*Journal of Forecasting*, *J. Forecast.* **34**, 315–336 (2015)
Published online 19 April 2015 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/for.2321

# When are Direct Multi-step and Iterative Forecasts Identical?†

Tucker McElroy*
Center for Statistical Research and Methodology, US Census Bureau, Washington, DC, USA

ABSTRACT

Although both direct multi-step-ahead forecasting and iterated one-step-ahead forecasting are two popular methods for predicting future values of a time series, it is not clear that the direct method is superior in practice, even though from a theoretical perspective it has lower mean squared error (MSE). A given model can be fitted according to either a multi-step or a one-step forecast error criterion, and we show here that discrepancies in performance between direct and iterative forecasting arise chiefly from the method of fitting, and is dictated by the nuances of the model's misspecification. We derive new formulas for quantifying iterative forecast MSE, and present a new approach for assessing asymptotic forecast MSE. Finally, the direct and iterative methods are compared on a retail series, which illustrates the strengths and weaknesses of each approach. Copyright © 2015 John Wiley & Sons, Ltd.

KEY WORDS ARIMA models; econometric models; long-term forecasting; seasonality; spectral analysis; time series

## INTRODUCTION

There is considerable interest among econometricians in forecasting time series, and both direct and iterated forecasting methods play a prominent role. Relevant literature includes Findley (1983, 1985), Weiss (1991), Tiao and Xu (1993), Lin and Granger (1994), Tiao and Tsay (1994), Clements and Hendry (1996), Bhansali (1996, 1997), Kang (2003), Chevillon and Hendry (2005), Chevillon (2007), and Schorfheide (2005). Much of the research focuses on autoregressions—Hoque *et al.* (1988), Ing (2003, 2004), Ing *et al.* (2009), Franses and Legerstee (2010), for example—although Tiao and Xu (1993) consider the impact of moving averages. A study by Marcellino *et al.* (2006)—henceforth MSW—made comparisons between the direct and iterated methods, with the conclusion that in practice the iterated method often performed better. Extensions of these results appear in Proietti (2011), which considers a particular class of ARIMA models as the basis of the forecast functions. The series in MSW were also revisited by Pesaran *et al.* (2011), who considered multivariate approaches as well.

To frame our discussion, we must highlight that either method—direct or iterated—involves not only the use of forecast weights (or filters) peculiar to each method, but also model parameters fitted accordingly. Thus the model parameters used in the direct method and the iterated method could differ in practice; for example, a practitioner could fit an AR($p$) model using a one-step -ahead criterion—say ordinary least squares (OLS) or maximum likelihood estimation (MLE) under a Gaussian assumption—or could fit the parameters so as to minimize $\ell$-step-ahead forecast mean squared error (MSE). Taking $p = \ell$ means that if we are interested in $\ell$-step-ahead forecasts we utilize an AR($\ell$) model; MSW uses a gap AR($\ell$) to directly regress each variable on itself lagged by $\ell$ time units.

Hence discrepancies in performance can arise from several sources: (i) different models are being used; (ii) different fitting methods are being used; (iii) different forecasting functions are being used. By the third point, we refer to forecasting functions that differ even when the same models and same parameter estimates are plugged in. To focus our results, we concentrate on the scenario that a (univariate) model of interest is fixed, but we entertain the possibility of different parameter estimates for the direct and iterated approaches. (Note that some authors have studied how model selection also impacts the comparison of direct and iterated forecasts—see Findley (1983), Liu (1996), Bhansali (1999), and Pesaran *et al.* (2011) for the AR case.) Initially, we might ask: when the parameters are the same, are there differences between the direct and iterated forecast functions? Secondly, we might ask: when the parameters are allowed to differ, how much is forecasting performance affected?

To make the problem well posed, we consider forecasting formulas arising from difference stationary linear time series models (including ARIMA models, for example) such that the resulting forecasts—under the assumption that the data generating process (DGP) has been correctly identified—have minimal MSE given an information set not involving future values of the time series. For difference stationary time series we provide explicit forecasting formulas for either procedure, for both a semi-infinite information set and a finite information set. We also derive the

* Correspondence to: Tucker McElroy, Center for Statistical Research and Methodology, US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA. E-mail: tucker.s.mcelroy@census.gov
† This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not necessarily those of the US Census Bureau.

forecast error processes and determine the MSEs for each case, allowing for misspecification of the model's stationary aspects (although the uncertainty in parameter estimates is not quantified); in the iterative case the formulas are new.

Our main results are summarized as follows. In the case of a semi-infinite past, the multi-step and iterative forecast filters are identical (assuming the same parameters are used to construct each)—the semi-infinite concurrent filters in each case are algebraically the same. Essentially this is due to a property of nested conditional expectations. When a finite past is utilized, the forecast error is mean zero in both cases, and explicit expressions for it give insight into the MSEs for either method. These main mathematical results are presented in the next section; as they are formulated for any difference stationary time series with a Wold representation, the formulas generalize previous work such as that of Stoica and Nehorai (1989), Tiao and Xu (1993), and Ing *et al.* (2009).

We also provide new results on the performance of direct and iterated methods when the parameter estimates differ. A discussion of multi-step-ahead parameter estimation is provided (the basis for this method is discussed in McElroy and Wildi, 2013) within the context of misspecified models—along with a replicable method for comparing asymptotic performance—in the the third section. The method and the numerical results generalize previous work (such as in Proietti, 2011, and other cited papers) from difference AR models to general linear time series models. The key conclusion here is that the parameter estimate—whether based on optimizing one-step-ahead forecasting or multi-step-ahead forecasting—is chiefly responsible for differences in forecast filters and forecast performance.

The fourth section focuses on a retail series, and we compute out-of-sample forecasts to graphically illustrate the main results of the paper. We also provide discussion of MSW, analyzing their conclusions in light of the results of the earlier sections. We draw our main conclusions in the fifth section: the objective function that is used to fit models is of vital importance, with multi-step-ahead criteria tending to generate more conservative forecasts than those based on the one-step-ahead criterion.

## MATHEMATICS OF DIRECT AND ITERATIVE FORECASTING

Some of the following material can be found in a variety of time series references, but we assemble the mathematics here with a coherent notation. A related treatment of direct multi-step-ahead forecasting can be found in McElroy and Findley (2010). We begin by focusing on the case of a semi-infinite past as the information set, and then treat a finite past in the following subsection.

### Semi-infinite past

With $B$ denoting the backshift operator, we suppose that the time series $\{X_t\}$ is difference stationary with operator $\delta(B)$ of degree $d$, such that $\{W_t\}$ satisfies $W_t = \delta(B)X_t$ and is a short-memory covariance stationary process with mean zero. We write its causal Wold representation

$$W_t = \sum_{j \geq 0} \psi_j \epsilon_{t-j} = \Psi(B)\epsilon_t \tag{1}$$

where the process $\{\epsilon_t\}$ is uncorrelated with variance $\sigma^2$. We assume that the power series $\Psi(B)$ is invertible and the coefficients $\psi_j$ are absolutely summable. Without loss of generality, assume also that $\psi_0 = 1$. This type of causal difference linear process includes all ARIMA and SARIMA processes, and thus is fairly broad. It also includes exponential models, but excludes long-memory models.

Suppose that at time $t$ we are interested in generating $h$-step-ahead forecasts based on present and past information, denoted by $X_{t:} = \{X_s : s \leq t\}$. The problem is to compute $\mathbb{E}[X_{t+h}|X_{t:}]$ for some $h > 0$ under a Gaussian assumption—or equivalently, to find the minimal MSE *linear* estimate of $X_{t+h}$ given data up to time $t$. This optimal direct estimate, denoted by $\widehat{X}^{(D)}_{t+h|t}$, can be expressed as a causal filter operating on the $\{X_t\}$ time series, called $\Upsilon_h(B) = \sum_{j \geq 0} \upsilon_j^{(h)} B^j$, namely $\widehat{X}^{(D)}_{t+h|t} = \Upsilon_h(B)X_t$. Because this filter works in an optimal fashion, it may be called the direct multi-step-ahead forecasting filter (cf. Proietti, 2011). In contrast, we might consider applying $\Upsilon_1(B)$ repeatedly, each time appending the previous forecasts to the end of the series, and thereby attaining an iterated multi-step-ahead forecasting filter. This will be denoted by $\Pi_h(B) = \sum_{j \geq 0} \pi_j^{(h)} B^j$, and is described below; both methods, direct or iterated, produce the same estimate when based upon the semi-infinite past $X_{t:}$ due to the nesting property of conditional expectations, as shown below.

We begin the treatment with some results from Bell (1984) on nonstationary stochastic processes. Let $\delta(z) = 1 - \sum_{j=1}^{d} \delta_j z^j$, and its reciprocal power series is $\xi(z) = 1/\delta(z) = \sum_{j \geq 0} \xi_j z^j$. One can recursively solve for the $\{\xi_j\}$ via $\xi_0 = 1$ and $\xi_j = \sum_{k=1}^{\min(d,j)} \delta_k \xi_{j-k}$ for $j \geq 1$. Moreover, certain time-dependent coefficient functions $A_{j,t}$ lying in the null space of $\delta(B)$ are defined via

$$A_{j,t} = \xi_{t-j} - \sum_{k=1}^{d-j} \delta_k \xi_{t-j-k}$$

 

for $j = 1, 2, \ldots, d$ and $t \geq 1$. The process $\{X_t\}$ can then be represented at time $t + h$ for any $h \geq 0$ via

$$X_{t+h} = \sum_{j=1}^{d} A_{j,d+h} X_{t+j-d} + \sum_{j=0}^{h-1} \xi_j W_{t+h-j}$$

It is common in time series analysis to assume that any initial values that generate $\{X_t\}$ are independent of $\{W_t\}$, from which we may conclude that $X_{t:}$ is uncorrelated with $\{\epsilon_s : s > t\}$. We introduce the following bracket notation to refer to that portion of a power series that is retained, namely $[\Psi]_a^b(B) = \sum_{j=a}^{b} \psi_j B^j$ for integers $a$ and $b$. We also study $[\Psi/\delta]_0^{h-1}(B)$, which refers to the first $h$ coefficients in the power series expansion of the rational function $\Psi(B)/\delta(B)$. Then we can state the following result.

**Proposition 1.** Assume that $\{W_t\}$ is covariance stationary, short memory, and mean zero with Wold representation (1), and assume that $X_{t:}$ is uncorrelated with $\{\epsilon_s : s > t\}$. Then the direct forecast filter is given by

$$\Upsilon_h(B) = \sum_{j=1}^{d} A_{j,d+h} B^{d-j} + \sum_{k=1}^{h} \xi_{h-k} [\Psi]_k^{\infty}(B) B^{-k} \delta(B) \Psi^{-1}(B) \tag{2}$$

We may consider a situation where we utilize the filter $\Upsilon_h(B)$ based upon a belief (or model) of $\Psi(B)$ that is incorrect. In other words, suppose that $\{W_t\}$ does not satisfy equation (1), and yet we apply the filter in equation (2). The proof of Proposition 1 also derives the forecast error process $\{\varepsilon_t\}$ in this case. It then follows that the forecast MSE is given by

$$\text{var}(\varepsilon_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\left|[\Psi/\delta]_0^{h-1}(z)\right|^2}{|\Psi(z)|^2} \tilde{f}(\lambda) \mathrm{d}\lambda \tag{3}$$

where $z = e^{-i\lambda}$ and $\tilde{f}$ is the true spectral density of $\{W_t\}$. This formula (3) can be used as the basis for fitting models so as to minimize $h$-step-ahead forecast MSE, as articulated in McElroy and Wildi (2013); special cases have been considered by other authors, e.g. Tiao and Xu (1993).

Now evaluating equation (2) for $h = 1$ produces the one-step-ahead direct forecast filter. Applying its iterative definition yields

$$\Pi_{p+1}(B) = \upsilon_0^{(1)} \Pi_p(B) + \upsilon_1^{(1)} \Pi_{p-1}(B) + \ldots + \upsilon_{p-1}^{(1)} \Pi_1(B) + F^p [\Upsilon_1]_p^{\infty}(B) \tag{4}$$

This is initialized with $\Pi_1(B) = \Upsilon_1(B)$. Iterative forecasting produces an $h$-step-ahead estimate $\widehat{X}_{t+h|t}^{(I)} = \Pi_h(B) X_t$. This filter $\Pi_h(B)$ can be expressed compactly in terms of $\Upsilon_1(B)$ as follows. Let $F = B^{-1}$, and define the degree $k$ polynomials $p_k(F)$ recursively via $p_0(F) = 1$ and

$$p_{k+1}(F) = \sum_{j=0}^{k} \upsilon_j^{(1)} p_{k-j}(F) + F^{k+1}$$

For example, $p_1(F) = \upsilon_0^{(1)} + F$ and $p_2(F) = [\upsilon_0^{(1)}]^2 + \upsilon_1^{(1)} + \upsilon_0^{(1)} F + F^2$, etc. Then

$$\Pi_h(B) = F^h + p_{h-1}(F) [\Upsilon_1(B) - F] \tag{5}$$

which is proved by induction; it also relates the iterative forecast error filter $F^h - \Pi_h(B)$ to the one-step-ahead forecast error filter $F - \Upsilon_1(B)$. Now because $\Pi_h(B) X_t$ and $\Upsilon_h(B) X_t$ both equal $\mathbb{E}[X_{t+h} | X_{t:}]$, the filters must be equal; we state this formally below.

