ISSN 1032-3813
ISBN 0 7326 1031 1

MONASH UNIVERSITY

AUSTRALIA

PREDICTION INTERVALS FOR ARIMA MODELS

R.D. Snyder
J.K. Ord
A.B. Koehler

Working Paper 8/97
October 1997

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

PREDICTION INTERVALS FOR ARIMA MODELS

R.D Snyder [1], J.K. Ord [1,2] and A.B. Koehler [3]

Abstract

The problem of constructing prediction intervals for linear time series (ARIMA) models is examined. The aim is to find prediction intervals which incorporate an allowance for sampling error associated with parameter estimates. The effect of constraints on parameters arising from stationarity and invertibility conditions is also incorporated. Two new methods, based to varying degrees on first-order Taylor approximations, are proposed. These are compared in a simulation study to two existing methods: a heuristic approach and the 'plug-in' method whereby parameter values are set equal to their maximum likelihood estimates

Keywords

ARIMA, Bayesian, forecasting, Holt-Winters, simulation, state space.

Correspondence

Associate Professor Ralph D. Snyder, Department of Econometrics, Monash University, Wellington Road, Clayton, VIC 3168, Australia.

1. Department of Econometrics, Monash University, Australia.
2. Departments of Management Science and Statistics, Pennsylvania State University, USA.
3. Department of Decision Science and MIS, Miami University of Ohio, USA.

INTRODUCTION

As indicated by Chatfield (1993) in his comprehensive state-of-the-art review, the construction of valid prediction intervals (PIs) for time series continues to present considerable difficulties. In particular, Chatfield notes a number of reasons why PIs may be too narrow; these include:

> " Model parameters may have to be estimated.
>
> Innovations may not be normally distributed.
>
> There may be outliers in the data.
>
> The wrong model may be identified.
>
> The underlying model may change, either during the period of fit or in the future."

In this paper, we focus upon the first of these issues. If the uncertainty relating to parameter estimation is not allowed for explicitly, the resulting prediction intervals will be too narrow. Further, the non-linear nature of the parameter estimates in time series makes the problem intractable as regards an exact analytic solution, so we develop various approximate solutions which are then explored in a simulation study. Only when we are confident of our ability to produce reliable PIs in the basic case can we address the remaining issues. Thus, in the present paper, we examine the construction of PIs when the parameters are unknown and the errors are assumed to be normal, leaving the other issues to be addressed in further research.

We identify four approaches to the construction of prediction intervals and report on an extensive simulation study of these alternatives. The particular model used in our simulations

is the additive Holt-Winters scheme; see Example 2.2 below. Yar and Chatfield (1990) provide PIs for this scheme based upon its ARIMA representation and setting the parameter values equal to their estimates (the 'plug-in' approach). These authors find the method to be superior to previous, albeit heuristic, approaches and the plug-in PI is one of the options considered in our study. However, rather than use an ARIMA framework, we have opted for a state space scheme; details are given in section 2.

The principal method considered in the paper is a Bayesian simulation scheme. Ansley and Kohn (1986) showed how to obtain the conditional mean squared error (MSE) for a time series in the state space framework and pointed out that the correction to the MSE has a Bayesian interpretation. Under appropriate conditions we can use the asymptotic sampling distribution developed by Ansley and Kohn(1986) to generate the predictive distribution, using simulation. De Jong and Whiteman (1994) followed this approach in developing PIs for AR(p) schemes; the resulting simulated distribution is shown to converge to the predictive distribution using a result of Geweke (1989) and the same justification may be employed here.

An alternative approach would be to use a complete Monte Carlo Markov Chain (MCMC) approach; see Barnett, Kohn and Sheather (1996, 1997) for the development of MCMC estimation procedures for ARMA models. Our method uses an analytic approximation to the posterior distribution of the parameters, which we then feed into the computation of the predictive distribution. Thus, our scheme may be viewed as a 'partial' MCMC technique, which should be less demanding computationally, an important consideration when a large number of series is to be analyzed.

Another approach would be to consider the non-parametric bootstrap; see, for example, Thombs and Schucany (1990), Kabaila (1993) and McCullough (1994). However, since our current focus is on getting the correct coverage with a known underlying error process, we have not pursued that line of inquiry at this time.

In the paper, we address three main issues:

1. the extension of the Bayesian simulation approach to state space schemes;
2. the use of approximations to simplify the computational task;
3. an extensive simulation study to determine whether the suggested approach provides PIs with the appropriate coverage.

The structure of the remainder of this paper is as follows. In section 2, we compare single source and multiple source state space schemes and justify our use of a single source model. In section 3, we describe the various approaches to be considered for the construction of prediction intervals. Section 4 describes the simulation study and summarises the conclusions from that study. The summary and outline of future directions appears in section 5.

2. STATE SPACE REPRESENTATIONS

We consider the usual autoregressive integrated moving average or ARIMA (p,d,q) representation for a time series given by $\phi(B)w_t = \theta(B)\varepsilon_t$ where $w_t = \nabla^d y_t$, ∇=(1-B), B is the lag operator and ϕ and θ represent polynomials in B of orders p and q respectively. The errors $\{\varepsilon_t\}$ are taken to be independent and identically distributed with zero means and

constant variances; that is, $\varepsilon_t \sim \text{IID}(0,\sigma^2)$. The polynomials $\phi(B)$ and $\theta(B)$ may be partitioned into regular and seasonal components, and we may add seasonal differences in the usual way.

Akaike (1974) showed that any ARMA(p, q) scheme has a Markovian state space representation consisting of the observation equation $y_t = \mathbf{h}'\mathbf{x}_t$ and the state equation $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{b}\varepsilon_t$ where $\mathbf{x}_t$ is the state vector of order k [k=max(p,q+1)], $\mathbf{h}' = (1,0,\ldots,0)$, $\mathbf{b}' = (1, \psi_1, \ldots, \psi_{k-1})$ and $\mathbf{F} = \begin{bmatrix} 0 & \mathbf{I}_p \\ \phi_p, \ldots & \ldots \phi_1 \end{bmatrix}$, the ψ_j being the psi-weights given by the coefficients of powers of B in $\theta(B)/\phi(B)$. It is a single-source model since only a single source of stochastic variation (ε_t) is included in the model specification.

