M^{(n)} e_n\right) \\ & + \left(\underline{e}_n' M^{(n+h)} \Delta_N' \Sigma_V^{-1} \left[1_{n+h-d_N}\, 0_k\right] - e_n' M^{(n)} \Delta_N' \Sigma_V^{-1} \left[1_{n-d_N}\, 0_{k+h}\right]\right) \widetilde{\Sigma}_V \\ & \left(\left[0_k\, 1_{n+h-d_N}\right]' \Sigma_V^{-1} \Delta_N M^{(n+h)} \underline{e}_n - \left[0_k\, 1_{n-d_N}\, 0_h\right]' \Sigma_V^{-1} \Delta_N M^{(n)} e_n\right).\end{aligned}$$

The dimension of the M matrices is indicated by the superscript, and the 1 *refers to an identity matrix of indicated dimension. The subscript on the* 0 *then indicates the number of zero columns. The other matrices, such as Σ_U, Δ_S, etc., have dimensions implied by the other matrices that multiply them.*

Proposition 1 will be useful for determining the statistical properties of our goodness-of-fit statistic. Our null hypothesis (stated below) states that the model used actually describes the true process, so that $\Sigma_U = \widetilde{\Sigma}_U$ and $\Sigma_V = \widetilde{\Sigma}_V$. Hence for implementation, one needs to compute $\gamma_\varepsilon(k)$ under this type of assumption, for a sufficient number of lags k. Below, we discuss the test statistic RV in more detail.

2.3 Goodness-of-Fit Test Statistic

Now we want to use the empirical within-sample revision error as a measure of goodness-of-fit. We suppose that unit root tests and model identification procedures have already been utilized (e.g., using TRAMO), and furthermore the identified models have been fitted, obtaining the parameter estimates via maximum likelihood or another consistent procedure. Having completed these stages of analysis, we are now interested in ascertaining the goodness of model fit by using the RV statistic as a diagnostic tool. Treatments of unit root testing, time series model identification, and parameter estimation can be found in the following references: Dickey and Fuller (1979), Findley et al. (1998), Maravall and Caporello (2004), Peña, Tiao, Tsay (2000), and Taniguchi and Kakizawa (2000).

Since the theoretical mean of the revisions is zero, we can compute an estimate of their variance via $\frac{1}{N}\sum_{t=0}^{N-1}\varepsilon_t^2$. More generally, let our Revision Variance statistic be defined as

$$RV(B) = \frac{1}{N}\varepsilon' B \varepsilon,$$

where B is a square matrix and $\varepsilon = (\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_{N-1})'$. Clearly, taking B equal to the identity matrix yields the sample second moment of the revisions, but other choices of B will grant better size and power properties. This $RV(B)$ has mean

$$\mathbb{E}RV(B) = \frac{1}{N} tr(B\widetilde{\Sigma}_\varepsilon),$$

where $\widetilde{\Sigma}_\varepsilon$ is the (true) covariance matrix of ε. Hence taking $B = \Sigma_\varepsilon^{-1}$ based on our model specification (using Proposition 1), the mean of the revision statistic will be equal to 1 under the null hypothesis. Moreover, if the data is Gaussian, the variance will be equal to $2/N$.

The goodness-of-fit statistic studied in this paper is defined as $RV(\Sigma_\varepsilon^{-1})$, or just RV for short. The normalized test statistic is then

$$\sqrt{N}\frac{RV-1}{\sqrt{2}}. \tag{5}$$

Note that if the data is Gaussian, $\varepsilon'\Sigma_\varepsilon^{-1}\varepsilon$ has a χ^2_N distribution. Suppose that we specify δ^S and δ^N correctly, so that by Proposition 1 the revision process is stationary; let f_ε be the spectral density corresponding to the given autocovariance sequence. If $\Sigma_U = \widetilde{\Sigma}_U$ and $\Sigma_V = \widetilde{\Sigma}_V$, then the model is correctly specified with correct parameter values as well. The corresponding spectral density is the true spectrum for the revision process, and is denoted by $\widetilde{f}_\varepsilon$. Likewise, let Σ_ε and $\widetilde{\Sigma}_\varepsilon$ be the associated covariance matrices. Then the statistical properties of RV follow from Theorem 1 of McElroy (2008b): the mean of RV is $tr(\Sigma_{\varepsilon}^{-1}\widetilde{\Sigma}_{\varepsilon})/N$, and if the third and fourth cumulants are zero the variance is $2tr([\Sigma_{\varepsilon}^{-1}\widetilde{\Sigma}_{\varepsilon}]^2)/N^2$. If $\widetilde{f}_{\varepsilon}$ and $1/f_{\varepsilon}$ are continuously differentiable, then

$$\mathbb{E}RV \to \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\widetilde{f}_{\varepsilon}(\lambda)}{f_{\varepsilon}(\lambda)}\,d\lambda$$
$$N\,VarRV \to \frac{2}{2\pi}\int_{-\pi}^{\pi}\frac{\widetilde{f}_{\varepsilon}^{2}(\lambda)}{f_{\varepsilon}^{2}(\lambda)}\,d\lambda$$

as $N \to \infty$. Some mild conditions on the data are required for the asymptotic theory; we follow the material in Taniguchi and Kakizawa (2000, Section 3.1.1). Condition (B), due to Brillinger (1981), states that the process is strictly stationary and condition (B1) of Taniguchi and Kakizawa (2000, p. 55) holds. Condition (HT), due to Hosoya and Taniguchi (1982), states that the process has a linear representation, and conditions (H1) through (H6) of Taniguchi and Kakizawa (2000, pp. 55 – 56) hold. If the revision process satisfies either condition (B) or (HT), then as $N \to \infty$

$$\frac{RV - \mathbb{E}RV}{\sqrt{VarRV}} \overset{\mathscr{L}}{\Longrightarrow} \mathscr{N}(0,1).$$

We note that the computations required for the variance of the empirical revision measure RV are considerable, since we must consider up to $N+h$ different MSE matrices of various dimensions. There is no straight-forward way to obtain the required quantities using a State Space smoother – one must use the direct matrix approach of McElroy (2008a).

Our null hypothesis is that the model is correctly specified with correct covariance structure for the components as well, i.e.,

$$H_0 : \delta^N = \widetilde{\delta}^N,\, \delta^S = \tilde{\delta}^S,\, \Sigma_U = \widetilde{\Sigma}_U,\, \Sigma_V = \tilde{\Sigma}_V.$$

The alternative hypothesis is that the model is incorrectly specified, which includes not only the case that the proposed differencing operators may be incorrect, but also that the models for U_t and/or V_t may be incorrect. Not only may the parameter values be faulty, but the model specifications for these components may be wrong as well. In general we may speak of over- and under-specification of differencing operators. This refers to assigning too many or too few unit root differencing factors in δ (which are then allocated among the signal and the noise). For example, if the true process is $I(1)$ and we use an $I(2)$ model, this corresponds to over-specification, whereas using an $I(0)$ model corresponds to under-specification. Generally speaking, our test is much more powerful for detection of under-specification, because in this case the revision process is nonstationary and the RV statistic explodes asymptotically. But with over-specification, the revision process will still be stationary; only now the variance normalization will be incorrect, leading us to reject H_0. There are many other interesting cases that arise, for example: $\widetilde{\delta}(z) = 1 - z^{12}$ is the true differencing operator, but our model specifies $\delta(z) = 1 - z$ instead; this is under-specification, because the operator $1 + z + \cdots + z^{11}$ has been omitted.