**Proposition 2.** Assume that $\{W_t\}$ is covariance stationary, short memory, and mean zero with Wold representation (1), and assume that $X_{t:}$ is uncorrelated with $\{\epsilon_s : s > t\}$. Then

$$\Upsilon_h(B) = \Pi_h(B)$$

From the proof of Proposition 2, we obtain the identity

$$[\Psi/\delta]_0^{h-1}(B) F^{h-1} = p_{h-1}(F)$$

In order to fit models via the criterion (3), we need to compute this polynomial; the LHS formula $[\Psi/\delta]_0^{h-1}(B)$ is more convenient to work with, since the power series representation of $[\Psi/\delta](z)$ up to a finite number of terms is

 

easily computed computed on a computer. In contrast, the recursive polynomials $p_{h-1}$ require knowledge of the first $h-2$ coefficients of $\Upsilon_1(B)$, requires a separate calculation.

We also mention an alternative approach to these topics using an AR representation in lieu of equation (1):

$$\Omega(B)X_t = \epsilon_t$$

with $\Omega(B) = \delta(B)/\Psi(B)$. Then defining $\Phi(B) = (1 - \Omega(B))F$ and noting that the leading coefficient of $\Omega(B)$ is unity, we obtain

$$X_{t+1} = \Phi(B)X_t + \epsilon_{t+1}$$

As a result, $\mathbb{E}[X_{t+1}|X_{t:}] = \Phi(B)X_t$ and $\Upsilon_1(B) = \Phi(B)$. We can use equation (5), the recursion for $p_k$, and Proposition 2 to obtain $\Pi_h(B)$ and hence $\Upsilon_h(B)$ as well. Formula (3) can then be written in terms of $\Omega$ via

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|[1/\Omega]_0^{h-1}(z)\right|^2 |\Omega(z)|^2}{|\delta(z)|^2}\, \tilde{f}(\lambda)\, \mathrm{d}\lambda$$

noting that $\Omega/\delta = 1/\Psi$ is bounded. Since $1/\Omega$ must be expanded in order to compute forecast MSE, there is no labor saved in approaching the topic through the AR representation. Therefore we will focus on the Wold approach for the rest of the paper.

**Finite past**

Here we maintain the same basic assumptions on the process, but suppose that we are interested in forecasts based on a finite information set $X_{1:n}$, where $n$ denotes the present observation time, as well as the sample size. The conditional expectation $\mathbb{E}[X_{n+h}|X_{1:n}]$ is the target of the direct approach, and we begin by presenting matrix formulas for forecasting and the autocovariance of the forecast error process. Although the treatment is standard (and some of the results can be found in McElroy, 2008, and other literature), we review all derivations for a cohesive treatment.

In the finite-sample treatment it is not necessary to utilize a causal Wold representation for the differenced data process; we only require that the covariance function $\gamma_h$ of the $\{W_t\}$ process be well defined. This allows us to extend our formulas to long-memory models, or any model for $\{W_t\}$ that is covariance stationary. We require the following notation. Let $\Delta_m$ be the (square) differencing matrix of dimension $m$ such that the upper left $d \times d$ block is an identity matrix and the lower $m - d$ rows are given by the coefficients of $\delta(z)$ appropriately shifted. Such matrices are standard in the literature on finite-sample projections (cf. McElroy, 2008). The $jk$th entry of $\Delta_m$ for $j > d$ is given by the $j - k$th coefficient of $\delta(z)$ (by convention, a coefficient of $\delta(z)$ with index less than zero or greater than $d$ is just equal to zero). This differencing matrix is unit lower triangular, as is its inverse. Likewise, the Toeplitz covariance matrix of dimension $m$ for $\{W_t\}$ is denoted by $\Sigma_m$, i.e. $[\Sigma_m]_{jk} = \gamma_{j-k}$.

As a preliminary, we have a representation of $X_{1:n}$ in terms of initial values given as follows. Interpreting $X_{1:n}$ as a column vector, we have $\Delta_n X_{1:n} = [X'_{1:d}, W'_{d+1:n}]'$. Here, subindices refer to the collection of corresponding random variables, collected into a column vector. The first $d$ values $X_{1:d}$ are referred to as initial values in the forecasting and signal extraction literature. Typically a time series model is specified conditionally on these initial values, and furthermore it is common to assume the initial values are uncorrelated with the increment process $\{W_t\}$. This assumption is ubiquitous in the time series forecasting literature, and is implicit in all state space smoothing algorithms for forecasting and signal extraction (Bell and Hillmer, 1991). When deriving forecasting results under this assumption, the filters do not depend on the distribution of the initial values, and moreover the forecast error process does not depend on the initial values themselves, which is a desirable feature.

Let $\widehat{X}^{(D)}_{n+h|1:n} = \mathbb{E}[X_{n+h}|X_{1:n}]$ be the optimal direct $h$-step-ahead forecast under a Gaussian assumption; for non-Gaussian data, $\widehat{X}^{(D)}_{n+h|1:n}$ denotes the minimal MSE linear estimate of $X_{n+h}$ given data $X_{1:n}$. We next derive its formula; while this is not novel (similar expressions can be found in McElroy, 2008), we present it for completeness. Let $e_{n+h}$ denote a vector of length $n + h$ that is zero except for a unit in the last entry, and let $\dagger$ denote the inverse transpose of a matrix. Also $1_m$ denotes an $m$-dimensional identity matrix, while $o_{m\times p}$ is an $m \times p$ dimensional matrix of zeros.

**Proposition 3.** Assume that $X_{1:d}$ is uncorrelated with $\{W_t\}$. Then

$$\widehat{X}^{(D)}_{n+h|1:n} = e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}\Delta_n\, X_{1:n} \tag{6}$$

 

and the covariance of the forecast error is

$$e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_{d\times d} & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d} - \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}'\Sigma_{n+h-d}\end{bmatrix}\Delta^{\dagger}_{n+h}e_{n+h} \tag{7}$$

Both equations (6) and (7) are very easy to program. We now proceed to discuss iterative forecasting. Let $\widehat{X}^{(I)}_{n+h|1:n}$ be the iterative forecast obtained by recursively applying the one-step-ahead direct forecast matrix, as described above. For example, $\widehat{X}^{(I)}_{n+1|1:n} = \widehat{X}^{(D)}_{n+1|1:n} = \eta' X_{1:n}$, where $\eta'$ corresponds to the length $n$ row vector multiplying $X_{1:n}$ in (6). Then $\widehat{X}^{(I)}_{n+2|1:n}$ is obtained by applying $\eta'$ to $X_{2:n}$ appended by $\widehat{X}^{(I)}_{n+1|1:n}$. Alternatively, we could have used the length $n+1$ row vector $\eta'$ corresponding to a sample of length $n+1$, and applied it to $X_{1:n}$ appended by $\widehat{X}^{(I)}_{n+1|1:n}$; this would use all of the data. However, this procedure requires us to recompute $\eta$ for each iterate, and the idea of iterative forecasting is to use the *same* method iteratively. For example, with AR($p$) models $\eta'$ consists of at most $p$ nonzero coefficients, no matter the length of the data (so long as $n \geq p$). In terms of conditional expectations, we have

$$\begin{aligned} \widehat{X}^{(I)}_{n+1|1:n} &= \mathbb{E}[X_{n+1}|X_{1:n}] \\ \widehat{X}^{(I)}_{n+2|1:n} &= \mathbb{E}[\,\mathbb{E}[X_{n+2}|X_{2:n+1}]\,|X_{1:n}] \end{aligned} \tag{8}$$

Note that the latter expression could be defined instead as

$$\mathbb{E}[\,\mathbb{E}[X_{n+2}|X_{1:n+1}]\,|X_{1:n}] \tag{9}$$

if we used all the data in each iteration; by nested conditional expectations, this reduces at once to $\mathbb{E}[X_{n+2}|X_{1:n}]$, i.e. the direct estimate. Instead, we focus on the iterative case, namely equation (8), since the definition by equation (9) immediately yields the direct estimate.

The formula for $\eta$ is given by

$$\eta' = e'_{n+1}\,\Delta^{-1}_{n+1}\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n+1-d\times d} & \Sigma_{n+1-d}\begin{bmatrix} 1_{n-d} \\ 0_{1\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}\Delta_n \tag{10}$$

The iterative procedure amounts to appending the most recent one-step-ahead forecast to past data and forecasts, which may be formalized in matrix notation as follows. Define

$$J = \begin{bmatrix} 0_{n-1\times 1} & 1_{n-1} \\ \eta' \end{bmatrix} \tag{11}$$

so that $JX_{1:n}$ consists of data values $X_{2:n}$ with the one-step-ahead forecast $\eta' X_{1:n}$ appended at the end. This process is then recursively repeated, so that $\widehat{X}^{(I)}_{n+h|1:n} = e'_n\, J^h\, X_{1:n}$. Note the similarity to results in Proietti (2011), although the matrix power here involves matrices of full dimension $n$ rather than just the model order. Our treatment generalizes the previous treatment of difference AR processes to generic difference stationary processes.

It is shown in the proof of Proposition 3 that the direct forecast error does not depend upon initial values $X_{1:d}$. This is important, so that the forecast errors have dynamics that are not contingent on the level of the series. The same property holds for the iterative forecasts, which is not obvious from the formula; in fact, the iterative forecast error can be expressed in terms of the one-step-ahead (direct) forecast errors, which depend upon the following matrix $K$:

$$K = \Sigma_{n+1-d}\begin{bmatrix} 1_{n-d} \\ 0_{1\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\,[1_{n-d} \quad 0_{n-d\times 1}] - 1_{n+1-d} \tag{12}$$

The next result summarizes the properties of the iterative forecasts.

**Proposition 4.** Assume that $X_{1:d}$ is uncorrelated with $\{W_t\}$, and that the iterative forecasts are defined via

$$\widehat{X}^{(I)}_{n+h|1:n} = e'_n\, J^h\, X_{1:n} \tag{13}$$

 

where $J$ is given by equation (11). Then the iterative forecast error can be written as

$$\widehat{X}^{(I)}_{n+h|1:n} - X_{n+h} = e_n' \sum_{k=0}^{h-1} J^k \begin{bmatrix} 0_{n-1\times n+h-d} \\ 0_{1\times h-k-1} \quad \beta' \quad 0_{1\times k} \end{bmatrix} W_{d+1:n+h}$$

$$\beta' = e_{n+1}' \, \Delta_{n+1}^{-1} \begin{bmatrix} 0_{d\times n+1-d} \\ K \end{bmatrix}$$

where $K$ is given in equation (12). Hence the MSE of the $h$th iterative forecast is

$$e_n' \sum_{k,\ell=0}^{h-1} J^k \begin{bmatrix} 0_{n-1\times n+h-d} \\ 0_{1\times h-k-1} \quad \beta' \quad 0_{1\times k} \end{bmatrix} \Sigma_{n+h-d} \left[ 0_{n+h-d\times n-1} \; [0_{1\times h-\ell-1} \quad \beta' \quad 0_{1\times \ell}]' \right] J'^{\ell} \, e_n$$

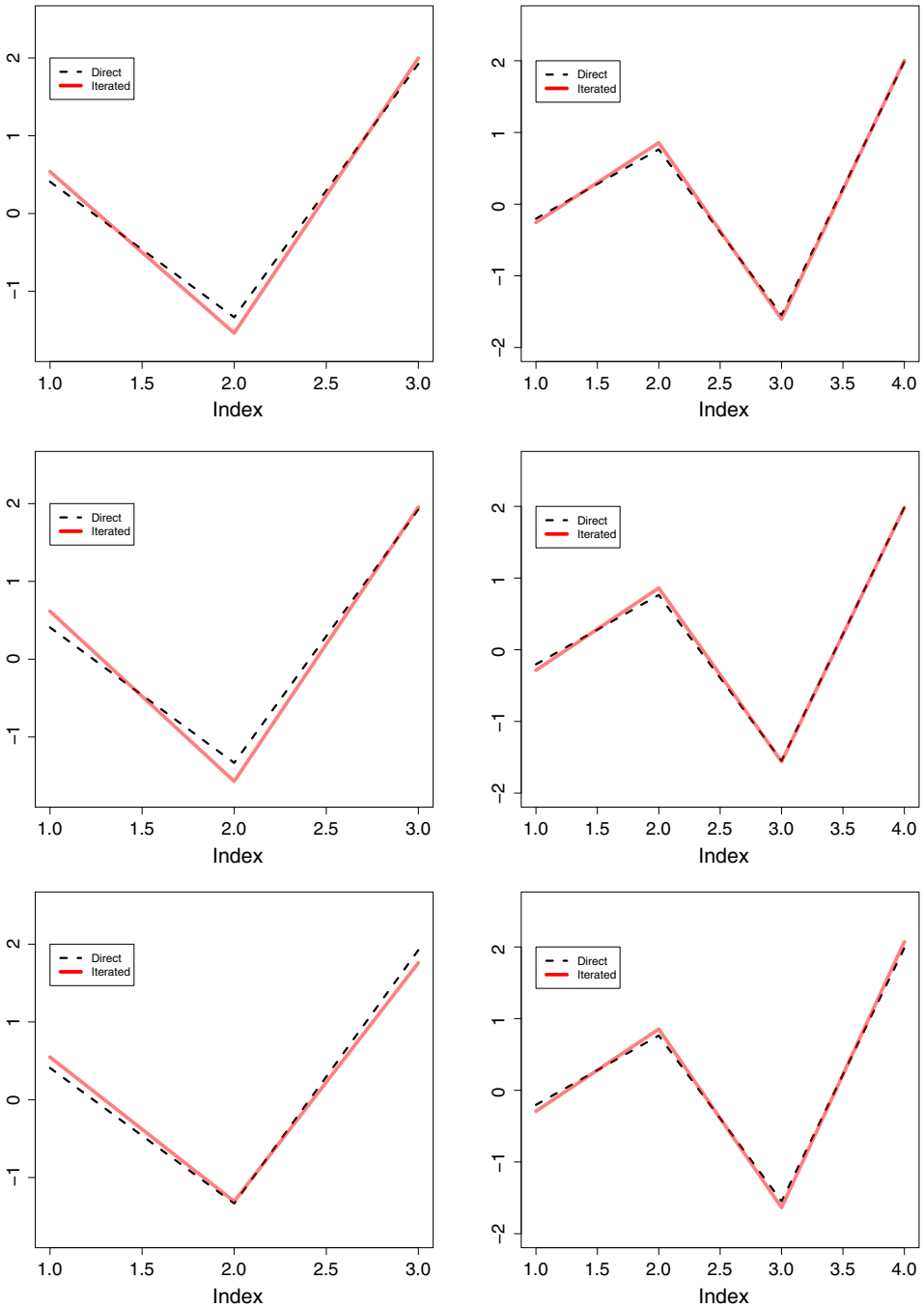


Figure 1. Forecast weights for the direct and iterated methods, for an ARIMA(0,1,2) model with sample sizes $n = 3$ (left panels) or $n = 4$ (right panels) data points. Forecast horizons $h = 2, 3, 6$ are given in the top, middle, and lower panels respectively

 

In general, direct and iterative forecasting utilize different formulas that are only identical in special cases. Given a model, both the direct and indirect filters can be calculated as a function of the model parameters—these might be parameters fitted to data, or any other numbers we choose. Then, for any such parameter choice, it is simple to compare forecast filters numerically. For example, for an MA(1) model we can compute both types of filters for any value of the MA parameters in $(-1, 1)$, and compare the resulting filters. We have investigated this, with the main conclusion that there is little discrepancy between direct and iterative forecasts even when the sample size is quite small, whenever *the same parameters are used in the forecast formulas*; the situation where different parameters are used for the competing methods is explored in the next two sections.