Multiple source representations are also available (c/f Harvey, 1990, chapter 2) with a measurement equation $y_t = \mathbf{h}'\mathbf{x}_t + \varepsilon_t$ and state equation $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{C}\delta_t$ where $\mathbf{C}$ is usually the identity matrix, δ_t is a vector of independent errors that are also independent over time and are independent of ε_{t+j} for all j. ARIMA schemes may be represented by such a process, but sometimes restrictions must be placed on the ranges of the parameters. If $\mathbf{C}$ is sufficiently general, these restrictions disappear, but the diagonal form of $\mathbf{C}$ is the only one used to any extent.

An alternate single source representation (Snyder, 1985) is

State Space Model: Single Disturbance Source (SSMS)

$$y_t = \mathbf{h}'\mathbf{x}_{t-1} + \varepsilon_t \text{ where } \mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \alpha\varepsilon_t \text{ and } \varepsilon_t \sim \text{IID}(0,\sigma^2), \qquad (2.1)$$

α being a $(k \times 1)$ parameter vector. A variation uses $\mathbf{x}_t$ instead of $\mathbf{x}_{t-1}$ in the measurement equation, but the present version is more convenient and the two schemes are formally equivalent. Aoki (1994) derives this structure as his 'forward-innovation representation' and develops estimators for the multivariate case. Any ARIMA(p, d, q) scheme may be

represented by an SSMS with $\mathbf{F} = \begin{bmatrix} \phi_1 & \mathbf{I}_{k-1} \\ \vdots & \\ \phi_k & \mathbf{0} \end{bmatrix}$ and $\alpha = \begin{bmatrix} \phi_1 + \theta_1 \\ \vdots \\ \phi_k + \theta_k \end{bmatrix}$ where $k = \max(p+d, q)$.

In the reverse direction, any SSMS can be expressed as an ARIMA model. We can write the state equation of (2.1) as

$$\mathbf{x}_t = (\mathbf{I} - \mathbf{F}B)^{-1} \alpha \varepsilon_t . \tag{2.2}$$

Substituting (2.2) into the measurement equation of (2.1),

$$y_t = \left(\mathbf{h}'(\mathbf{I} - \mathbf{F}B)^{-1} \alpha B + 1\right)\varepsilon_t \equiv \psi(B)\varepsilon_t . \tag{2.3}$$

Equation (2.3) is the moving average form of the state space model. If we write $\mathbf{F}^j = \mathbf{U}\Lambda^j\mathbf{V}$ where Λ is the diagonal matrix of eigenvalues and $(\mathbf{U}, \mathbf{V})$ are the matrices of eigenvectors, (2.3) becomes

$$y_t = \left(\mathbf{h}'\mathbf{U}(\mathbf{I} - \Lambda B)^{-1}\mathbf{V}\alpha B + 1\right)\varepsilon_t .$$

If all eigenvalues of $\mathbf{F}$ lie inside the unit circle

$$y_t = \left(1 + \mathbf{h}'\mathbf{U}\left(\sum_{j=0}^{\infty} \Lambda^j B^{j+1}\right)\mathbf{V}\alpha\right)\varepsilon_t .$$

The convergence of the coefficients in the infinite polynomial $\psi(B)$ corresponds to the roots of $(\psi(B))^{-1} = 0$ lying outside the unit circle. Thus y_t will be stationary if and only if the eigenvalues of F are inside the unit circle.

In the case where $\mathbf{F}$ has unit eigenvalues, we can write the state equation of (2.1) as

$$(\mathbf{I}-\mathbf{F}B)\mathbf{x}_t = \alpha\varepsilon_t . \tag{2.4}$$

We can multiply both sides of (2.4) by the adjoint of $\mathbf{I}-\mathbf{F}B$, $\mathbf{W}(B)$, to obtain

$$\det(\mathbf{I}-\mathbf{F}B)\mathbf{x}_t = \mathbf{W}(B)\alpha\varepsilon_t . \tag{2.5}$$

If the eigenvalues do not exceed one, then

$$\det(\mathbf{I}-\mathbf{F}B) = G(B)H(B)$$

where $G(B)$ is a polynomial whose roots are all the unit eigenvalues of $\mathbf{F}$, and $H(B)$ is a polynomial that has an inverse. Then (2.4) can be written as

$$G(B)\mathbf{x}_t = \frac{\mathbf{W}(B)}{H(B)}\alpha\varepsilon_t . \tag{2.6}$$

The new substitution into the measurement equation of (2.1) will produce the following ARIMA model in place of (2.3):

$$G(B)y_t = \left(\mathbf{h}'\frac{\mathbf{W}(B)}{H(B)}\alpha B + G(B)\right)\varepsilon_t = \psi(B).$$

If an eigenvalue of $\mathbf{F}$ exceeds 1, then the roots of $[\psi(B)]^{-1}$ lie within the unit circle and the process is not stationary and cannot be made stationary by applying unit root operators.

In a similar manner we can derive the requirements for invertibility. We may write the transition equation of (2.1) as $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \alpha(y_t - \mathbf{h}'\mathbf{x}_{t-1}) = \alpha y_t + \mathbf{D}B\mathbf{x}_t$ where $\mathbf{D} = \mathbf{F} - \alpha\mathbf{h}'$. Thus

$$\mathbf{x}_t = (\mathbf{I}-\mathbf{D}B)^{-1}\alpha y_t \tag{2.7}$$

Substituting (2.7) into the measurement equation of (2.1)

$$y_t = \mathbf{h}'(\mathbf{I}-\mathbf{D}B)^{-1}\alpha y_{t-1} + \varepsilon_t .$$

Hence

$$\left(1+\mathbf{h}'(\mathbf{I}-\mathbf{D}B)^{-1}\alpha B\right)y_t \equiv \pi(B)y_t = \varepsilon_t \qquad (2.8)$$

Equation (2.8) is the autoregressive form of the state space model. This model will be invertible (ie roots of $(\pi(B))^{-1} = 0$ lie outside the unit circle) if and only if the eigenvalues of $\mathbf{D}$ lie inside the unit circle. An important observation for later use is that $\mathbf{D}^t \to O$ as $t \to \infty$ if and only if the model is invertible.