In practice, since the DGP is not known, the parameter values are obtained by using MLEs (or other parameter estimates), pretending that these are fixed and non-random. One could base these estimates either on the whole span of data, or only on the first window of size n. Since the null hypothesis is so broad, it is difficult to determine which part of the model is wrong when a significant RV statistic is obtained. From empirical studies we know that RV is much more sensitive to misspecification (i.e., a wrong polynomial for δ^N or δ^S, or a wrong specification of the models for U_t and V_t) than to parameter error (i.e., parameter values that differ from those of the DGP). Note that when unit roots are misspecifed, the resulting component models are also misspecified and therefore the signal extraction filters are faulty, resulting in a signal extraction revision process that does not, in general, have the properties of Proposition 1; this should lead to rejection of the null hypothesis, as desired. Our focus in this paper is on unit root identification; with a significant RV, the practitioner should first seek to adjust the unit root specification (if multi-step ahead forecasting performance is important to the application), and then afterwards see to the modeling of the stationary aspects of the data.

3 Implementation

The previous section discussed the theoretical properties of the revision diagnostic RV, given that we compute signal extraction estimates using (3). We now discuss the details of implementing these ideas. In order to construct the signal extraction matrix F, we must specify the matrices Σ_U and Σ_V (as well as δ^S and δ^N) – or equivalently, their spectral densities f_U and f_V. For notation, let f_U, f_V, and f_W denote spectral densities for the differenced signal, noise, and data processes U_t, V_t, and W_t discussed in Section 2.1, which we assume to be stationary processes. This assumption involves no loss of generality, since we only need to implement RV under the null hypothesis, which stipulates that the model (including unit roots) is correctly specified.

Thus under H_0 we have an explicit form for f_W. Typically f_U and f_V are in turn determined from f_W in a variety of ways: (1) decomposition, (2) structural, (3) direct. The first two techniques are widely used in the econometrics community, whereas the third requires more exposition due to its relative obscurity. We here briefly review these approaches to obtaining component models, providing references and a short example.

If f_W is the spectral density of an ARIMA or SARIMA process, it may be possible to mathematically solve for the spectra f_U and f_V using the canonical decomposition approach of Hillmer and Tiao (1982). No *a priori* restrictions are placed on the form of f_W; fairly simple algebra is utilized to obtain f_U and f_V, although the solution is not guaranteed to exist, and is typically not unique. This is the procedure adopted in SEATS, the seasonal adjustment portion of program TSW of the Bank of Spain; further details can be found in Maravall and Caporello (2004). Also see the extended discussion in McElroy (2008a).

In contrast, the structural approach first specifies models for f_U and f_V, from which an implied spectral density f_W is obtained by summation, referred to in the literature as the reduced form of the data model. Then the model parameters of f_U and f_V enter into the likelihood for f_W through the formula

$$f_W(\lambda) = |\delta^N(e^{-i\lambda})|^2 f_U(\lambda) + |\delta^S(e^{-i\lambda})|^2 f_V(\lambda). \tag{6}$$

This follows at once from (2) and the other definitions in Section 2.1. The structural approach was first developed by Gersch and Kitagawa (1983), and later was popularized in Harvey (1989). This latter work develops the structural approach with so-called structural models, which are parameter-restricted ARIMA models that are convenient for state space calculations, thereby facilitating efficient Gaussian maximum likelihood estimation. One drawback of the structural approach is that the pre-specified model form of f_U and f_V restrains f_W from attaining a model deemed optimal according to standard identification techniques, such as unit-root testing and model specification methods.

The direct approach is similar to the decomposition method, in that it starts with the specified form of f_W dictated by prior unit-root tests and model specification techniques. However, the spectra f_U and f_V are given as a fixed function of f_W, where this mapping does not at all depend on model parameters – in this sense it is direct. It is easiest to describe through the pseudo-spectra, which are given by $f_S(\lambda) = f_U(\lambda)|\delta^S(e^{-i\lambda})|^{-2}$, $f_N(\lambda) = f_V(\lambda)|\delta^N(e^{-i\lambda})|^{-2}$, and $f_Y(\lambda) = f_W(\lambda)|\delta(e^{-i\lambda})|^{-2}$, corresponding to signal S_t, noise N_t, and data Y_t respectively. Then the direct approach relates signal and noise pseudo-spectra to the data pseudo-spectrum via multiplication by a fixed function g as follows:

$$f_S(\lambda) = g(\lambda) f_Y(\lambda) \qquad f_N(\lambda) = (1 - g(\lambda)) f_Y(\lambda). \tag{7}$$

Here $g : [-\pi, \pi] \rightarrow [0, 1]$ is a user-defined function. An intuitive example is given by $g(\lambda) = 1$ if and only if $|\lambda| \leq \pi/60$, and zero otherwise; this defines the signal to consist of only those frequencies of the data spectrum lying in the low frequency range $[-\pi/60, \pi/60]$, while the noise contains all other frequencies. This direct approach is formally developed and utilized in Kaiser and Maravall (2005), where the idea is that g is the frequency response function of the Hodrick Prescott filter, thereby defining the signal to be a cycle.

Because the structural approach interferes with a direct diagnosis of model misspecification of the data, we will focus on the decomposition and direct approaches in the rest of the paper. Now of course the decomposition approach can be conceived in terms of a function g in (7), but in this case g depends on the parameters of f_W and takes its general form from the data model; in contrast, the direct approach utilizes a user-determined g that is dependent upon neither the selected model nor the parameters. The choice of g dictates the form of signal and noise, and thus can be defined to correspond with the analyst's particular interests. The basic conditions on g are that $g|\delta^N(e^{-i\cdot})|^{-2}$ and $(1-g)|\delta^S(e^{-i\cdot})|^{-2}$ are bounded functions. (These conditions ensure that f_S and f_N only have poles at the appropriate signal and noise frequencies.) We next provide a more detailed illustration, which shall be used in Section 4.

Suppose that $\delta^S(z) = 1 - z$ and $\delta^N(z) = 1 + z + \cdots + z^{11}$, which correspond to trend signal and seasonal noise processes respectively. Note that this implicitly defines an $I(1)$ model for the data, in contrast with the Airline model discussed in Section 1. Let $g(\lambda) = |\delta^N(e^{-i\lambda})|^2/144$; this choice is the simplest form for g such that $g|\delta^N(e^{-i\cdot})|^{-2}$ is bounded (the scaling factor of 144 ensures that the range of g is contained in $[0,1]$). Then it follows that

$$\frac{1-g(\lambda)}{|\delta^S(e^{-i\lambda})|^2} = |h(e^{-i\lambda})|^2/144,$$

where $h(z) = 10.787 + 8.570z + 6.672z^2 + 5.070z^3 + 3.738z^4 + 2.652z^5 + 1.788z^6 + 1.123z^7 + .634z^8 + .297z^9 + .093z^{10}$. Hence we have that $(1-g)|\delta^S(e^{-i\lambda})|^{-2}$ is also a bounded function. This choice of g is therefore a very simple and natural candidate, and also satisfies the basic stipulated requirements. We easily obtain that

$$f_U(\lambda) = f_W(\lambda)/144 \qquad f_V(\lambda) = |h(e^{-i\lambda})|^2 f_W(\lambda)/144.$$

These equations represent a very direct and clear relationship between f_W and f_U, f_V; this relationship only depends on the properties of the data through f_W. For more guidance in general on the selection of g for other problems, see Kaiser and Maravall (2005).