For example, consider the forecast coefficients as a function of forecast lead $h$ and sample size $n$, for both the direct and iterated methods. In order to isolate different aspects of the comparison, we utilize the same model parameters (set equal to the true parameters corresponding to our simulation) so that we may focus on the impact of sample size $n$. Recall that, from the results above ('Semi-infinite past'), as $n \to \infty$ the direct and iterated forecast coefficients will converge to one another, because their model parameters are the same. In fact, for simple cases like an ARIMA(1,0,0) or ARIMA(1,1,0), the forecast coefficients for the iterated and direct methods are identically the same—this is easily seen from the formula for $\Upsilon_h(B)$, which only requires a few data points ($n = 1, 2$ respectively for the above ARIMA models), and hence the finite-sample formulas agree with the semi-infinite case, and the earlier result applies. More generally, subtle differences can arise between forecast coefficients of the direct and iterated methods (again, using the same parameters); Figure 1 displays results for an ARIMA(0,1,2) model with (non-invertible) MA polynomial $1 + 0.8B + 0.2B^2$. Discrepancies are minute in this case, even with $n = 3, 4$; under 'Empirical illustration' below we demonstrate these types of comparisons on a real time series.

## ASYMPTOTIC FORECAST MSE PERFORMANCE

Given that direct and iterated forecast formulas are extremely similar when the parameters in the respective formulas are identical, how much do they differ when the parameters differ? In order to answer this question, such that finite-sample effects are isolated, we consider the following paradigm: the parameters for both the direct and iterative methods respectively will correspond to asymptotic values (called pseudo-true values) for the pertinent estimates, and we then calculate the asymptotic forecast MSE arising in either case. First we describe the mathematics behind this asymptotic forecast MSE, and then apply it to two different types of data processes.

**Pseudo-true values and asymptotic forecast MSE**

The parameter estimates for the direct forecasting method arise from minimizing an in-sample empirical $\ell$-step-ahead forecasting criterion, whereas the iterative method uses one-step-ahead estimates—or equivalently, the MLEs when the likelihood is Gaussian. Although the formula for empirical multi-step forecasting is simple for an AR or difference AR model, the general ARIMA case is more complicated. A full treatment is given in McElroy and Wildi (2013), which we briefly summarize here. It turns out that formula (3) can be used as a model-fitting criterion when we replace the true data spectrum $\tilde{f}$ by the periodogram $I$ of the differenced data $\{W_t\}$, which is defined via

$$I(\lambda) = (n-d)^{-1} \left| \sum_{t=d+1}^{n} W_t \, e^{-i\lambda t} \right|^2$$

for $\lambda \in [-\pi, \pi]$. Suppose that we are studying a model with parameter vector $\vartheta$, so that the Wold representation of that model involves a causal filter $\Psi_\vartheta(B)$. Throughout this section, $\Psi_\vartheta$ denotes the model's Wold filter, while $\tilde{f}$, in contrast, is the true process's spectrum. Then equation (3) tells us that the asymptotic forecast MSE—when we generate forecasts from the given model with (deterministic) parameter $\vartheta$—is given by $J_\ell(\vartheta, \tilde{f})$, where

$$J_\ell(\vartheta, g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|[\Psi_\vartheta/\delta]_0^{\ell-1}(z)|^2}{|\Psi_\vartheta(z)|^2} \, g(\lambda) \mathrm{d}\lambda \tag{14}$$

and $z = e^{-i\lambda}$. However, $J_\ell(\vartheta, \tilde{f})$ does not depend on the observed sample of data, and thus cannot be used as an empirical criterion. But it turns out—demonstrated in McElroy and Wildi (2013)—that $J_\ell(\vartheta, I)$ is approximately equal to the in-sample $\ell$-step-ahead empirical forecast MSE. Note that $J_\ell(\vartheta, I)$ is well defined for all $n$ and $\ell$, because we take the periodogram at all frequencies, not just the Fourier frequencies; moreover, it can be shown that $J_\ell(\vartheta, I) = (n-d)^{-1} W' \Sigma W$ for $W' = [W_{d+1}, \ldots, W_n]$ and $\Sigma$ the $n-d$ dimensional covariance matrix corresponding to a time series with spectral density $|[\Psi/\delta]_0^{\ell-1}(z)|^2 |\Psi(z)|^{-2}$.

Hence minimization of $J_\ell(\vartheta, I)$ with respect to $\vartheta$ is the appropriate way to fit models so as to minimize $\ell$-step-ahead forecast error. The resulting minima will be denoted $\vartheta_I^\ell$. Similarly, the minima of $J(\vartheta, \tilde{f})$ will be denoted $\vartheta_{\tilde{f}}^\ell$, and are called the $\ell$-step-ahead pseudo-true values (PTVs). When $\ell = 1$, these are the usual PTVs arising from the

 

Whittle likelihood (Taniguchi and Kakizawa, 2000). The $\ell$-step-ahead estimates $\vartheta_I^\ell$ are consistent for the $\ell$-step-ahead PTVs under some regularity conditions described in McElroy and Wildi (2013).

Thus PTVs may be conceived of as the numbers to which estimates converge when a model is misspecified. When the model is correctly specified, it is true that $\tilde{f}(\lambda)$ is proportional to $|\Psi_{\tilde{\vartheta}}(e^{-i\lambda})|^2$ for some true $\tilde{\vartheta}$. In this case, it can be shown that the PTVs are identical with the true parameter, i.e. $\vartheta_{\tilde{f}}^\ell = \tilde{\vartheta}$. But when the model is misspecified, there is no such thing as a true parameter, and the PTVs are instead the pertinent asymptotic quantity.

Thus the empirical $\ell$-step-ahead forecast MSE is approximated by $J_\ell(\vartheta_I^\ell, I)$; noting that $\vartheta_I^\ell \xrightarrow{P} \vartheta_{\tilde{f}}^\ell$, by Slutsky's theorem and Lemma 3.1.1 of Taniguchi and Kakizawa (2000) we have

$$J_\ell\left(\vartheta_I^\ell, I\right) \xrightarrow{P} J_\ell\left(\vartheta_{\tilde{f}}^\ell \tilde{f}\right)$$

as $n \to \infty$. This result requires that the Wold filter of the model be invertible, as discussed above ('Mathematics of direct and iterative forecasting'), in addition to technical assumptions on the process $\{W_t\}$ such as Gaussianity—see full discussion in McElroy and Wildi (2013) for details.

Thus, from an asymptotic viewpoint, $J_\ell\left(\vartheta_{\tilde{f}}^\ell, \tilde{f}\right)$ is the quantity of interest for direct $\ell$-step-ahead forecasting, because it equals the asymptotic forecast MSE that arises from using parameters fitted so as to minimize forecast error, such that the case of a misspecified model is included. Given that the PTV $\vartheta_{\tilde{f}}^\ell$ arises from an $\ell$-step-ahead criterion, we might be interested in asymptotic performance at some other forecast lead $h \neq \ell$, in which case we compute $J_h\left(\vartheta_{\tilde{f}}^\ell, \tilde{f}\right)$. In particular, the iterative technique obtains parameters using $\ell = 1$; because direct and iterative *filters* are identical asymptotically, the $h$-step-ahead performance is $J_h\left(\vartheta_{\tilde{f}}^1, \tilde{f}\right)$.

This treatment generalizes the focused discussion in Proietti (2011), maintaining a similar notation—here we use $\ell$ for the steps ahead for the direct forecasts, whereas in Proietti (2011) the models contain an AR factor of order $p$, so that considering $p$-step-ahead forecasting is natural. Thus $\ell$ here serves the role that $p$ played in Proietti (2011). As in that paper, the quantities of chief interest from an asymptotic perspective are $J_h\left(\vartheta_{\tilde{f}}^\ell, \tilde{f}\right)$ with $\ell = 1$ for the iterative, and any other value exceeding one for $\ell$-step-ahead direct forecast fitting, and for a variety of $h \geq 1$. The minimal possible such asymptotic MSE is obtained in the case of a correct model, where $\tilde{f}(\lambda) = |\Psi_{\tilde{\vartheta}}(e^{-i\lambda})|^2$ (assuming a unit innovation variance), which by equation (3) is

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |[\Psi_{\tilde{\vartheta}}/\delta]_0^{h-1}(z)|^2 \, \mathrm{d}\lambda$$

This quantity is always a lower bound on the other asymptotic forecast MSEs, and we denote it by $J_h(\tilde{\vartheta}, \tilde{f})$—it being understood by this notation that we do the calculation under the assumption that the model is correctly specified. Any of these quantities ultimately depends upon the data process only through $\tilde{f}$. So if the data process can be described through a single parameter (e.g. an AR(1)) then the resulting quantities of interest can be graphed as curves. If the data process is described by two parameters (e.g. an ARMA(1,1)), then the resulting visualization would be a surface. In what follows, we use contour plots to visualize PTVs and forecast MSE surfaces.

**The ARIMA(1,1,1) data process**

As an application of the preceding discussion, we suppose the true data process is an ARIMA(1,1,1) that is fitted by an ARIMA(0,1,2) model. Suppose we fit using both a one-step- and two-step-ahead forecasting criterion, so $\ell = 1, 2$. For the one-step-ahead fitting we minimize $J_1(\vartheta, I)$ and compute PTVs by minimizing $J_1(\vartheta, \tilde{f})$. The differencing operator is $\delta(B) = 1 - B$, although this is unimportant in the criterion because $\ell = 1$. The Wold filter for $W_t = (1 - B)X_t$ is $\Psi(B) = 1 + \vartheta_1 B + \vartheta_2 B$, and $\tilde{f}(\lambda) = |1 + \theta z|^2 |1 - \phi z|^{-2}$. The true data process is governed by the parameters $\theta$ and $\phi$, which each belong to $(-1, 1)$ to guarantee stability and invertibility. This describes a manifold $(-1, 1) \times (-1, 1)$, whose counter-diagonal element (where $\theta = -\phi$) describes a white noise process. On the other hand, when $\ell = 2$, $\delta(B) = 1 - B$ has a presence in the objective function, as equation (14) shows. Hence the ARMA$(p, q)$ and ARIMA$(p, d, q)$ have different objective functions for their estimates and PTVs.

The exercise resolves into two steps: first to compute the PTVs for each $\ell = 1, 2$, and secondly to compute $J_h\left(\vartheta_{\tilde{f}}^\ell, \tilde{f}\right)$, as well as the minimal MSE $J_h(\tilde{\vartheta}, \tilde{f})$. For the data process we consider all invertible ARIMA(1,1,1) models, so that $\phi, \theta \in (-1, 1)$ (we consider values in increments of 1/10). For the fitted models, we consider the ARIMA(0,1,2) as Model 1, the ARIMA(2,1,0) as Model 2, and the correct ARIMA(1,1,1) as Model 0. Note that for some special values of the process parameters, Models 1 and 2 can also be correctly specified. For example, if $\theta = 0$ then Model 2 is correctly specified, because it nests the resulting ARIMA(1,1,0) process. Each PTV involves two components, each of which is a function of $(\phi, \theta)$, so these can be plotted as pairs of surfaces. For Model 1, these

 

two surfaces are the first and second MA coefficients, whereas for Model 2 the two surfaces are the first and second AR coefficients (in the graphs the first coefficient PTVs are green, and the second are red). Thus each component of a PTV is graphed as a two-dimensional manifold, and depends on whether $\ell = 1$ or $\ell = 2$. Recall that when the model is correctly specified, PTVs are equal to the true parameter—so on the subset where $\theta = -\phi$, the true process is white noise, and so the PTVs should be equal to zero.