Example 2.1: AR(1) Scheme

The state space scheme $y_t = \mu_{t-1} + \varepsilon_t$ where $\mu_t = \phi\mu_{t-1} + \alpha\varepsilon_t$ corresponds to the ARIMA(1,0,1) scheme $y_t = \phi y_{t-1} + \varepsilon_t - (\phi-\alpha)\varepsilon_{t-1}$ with invertibility condition $|\phi-\alpha|<1$ and stationarity condition $|\phi|<1$. Referencing the SSMS equivalent, $\mathbf{D}=(\phi-\alpha)$ and $\mathbf{F}=\phi$, yielding the same invertibility and stationarity requirements.

The primary focus in section 4 is on the following special case of the SSMS which underlies the Holt-Winters method of forecasting (Winters, 1960).

Example 2.2: Additive Holt Winters Model

The observation equation $y_t = \ell_{t-1} + b_{t-1} + c_{t-m} + \varepsilon_t$ is accompanied by state equations for the level $\ell_t = \ell_{t-1} + b_{t-1} + \alpha_1\varepsilon_t$, the growth rate $b_t = b_{t-1} + \alpha_2\varepsilon_t$ and the seasonal factors $c_t = c_{t-m} + \alpha_3\varepsilon_t$. By eliminating the state variables, it can be shown that this example reduces to the seasonal ARIMA scheme found by McKenzie (1984) for the Holt-Winters method. It may then be established that the conditions $0<\alpha_1<2$,

$0<2\alpha_1+\alpha_2<4$ and $0<\alpha_3<1$ are necessary for invertibility. The full conditions are given by Archibald (1990).

An empirical study by Garcia-Ferrer and Del Hoyo (1992) contrasted the multiple source scheme (Harvey's basic structural model or BSM; Harvey, 1990) with ARIMA modelling for a number of series. Garcia-Ferrer and del Hoyo conclude that the ARIMA formulation generally produces better predictions than BSM, a result they attribute to the lack of orthogonality among the components of the state vector. Given the equivalence of the ARIMA and SSMS representations, their conclusions imply that the SSMS form of the BSM is superior to its traditional multiple source counterpart. Note that the issue of orthogonality does not arise with SSMS.

3. MODEL ESTIMATION AND PREDICTION INTERVALS

3.1 Kalman Filter Approach

Maximum likelihood estimates of the parameters α and σ may be obtained using a procedure that incorporates the Kalman filter to expedite the evaluation of the likelihood function (Schweppe, 1965). The Kalman filter for the SSMS (Snyder, 1985) includes the equations

$$e_t = y_t - \mathbf{h}'\mathbf{x}_{t-1|t-1} \tag{3.1}$$

and

$$\mathbf{x}_{t|t} = \mathbf{F}\mathbf{x}_{t-1|t-1} + \mathbf{a}_t(y_t - \mathbf{h}'\mathbf{x}_{t-1|t-1}) \tag{3.2}$$

where $x_{t|s}$ denotes the estimator for $\mathbf{x}_t$ based upon the sample $\mathbf{Y}_s=(y_1,...,y_S)$ and where $\mathbf{a}_t$ is the Kalman gain. From SSMS and (3.2), the estimation error satisfies the recurrence

relationship $\mathbf{x}_{t|t} - \mathbf{x}_t = (\mathbf{F} - \mathbf{a}_t\mathbf{h}')(\mathbf{x}_{t-1|t-1} - \mathbf{x}_{t-1}) + (\mathbf{a}_t - \alpha)\varepsilon_t$ from which the variance of $\mathbf{x}_{t|t}$, $\sigma^2\mathbf{V}_t$ say, may be obtained. Minimization of the variance with respect to $\mathbf{a}_t$ yields the expression for the Kalman gain $\mathbf{a}_t = (\mathbf{F}\mathbf{V}_{t-1}\mathbf{h} + \alpha)/(1 + \mathbf{h}'\mathbf{V}_{t-1}\mathbf{h})$. Provided the process is invertible (see section 2 and 3.2) V_t will converge to zero as the sample size increases, so that $\mathbf{a}_t$ tends to α. In the steady state the updating equations correspond to those of exponential smoothing.

It should be emphasised that it only makes sense to use the Kalman filter for normally distributed disturbances. In other cases, the Kalman filter still yields the best *linear* filter, but this may not be compatible with maximization of the likelihood function.

3.2 Exponential Smoothing

An inherently simpler strategy is to bypass the Kalman filter and use exponential smoothing from the outset. Conditioning on a trial value for $\mathbf{x}_0$ and assuming the sample $Y_n = (y_1, \ldots y_n)$ is known, SSMS implies that fixed successive values of the state vector $\mathbf{x}_t$ can be computed recursively with the error correction form of the exponential smoothing equation

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \alpha(y_t - \mathbf{h}'\mathbf{x}_{t-1}). \tag{3.3}$$

Strictly speaking the $\mathbf{x}_t$ should be read as $\mathbf{x}_t|\mathbf{Y}_{t-1}, \theta, \mathbf{x}_0$ in (3.3) where θ denotes the vector of unknown parameters contained in ($\mathbf{h}$, $\mathbf{F}$, α, σ). The SSMS then implies that $y_t|\mathbf{Y}_{t-1}, \theta, \mathbf{x}_0 \sim IID(\mathbf{h}'\mathbf{x}_{t-1}, \sigma^2)$, from which it follows that the likelihood function has the form $L(\theta, \mathbf{x}_0|Y_n) = \prod_{t=1}^{n} p(y_t|\mathbf{Y}_{t-1}, \theta, \mathbf{x}_0)$ where $p()$ is the pdf for ε_t. Since $\mathbf{x}_0$ is treated as a fixed vector of unknown parameters we have here a conditional, rather than a marginal, likelihood function.

In the case where the likelihood is based on a normal distribution, the maximum likelihood estimate of $\mathbf{x}_0$, denoted by $\mathbf{x}_{0|n}$, for given θ, is a linear least squares estimate. Estimates of successive state vectors, given a sample of size *n*, are then obtained with the recurrence relationship based on (3.3)

$$\mathbf{x}_{t|n} = \mathbf{F}\mathbf{x}_{t-1|n} + \alpha\left(y_t - \mathbf{h}'\mathbf{x}_{t-1|n}\right). \tag{3.4}$$

This is similar to the updating equation (3.2) of the Kalman filter, there being two differences:

a) the Kalman gain is replaced by the vector of smoothing parameters;

b) filtered values of the state vectors are replaced by corresponding smoothed values.