So given the component spectra f_U and f_V – obtained via either through the decomposition or direct approaches – we can immediately compute Σ_U and Σ_V, their associated covariance matrices. The general procedure for computing RV is the following:

1. Begin with a proposed model f_Y, which consists of signal and noise differencing operators δ^S and δ^N, and the spectrum of the differenced process f_W.
2. Obtain f_U and f_V from f_W. In the decomposition approach there are algorithms for computing f_U and f_V from f_W, whereas in the direct approach we use (7).
3. Construct the filter matrix F via (3) and the revision process ε by applying the appropriate rows of F to the data.
4. Obtain the covariance matrix of ε under the null hypothesis (by using Proposition 1). Compute the normalized RV via (5) and get the p-value using the χ^2_N distribution.

In the context of model-based seasonal adjustment or trend estimation of economic data, typically steps 1 and 2 (and part of 3) are already performed by the analyst. The implementation challenge lies in the correct construction of Σ_ε based on Proposition 1; this formula is complicated, but the algebraic operations are all standard. Pseudo-code is available in the extended version of the paper (McElroy and Wildi, 2008). Also, as noted in the previous section, the computation of RV requires a choice of revision lead h and window size n. We have written our implementation (of the decomposition and direct approaches) in Ox (Doornik, 2006), utilizing SsfPack routines (Koopman, Shephard, and Doornik, 1999).

4 Empirical Studies

In this section, we focus on the finite-sample statistical properties of the empirical revision measure RV, considering both the decomposition and direct approaches (for the direct approach, we take g as defined in Section 3). In 4.1 we summarize various size and power studies, and in 4.2 we examine the method on several series from the KOF Economic Barometer, as well as some unemployment and manufacturing series.

4.1 Simulations

The DGPs considered in this section were chosen such that they correspond with the KOF empirical studies of Section 4.2. The series we consider are of length 322, so we take three window sizes $n = 120, 150, 180$ – hence the number of windows is $N = 202 - h, 172 - h, 142 - h$, where h is the revision lead. We consider several values of h, up to five years out (the data is monthly): $h = 12, 24, 36, 48, 60$. For our first study, we employ the decomposition approach applied to the Box-Jenkins Airline model. Our second study employs the direct approach, but in this case the model is only $I(1)$ plus seasonal. Details on these two approaches are provided below.

In the decomposition study (Study 1), there are three components: trend, seasonal, and irregular. The Airline model is given by the SARIMA equation

$$(1-B)(1-B^{12})X_t = (1-\theta B)(1-\Theta B^{12})\varepsilon_t, \tag{8}$$

where X_t is the time series, ε_t is white noise of variance σ^2, and θ and Θ are the parameters. Both the trend and seasonal are typically nonstationary in economic data, and thus are the components of greatest interest for our purposes. Here the trend differencing operator is $(1-B)^2$, whereas the seasonal differencing operator is $U(B) = 1+B+\cdots+B^{11}$. Hence we will consider either the trend or the seasonal as the signal of interest – note that the revision process for the associated noise is always that of the signal multiplied by -1. So the RV for the seasonal component and the seasonally adjusted component will be identical. We consider a null hypothesis of a Box-Jenkins Airline model with various specifications of the parameters θ, Θ. Given the specification of a null model via a choice θ, Θ, we can determine RV for either the trend or seasonal components as discussed in Section 3.

In the direct approach (Study 2), there are two components: the seasonal and the nonseasonal. The spectra of these components are defined through $g(\lambda) = \left|U(e^{-i\lambda})\right|^2/144$, as discussed in Section 3. In that section, S_t is nonseasonal and N_t is seasonal; note that if we swap roles and let the seasonal be the signal instead, the revision measure RV will yield identical results (again, since the revision process for noise is related to the revision process of signal via multiplication by -1). So, we only report results for the nonseasonal. The model for the data process is

$$(1-B^{12})X_t = (1-\Theta B^{12})\varepsilon_t, \tag{9}$$

which can be viewed as a subset model of the Airline model when $\theta = 1$ (after cancelation).

These clearly do not reflect a comprehensive study, but nevertheless will reveal some useful observations. First, Airline models form a fairly basic trend-seasonal model, and thus are a good starting place. The window sizes are chosen to reflect common data lengths – typically monthly seasonal time series at many statistical agencies are between 10 and 15 years long. Of course, the number of revisions N is much larger than it would be in practice, though in our case the length of the KOF series facilitates a large N. The asymptotics discussed in Section 2.3 are with respect to increasing N, so decreasing n and h should provide a RV that is more normally distributed. The revision leads h are fairly typical – in practice the revisions from model-based seasonal adjustments (e.g., using SEATS) are usually negligible after 5 years.

In order to investigate the power of the diagnostic tests in both studies, we consider the following alternative models:

$$(1-\phi B)(1-\Phi B^{12})X_t = (1-.6B)(1-.6B^{12})\varepsilon_t$$

with $\phi, \Phi = .6, .9, 1$. Therefore, taking all possible combinations and making cancelations where appropriate, we obtain the following 9 DGPs: $\phi = \Phi = 1$ (DGP 0, SARIMA(011)(011)); $\phi = .9$, $\Phi = 1$ (DGP 1, SARIMA(101)(011)); $\phi = .6$, $\Phi = 1$ (DGP 2, SARIMA(101)(011)); $\phi = 1$, $\Phi = .9$ (DGP 3, SARIMA(011)(101)); $\phi = 1$, $\Phi = .6$ (DGP 4, SARIMA(011)(101)); $\phi = .9 = \Phi$ (DGP 5, SARIMA(101)(101)); $\phi = .9$, $\Phi = .6$ (DGP 6, SARIMA(101)(101)); $\phi = .6$, $\Phi = .9$ (DGP 7, SARIMA(101)(101)); $\phi = .6 = \Phi$ (DGP 8, SARIMA(101)(101)).

These alternative DGPs have different unit root structures. In the empirical studies we simulated Gaussian data from the nine models but applied the signal extraction filters associated with the null model, which for Study 1 was an Airline model with parameters $\theta = .6$, $\Theta = .6$ in (8) – this is DGP 0; we consider both the trend and seasonal signals. For Study 2 (the direct approach), the null model corresponds to the choice $\Theta = .6$ in the data process (9), which is actually DGP 2. In Study 1 all 8 alternative DGPs (i.e., DGPs 1 through 8) correspond to over-specification of the order of nonstationarity. However, in Study 2 under-specification occurs with DGP 0, and DGPs 3 through 8 correspond to over-specification (i.e., the null model over-differences these processes).