Although PTVs are not a primary focus of this paper, we display some of the PTV surfaces to allow visualization of the impact of model misspecification (Figures 2 and 3). In our implementation, we minimize $J_\ell(\vartheta, \tilde{f})$ numerically,

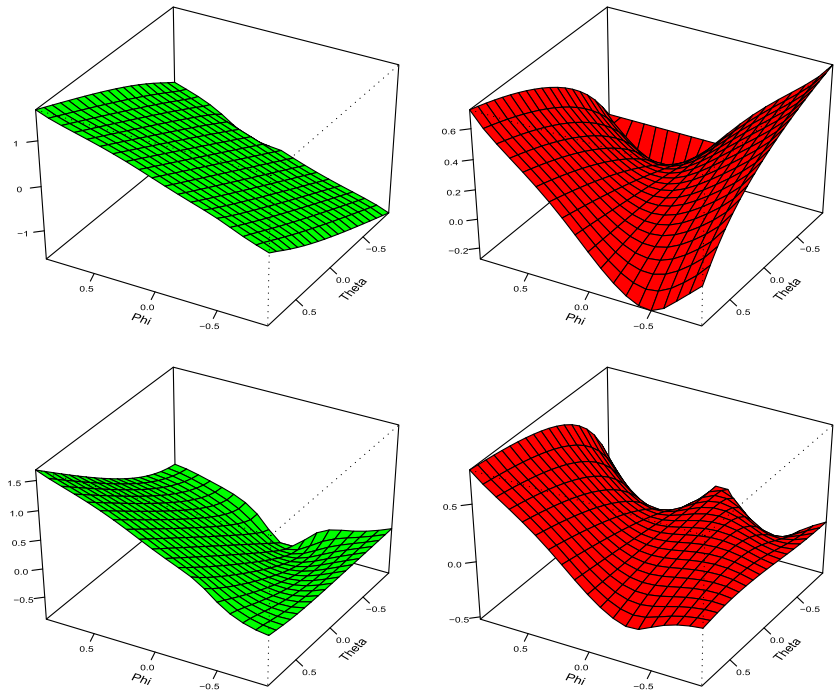


Figure 2. PTVs for the ARIMA(0,1,2) model (Model 1) as a function of the $\phi$ and $\theta$ parameters of the ARIMA(1,1,1) data process. Left panels (green) are for the first MA parameter, and right panels (red) are for the second MA parameter. Top panels are for $\ell = 1$ (iterated) and bottom panels are for $\ell = 2$ (direct)

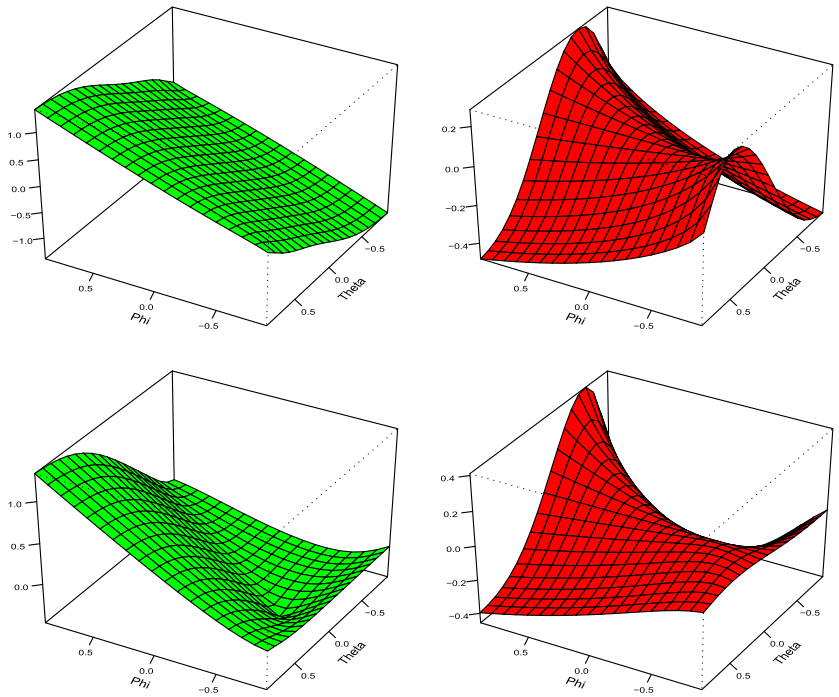


Figure 3. PTVs for the ARIMA(2,1,0) model (Model 2) as a function of the $\phi$ and $\theta$ parameters of the ARIMA(1,1,1) data process. Left panels (green) are for the first AR parameter, and right panels (red) are for the second AR parameter. Top panels are for $\ell = 1$ (iterated) and bottom panels are for $\ell = 2$ (direct)

using a root-flipping technique that ensures the optima correspond to stable/invertible ARMA processes. Then for each PTV (which is a function of the process parameters, $\ell$, and the model) we can plug into $J_h\left(\vartheta^{\ell}_{\tilde{f}}, \tilde{f}\right)$ and compute the resulting surface, for any $h$ desired. In order to standardize these surfaces for comparison, we plot

$$S_{\ell,h,j}(\theta,\phi) = \frac{J_h\left(\vartheta^{\ell}_{\tilde{f}}, \tilde{f}\right)}{J_h(\tilde{\vartheta}, \tilde{f})}$$

for $\ell = 1, 2$ and $h = 1, 2, 3, 6$, and $j = 1, 2$ indexing Model $j$. If we fix $h$ and consider $\ell = 1, 2$ together, it provides a comparison of the iterative and direct methods (of fitting) assessed in terms of asymptotic $h$-step-ahead forecast performance, each made relative to the (common) minimal possible MSE. These surfaces $S_{\ell,h,j}$ are presented as contour plots, with a minimal value of one due to the normalization, in Figures 4–7 for $h = 1, 2, 3, 6$ respectively. In each case, the top panels are $S_{\ell,h,1}$, while bottom panels are $S_{\ell,h,2}$; left panels are $S_{1,h,j}$ and right panels are $S_{2,h,j}$.

The gains to performance obtained in the direct approach are slight in those particular cases where they are to be expected (e.g. when $h = \ell$), but the iterated performance can be quite a bit superior otherwise. This reflects the idea that direct estimation is customized to a particular forecast lead, i.e. with $\ell$ chosen to be equal to a particular $h$ of interest; but the iterative approach with $\ell = 1$ is better as an all-purpose fit, when one might be concerned about performance across a variety of $h$ horizons.

One step ahead ($h = 1$), we expect the iterative ($\ell = 1$) method to be superior (because here $h = \ell = 1$, so that PTVs have been optimized for one-step-ahead performance); comparing left panels to right panels, this is indeed the case. In comparing Model 1 and Model 2, up to scale changes the latter appears to be roughly a 90° clockwise rotation of the latter, for the iterative method. Moving to $h = 2$ (Figure 5), the gains of the direct method are minuscule, and hardly apparent. Also for $h = 3$ and $h = 6$ the MSE surfaces are quite similar for the iterative and direct methods; the more interesting discrepancies lie between Model 1 and Model 2, the latter tending to have lower MSE overall.

This would indicate that the autoregressive formulation is superior for short- and long-term forecasting when integration is present. While the role of the forecast model, and its form of misspecification, has a notable impact upon forecast performance, this is not the primary focus of this paper; nevertheless, the tools introduced above ('Pseudo-true values and asymptotic forecast MSE') can be profitably utilized to study this role.

**The airline data process**

We now repeat the exercise of the previous subsection, but with a slightly expanded orientation. We now consider the true data process to be a Box–Jenkins airline model, which is the SARIMA(0,1,1)(0,1,1) with monthly frequency.

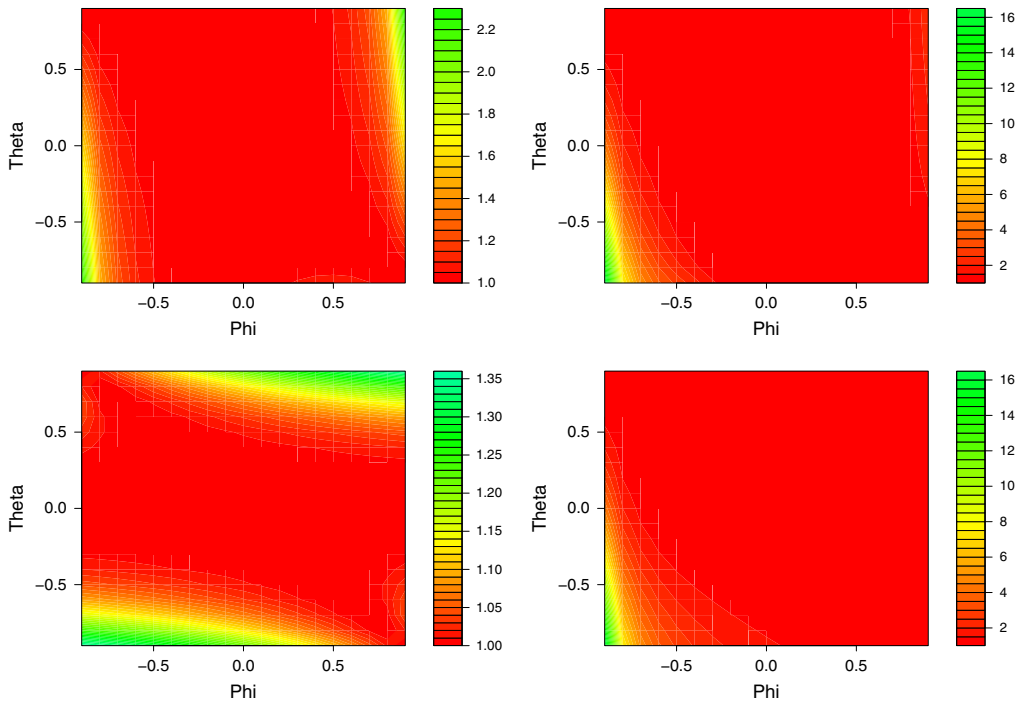


Figure 4. MSE ratio surfaces $S_{\ell,1,j}$ for the direct and iterative methods for misspecified models, as a function of the $\phi$ and $\theta$ parameters of the ARIMA(1,1,1) data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 2$). Top panels are for Model 1 (the ARIMA(0,1,2)) and bottom panels are for Model 2 (the ARIMA(2,1,0)). The forecast lead considered is $h = 1$

 

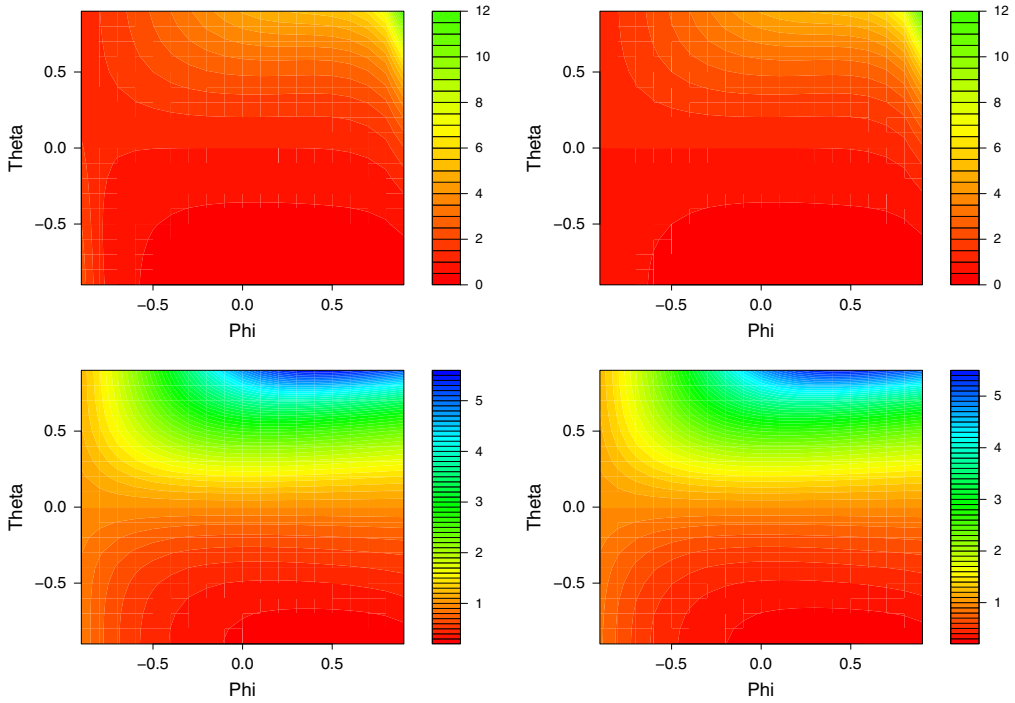


Figure 5. MSE ratio surfaces $S_{\ell,2,j}$ for the direct and iterative methods for misspecified models, as a function of the $\phi$ and $\theta$ parameters of the ARIMA(1,1,1) data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 2$). Top panels are for Model 1 (the ARIMA(0,1,2)) and bottom panels are for Model 2 (the ARIMA(2,1,0)). The forecast lead considered is $h = 2$