The links, for given θ, are further highlighted by the relationship

$$\mathbf{x}_{n|n} = \mathbf{F}\mathbf{x}_{n-1|n-1} + \alpha\left(y_n - \mathbf{h}'\mathbf{x}_{n-1|n-1}\right) + \mathbf{D}^n\left(\mathbf{x}_{0|n} - \mathbf{x}_{0|n-1}\right), \tag{3.5}$$

This equation is obtained by rewriting equation (3.4) as $\mathbf{x}_{t|n} = \mathbf{D}\mathbf{x}_{t-1|n} + \alpha y_t$. Lagging this recurrence relationship with respect to the sample size *n* rather than *t*, and then subtracting the result we obtain $\left(\mathbf{x}_{t|n} - \mathbf{x}_{t|n-1}\right) = \mathbf{D}\left(\mathbf{x}_{t-1|n} - \mathbf{x}_{t-1|n-1}\right)$ with a solution $\left(\mathbf{x}_{t|n} - \mathbf{x}_{t|n-1}\right) = \mathbf{D}^t\left(\mathbf{x}_{0|n} - \mathbf{x}_{0|n-1}\right)$. Solving for $\mathbf{x}_{t|n}$, using (3.4) to substitute for $\mathbf{x}_{t|n-1}$ and letting $t = n$, we get (3.5).

This is a decomposition of the updating relationship (3.2) of the Kalman filter into three components: an origin shift effect, a structural change effect, and a learning effect. Although never explicitly considered beyond period 1, the learning effect indicates that the Kalman filter implicitly revises the least squares estimates of the seed vector $\mathbf{x}_0$ in the light of new information provided by the latest observation y_n. If the process is invertible (see section 2) the learning effect disappears in large samples and we are left with exponential smoothing.

As indicated earlier, the Kalman filter can only be used as part of the maximum likelihood procedure when the disturbances in the SSMS are normally distributed. The exponential smoothing method outlined in this section, however, can be applied for any disturbance distribution. The seed vector estimates no longer correspond to linear least squares estimates and the links with the Kalman filter disappear. This is of little consequence, however, because the method continues to yield maximum likelihood estimates.

For the rest of the paper the maximum likelihood estimates of θ and $\mathbf{x}_0$ will be denoted by $\hat{\theta}$ and $\hat{\mathbf{x}}_0$ (replaces $\mathbf{x}_{0|n}$). Likewise $\hat{\mathbf{x}}_t$ will replace $\mathbf{x}_{t|n}$ as the estimate for $\mathbf{x}_t$, obtained by applying (3.3). At some points of the paper σ will not be part of θ and its maximum likelihood estimate will be denoted by $\hat{\sigma}$.

3.3 Point Predictions and Prediction Intervals

Once the maximum likelihood (ML) estimators have been found, we construct both point predictions and prediction intervals (PI) for up to j periods ahead; $j=1,2,...,r$. The point predictions for $t = n+1,\ldots,n+r$ are the conditional expectations of the model (2.1): $\hat{\mathbf{y}}_t = \mathbf{h}'\hat{\mathbf{x}}_{t-1}$ and $\hat{\mathbf{x}}_t = \mathbf{F}\hat{\mathbf{x}}_{t-1}$. Three principal approaches for the PI's will be considered: the 'plug-in' method, a linear approximation and a Bayesian simulation scheme. For completeness, we include the heuristic PI construction outlined in Bowerman and O'Connell (1993, Chapter 8) and call it the heuristic method (HEM). To simplify the notation we will use $\mathbf{y}_p = (y_1 \quad ,\ldots, \quad y_n)$ for the sample of past values and $\mathbf{y}_f = (y_{n+1} \quad ,\ldots, \quad y_{n+r})$ for future values.

'Plug-in' Method (PI)

For the construction of the PI by the 'plug-in' method, the density function $p(\mathbf{y}_f|\theta,\mathbf{x}_0,\sigma,\mathbf{y}_p)$ of the future time series is approximated by the Gaussian density $\phi(\mathbf{y}_f|\hat{\theta},\hat{\mathbf{x}}_0,\hat{\sigma},\mathbf{y}_p)$. The predictions are determined in the usual way from this distribution. This method is equivalent to the usual procedure where we assume the psi-weights, ψ_j, which may be determined from the parameters, are known. We anticipate that this method will yield intervals which understate the true width.

Linear Approximation Method (LA)

The Linear Approximation Method accounts for the sampling error that is associated with θ, where σ is not included in θ. We expect that this method will produce intervals with better coverage than the PI's from HEM and PIM. Let e_t denote the t th residual estimated from a sample of size n obtained during the calculations with (3.3) and write $\mathbf{e}'=[\mathbf{e}'_p,0']$, where $\mathbf{e}_p'=[e_1,\ldots,e_n]$ and $\varepsilon'=[\varepsilon'_p,\varepsilon'_f]$ is the corresponding $(n+r)$x1 vector of error terms; the vector of zeros corresponds to the predicted values for ε_f. For an invertible process, as noted in section 3.1, the dependence of $\mathbf{y}_f$ on $\mathbf{x}_0$ will be slight; thus, we assume that $\mathbf{y}$ is approximately a linear function of θ and ε only and write $\mathbf{y}^* = \mathbf{Z}\theta + \mathbf{L}\varepsilon$ where $\mathbf{y}^*$ denotes $\mathbf{y}$ minus the constant term from the Taylor series expansion where the matrices $\mathbf{Z}$ and $\mathbf{L}$ contain the derivatives of $\mathbf{y}$ with respect to θ and ε, evaluated at $[\hat{\theta}, \mathbf{e}]$. Note that $\mathbf{L}$ is a unit lower triangular matrix because the typical y_t cannot depend on future values of the disturbances.

This linear approximation can be expressed as the following equations for the past (p) and future (f) values

$$\mathbf{y}_p^* = \mathbf{Z}_p\theta + \mathbf{L}_{pp}\varepsilon_p \tag{3.6}$$

$$\mathbf{y}_f^* = \mathbf{Z}_f\theta + \mathbf{L}_{pf}\varepsilon_p + \mathbf{L}_{ff}\varepsilon_f \tag{3.7}$$

Assuming a diffuse prior, we approximate the posterior by a Gaussian distribution with mean $\hat{\theta}$ and variance

$$Var\left(\theta|\sigma,\mathbf{x}_0,\mathbf{y}_p\right) = \sigma^2\left(\overline{\mathbf{Z}}_p\overline{\mathbf{Z}}_p\right)^{-1} \tag{3.8}$$

where $\mathbf{L}_{pp}\overline{\mathbf{Z}}_p = \mathbf{Z}_p$.