The results are reported in Tables 1, 2, and 3 below. In Study 1 the power is below 50% for DGPs 1, 3, and 5, which most closely resemble the null model. DGPs 2 and 4 have power exceeding 50% in some cases (i.e., smaller h and n), whereas power is higher for DGPs 6, 7, and 8. These latter three models are stationary, not having the unit root structure of the null model, so it is reasonable to expect higher power in these cases. Results are similar for the trend and seasonal signals. In Study 2, DGP 0 produces 100% power – in this case the normalized RV statistic was explosive, taking on values in the thousands; this is an expected outcome of the under-specification case. DGP 2 just provides the size. Power was at or close to 100% for DGPs 3 and 4, which have no nonstationary seasonality. DGPs 1, 5 and 6 provide power exceeding 70%. Power for DGP 7 does not exceed 10% and DGP 8 has power around the 50% level.

In order to provide a reference frame for these results, we also discuss LB statistics applied to fits of the Airline model to the same DGPs. As with the RV study, we keep the parameters in the fitted model fixed at .6, .6. We do this, rather than using the MLEs in each simulation, in order that comparisons with the RV

Table 1: Entries indicate empirical size and power as a percentage, computed via 10,000 Monte Carlo simulations, of the RV statistic for Study 1 (T). The DGPs 0 through 8 indicate the data generating process that was simulated, with the Lead across the top. Study 1 (T) refers to revision statistics based on the trend signal in (8). For these studies DGP 0 corresponds to the null hypothesis, so this row gives size, whereas the other rows give power. The three numbers in each cell are size/power for window sizes 120, 150, and 180 respectively, from left to right.

DGPs	Lead 12	Lead 24	Lead 36	Lead 48	Lead 60
0	.05 .05 .05	.05 .05 .05	.05 .05 .05	.05 .04 .05	.05 .04 .05
1	.09 .08 .07	.08 .08 .07	.08 .08 .07	.08 .08 .07	.08 .07 .07
2	.59 .53 .45	.56 .50 .42	.54 .47 .39	.51 .44 .35	.48 .40 .32
3	.08 .08 .07	.08 .08 .07	.08 .08 .07	.08 .07 .07	.07 .07 .07
4	.60 .53 .46	.57 .50 .42	.54 .47 .39	.52 .43 .36	.49 .41 .32
5	.19 .17 .15	.18 .16 .14	.18 .15 .13	.16 .14 .12	.15 .13 .11
6	.78 .71 .63	.75 .68 .59	.72 .65 .55	.69 .61 .50	.66 .57 .46
7	.75 .68 .60	.73 .65 .56	.70 .62 .52	.67 .58 .48	.64 .55 .43
8	.98 .96 .92	.97 .95 .90	.96 .93 .87	.96 .91 .83	.94 .89 .79

Table 2: Entries indicate empirical size and power as a percentage, computed via 10,000 Monte Carlo simulations, of the RV statistic for Study 1 (S). The DGPs 0 through 8 indicate the data generating process that was simulated, with the Lead across the top. Study 1 (S) refers to revision statistics based on the seasonal signal in (8). For these studies DGP 0 corresponds to the null hypothesis, so this row gives size, whereas the other rows give power. The three numbers in each cell are size/power for window sizes 120, 150, and 180 respectively, from left to right.

DGPs	Lead 12	Lead 24	Lead 36	Lead 48	Lead 60
0	.05 .05 .05	.05 .05 .05	.05 .05 .05	.05 .05 .05	.05 .05 .05
1	.09 .08 .08	.08 .08 .08	.08 .08 .07	.08 .08 .07	.08 .07 .07
2	.59 .53 .46	.57 .50 .43	.54 .47 .40	.51 .44 .37	.48 .40 .33
3	.08 .08 .07	.08 .08 .07	.07 .07 .07	.07 .07 .06	.07 .07 .06
4	.59 .53 .45	.52 .45 .37	.47 .39 .31	.43 .34 .26	.39 .30 .21
5	.20 .17 .16	.18 .16 .13	.17 .14 .12	.15 .13 .11	.14 .12 .10
6	.78 .71 .62	.71 .63 .53	.66 .56 .45	.61 .51 .38	.57 .45 .32
7	.75 .68 .61	.72 .64 .56	.68 .60 .51	.65 .57 .46	.62 .52 .41
8	.98 .96 .92	.97 .93 .87	.95 .90 .81	.93 .86 .75	.91 .82 .67

Table 3: Entries indicate empirical size and power as a percentage, computed via 10,000 Monte Carlo simulations, of the RV statistic for Study 2. The DGPs 0 through 8 indicate the data generating process that was simulated, with the Lead across the top. Study 2 refers to revision statistics based on the signal in (9), and the null hypothesis corresponds to DGP 2, so this row gives size and the other rows give power. The three numbers in each cell are size/power for window sizes 120, 150, and 180 respectively, from left to right.

DGPs	Lead 12	Lead 24	Lead 36	Lead 48	Lead 60
0	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0
1	.87 .83 .77	.86 .80 .75	.84 .78 .72	.82 .76 .70	.80 .73 .66
2	.05 .05 .05	.05 .05 .05	.05 .05 .05	.05 .05 .05	.05 .05 .05
3	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 .99	1.0 1.0 .99
4	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 1.0	1.0 1.0 .99	1.0 1.0 .99
5	.91 .87 .83	.90 .85 .803	.88 .83 .78	.87 .81 .74	.85 .79 .71
6	.98 .97 .94	.98 .96 .93	.97 .95 .91	.97 .94 .89	.96 .92 .86
7	.09 .09 .08	.09 .09 .08	.09 .09 .08	.09 .08 .08	.09 .09 .08
8	.62 .57 .49	.60 .54 .46	.58 .51 .43	.56 .48 .40	.53 .45 .37

statistic – which uses fixed parameters – will be meaningful. Table 4 summarizes the results; we consider the LB at lags 12, 24, and 36, since these are multiples of the seasonal lag. Although there are some problems with the size (DGP 0 for Study 1 and DGP 2 for Study 2), the power exceeds 50% in most cases. In terms of comparing the RV and LB methods, we note that for Study 1 our RV statistic is slightly more powerful for DGPs 4, 5, and 6 (note that we can maximize power by taking h and n smaller, but there is no *a priori* reason to consider one of the lags 12, 24, or 36 as preferable to the others in the LB statistics), but the LB statistics are superior in the other cases. In Study 2, only DGP 8 provides greater power than the LB statistics. Therefore, according to the simulation studies the RV statistic is not superior to LB, but has similar overall performance.