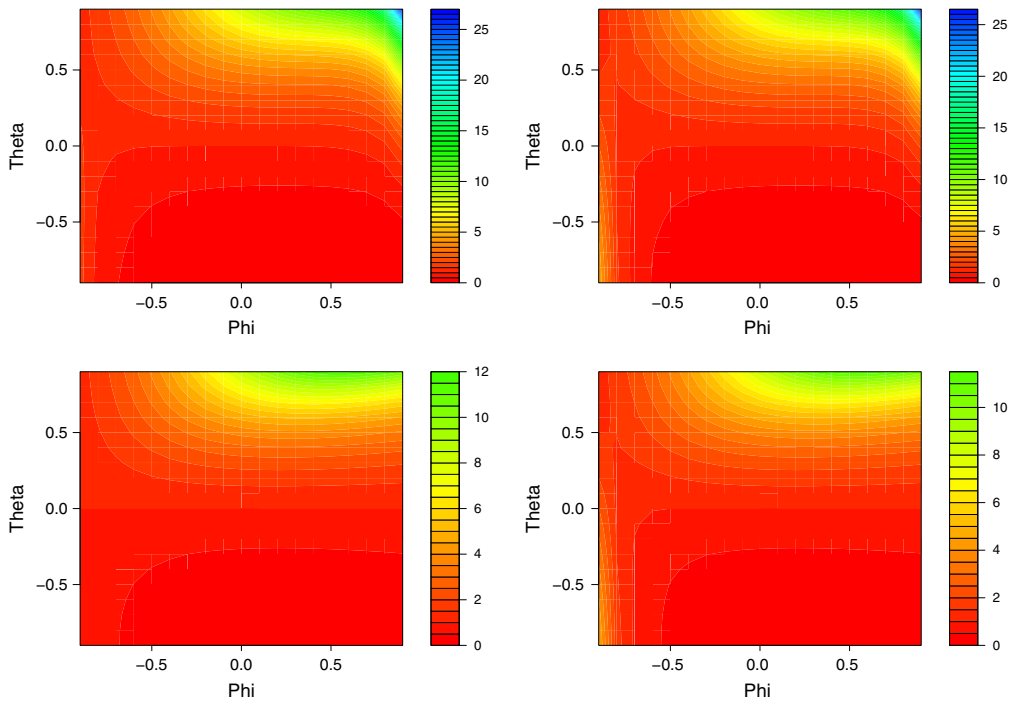


Figure 6. MSE ratio surfaces $S_{\ell,3,j}$ for the direct and iterative methods for misspecified models, as a function of the $\phi$ and $\theta$ parameters of the ARIMA(1,1,1) data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 2$). Top panels are for Model 1 (the ARIMA(0,1,2)) and bottom panels are for Model 2 (the ARIMA(2,1,0)). The forecast lead considered is $h = 3$

Multi-step-ahead forecasting generated from SARIMA models is utilized in seasonal adjustment procedures—such as X-12-ARIMA (Findley *et al.*, 1998)—that are used to adjust millions of time series each month at statistical agencies across the world. While the SARIMA models are typically fitted to minimize one-step-ahead forecast error (i.e. maximum likelihood estimation is used), there is some literature on multi-step-ahead fitting of seasonal time

 

series; see Gersch and Kitagawa (1983) and, more recently, Haywood and Tunnicliffe-Wilson (1997) and McElroy and Wildi (2013). These considerations motivate our study of such models.

We are chiefly interested in $\ell = 1, 12$, with a focus on $h = 1, 12$. The airline process has two parameters: the nonseasonal moving average parameter $\theta_{\mathrm{NS}}$ and the seasonal moving average parameter $\theta_{\mathrm{S}}$—denoted by thetaNS and

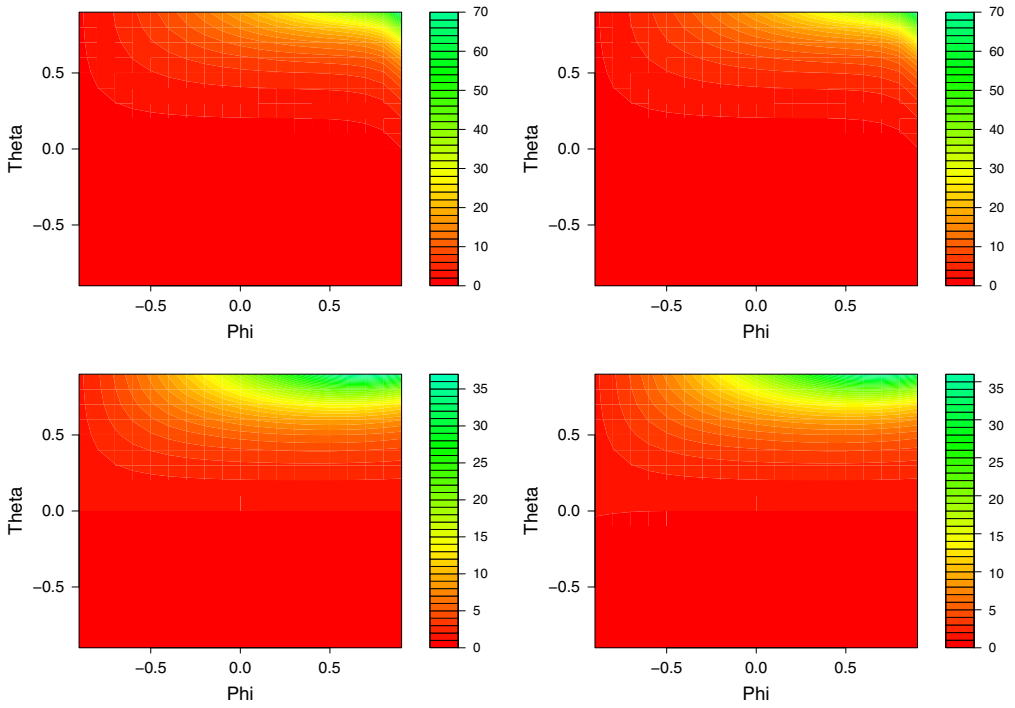


Figure 7. MSE ratio surfaces $S_{\ell,6,j}$ for the direct and iterative methods for misspecified models, as a function of the $\phi$ and $\theta$ parameters of the ARIMA(1,1,1) data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 2$). Top panels are for Model 1 (the ARIMA(0,1,2)) and bottom panels are for Model 2 (the ARIMA(2,1,0)). The forecast lead considered is $h = 6$

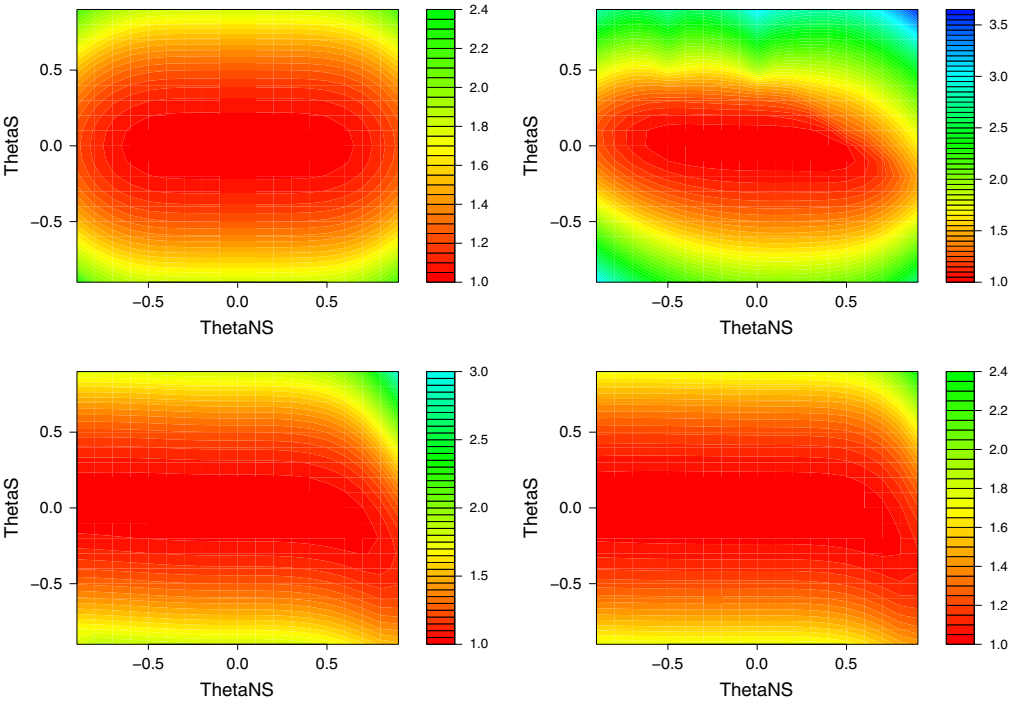


Figure 8. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (2,1,0)(0,1,0) model, as a function of the $\theta_{\mathrm{NS}}$ and $\theta_{\mathrm{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

thetaS in the figures—which each take values in $(-1, 1)$ to ensure invertibility. We consider a broad class of SARIMA models with correctly specified differencing operator here is $\delta(B) = (1 - B)(1 - B^{12})$, including the following types (in addition to the airline model): (2,1,0)(0,1,0); (0,1,2)(0,1,0); (1,1,1)(0,1,0); (0,1,1)(1,1,0); (1,1,0)(0,1,1); (1,1,0)(1,1,0); (0,1,0)(1,1,1). All of these models involve only two parameters, so the PTVs can be displayed in each case by a pair of surfaces. These PTV plots tend to be less interesting than in the previous ARMA case, and for concision are not displayed.

Figures 8–14 display the surfaces $S_{\ell,h}(\theta_{\text{NS}}, \theta_{\text{S}})$ for $\ell = 1, 2$ and $h = 1, 2$, for each of the seven misspecified models. The top panels of each figure correspond to $h = 1$, whereas bottom panels are for $h = 12$; left panels are for $\ell = 1$, while right panels are for $\ell = 2$. In some cases there is a substantial discrepancy between performance of the direct and iterative methods when $h = 12$, which is of some interest. Note that the $z$-scales are not common in the plots. In one case (Figure 10) the values were quite large in some areas of the surface, and these are colored white.

Each of Figures 8–14 considers a particular model, and for $h = 1$ (top panels) the iterative method is superior to the direct, as expected. In some cases, e.g. Figures 9 and 10, the discrepancy to performance is dramatic. On the other hand, for $h = 12$ (bottom panels) the direct method is superior, although in most cases the improvement is not dramatic. For example, Figures 13 and 14 show a rough 15–25% improvement to MSE, with the pattern being quite similar for the iterative and direct methods.

In terms of the best models to utilize, obviously the (011)(011) is the optimal (not shown) because it is correctly specified. But if using misspecified models, for $h = 1$ the (011)(110) of Figure 11 and the (110)(011) of Figure 12 have the best performance; it is interesting that their seasonal and nonseasonal AR and MA portions are interchanged in these models, and that their MSE surfaces appear to be a 90° rotation of one another. For $h = 12$, the (110)(011) and (010)(111) models of Figures 12 and 14 appear to have the lowest overall MSE, and the patterns of the surfaces are similar. It would be of interest to conduct a large study that seeks to discover which models have the best forecast performance across a whole range of DGP types (so instead of just considering the airline DGP, repeat the studies with other SARIMA DGPs). This question is beyond our scope here, but could be investigated numerically using the above tools ('Pseudo-true values and asymptotic forecast MSE').

## EMPIRICAL ILLUSTRATION

We have already claimed that forecasts generated from the direct and iterative approaches—*when the same parameter estimates are used in the formulas*—produce indistinguishable results. However, when parameter estimates differ the performance can be quite different. We proceed to illustrate these concepts through the monthly time series

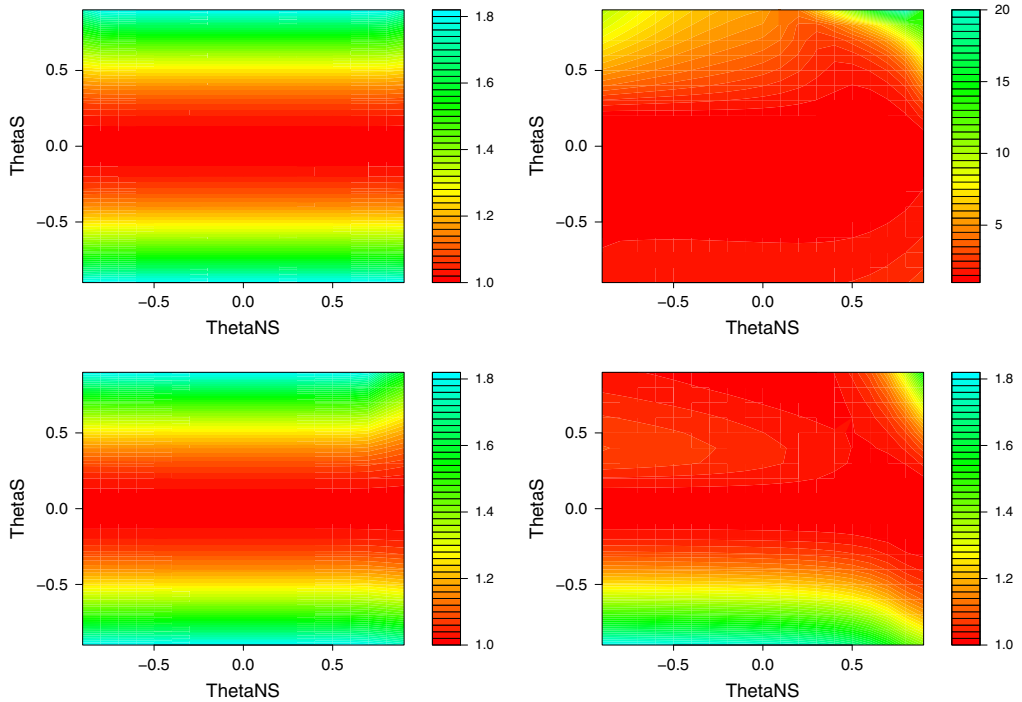


Figure 9. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (0,1,2)(0,1,0) model, as a function of the $\theta_{\text{NS}}$ and $\theta_{\text{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

 