To construct PI's we need the variance of the forecast error for future time periods. In the development of this variance first solve for ε_p in (3.6) to get

$$\varepsilon_p = \mathbf{L}_{pp}^{-1}\left(\mathbf{y}_p^* - \mathbf{Z}_p\theta\right). \tag{3.9}$$

Then substitute (3.9) into (3.7) to find

$$\mathbf{y}_f^* = \left(\overline{\mathbf{Z}}_f - \mathbf{L}_{pf}\overline{\mathbf{Z}}_p\right)\theta + \mathbf{L}_{ff}\varepsilon_f + \mathbf{L}_{pf}\mathbf{L}_{pp}^{-1}\mathbf{y}_p^*.$$

This equation can be rewritten in the following form

$$\mathbf{y}_f^* = \overline{\mathbf{Z}}_f\theta + \mathbf{L}_{ff}\varepsilon_f + const \tag{3.10}$$

where $\overline{\mathbf{Z}}_f = \mathbf{Z}_f - \mathbf{L}_{pf}\overline{\mathbf{Z}}_p$. As a result, the distribution $p(\mathbf{y}_f|\mathbf{x}_0,\sigma,\mathbf{y}_p)$ is approximated by $p(\mathbf{y}_f|\hat{\mathbf{x}}_0,\hat{\sigma},\mathbf{y}_p)$ which, in turn, is approximated by a multivariate Gaussian distribution with mean $(\mathbf{h}'\hat{\mathbf{x}}_{n+1},\ldots,\mathbf{h}'\mathbf{x}_{n+r})'$ and variance matrix $\sigma^2\left(\overline{\mathbf{Z}}_f\left(\overline{\mathbf{Z}}_p'\overline{\mathbf{Z}}_p\right)^{-1}\overline{\mathbf{Z}}_f + \mathbf{L}_{ff}\mathbf{L}_{ff}'\right)$. The prediction intervals, for a specified coverage probability, are determined from a standard Gaussian distribution.

In our context, some of the elements of θ must be nonnegative. Thus whenever $\hat{\theta}_i < 0$, we replace it by 0 for the estimation of $Var(\mathbf{y}_f|\theta,\mathbf{x}_0,\sigma,\mathbf{y}_p)$. Since we still have yet to include the sampling error for $\hat{\sigma}$ in the variance matrix, it may be possible to improve the coverage of the PI's further. We investigate the addition of this sampling error in the next method.

Bayesian Simulation Method (BS)

We may specify the sampling distributions for $\hat{\theta}$, $\hat{\mathbf{x}}_0$ and $\hat{\sigma}$ in terms of the joint pdf: $p(\hat{\theta},\mathbf{x}_0,\hat{\sigma}|\theta,\mathbf{x}_0,\sigma)$; where $p()$ denotes a generic pdf. We may then develop the predictive distribution, in the Bayesian framework of Aitchison and Dunsmore (1975), as

$$p(\mathbf{y}_f|\mathbf{y}_p) = \iiint p(\mathbf{y}_f|\mathbf{y}_p,\theta,\mathbf{x}_0,\sigma)p(\theta,x_0,\sigma|\mathbf{y}_p)(d\theta d\mathbf{x}_0 d\sigma) \qquad (3.11)$$

where the differential element covers all the items in θ and in $\mathbf{x}_0$, so that the triple integral represents $k+a+1$ dimensions in all, a being the dimensionality of θ and k being the number of states. The posterior density for $(\theta,\mathbf{x}_0,\sigma)$ is determined from the sampling distribution and a suitable prior in the usual way as

$$p(\theta,\mathbf{x}_0,\sigma|\mathbf{y}_p) = p(\theta|\sigma,\mathbf{x}_0,\mathbf{y}_p)p(\sigma|\mathbf{x}_0,\mathbf{y}_p)p(\mathbf{x}_0|\mathbf{y}_p) \qquad (3.12)$$

In our numerical work, we found that $\hat{\mathbf{x}}_t$ tended to $\mathbf{x}_t$ quite quickly whenever the estimates of the smoothing parameters were non-zero; this observation has two main consequences: first of all, the results were largely unaffected by the variations in the seed state vector $\hat{\mathbf{x}}_0$; secondly, this relative insensitivity led to some numerical instabilities, particularly when m=12. For both reasons, we decided to focus attention upon the variations in $\hat{\theta}$ and $\hat{\sigma}$ only. Thus, we now reformulate (3.11) for the current problem as:

$$p(\mathbf{y}_f|\mathbf{y}_p) = \iint p(\mathbf{y}_t|\mathbf{y}_p,\theta,\hat{\mathbf{x}}_0,\sigma)p\left(\theta,\sigma|\mathbf{y}_p,\hat{\mathbf{x}}_0\right)(d\theta d\sigma) \tag{3.13}$$

that is, we perform a simulation to arrive at the predictive distribution, which has the form:

$$p(\mathbf{y}_f|\mathbf{y}_p) = \iint p(\mathbf{y}_f|\mathbf{y}_p,\theta,\hat{x}_0,\sigma)p(\theta|y_p,\hat{x}_0,\sigma)p(\sigma|\mathbf{y}_p,\hat{\mathbf{x}}_0)(d\theta d\sigma) \tag{3.14}$$

A Monte Carlo integration method is employed to evaluate $p\left(\mathbf{y}_f|\mathbf{y}_p\right)$ as follows; the steps are entirely the same as those described in Ord, Koehler and Snyder (1997):

a) $p\left(\sigma|\hat{\mathbf{x}}_0,\mathbf{y}_p\right)$ is approximated by an inverted gamma distribution. A value of σ^2 is randomly generated from the approximating distribution.

b) $p\left(\theta|\sigma,\hat{\mathbf{x}}_0,\mathbf{y}_p\right)$ is approximated by a Gaussian distribution with mean $\hat{\theta}$ and variance matrix (3.8). A value of θ is randomly generated from the approximating distribution. Those elements of θ which violate the invertibility conditions are adjusted. For example, for the additive structural model in Example 2.2, we saw that the smoothing parameters α must be nonnegative. Negative values are truncated to zero. (We ignore the other restrictions on the smoothing parameters for the Holt-Winters method because we rarely found them to be binding in practice.)