In summary, we note that the RV procedure is flexible, as any combination of unit roots can be specified in the null hypothesis, and tested against an alternative where some or all of the roots no longer lie on the unit circle. The size is at the nominal level, and the power exceeds 50% in many cases. Generally speaking, the power for RV is lower than that of LB, although the LB is slightly over-sized. We observe that our statistic emphasizes signal extraction problems so that it cannot detect misspecifications that do not affect real-time filter performances. In terms of a recommendation for the choice of h and n, it is noted that smaller values effectively

Table 4: Entries indicate empirical size and power as a percentage, computed via 10,000 Monte Carlo simulations, of the LB statistic (computed using fixed parameters). The DGPs 0 through 8 indicate the data generating process that was simulated, with the type of Study (Study 1, Study 2) across the top. DGP 0 corresponds to the null hypothesis for Study 1, so this entry gives size, whereas the other entries give power. For Study 2 the null hypothesis corresponds to DGP 2, so this entry gives size and the other entries give power. The three numbers in each cell are size/power for LB lags 12, 24, and 36 respectively, from left to right.

DGPs	Study 1	Study 2
0	.05 .06 .07	1.0 1.0 1.0
1	.15 .14 .13	1.0 1.0 1.0
2	.97 .83 .72	.06 .06 .07
3	.07 .08 .10	1.0 1.0 1.0
4	.57 .60 .56	1.0 1.0 1.0
5	.18 .18 .18	1.0 1.0 1.0
6	.72 .72 .68	1.0 1.0 1.0
7	.97 .83 .77	.08 .09 .10
8	1.0 .99 .98	.56 .60 .56

increase the sample size N used in the RV statistic, and thus increase the power; therefore, these should be taken as small as practicable.

4.2 Revisions of the KOF, Unemployment, and Manufacturing Data

We next applied these diagnostic tests to the KOF series mentioned in the Introduction. To focus the discussion, we concentrated on four series that were all identified by X-12-ARIMA as having seasonality and an $I(2)$ trend. Based on *a priori* beliefs of boundedness for these series, it would seem that an $I(2)$ trend is a misspecification – and this is confirmed by sample ACF plots. So we expected our diagnostic tests to reject these models. We applied both revision diagnostic tests discussed above – namely the one based on the decomposition approach (Study 1) and the one based on the direct approach (Study 2); the first was used to show that the $I(2)$ trend was over-specified, and the second showed that the $I(1)$ trend was under-specified. Now when utilizing the automatic modeling procedure of X-12-ARIMA, the four series KOF9, KOF25, KOF27, and KOF29 were specified with Airline models (8). Values of the standardized RV statistic are reported in Table 5.

Table 5: Normalized RV test statistics for KOF series 9, 25, 27, and 29. For St1 an Airline model was fitted to the data using Maximum Likelihood Estimation, and the corresponding parameter values were used to determine the null hypothesis. For St2 model (9) was fitted instead. The three numbers in each cell are normalized RV at window sizes 120, 150, and 180 respectively, from left to right.

	Series			
St1 (T)	KOF9	KOF25	KOF27	KOF29
Ld 12	-.32 -.59 .45	-1.10 -1.15 -1.10	-1.35 -1.38 -1.06	-.86 -.55 -.61
Ld 24	-1.40 -1.77 -.85	-.78 -.80 -0.72	-1.25 -1.25 -.88	-.55 -.22 -.24
Ld 36	-1.48 -1.87 -.96	-1.11 -1.16 -1.15	-1.19 -1.20 -.84	-.55 -.21 -.24
Ld 48	-2.20 -2.70 -1.88	-1.39 -1.45 -1.52	-1.47 -1.46 -1.16	-.99 -.67 -.78
Ld 60	-2.56 -3.10 -2.33	-1.26 -1.32 -1.40	-1.66 -1.70 -1.50	-1.16 -.86 -1.02
St1 (S)	KOF9	KOF25	KOF27	KOF29
Ld 12	-.24 -.65 .41	-1.09 -.95 -.67	-1.58 -1.19 -.96	-.90 -.71 -.50
Ld 24	-.15 -.65 .34	-1.51 -1.49 -1.23	-1.17 -1.11 -.99	-.59 -.34 -.08
Ld 36	-.88 -1.45 -.47	-1.23 -1.27 -1.15	-2.19 -1.88 -1.75	-1.00 -.78 -.62
Ld 48	-1.45 -2.11 -1.21	-1.22 -1.04 -.83	-.49 -.42 .03	-.72 -.41 -.16
Ld 60	-1.19 -2.18 -1.33	-1.19 -1.04 -.82	-1.40 -1.17 -.88	-1.13 -.89 -.71
St2	KOF9	KOF25	KOF27	KOF29
Ld 12	2.60 2.20 1.86	1.74 1.14 .76	3.74 3.04 2.76	2.07 1.72 .59
Ld 24	2.50 2.08 1.72	2.21 1.63 1.30	4.01 3.32 3.07	2.45 2.13 1.00
Ld 36	2.38 1.94 1.56	1.72 1.07 .66	3.67 2.93 2.64	2.67 2.37 1.22
Ld 48	2.42 1.97 1.59	1.96 1.32 .93	4.00 3.28 3.04	2.85 2.57 1.39
Ld 60	2.45 2.00 1.61	2.03 1.38 .98	3.68 2.90 2.62	2.49 2.16 .86

All of the RV statistics were computed with the null model given by the maximum likelihood parameter estimates, for each given model specification, when fitted to the entire data set. Recall from the introduction that LB statistics were generally not significant for all of the KOF series; in particular, at lag 12 the LB statistics are above the 5% level for all four series, though KOF9 and KOF29 have a few significant LB statistics at other lags. (Some rejections due to pure chance are to be expected due to multiple testing.) For KOF9 the Airline model was over-specified, which is apparent from the negative RV statistics at leads 48 and 60; results were more extreme for Trend signal than Seasonal signal. However, the positive values under St2 corresponding to (9) indicate some evidence that this $I(1)$ process is under-specified. This may indicate that a nonstationary long memory model with order of integration between 1 and 2 would be more suitable. Results for KOF25, KOF27, and KOF29 indicate that the Airline model is not unsuitable;

conversely, there are fairly strong indications that model (9) is under-specified.

We next studied three manufacturing series from Great Britain (PPIPFU01, PRMNVE02, PRMNVE03) of length 577, 571, and 571 respectively. These series have the titles “GBR PPI Manufacturing input fuel,” “GBR Production of commercial vehicles,” and “GBR Production of passenger cars.” For these series an Airline model was selected by the automatic modeling procedure of X-12-ARIMA, with significant LB statistics at lag 17 for PRMNVE02 and at lags 14, 15, and 16 for PPIPFU01, both of which required a log transformation. Since these monthly series are not bounded, we have no *a priori* grounds to disbelieve an $I(2)$ unit root hypothesis. The results in Table 6 indicate that the Airline model is actually under-specified for PPIPFU01, so that one may want to consider an $I(3)$ model. The large RV statistics for (9) for this series indicate an explosive revision process corresponding to a severe under-specification of the unit root structure. For PRMNVE02 and PRMNVE03 the Airline model seems to be adequate, whereas model (9) is under-specified for the former series.