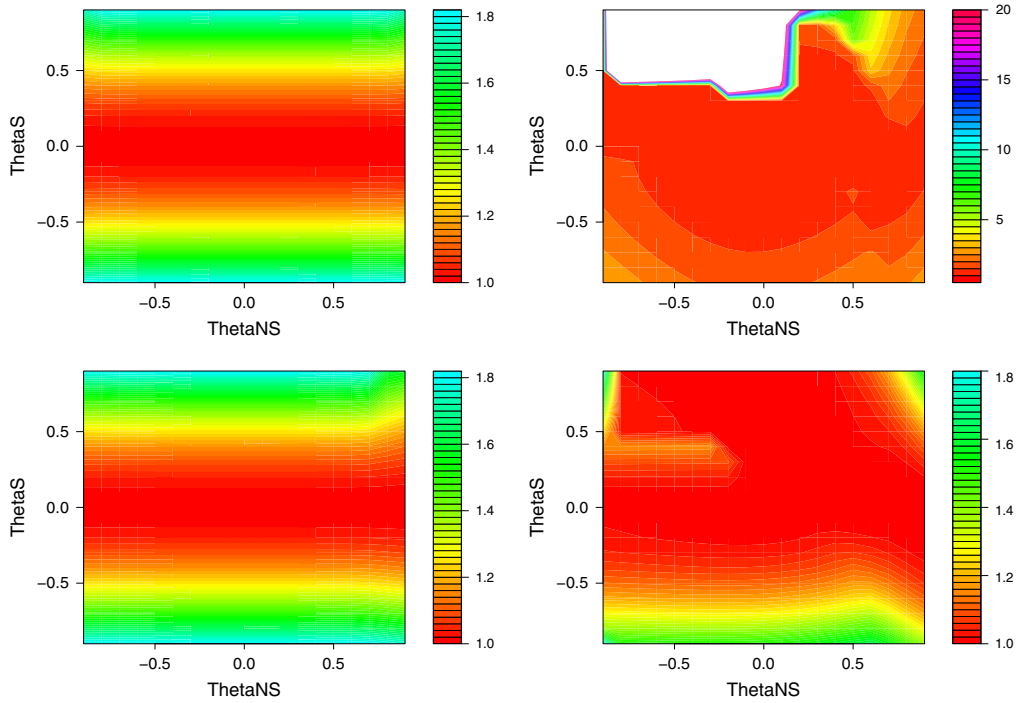


Figure 10. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (1,1,1)(0,1,0) model, as a function of the $\theta_{\text{NS}}$ and $\theta_{\text{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

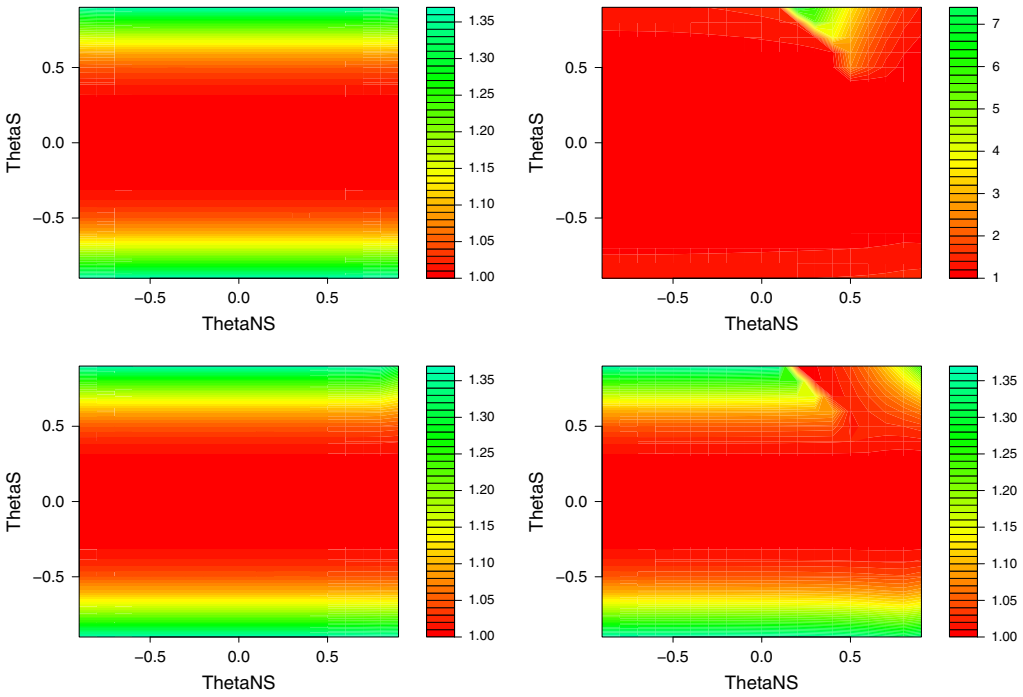


Figure 11. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (0,1,1)(1,1,0) model, as a function of the $\theta_{\text{NS}}$ and $\theta_{\text{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

of women's clothing retail sales, 1992 up to 2008, denoted by WomCloth for short. The program X-12-ARIMA automatically identified a SARIMA(011)(011) model for the trading day, Easter holiday, and outlier-adjusted series, in logs, and standard diagnostics indicate that this specification is adequate to model the series. (The model identification proceeds via information criteria comparisons used to discriminate between a default set of models—see US Census Bureau, 2014, for more details.) The nonseasonal and seasonal moving average parameters $\vartheta_1$ and $\vartheta_2$ were estimated

 

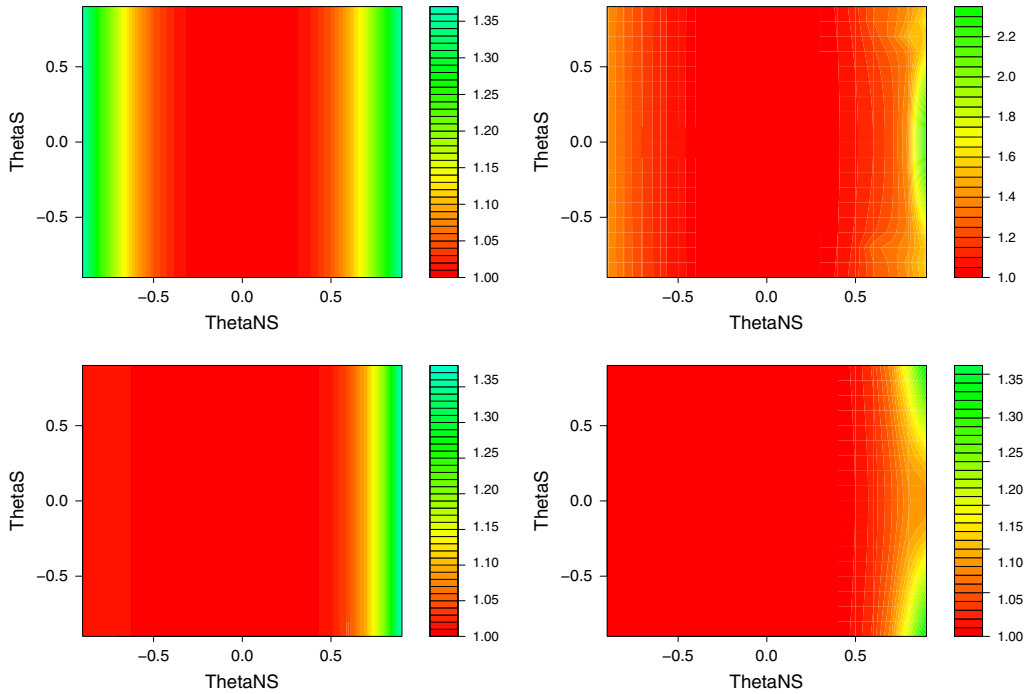


Figure 12. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (1,1,0)(0,1,1) model, as a function of the $\theta_{\text{NS}}$ and $\theta_{\text{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

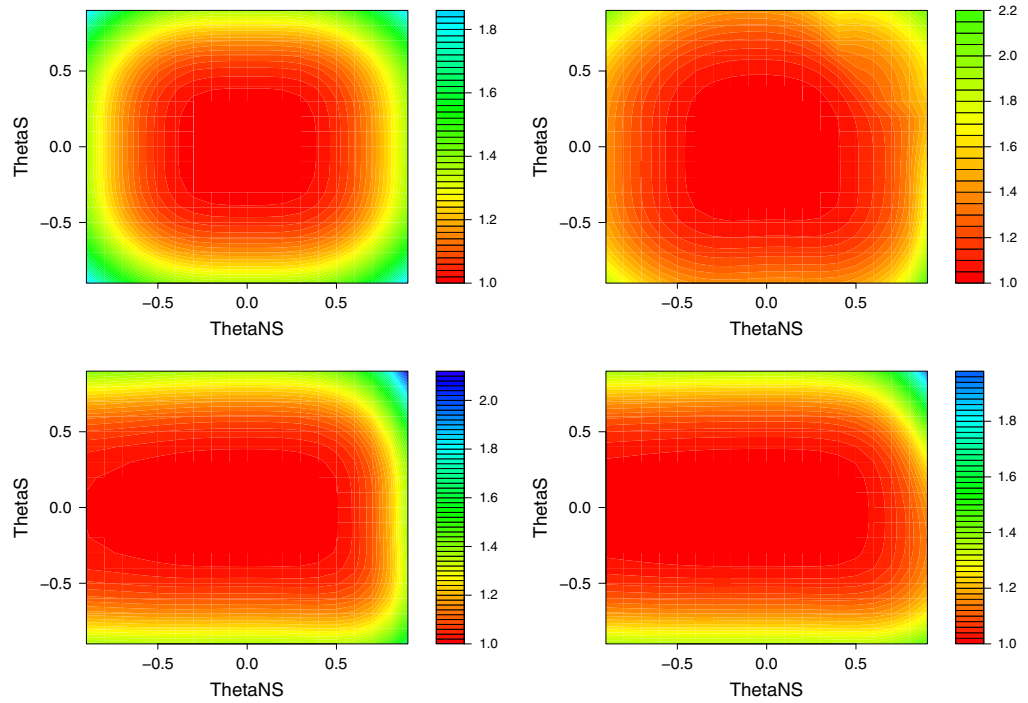


Figure 13. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (1,1,0)(1,1,0) model, as a function of the $\theta_{\text{NS}}$ and $\theta_{\text{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

via the one-step-ahead forecast criterion to be 0.441 and 0.467 respectively. Using this model will also allow us to make a connection with the results above ('The airline data process').

In this section we compare the competing forecast methods via an out-of-sample forecasting exercise. In order to observe some divergence between forecasts and truth, it was necessary to consider a hold-out span exceeding 2 years; we chose to use years 1992–2001 for parameter fitting, and years 2002 up to 2008 for forecast evaluation.

 

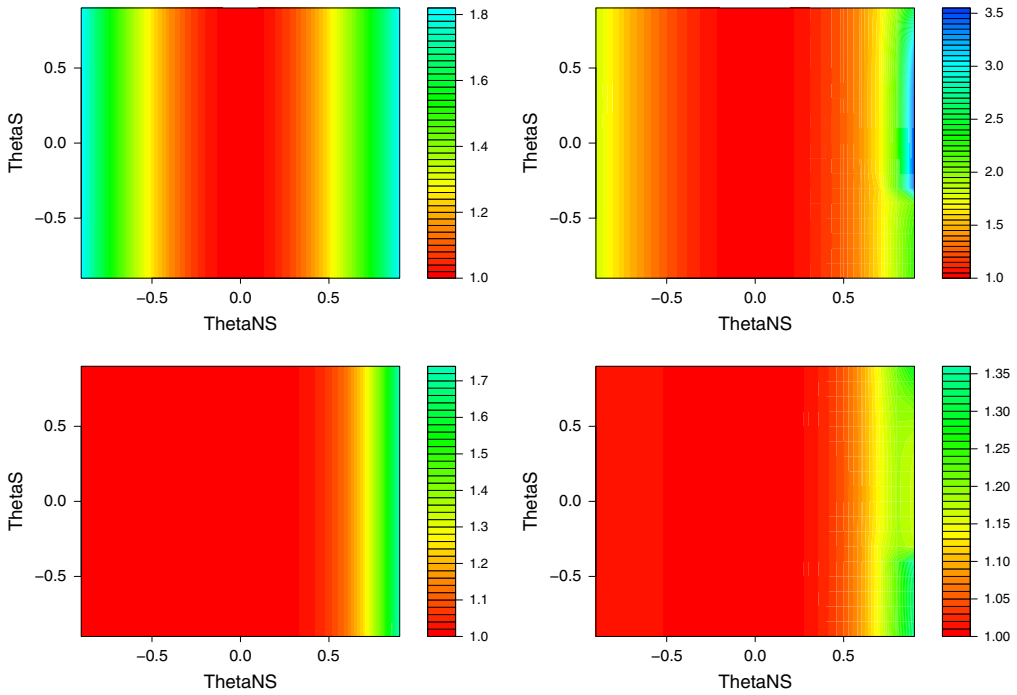


Figure 14. MSE ratio surfaces $S_{\ell,h}$ for the direct and iterative methods for the misspecified SARIMA (0,1,0)(1,1,1) model, as a function of the $\theta_{\text{NS}}$ and $\theta_{\text{S}}$ parameters of the airline data process. Left panels are for comparing the iterative forecast MSE ($\ell = 1$), whereas right panels are for comparing the direct forecast MSE ($\ell = 12$). Top panels are for forecast horizon $h = 1$, while bottom panels are for $h = 12$

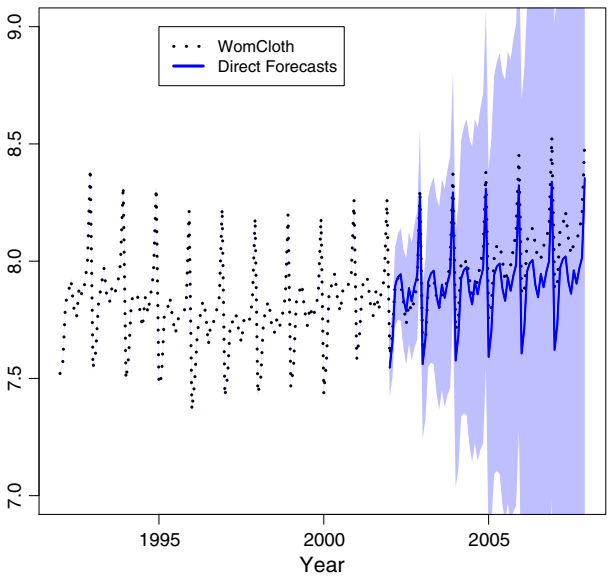


Figure 15. WomCloth monthly time series (black) together with out-of-sample forecasts generated from parameters minimizing one-step-ahead forecast MSE, with confidence intervals in shaded blue

Our sample design utilizes only present and past data for parameter fitting, so that all forecasts, whether iterative or direct, are out-of-sample; this is the same out-of-sample design as in MSW (2006). On the sub-span of 1992–2001, the parameter estimates were 0.389 and 0.445. Direct one-step-ahead forecasts $\widehat{X}^{(D)}_{120+h|1:120}$ were then computed for $1 \leq h \leq 72$ (corresponding to years 2002 up to 2008), plotted in Figure 15, along with forecast confidence intervals shaded blue. We also calculated $\widehat{X}^{(I)}_{120+h|1:120}$ for $1 \leq h \leq 72$, but these were visually indistinguishable from the direct forecasts. The maximum discrepancy between direct and iterative forecasts was less than 0.0002, relative to a series that numerically ranges between 7.5 and 8.5.