c) The distribution $p(\mathbf{y}_f|\theta,\hat{\mathbf{x}}_0,\sigma,\mathbf{y}_p)$ is approximated by a synthetic sample of M values of the vector $\mathbf{y}_f$ generated from the model in Example 2.2. The future values of e_t required for the calculation of each instance of $\mathbf{y}_f$ are themselves generated from an $N(0,\sigma^2)$ distribution. Thus, for each $\mathbf{y}_f$ we estimate the probability density function by

$$p(\mathbf{y}_f|\mathbf{y}_p) = N^{-1}\sum_i p(\mathbf{y}_f|\mathbf{y}_p,\theta_i,\hat{\mathbf{x}}_0,\sigma_i) \qquad (3.15)$$

where the N sets of values of θ_i and σ_i are generated in accordance with steps a-b above.

d) Prediction intervals are constructed directly from the sample of the $\mathbf{y}_f$, for a specified confidence interval P, by deleting those $M(1-P)$ sample values that are furthest from the associated point prediction for each period t. The smallest and largest values that remain in the culled sample are used as the lower and upper boundaries of the prediction intervals.

This method is similar to one proposed by Thompson and Miller (1986) for stationary AR(p) processes. They do not impose constraints on the parameters for stationarity. Nor do they employ maximum likelihood estimates.

4. THE SIMULATION STUDY

A simulation study was undertaken to compare the above prediction interval methods. We constrained the scope of the study to the additive Holt-Winters method of forecasting. Our choice was motivated by the fact that the additive Holt-Winters method is widely used in practice, that traditionally users of this method have relied on heuristics rather than sound statistical methods to compute associated prediction intervals, and that the structural model

underpinning it is a non-trivial example from the ARIMA class (as shown in Example 2.2). Any simulation study is necessarily limited by the range of model options selected, but we believe that the results of our study are reasonably representative of more complex models, and likely to provide greater insights than special cases such as AR(1) or MA(1). The case of the multiplicative HW scheme has been considered in Ord, Koehler and Snyder (1997).

4.1 Design

Each original series was generated using the additive Holt-Winters scheme described in Example 2.2. The initial conditions were: $\ell_0 = 100$, b_0=2, $c_{j-m} = A \sin(2\pi j/m), j$=1,2...,$m$ where A denotes the seasonal amplitude . Clearly any distribution may be used in the simulations, although our emphasis in this paper has been on the normal distribution. We considered a mixture of normals (1-q)*N(0,1) + q*N(0,4) for ε_t/σ, q representing the proportion of outliers. Prediction intervals were generated for probability levels 0.90, 0.95 and 0.99. The specified factors in the design are shown in Table 4.1. Note that one set of values for α is (0,0,0) so that we may examine the impact of boundary values upon the Bayesian simulation method.

-- Insert Table 4.1 ----

The design generated a total of 36 scenarios or factor combinations for quarterly data and 24 for monthly data. Each scenario corresponds to a single choice for (n, A, σ, α) from Table 4.1. Each scenario was replicated 10 times. For each replication, we executed the following sequence of steps:

a) generate estimates of $\mathbf{x}_0$, denoted by $\tilde{\mathbf{x}}_0$, using Winters' approach (see Bowerman and O'Connell, 1993, Chapter 8), based upon the first three years of data.

b) determine the ML estimators for α conditional upon $\tilde{\mathbf{x}}_0$;

c) generate the forecasts for 8 lead times with quarterly data and 24 lead times with monthly data.

d) generate the PI for each of the four approaches described earlier: heuristic, plug-in, Bayesian simulation, and linear approximation.

e) for each of the four methods generate intervals for 0.90, 0.95 and 0.99 probability levels.

This entire process was done twice: once when the error terms are normally distributed (8,640 intervals for quarterly data and 17,280 for monthly data) and once when they have a mixed distribution. We used N=M=1000 and based the analysis upon expression (3.15). The simulations were carried out using GAUSS on a Pentium PC.

4.2 Analysis

A so-called *coverage index*, defined by

CI = (proportion of original series that fall within constructed PIs)/(nominal coverage),

was used to measure the performance of the prediction intervals. Thus, when the nominal coverage is (1-α), CI has an upper limit of 1/(1-α) or 1.11, 1.05, 1.01 for the values 0.90, 0.95 and 0.99 respectively. Since we want CI to be concentrated around 1.0, it comes as no surprise that the measure exhibits marked negative skewness. In general, we found the median to be a more reliable guide to performance than the mean, although both sets of values are recorded in Tables 4.2-4.3. The left interquartile range(LQ=median - lower quartile) was used to describe variation in the lower tail, and PUL, the percentage of CI

values hitting the upper limit was used to describe the behaviour of the upper tail. The standard deviation was found to be unreliable as a summary of variability because of the asymmetric nature of the distribution and the severe impact of a few extreme values.

-- Insert tables 4.2 and 4.3 –

The results presented in these tables as percentages, are averaged across replications and scenarios; that is, we report 'main effects'. Thus, for example, the CI values in Table 4.2 on the impact of sample size for quarterly series with q=0 are each based upon 2880 values [={36 scenarios} X {10 replicates} X {8 lead times} X{3 levels of PI} / {3 levels of n}]. As observed earlier, each value is based upon $M = N = 1000$ iterates. With the exception of the row labeled 'Mixed', all results are based on normally distributed disturbances. The single row for a non-normal distribution shows results for a mixture of normal distributions with q = 0.2.

4.3 Interpretation of the results

An examination of Tables 4.2-4.3 leads to the following conclusions, grouped by method.

Heuristic (HE)

1. Coverage is below nominal in all major categories, but generally close to that of the plug-in method.
2. Underestimation appears to be most serious for small sample sizes, longer lead times, a confidence level of 90 percent and the intermediate level of smoothing constants.
3. Changes in the standard deviation, the seasonal amplitude, and the presence/absence of outliers seem to have a negligible impact on performance.

4. Coverage tends to be lower for monthly than for quarterly series.
5. The values of LQ and PUL tend to be about the same as for the plug-in method. In each case the LQ is somewhat higher than for the other two methods, and the PUL is very small, consistent with the persistent tendency for the intervals to have coverage below their nominal levels.