We also looked at three series of unemployment rates for Hungary, Brazil, and Japan (HUN.UNRTSUTT, BRA.UNRTSUTT, JPN.UNRTSUTT) of length 95, 325, and 577 respectively. Since these series are bounded by construction (unemployment cannot exceed 100%), we were sceptical about the correctness of an $I(2)$ specification. Airline models were fitted to all three monthly series, with a log transformation needed for Brazil and Hungary. While there were no LB problems with Brazil, Hungary had one significant LB at lag 4, while Japan had several at lags 3, 23, 24, 25, 26, and 27. Table 7 reveals that the Airline model was adequate for Hungary and Japan, with explosive RV statistics obtained for the (9) model. For Brazil, there is evidence of over-specification of the Airline model at shorter revision leads; interestingly, the evidence of over-specification is even stronger with the $I(1)$ model! This is admittedly a puzzling result, but shows up some potential problems with both models.

Finally, we examined four shorter manufacturing series from the U.S. Census Bureau (X3, X3020, X3022, X10140) of length 155. These are monthly series for which the Airline model was again selected by X-12-ARIMA (a log transformation was needed for X10140). Like the GBR series we have no prior grounds for rejecting an $I(2)$ specification; in contrast to those series, these ones are much shorter. For this reason the window sizes were taken as $n = 60, 72, 84$. Results are reported in Table 8, and indicate that the Airline model is generally suitable, although there is evidence for over-specification for X10140 and under-specification for X3022. Explosive revisions were present for X3 and X3022 for model (9), indicating that $I(1)$ is under-specified; X3020 also had some large RV statistics.

In summary, we found that RV was effective at identifying under- specifications (e.g., model (9)), and was able to reject $I(2)$ models for some of the KOF

Table 6: Normalized RV test statistics for GBR series PPIPFU01, PRMNVE02, and PRMNVE03. For St1 an Airline model was fitted to the data using Maximum Likelihood Estimation, and the corresponding parameter values were used to determine the null hypothesis. For St2 model (9) was fitted instead. The three numbers in each cell are normalized RV at window sizes 120, 150, and 180 respectively, from left to right.

	Series		
St1 (T)	PPIPFU01	PRMNVE02	PRMNVE03
Ld 12	2.27 3.25 3.03	.81 .70 .81	.80 .77 .99
Ld 24	2.30 3.29 3.06	1.04 .94 1.07	1.09 1.04 1.30
Ld 36	1.34 2.32 2.05	1.28 1.19 1.34	1.37 1.34 1.62
Ld 48	1.08 2.06 1.77	1.54 1.46 1.62	1.75 1.73 2.03
Ld 60	.79 1.77 1.47	1.77 1.70 1.87	1.75 1.74 2.04
St1 (S)	PPIPFU01	PRMNVE02	PRMNVE03
Ld 12	2.66 3.65 3.44	1.06 1.03 1.34	.30 .57 .76
Ld 24	1.94 2.93 2.68	1.04 1.06 1.40	.57 .85 1.05
Ld 36	1.43 2.41 2.14	.75 .83 1.14	.87 1.17 1.38
Ld 48	.98 1.96 1.67	1.03 1.17 1.52	1.19 1.51 1.74
Ld 60	.86 1.85 1.55	1.52 1.57 1.93	1.44 1.77 2.01
St2	PPIPFU01	PRMNVE02	PRMNVE03
Ld 12	405.12 412.44 412.16	2.26 2.75 3.43	-.21 -.20 -.44
Ld 24	404.43 412.04 411.99	2.53 3.04 3.75	-.10 -.09 -.34
Ld 36	381.64 388.73 387.98	2.87 3.40 4.14	.20 .23 -.01
Ld 48	372.82 379.90 379.04	3.17 3.73 4.49	.57 .61 .38
Ld 60	378.68 386.34 386.01	3.52 4.10 4.90	.75 .79 .57

series even when the LB statistics were satisfactory. The performance on shorter series (like the Manufacturing series and Hungarian unemployment) was adequate, although the revision lead h and window size n had to be adjusted downwards. Likewise, the results on longer series (like the Great Britain series and Japanese unemployment) were as expected; it is surprising that the RV did not detect over-specification for Japanese unemployment, given the great length, but this may be due to the choice of parameters. For purposes of comparison we kept h and n the same across series (except for the shorter series, where this was impossible), but noted that over-specification for the KOF series tended to increase with $h > 60$.

Given that the RV diagnostics indicate rejection of a given model, what is to be done next? If one-step ahead forecasting is the practitioner's goal, then nothing should be done. If the purpose is real-time signal extraction, then one should either

Table 7: Normalized RV test statistics for Unemployment series for Hungary, Brazil, and Japan. For St1 an Airline model was fitted to the data using Maximum Likelihood Estimation, and the corresponding parameter values were used to determine the null hypothesis. For St2 model (9) was fitted instead. The three numbers in each cell are normalized RV at window sizes 120, 150, and 180 respectively for BRAZIL and JAPAN, from left to right. The window sizes for HUNGARY are 60, 66, and 72; also the HUNGARY revision leads are in parentheses: 1, 2, 3, 6, and 12.

	Series		
St1 (T)	HUNGARY	BRAZIL	JAPAN
Ld 12 (1)	-.51 -.65 -.03	-.77 -2.05 -2.32	.40 .41 .45
Ld 24 (2)	-.48 -.62 -.00	-.40 -1.72 -1.93	-.16 -.16 -.14
Ld 36 (3)	-.43 -.56 .03	.03 -1.34 -1.58	-.24 -.25 -.24
Ld 48 (6)	-.75 -.93 -.34	.42 -.90 -1.07	-.82 -.84 -.86
Ld 60 (12)	-1.00 -.91 -.21	.92 -.39 -.45	-1.32 -1.37 -1.40
St1 (S)	HUNGARY	BRAZIL	JAPAN
Ld 12 (1)	-.51 -.65 -.03	-.74 -2.12 -2.36	.23 .52 .57
Ld 24 (2)	-.16 -.64 -.12	-.34 -1.78 -1.92	.12 .37 .33
Ld 36 (3)	-.36 -.42 .16	.24 -1.50 -1.64	-.16 .12 .08
Ld 48 (6)	-.51 -.64 -.10	.60 -1.02 -1.13	-.29 -.04 -.05
Ld 60 (12)	-1.23 -.90 -.22	1.23 -.76 -.70	-1.28 -1.10 -1.21
St2	HUNGARY	BRAZIL	JAPAN
Ld 12 (1)	7.01 5.14 2.04	-2.38 -3.37 -2.62	30.39 32.10 34.06
Ld 24 (2)	7.23 5.36 2.23	-1.99 -2.99 -2.17	28.93 30.63 32.59
Ld 36 (3)	7.45 5.58 2.41	-1.79 -2.81 -1.95	28.03 29.73 31.70
Ld 48 (6)	8.05 6.19 2.90	-1.56 -2.62 -1.70	27.57 29.29 31.30
Ld 60 (12)	9.56 7.73 4.38	-1.04 -2.10 -1.05	27.98 29.77 31.86

change the model to one that allows for mean-reversion by removing unit roots (in the case of over-specification) or dispense with model-based approaches altogether (e.g., one could implement the Direct Filter Approach to real-time signal extraction developed in Wildi (2004, 2008)). For the case of under-specification (e.g., an explosive revisions process), one should add more unit roots to fix the model.