Figure 15 shows the logged series along with the direct forecasts, with the iterative forecasts omitted. The $\alpha$ value used to compute the $1 - \alpha$ double-sided confidence intervals was made extremely large ($\alpha = 0.90$) merely to allow the intervals to be visible on the plot. Taking a more conventional $\alpha = 0.05$ makes the intervals so wide that the upper

 

and lower boundaries cannot be portrayed without altering the $y$-axis. In terms of the forecasts, it is noteworthy that in the first 2 years the forecasts are quite accurate, but over a longer horizon fail to capture the upward drift of the pre-recession sales data.

Figure 16 displays the MSE curves for both the direct forecasts, beginning with the first month of 2002; the iterative forecast MSE was indistinguishable (the forecast error process is approximately identical for the direct and iterative methods, and hence the MSE is approximately the same). In order to visualize the MSE curve more easily, it is presented in log scale. The dramatic size of forecast MSE is due to the $I(2)$ structure of the airline model—for stationary series the forecast MSEs for direct and iterative methods are much more stable.

The fact discussed above—that iterative and direct forecasts of WomCloth are indistinguishable when the same parameters are used in the formulas—reinforces the first point of the paper. However, it is more realistic to consider different parameters in the two methods' forecast formulas. For the iterative method, which relies on iteratively applying one-step-ahead forecasting, it makes sense to take the one-step-ahead forecasting parameters. For the direct method, we could compute $h$-step-ahead parameters for any $1 \leq h \leq 72$. We computed these parameter estimates for $h = 12, 24$, but generated the entire sweep of 72 forecasts using either of these parameters. We refer to these forecast paths as 'Direct 12' and 'Direct 24' forecasts in Figures 17 and 18. The method used for fitting is the minimization of $J_\ell(\vartheta, I)$ for $\ell = 12, 24$, as discussed above ('Asymptotic forecast MSE performance'). The 12-step-ahead parameters were estimated to be .073 and .912, whereas for 24 steps ahead we obtained .025 and .925. Observe that the seasonal moving average behavior becomes increasingly stable as we increase $h$, which was also an effect observed in McElroy and Wildi (2013) on housing starts data.

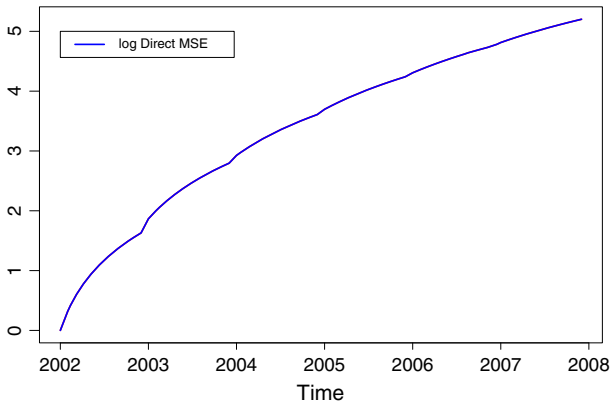


Figure 16. Log MSE for the direct forecast method generated from parameters minimizing one-step-ahead forecast MSE. The forecast horizon $1 \leq h \leq 72$ corresponds to the number of months after December 2001

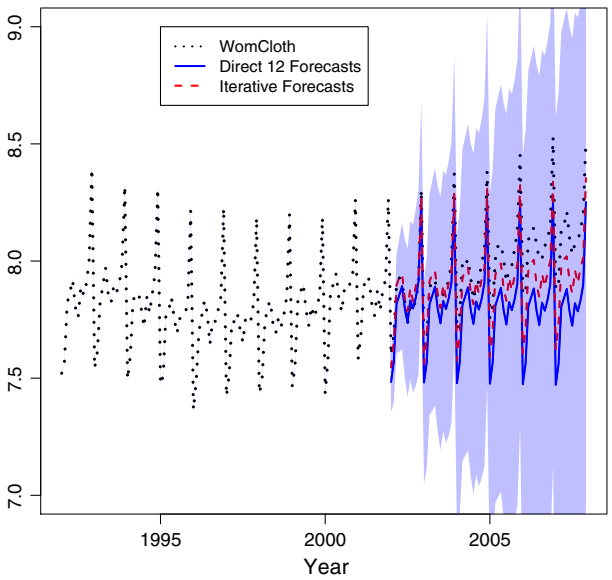


Figure 17. WomCloth monthly time series (black) together with out-of-sample forecasts generated from parameters minimizing one-step-ahead forecast MSE (red) and 12-step-ahead forecast MSE (blue), with confidence intervals in shaded blue

 

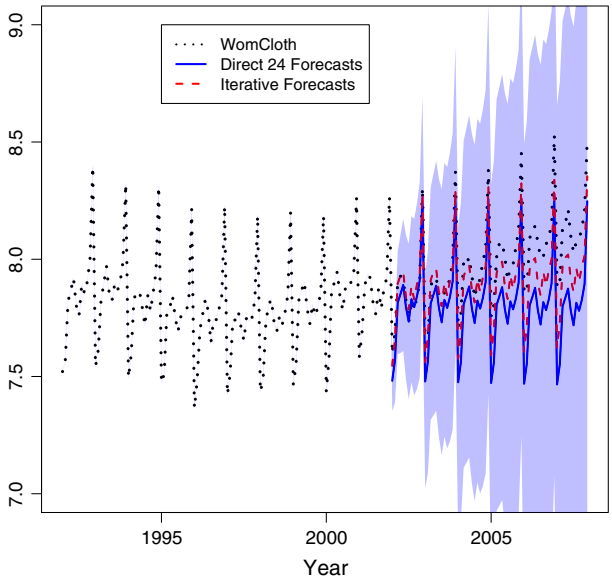


Figure 18. WomCloth monthly time series (black) together with out-of-sample forecasts generated from parameters minimizing one-step-ahead forecast MSE (red) and 24-step-ahead forecast MSE (blue), with confidence intervals in shaded blue

From Figures 17 and 18, it is not clear if the direct forecast is really superior to the iterative at leads $h = 12$ and $h = 24$ respectively, though we do not expect it to do this out-of-sample. Overall, the Direct 12 and Direct 24 forecast paths are more stable (and thus worse in performance) than the Iterative forecasts, which recall are numerically equivalent to Direct 1 forecasts. There is also little difference between the Direct 12 and Direct 24 forecasts, because their parameters are so similar. The same $\alpha = 0.90$ was used for the uncertainty bands of both Direct 12 and Direct 24, with the latter being slightly wider. One might be tempted to conclude that long-term forecasts based on one-step-ahead criteria are superior to those generated from 12-step-ahead or 24-step-ahead criteria, but this would be an unwarranted generalization from a single dataset. In fact, by altering the hold-out period, we can change the story around quite easily (see McElroy and Wildi, 2013, for an example where $h$-step-ahead criteria offer improvements long term). The key conclusion is that direct and iterative forecast paths differ chiefly due to the different parameter values that are used.

## CONCLUSIONS

The work of MSW (2006) provides the interesting conclusion that iterative forecasting actually works better out-of-sample for many time series. Several explanations of this performance could be offered; it could be due to differences between models, differences between parameters, or differences between formulas. This paper attempts to settle the question of whence the benefits of iteration chiefly arise. Along the way, we derive some new formulas for iterative forecasting, allowing for quantification of forecast MSE. While many papers treat the topic for autoregressive models, our results are the most general to date, covering all difference stationary models that have short memory and an invertible Wold representation.

If different models are utilized, the question becomes one of model specification, which we do not treat here. If the model is fixed, and the *same parameters* are used for the direct and iterative forecasting formulas, then the forecasts are virtually indistinguishable; the discrepancies vanish as sample size (the past) increases to infinity. If differing parameters are used, we suppose that direct parameters are estimated by minimizing an $h$-step-ahead direct forecast error, while the iterative parameters are estimated by minimizing a one-step-ahead forecast error (i.e. MLE). The theory of estimation according to such criteria is reviewed and extended by our work, and we study numerically the impact on forecast MSE when using a misspecified model fitted so as to minimize $h$-step-ahead forecast error.

So if we now consider iterative forecasts with one-step-ahead optimized parameters versus direct forecasts with $h$-step-ahead optimized parameters, which is superior? Again, the question really reduces to: which parameters should I use? There is no universal champion, the victor in each case being determined by the data (its past dynamics used for fitting, as well as the nature of future dynamics yet unseen). We have observed that forecasts generated from $h$-step-ahead optimized parameters tend to be more conservative over the longer term, reaching back further into the past to generate future predictions; this was illustrated on the WomCloth series. Whether or not this is an advantage depends on the nature of the series' current trend: if we attempt to forecast at the initiation of new dynamics (e.g. beginning or end of an expansion) then one-step-ahead is better, assuming the new trend has just begun. If instead

 

we have a sustained pattern and need to ignore minor swings, the $h$-step-ahead is superior, being more conservative. Of course, in real-time forecasting problems the practitioner can never know whether or not new dynamics have just been initiated!

The results of MSW (2006) indicate that one-step-ahead parameter estimates have a better overall performance (their sample design is concerned with multiple leads $h = 3, 6, 12, 24$). Our results confirm this finding: if we optimize with respect to a particular lead $\ell$, we may expect somewhat better performance at leads $h$ that are a multiple of $\ell$, but possibly much worse performance at other leads $h$. Since $\ell = 1$ has the property that all forecast leads $h$ are a multiple of it, it tends to have the best *overall* performance. We can think of the direct $\ell$-step-ahead parameter estimates as customized for $\ell$-step-ahead forecasting, whereas one-step-ahead parameter estimates perform better when a range of forecast leads is considered. These findings should be balanced against the experience of real-time forecasting competitions, where the more conservative (and less hasty) methods and parameter settings tend to perform better over the longer term (see the discussion in McElroy and Wildi, 2013).

To summarize our findings, one may fix model and model estimation technique to be the same between the forecasting methods, and then their performance is extremely similar, even in finite sample—this provides an answer to the question in the title of this article. However, it is more common to use different fitting criteria (and the same model) in econometric practice, in which case the forecast weights can differ substantially. Which method performs better then depends upon the forecast lead desired, the dynamics of the process (is the future like the past?), and also the features of the model being employed. We emphasize that unit root specification plays an important role in the parameter estimates when the forecast lead is greater than one, and therefore in practice has an impact on parameter estimates. In future work we plan to explore this facet, and also extend the methods herein to a multivariate framework.

## ACKNOWLEDGEMENTS

The preliminary version has benefited from comments by participants at the 11th OxMetrics User Conference, as well as the faculties of the George Washington University Economics Department and the University of Maryland Baltimore Campus Mathematics Department. The author thanks Marc Wildi for stimulating discussions of the topic.

## REFERENCES

Bell W. 1984. Signal extraction for nonstationary time series. *Annals of Statistics* **12**: 646–664.

Bell W, Hillmer S. 1991. Initializing the Kalman filter for nonstationary time series models. *Journal of Time Series Analysis* **12**: 283–300.

Bhansali R. 1996. Asymptotically efficient autoregressive model selection for multistep prediction problems. *Annals of the Institute of Statistical Mathematics* **48**: 577–602.

Bhansali R. 1997. Direct autoregressive predictions for multistep prediction: order selection and performance relative to the plug in predictors. *Statistica Sinica* **7**: 425–449.

Bhansali R. 1999. Parameter estimation and model selection for multistep prediction of time series: a review. In *Asymptotics, Nonparametrics and Time Series*, Ghosh S (ed). Marcel Dekker: New York; 201–225.

Chevillon G. 2007. Direct multi-step estimation and forecasting. *Journal of Economic Surveys* **21**: 746–785.

Chevillon G, Hendry D. 2005. Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting* **21**: 201–218.

Clements M, Hendry D. 1996. Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics* **58**: 657–684.

Durrett R. 1996. *Probability: Theory and Examples*. Duxbury Press: New York.

Findley F. On the use of multiple models for multi-period forecasting, *Proceedings of the Business and Economics Statistics Section, American Statistical Association,* Toronto, Ontario, 1983; 528–531.

Findley F. 1985. Model selection for multi-step-ahead forecasting. In *Proceedings of the Seventh Symposium on Identification and System Parameter Estimation*, Baker HA, Young PC (eds). Pergamon Press: Oxford; 1039–1044.

Findley DF, Monsell BC, Bell WR, Otto MC, Chen BC. 1998. New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics* **16**: 127–177.