'Plug-In' (PI)

1. Coverage again is below nominal, but slightly better than the heuristic method for longer lead times.
2. Underestimation occurs when the sample size is small (especially for monthly data), for longer lead times, when confidence level is 90 percent, for the highest level of the smoothing constants (especially for monthly data).
3. Changes in the standard deviation, the seasonal amplitude, and the presence/absence of outliers seem to have a negligible impact on performance.
4. Coverage tends to be lower for monthly than for quarterly series.
5. Comments for LQ and PUL as for the heuristic method.

Bayesian simulation (BS)

1. Coverage is a little below nominal across all major categories for quarterly data primarily because of the effects of the smaller sample size, which has spill-over effects into the other summary classifications.
2. For monthly data, the average coverage is quite close to the nominal level.
3. In the case of quarterly data, some underestimation occurs for the highest level of smoothing constants, but this is much less marked than for the heuristic and plug-in approaches.

4. The LQ values are somewhat lower than for the heuristic and plug-in methods, but are accompanied by higher values for PUL; the linear approximation has slightly higher LQ values, but much lower PUL levels. There are modest improvements in these measures with increases in sample size.

5. Including the sampling error for $\hat{\theta}$ and $\hat{\sigma}$ in the variance of y_f improves coverage.

Linear Approximation (LA)

1. The intervals tend to be too narrow although the results are quite reasonable in larger samples.
2. LQ is higher than the Bayesian approach, but PUL is lower.
3. Coverage is much better than heuristic and 'plug-in' methods. Taking the sampling error for $\hat{\theta}$ into account makes a difference.

Overall, we regard the performance of the Bayesian simulation as superior to the other three methods, unless the sample size is large. For large samples, the linear approximation method is reasonable but not quite as good as the Bayesian simulation method. Although the Bayesian simulation method is computationally quite intensive, it takes less than a minute on a Pentium computer for application to a single series and provides an estimate of the complete predictive distribution, not just the PI; Tsay(1993) and others have argued that such an approach is more desirable. In particular, the entire predictive distribution is required when we consider cost-based loss functions, rather than measures such as squared-error loss. Also, Bayesian simulation is readily extendable to non-normal errors, although its performance remains to be explored.

Finally, we note that the results for the Bayesian simulation hold up even when $\alpha = 0$. As a practical matter, we recommend that the model be re-estimated whenever one or more (close to) zero estimates arise; the PI should then be constructed from the revised model.

5. SUMMARY

We have identified four approaches to the construction of prediction intervals for linear time series processes. Using the additive Holt-Winters method as an example, we conducted an extensive simulation study to examine the coverage provided by these methods and found the Bayesian simulation approach to be superior to the others, at least in the case of normally distributed errors. We found considerable gains in the accuracy of coverage by taking the estimation of θ and σ into account in the Bayesian simulation method.

Chatfield (1993) noted that several problems remain in the construction of valid prediction intervals, as discussed in the Introduction. We believe that the present framework will provide a sound basis for examining a number of these issues.

REFERENCES

Aitchison, J. and Dunsmore, I.R., *Statistical Prediction Analysis.* London: Cambridge University Press, 1975.

Akaike, H., 'Markovian representation of stochastic processes and its application to the analysis of autoregresssive moving average processes', *Annals of the Institute of Statistical Mathematics*, **26** (1974), 363-87.

Ansley, C.F. and Kohn, R., 'Prediction mean square error for state space models with estimated parameters', *Biometrika,* **73** (1986), 467-73.

Aoki, M., 'Two complementary representations of multiple time series in state-space innovation forms', *Journal of Forecasting*, **13** (1994), 69-90.

Archibald, Blyth C., 'Parameter Space of the Holt-Winters Model', *International Journal of Forecasting*, **6** (1990), 199 - 209.

Barnett, G. Kohn, R., and Sheather, S.J. 'Robust estimation of an autoregressive model using Markov chain Monte Carlo', *Journal of Econometrics*, **74** (1996), 237-254.

Barnett, G. Kohn, R., and Sheather, S.J., 'Robust Bayesian estimation of autoregressive-moving average models', *Journal of Time Series Analysis,* **18** (1997), 11-28.

Bowerman, B.L. and O'Connell, R.T., *Time Series Forecasting.* Boston: Duxbury, 1993.

Chatfield, C., 'Calculating interval forecasts', *Journal of Business and Economic Statistics,* **11** (1993), 121-35.

De Jong, D.N. and Whiteman, C.H., 'The forecasting attributes of trend- and difference-stationary representations for macroeconomic time series', *Journal of Forecasting,* **13** (1994), 279-97.

Garcia-Ferrer, A. and Del Hoyo, J., 'On trend extraction models: Interpretation, empirical evidence and forecasting performance', *Journal of Forecasting,* **11** (1992), 645-65.

Geweke, J., 'Bayesian inference in econometric models using Monte Carlo integration', *Econometrica*, **57** (1989), 1317-39.

Harvey, A.C., *Forecasting, Structural Time Series Models, and the Kalman Filter*. London: Cambridge University Press, 1990.

Kabaila, P., 'On bootstrap predictive inference for autoregressive processes', *Journal of Time Series Analysis,* **14** (1993), 473-84.

McCullough, B.D., 'Bootstrapping forecast interval: An application to AR(p) models', *Journal of Forecasting*, **13** (1994), 51-66.

McKenzie, E., 'Concurrent seasonal adjustment with Census X-11', *Journal of Business and Economic Statistics*, **2** (1984), 235-49.

Ord, J.K., Koehler, AB. and Snyder, R.D., 'Estimation and prediction for a class of dynamic nonlinear statistical models', Journal of the American Statistical Association, **92** (1997), to appear.

Schweppe, F., 'Evaluation of likelihood functions for Gaussian signals', *IEEE Transactions on Information Theory'*, 11, 61-70.

Snyder, R.D., 'Recursive estimation of dynamic linear statistical models', *Journal of the Royal Statistical Society, Series B,* **47** (1985), 272-76.

Snyder, R.D., *Statistical Foundations of Exponential Smoothing.* Working Paper 5/88, Department of Econometrics, University of Monash. 1988.

Thombs, L.A. and Schucany, W.R., 'Bootstrap prediction intervals for autoregression', *Journal of the American Statistical Association*, **85** (1990), 486-92.