5 Conclusion

It is well-known that models that pass traditional one-step ahead diagnostic tests may perform rather poorly in a multi-step ahead perspective – recall the discussion

Table 8: Normalized RV test statistics for the Manufacturing series X3, X3020, X3022, and X10140. For St1 an Airline model was fitted to the data using Maximum Likelihood Estimation, and the corresponding parameter values were used to determine the null hypothesis. For St2 model (9) was fitted instead. The three numbers in each cell are normalized RV at window sizes 60, 72, and 84 respectively, from left to right.

	Series			
St1 (T)	X3	X3020	X3022	X10140
Ld 12	1.08 1.55 1.45	.73 .88 1.50	1.23 1.92 2.29	-1.38 -1.38 -.87
Ld 24	1.11 1.64 1.63	.33 .40 1.05	-.35 .20 .43	-2.15 -2.27 -1.72
Ld 36	1.09 1.67 1.61	.37 .49 1.24	-.69 -.11 .18	-1.70 -1.82 -1.11
Ld 48	.68 1.47 1.24	-.70 -.55 .05	-.56 .05 .51	-2.48 -2.61 -2.06
Ld 60	.12 1.07 .80	-.47 -.28 .65	-1.34 -.77 -.29	-2.10 -2.23 -1.53
St1 (S)	X3	X3020	X3022	X10140
Ld 12	1.27 1.00 1.13	1.14 1.50 1.73	2.19 2.46 3.10	-.57 -.08 .77
Ld 24	1.16 1.04 1.47	.45 .87 .99	1.59 1.92 2.72	-1.00 -.41 .47
Ld 36	.84 .95 1.21	.89 1.57 2.01	.47 .74 2.07	-2.00 -1.45 -.68
Ld 48	1.61 1.14 1.50	.38 1.01 1.20	.59 1.45 2.64	-1.73 -1.19 -.36
Ld 60	.86 .95 .66	-.75 -.20 -.93	-.78 .08 1.09	-2.35 -1.66 -1.10
St2	X3	X3020	X3022	X10140
Ld 12	15.41 16.18 18.07	5.72 6.91 8.30	11.11 12.17 11.91	-1.20 -.75 .01
Ld 24	13.29 14.12 16.24	5.62 6.96 8.66	11.20 12.49 12.40	-1.54 -1.07 -.31
Ld 36	12.17 13.30 15.89	1.35 2.55 4.00	5.22 6.28 5.44	-1.04 -.47 .48
Ld 48	12.14 13.80 17.48	-2.88 -1.96 -1.30	3.12 4.24 2.87	-1.73 -1.19 -.13
Ld 60	4.17 4.61 6.74	-2.64 -1.64 -.82	2.84 4.21 2.63	-1.92 -1.24 .11

in Section 1. It is therefore necessary to account for the purpose of a particular application when selecting and checking model performance. We have proposed a test for model misspecification that fits a general class of forecasting problems.

Although we restricted attention to real-time signal-extraction problems, the scope of the proposed approach is more general because we allowed for arbitrary signals. Therefore, revision errors can be "designed" by choosing suitable (artificial) signal definitions. As an example, assume that a signal is defined by a symmetric MA(3)-filter with coefficients $\gamma_{-1}, \gamma_0, \gamma_1$ where $\gamma_{-1} = \gamma_1$. If $\gamma_1 = 1$, then the revision error would correspond to the one-step ahead forecasting error. Thus, traditional (one-step ahead) diagnostics can be replicated in our framework by choosing the above artificial filter. More generally, revision errors relying on arbitrary linear combinations of one- and multi-step ahead forecasts can be derived by specifying a corresponding symmetric MA-filter. (Note that the central weight γ_0 is not important here.) Therefore, a diagnostic test can be set up that accounts for performances involving any linear combination of forecasts. As a consequence, the proposed diagnostic test can fit a variety of practically relevant estimation problems whose precise structures can be accounted for explicitly.

Our simulation results confirm a good concordance between asymptotic and finite sample test distributions; although in our Monte Carlo simulations the LB statistics generally out-performed RV (in terms of power), the latter is still fairly successful in real-data studies, particularly if the misspecification directly affects filter performances. Results in the context of the KOF Economic Barometer suggest stronger rejection of false unit roots hypotheses, in both seasonal and trend roots. This is due to the fact that the above series exhibit mid-term trend reversion that are difficult to detect with statistics relying exclusively on short-term forecasting performances.

Traditional model-fitting diagnostics are based on computing model residuals and testing them for whiteness, i.e., whether or not they are serially uncorrelated. The signal extraction revision process is similar in many ways to model residuals, although under the null hypothesis of correct model and covariance specification they do not behave as white noise, but rather have another covariance structure (as given in Proposition 1). Both model residuals and signal extraction revisions can be used to assess poorness of model fit, but each examines different aspects of the data's dynamics. For the KOF series, model residuals appear to be white and hence no problems with the over-specified model are indicated, whereas the signal extraction revisions tend to be less than what one would expect from the model, indicating an over-specification of the fitted model in some cases. Hence, model residuals and signal extraction revisions present different information about a series[7]. Essentially, signal extraction revisions allow the practitioner to focus on particular aspects or sections of the data's pseudo-spectrum, whereas model residuals look at the spectrum as a whole.

Given these findings, we present the RV statistic as a useful tool to complement standard goodness-of-fit statistics such as LB and unit root tests. One drawback of the RV statistic is that it takes some time and effort to encode the formulas of Proposition 1, and some thought must also be given to how the models for signal and noise are related to the data process. To assist readers, we have provided pseudo-code in an extended version of the paper (McElroy and Wildi, 2008). A second caution is the finite-sample power of the RV statistic will tend to be lower than desired in the over-specification case (Tables 1, 2, and 3). However, we argue

[7] Even though signal extraction residuals can be written as a linear combination of model residuals, the weights in this linear combination provide a different view of the model performance.

that these caveats are offset by the great flexibility of the RV diagnostics, which essentially allow one to assess the model over several forecasting leads simultaneously.

Disclaimer This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Appendix

Proof of Proposition 1. For the first assertion, we write out ε_t in vector form.

$$\begin{aligned}
\varepsilon_t &= \underline{e}_n' F^{(n+h)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n+h} \end{bmatrix} - e_n' F^{(n)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n} \end{bmatrix} \\
&= \underline{e}_n' \left(F^{(n+h)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n+h} \end{bmatrix} - \begin{bmatrix} S_{t+1} \\ \vdots \\ S_{t+n+h} \end{bmatrix} \right) - e_n' \left(F^{(n)} \begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+n} \end{bmatrix} - \begin{bmatrix} S_{t+1} \\ \vdots \\ S_{t+n} \end{bmatrix} \right) \\
&= \underline{e}_n' E_t^{(n+h)} - e_n' E_t^{(n)},
\end{aligned}$$

where $E_t^{(n)}$ denotes the error process at time t based on the sample from time $t+1$ to $t+n$. Such an error process is simply a linear combination of U_s and V_s – the differenced signal and noise processes – at times $t+1 \leq s \leq t+n$. The same goes for $E_t^{(n+h)}$, so ε_t is a linear combination of $\{U_s\}$ and $\{V_s\}$, which are weakly stationary and uncorrelated with one another. Thus, the revisions are weakly stationary, too (and if the $\{U_s\}$ and $\{V_s\}$ processes are strictly stationary, then so is the revision process). Since these error processes have mean zero, so does the revision process.