Franses P, Legerstee R. 2010. A unifying view on multi-step forecasting using an autoregression. *Journal of Economic Surveys* **24**: 389–401.

Gersch W, Kitagawa G. 1983. The prediction of time series with trends and seasonalities. *Journal of Business and Economics Statistics* **1**: 253–264.

Haywood J, Tunnicliffe-Wilson G. 1997. Fitting time series models by minimizing multi-step-ahead errors: a frequency domain approach. *Journal of the Royal Statistical Society, Series B* **59**: 237–254.

Hoque A, Magnus J, Pesaran B. 1988. The exact multiperiod mean square forecast error of the first-order autoregressive model. *Journal of Econometrics* **39**: 327–346.

Ing C. 2003. Multistep prediction in autoregressive processes. *Econometric Theory* **19**: 254–279.

Ing C. 2004. Selecting optimal multistep predictors for autoregressive process of unknown order. *Annals of Statistics* **32**: 693–722.

Ing C, Lin J, Yu S. 2009. Toward optimal multistep forecasts in non-stationary autoregressions. *Bernoulli* **15**: 402–437.

Kang I. 2003. Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. *International Journal of Forecasting* **19**: 387–400.

Lin J, Granger C. 1994. Forecasting from non-linear models in practice. *Journal of Forecasting* **13**: 1–9.

Liu S. 1996. Model selection for multiperiod forecasts. *Biometrika* **83**: 861–873.

 

Marcellino M, Stock J, Watson M. 2006. A comparison of direct and iterated multistep AR methods for forecasting microeconomic time series. *Journal of Econometrics* **135**: 499–526.
McElroy T. 2008. Matrix formulas for nonstationary ARIMA signal extraction. *Econometric Theory* **24**: 1–22.
McElroy T, Findley D. 2010. Discerning between models through multi-step-ahead forecasting errors. *Journal of Statistical Planning and Inference* **140**: 3655–3675.
McElroy T, Wildi M. 2013. Multi-step-ahead estimation of time series models. *International Journal of Forecasting* **29**: 378–394.
Pesaran M, Pick A, Timmermann A. 2011. Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics* **164**: 173–187.
Proietti T. 2011. Direct and iterated multistep AR methods for difference stationary processes. *International Journal of Forecasting* **27**: 266–280.
Schorfheide F. 2005. VAR forecasting under misspecfication. *Journal of Econometrics* **128**: 99–136.
Stoica P, Nehorai A. 1989. On multi-step prediction errors methods for time series models. *Journal of Forecasting* **13**: 109–131.
Taniguchi M, Kakizawa Y. 2000. *Asymptotic Theory of Statistical Inference for Time Series*. Springer: New York.
Tiao G, Tsay R. 1994. Some advances in non-linear and adaptive modeling in time series. *Journal of Forecasting* **13**: 109–131.
Tiao G, Xu D. 1993. Robustness of MLE for multi-step predictions: the exponential smoothing case. *Biometrika* **80**: 623–641.
US Census Bureau. 2014. X-12-ARIMA Reference Manual. Available: http://www.census.gov/ts/x12a/v03/x12adocV03.pdf [accessed on 2014].
Weiss A. 1991. Multi-step estimation and forecasting in dynamic models. *Journal of Econometrics* **48**: 135–149.

## APPENDIX

*Proof of Proposition 1.* The derivation of equation (2) is sketched in McElroy and Findley (2010), but to prove its optimality it suffices to show that the error process is orthogonal to $X_{t:}$. Letting $F = B^{-1}$ denote the forward shift operator, the forecast error is

$$\begin{aligned}\varepsilon_t = X_{t+h} - \Upsilon_h(B)X_t &= \sum_{j=0}^{h-1} \xi_j \left( F^{h-j} - [\Psi]_{h-j}^{\infty}(B) F^{h-j} \Psi^{-1}(B) \right) W_t \\ &= F^h \sum_{j=0}^{h-1} \xi_j B^j [\Psi]_0^{h-j-1}(B) \Psi^{-1}(B) W_t\end{aligned}$$

It can be shown that $\sum_{j=0}^{h-1} \xi_j B^j [\Psi]_0^{h-j-1}(B) = [\Psi/\delta]_0^{h-1}(B)$ using simple algebra. Now when the filter exactly matches the DGP, we have $\Psi^{-1}(B)W_t = \epsilon_t$, so that the error process is $\varepsilon_t = [\Psi/\delta]_0^{h-1}(B)\epsilon_{t+h}$, which only depends on future innovations; hence the error process is orthogonal to $X_{t:}$. □

*Proof of Proposition 2.* From equation (5) the iterated forecast error is

$$\eta_t = X_{t+h} - \Pi_p(B)X_t = \left( F^h - \Pi_p(B) \right) X_t = p_{h-1}(F) F \Psi^{-1}(B) W_t$$

Because we assume that the filter $\Psi(B)$ corresponds to the process's Wold filter, the error process is $\eta_t = p_{h-1}(F)\epsilon_{t+1}$, which is orthogonal to $X_{t:}$. Hence the iterated forecasts are also optimal, and by uniqueness of the Gaussian conditional expectation (i.e. the MSE optimal linear estimate) we must have $\Upsilon_h(B) = \Pi_h(B)$. In fact, we have

$$\varepsilon_t = \eta_t + (\Upsilon_h(B) - \Pi_h(B))\, X_t$$

with the two quantities on the RHS orthogonal (this is because $\eta_t$ is orthogonal to all linear functions of $X_{t:}$). Thus the optimal MSE is equal to $\mathrm{var}(\eta_t)$ plus a non-negative quantity; by optimality, this quantity must be zero, and it follows that $\Upsilon_h(z) = \Pi_h(z)$ almost everywhere.

A second derivation of the result stems from probability theory alone. From Theorem 1.2 of Durrett (1996, p. 226) within a conditional expectation we can always additionally condition on a larger information set, since 'the smaller $\sigma$-field always wins'. Thus for $h > 1$

$$\begin{aligned}\mathbb{E}[X_{t+h}|X_{t:}] &= \mathbb{E}[\,\mathbb{E}[X_{t+h}|X_{t+h-1:}]\,|X_{t:}] \\ &= \mathbb{E}[\Upsilon_1(B)X_{t+h-1}|X_{t:}] \\ &= \mathbb{E}[\sum_{j=0}^{h-2} \upsilon_j^{(1)} X_{t+h-1-j} + \sum_{j=h-1}^{\infty} \nu_j^{(1)} X_{t+h-1-j}|X_{t:}] \\ &= \sum_{j=0}^{h-2} \upsilon_j^{(1)} \mathbb{E}[X_{t+h-1-j}|X_{t:}] + [\Upsilon_1]_{h-1}^{\infty}(B)X_t\end{aligned}$$

From here we use induction on $h$ and equation (4) to prove that $\Pi_h(B)X_t = \mathbb{E}[X_{t+h}|X_{t:}]$. □

 

*Proof of Proposition 3.* We proceed to compute the forecast error $\widehat{X}^{(D)}_{n+h|1:n} - X_{n+h}$. We first note that due to the special structure of $\Delta_{n+h}$, we have $\Delta_n[1_n\ 0_{n\times h}] = [1_n\ 0_{n\times h}]\Delta_{n+h}$. Therefore the direct forecast error $\widehat{X}^{(D)}_{n+h|1:n} - X_{n+h}$ equals

$$
\begin{aligned}
&e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}[\Delta_n\ 0_{n\times h}]X_{1:n+h} - e'_{n+h}X_{1:n+h} \\
&= e'_{n+h}\Delta^{-1}_{n+h}\left(\begin{bmatrix} 1_d & 0_{d\times n-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\end{bmatrix}[1_{n-d}\ 0_{n-d\times h}] - 1_{n+h}\right)\begin{bmatrix} X_{1:d} \\ W_{d+1:n+h}\end{bmatrix} \\
&= e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_{d\times d} & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}[1_{n-d}\ 0_{n-d\times h}] - 1_{n+h-d}\end{bmatrix}\begin{bmatrix} X_{1:d} \\ W_{d+1:n+h}\end{bmatrix} \\
&= e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_{d\times 1} \\ \left(\Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}[1_{n-d}\ 0_{n-d\times h}] - 1_{n+h-d}\right)W_{d+1:n+h}\end{bmatrix}
\end{aligned}
$$

which shows that the initial values are not present in the forecast error. To show optimality, consider the covariance of this forecast error with the data $\Delta_n^{-1}[X'_{1:d}, W'_{d+1:n}]'$; we obtain

$$
\begin{aligned}
&e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_{d\times d} & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \left(\Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}[1_{n-d}\ 0_{n-d\times h}] - 1_{n+h-d}\right)\Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\end{bmatrix}\Delta^{\dagger}_n \\
&= e'_{n+h}\Delta^{-1}_{n+h}\begin{bmatrix} 0_{d\times d} & 0_{d\times n+h-d} \\ 0_{n+h-d\times d} & \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix}\Sigma^{-1}_{n-d}\left([1_{n-d}\ 0_{n-d\times h}]\ \Sigma_{n+h-d}\begin{bmatrix} 1_{n-d} \\ 0_{h\times n-d}\end{bmatrix} - \Sigma_{n-d}\right)\end{bmatrix}\Delta^{\dagger}_n
\end{aligned}
$$

which is identically zero. The expression (7) for the forecast error covariance follows from the above expression for the forecast error. □

*Proof of Proposition 4.* By use of a telescoping sum, we can write

$$
\begin{aligned}
\widehat{X}^{(I)}_{n+h|1:n} - X_{n+h} &= e'_n \sum_{\ell=1}^{h} J^{h-\ell}\ (J\ X_{\ell:n+\ell-1} - X_{\ell+1:n+\ell}) \\
&= e'_n \sum_{\ell=1}^{h} J^{h-\ell}\ (J\ [1_n\ 0_{n\times 1}] - [0_{1\times n}\ 1_n])\ [0_{n+1\times \ell-1}\ 1_{n+1}\ 0_{n+1\times h-\ell}]\ X_{1:n+h}
\end{aligned}
$$

where the one-step-ahead forecast error yields the relation

$$
J\ [1_n\ 0_{n\times 1}] - [0_{1\times n}\ 1_n] = \begin{bmatrix} 0_{n-1\times n+1} \\ \eta',\ -1\end{bmatrix}
$$

Now we claim that

$$
[\eta',\ -1] = e'_{n+1}\ \Delta^{-1}_{n+1}\begin{bmatrix} 0_{d\times d} & 0_{d\times n+1-d} \\ 0_{n+1-d\times d} & K\end{bmatrix}\Delta_{n+1},
$$

which follows by writing $\eta' = e'_{n+1}\ \Delta^{-1}_{n+1}\ F\ \Delta_n$, with $F$ the matrix defined in square brackets implicitly in formula (10). Then the claim is equivalent to

$$
\begin{aligned}
\left[e'_{n+1}\ \Delta^{-1}_{n+1}\ F\ \Delta_n,\quad -1\right]\ \Delta^{-1}_{n+1} &= \left[e'_{n+1}\ \Delta^{-1}_{n+1}\ F\ \Delta_n,\quad -1\right]\begin{bmatrix} [1_n\ 0_{n\times 1}]\ \Delta^{-1}_{n+1} \\ e'_{n+1}\ \Delta^{-1}_{n+1}\end{bmatrix} \\
&= e'_{n+1}\ \Delta^{-1}_{n+1}\ (F\ [1_n\ 0_{n\times 1}] - 1_{n+1}) \\
&= e'_{n+1}\ \Delta^{-1}_{n+1}\begin{bmatrix} 0_{d\times d} & 0_{d\times n+1-d} \\ 0_{n+1-d\times d} & K\end{bmatrix}
\end{aligned}
$$

 

which is true. Then the iterative forecast error is

$$e_n' \sum_{\ell=1}^{h} J^{h-\ell} \begin{bmatrix} 0_{n-1\times n+1} \\ \eta',\ -1 \end{bmatrix} \Delta_{n+1}^{-1} \begin{bmatrix} X_{\ell:d+\ell-1} \\ W_{d+\ell:n+\ell} \end{bmatrix}$$

$$= e_n' \sum_{\ell=1}^{h} J^{h-\ell} \begin{bmatrix} 0_{n-1\times n+1} \\ e_{n+1}' \Delta_{n+1}^{-1} \begin{bmatrix} 0_{d\times d} & 0_{d\times n+1-d} \\ 0_{n+1-d\times d} & K \end{bmatrix} \end{bmatrix} \begin{bmatrix} X_{\ell:d+\ell-1} \\ W_{d+\ell:n+\ell} \end{bmatrix}$$

$$= e_n' \sum_{\ell=1}^{h} J^{h-\ell} \begin{bmatrix} 0_{n-1\times 1} \\ e_{n+1}' \Delta_{n+1}^{-1} \begin{bmatrix} 0_{d\times n+1-d} \\ K \end{bmatrix} W_{d+\ell:n+\ell} \end{bmatrix}$$

$$= e_n' \sum_{\ell=1}^{h} J^{h-\ell} \begin{bmatrix} 0_{n-1\times 1} \\ \beta' W_{d+\ell:n+\ell} \end{bmatrix}$$

$$= e_n' \sum_{k=0}^{h-1} J^{k} \begin{bmatrix} 0_{n-1\times n+h-d} \\ 0_{1\times h-k-1}\ \beta'\ 0_{1\times k} \end{bmatrix} W_{d+1:n+h}$$

The MSE follows immediately. □

*Authors' biography*:
**Tucker McElroy** is a mathematical statistician at the U.S. Census Bureau, where he has been working since 2003. He completed his PhD in mathematics at the University of California, San Diego in 2001. His research interests include time series, signal extraction, and forecasting.

*Authors' address*:
**Tucker McElroy**, Center for Statistical Research and Methodology, US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA.

 