Thompson, P.A. and Miller, R.B., 'Sampling the future: A Bayesian approach to forecasting from univariate time series models', *Journal of Business and Economic Statistics*, **4** (1986), 427-436.

Tsay, R.S., Discussion of 'Calculating interval forecasts' by C. Chatfield, *Journal of Business and Economic Statistics*, **11** (1993), 140-42.

Winters, P.R., Forecasting sales by exponentially weighted moving averages', *Management Science*, **6** (1960), 324-42.

Yar, M. and Chatfield, C., 'Prediction intervals for the Holt-Winters forecasting procedure', *International Journal of Forecasting*, 6 (1990), 1-11.

Table 4.1 Design of Simulation Experiment

factor	symbol	levels
number of seasons in year	m	4 12
sample size	n	16* 36 72
seasonal amplitude	A	0 30
proportion of outliers	q	0.0 0.2
standard deviation	σ	5 20
forecasting horizon	h	[=2m]
smoothing constants	α	(0,0,0), (0.2,0.1,0.1), (0.8,0.1,0.1)

* only used for m=4

Table 4.2 Coverage index: quarterly data: normal disturbances

	Average				Median				LQ				PUL				
Method	HE	PI	BS	LA	HE	PI	BS	LA	HE	PI	BS	LA	HE	PI	BS	LA	Count
Effect of prediction lead time																	
1	89	88	97	95	94	93	99	99	16	16	13	13	1	1	13	7	1080
2	86	85	94	92	92	91	98	97	16	16	14	15	2	1	11	5	1080
3	84	84	93	91	90	89	97	96	16	16	15	15	2	1	9	5	1080
4	81	82	91	90	88	88	97	95	18	16	15	16	3	2	10	6	1080
5	76	79	91	88	83	86	96	94	17	17	15	16	2	3	12	6	1080
6	73	78	89	87	80	85	96	93	17	17	15	16	3	3	12	7	1080
7	71	78	89	86	77	86	97	94	17	17	15	16	2	3	14	8	1080
8	69	78	88	86	74	86	97	94	17	17	15	16	3	4	15	9	1080
Effect of nominal interval probability																	
90	74	77	88	86	81	83	94	92	17	17	16	16	0	1	3	1	2880
95	78	80	91	89	86	88	97	95	17	16	15	16	1	1	6	3	2880
99	84	87	96	93	93	94	99	98	16	16	13	14	6	6	27	17	2880
Effect of sample size																	
16	65	68	89	84	68	72	96	90	16	16	15	16	2	2	20	8	2880
36	81	83	91	89	89	89	96	95	16	16	15	16	3	2	9	6	2880
72	90	92	95	95	96	96	99	99	15	15	13	13	3	3	8	6	2880
Effect of distribution																	
Normal	79	81	91	89	87	89	97	96	17	16	15	15	2	2	12	7	8640
Mixed	79	83	92	90	88	91	98	96	16	16	13	15	1	2	9	4	8640
Effect of amplitude of seasonal cycle																	
0	77	81	91	89	86	88	97	96	17	16	15	15	2	2	11	6	4320
30	80	82	92	90	88	90	98	96	16	16	14	15	3	2	13	7	4320
Effect of standard deviation of disturbances																	
5	78	81	91	89	88	88	97	95	16	16	15	16	2	2	11	6	4320
20	79	82	92	90	86	89	98	96	17	16	14	15	2	3	13	8	4320
Effect of smoothing constants																	
0	85	86	97	95	95	94	100	99	16	16	13	13	2	2	18	10	2880
20	77	82	92	90	80	86	96	94	17	17	15	16	0	1	8	4	2880
80	74	76	86	84	86	84	93	90	17	17	16	16	5	4	11	6	2880

Table 4.3: Coverage index: monthly data: normal distribution

	Average				Median				LQ				PUL				
Method	HE	PI	BS	LA	HE	PI	BS	LA	HE	PI	BS	LA	HE	PI	BS	LA	Count
Effect of prediction interval																	
1	83	82	97	91	89	88	100	96	26	26	17	21	0	0	21	3	720
2	80	79	96	88	87	86	99	95	27	27	18	22	0	0	21	2	720
3	77	77	95	86	84	85	99	93	28	28	18	24	1	1	23	3	720
4	74	77	95	84	84	85	100	92	29	28	17	23	1	1	26	3	720
5	71	77	95	84	79	84	100	92	29	28	17	23	2	1	30	3	720
6	68	77	95	84	76	85	100	92	30	28	17	23	2	2	30	5	720
12	56	84	95	88	50	92	101	95	23	23	16	22	2	11	38	13	720
18	48	87	95	90	35	95	101	97	15	22	16	21	2	16	42	20	720
24	45	88	95	91	26	97	101	99	11	21	16	18	3	22	44	24	720
Effect of nominal interval probability																	
90	54	79	95	84	54	85	101	91	25	28	16	24	0	6	20	6	5760
95	57	82	95	87	60	90	101	95	27	25	16	22	1	9	33	10	5760
99	64	88	96	92	73	96	101	98	30	21	16	20	5	17	54	23	5760
Effect of sample size																	
36	41	76	96	83	32	83	101	91	14	28	16	24	0	11	50	13	8640
72	76	90	95	93	89	95	100	98	26	22	17	20	4	10	22	13	8640
Effect of distribution																	
Normal	58	83	95	88	62	91	101	95	28	24	16	22	2	10	36	13	17280
Mixed	60	84	95	88	66	92	100	96	29	23	14	20	1	7	30	9	17280
Effect of seasonal amplitude																	
0	57	81	94	86	58	88	101	94	26	26	16	23	2	10	34	12	8640
30	59	85	96	89	65	93	101	97	28	23	16	21	1	11	38	14	8640
Effect of standard deviation of disturbances																	
5	57	83	96	88	60	91	101	95	27	24	16	22	2	10	36	13	8640
20	59	83	94	88	65	91	101	96	28	24	16	21	2	10	35	13	8640
Effect of smoothing constants																	
0	65	87	99	95	84	95	101	100	28	22	16	17	2	22	45	28	5760
20	49	87	97	91	45	92	101	96	20	23	16	21	0	5	37	6	5760
80	60	75	90	78	74	82	99	85	30	28	18	28	4	4	26	5	5760