Finally, we consider the autocovariance at lag k; considering $k \geq 0$, we have

$$\begin{aligned}
\varepsilon_t \varepsilon_{t+k} &= \underline{e}_n' E_t^{(n+h)} E_{t+k}^{(n+h)\prime} \underline{e}_n - \underline{e}_n' E_t^{(n+h)} E_{t+k}^{(n)\prime} e_n \\
&\quad - e_n' E_t^{(n)} E_{t+k}^{(n+h)\prime} \underline{e}_n + e_n' E_t^{(n)} E_{t+k}^{(n)\prime} e_n.
\end{aligned}$$

Next, we compute each of the error processes:

$$\begin{aligned}
E_t^{(n)} &= -M^{(n)}\Delta_S'\Sigma_U^{-1}U_{t+1+d_S:t+n} + M^{(n)}\Delta_N'\Sigma_V^{-1}V_{t+1+d_N:t+n}\\
E_{t+k}^{(n)} &= -M^{(n)}\Delta_S'\Sigma_U^{-1}U_{t+k+1+d_S:t+k+n} + M^{(n)}\Delta_N'\Sigma_V^{-1}V_{t+k+1+d_N:t+k+n}\\
E_t^{(n+h)} &= -M^{(n+h)}\Delta_S'\Sigma_U^{-1}U_{t+1+d_S:t+n+h} + M^{(n+h)}\Delta_N'\Sigma_V^{-1}V_{t+1+d_N:t+n+h}\\
E_{t+k}^{(n+h)} &= -M^{(n+h)}\Delta_S'\Sigma_U^{-1}U_{t+k+1+d_S:t+k+n+h} + M^{(n+h)}\Delta_N'\Sigma_V^{-1}V_{t+k+1+d_N:t+k+n+h}
\end{aligned}$$

We can conceive of a vector U of dimension $k+n+h-d_S$, which contains the U_j for $t+1+d_S \leq j \leq t+k+n+h$. Then we can substitute selection matrices into the above expressions, such as $[1_{n+h-d_S}\, 0]U$, and so forth. Similarly, we can do the same with the vector V. These expressions may be substituted into the formula for $\varepsilon_t\varepsilon_{t+k}$ above, and the expectation of UU' is Σ_U of appropriate dimension. The same holds for V, though note that $\mathbb{E}UV'$ is a zero matrix due to our assumptions on the components. Then by rearranging terms, we arrive at the stated formula. □

References

[1] Bell, W. (1984) Signal Extraction for Nonstationary Time Series. *The Annals of Statistics* **12**, 646 – 664.

[2] Bell, W. and Hillmer, S. (1988) A Matrix Approach to Likelihood Evaluation and Signal Extraction for ARIMA Component Time Series Models. *SRD Research Report No. RR* – 88/22, U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/rr88-22.pdf

[3] Box, G. and Jenkins, G. (1976) *Time Series Analysis: Forecasting and Control, Revised Edition*. San Francisco: Holden-Day.

[4] Brillinger, D. (1981) *Time Series Data Analysis and Theory,* San Francisco: Holden-Day.

[5] Brockwell, P. and Davis, R. (1991) *Time Series: Theory and Methods, 2nd Ed.* New York: Springer.

[6] Dickey, D. and Fuller, W. (1979) Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association* **74**, 427–431.

[7] Doornik, J. (2006) *Object-Oriented Matrix Programming using Ox*, 5th Edition. London: Timberlake Consultants Press.

[8] Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C. and Chen, B. C. (1998) New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program. *Journal of Business and Economic Statistics* **16**, 127–177 (with discussion).

[9] Gersch, W. and Kitagawa, G. (1983) The Prediction of Time Series with Trends and Seasonalities. *Journal of Business and Economics Statistics* **1**, 253–264.

[10] Harvey, A. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter,* Cambridge: Cambridge University Press.

[11] Hillmer, S. and Tiao, G. (1982) An ARIMA-model-based Approach to Seasonal Adjustment. *Journal of the American Statistical Association* **77**, 63–70.

[12] Hosoya, Y., and Taniguchi, M. (1982) A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes. *Annals of Statistics* **10**, 132–153.

[13] Kaiser, R. and Maravall, A. (2005) Combining Filter Design with Model-based Filtering: An Application to Business-cycle Estimation. *International Journal of Forecasting* **21**, 691–710.

[14] Koopman, S., Shephard, N., and Doornik, J. (1999) Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal* **2**, 113 – 166.

[15] Ljung, G. and Box, G. (1978) On a Measure of Lack of Fit in Time Series Models. *Biometrika* **65**, 297–303.

[16] Maravall, A. (1986) Revisions in ARIMA Signal Extraction. *Journal of the American Statistical Association* **81**, 736–740.

[17] Maravall, A. and Caporello, G. (2004) Program TSW: Revised Reference Manual. *Working Paper 2004, Research Department, Bank of Spain.* http://www.bde.es

[18] McElroy, T. (2008a) Matrix Formulas for Nonstationary ARIMA Signal Extraction. *Econometric Theory* **24**, 1-22.

[19] McElroy, T. (2008b) Statistical Properties of Model-Based Signal Extraction Diagnostic Tests. *Communications in Statistics, Theory and Methods* **37**, 591–616.

[20] McElroy, T. and Gagnon, R. (2008) Finite Sample Revision Variances for ARIMA Model-Based Signal Extraction. *Journal of Official Statistics* **24**, 451–467.

[21] McElroy, T., and Wildi, M. (2008) Signal Extraction Revision Variances as a Goodness-of-Fit Measure. *SRD Research Report No.* $RRS2010-06$, U.S. Census Bureau.

[22] Peña, D., Tiao, G., and Tsay, R. (2000) *A Course in Time Series Analysis,* New York: John Wiley and Sons.

[23] Pierce, D. (1980) Data Revisions with Moving Average Seasonal Adjustment Procedures. *Journal of Econometrics* **14**, 95–114.

[24] Taniguchi, M. and Kakizawa, Y. (2000) *Asymptotic Theory of Statistical Inference for Time Series,* New York City, New York: Springer-Verlag.

[25] Wildi, M. (2004) Signal Extraction: How (In)efficient Are Model-Based Approaches? An Empirical Study Based on TRAMO/SEATS and Census X-12-ARIMA. *KOF-Working Paper Nr. 96*, ETH-Zurich.

[26] Wildi, M. (2008) *Real-Time Signal-Extraction: Beyond Maximum Likelihood Principles,* Berlin: Springer. http://www.idp.zhaw.ch/de/engineering/idp/forschung/finance-risk-management-and-econometrics/signal-extraction-and-forecasting/signal-extraction.html