A test for improved forecasting performance at higher lead times

John Haywood* and Granville Tunnicliffe Wilson†

20 September 2003

Abstract

Tiao and Xu (1993) proposed a test of whether a time series model, estimated by maximum likelihood, was robust with respect to multi-step prediction. Their statistic looks explicitly at the change in parameter estimates when the model parameters are chosen to minimise the L-step forecast error sum of squares (for some lead-time $L > 1$), rather than the 1-step forecast error sum of squares. For a model that correctly represents the series this leads to no significant change in the estimates. A significant change suggests that the model is incorrect, and better L-step forecasts could be obtained using a different model. Tiao and Xu reported positive size distortions for their statistic at lead time $L = 2$, and suggested that more work was required before their test could be used at lead times greater than 2. We consider a score version of this test, but our statistic is based on an approximation to the reduction in the L-step forecast error sum of squares, rather than the change in parameters. The test can be applied following maximum likelihood parameter estimation without re-estimating the model for L-step prediction. We show that the test has accurate (or conservative) size when applied to higher lead times, and present cases where it has greater power for detecting model inadequacy than the standard portmanteau test. The statistic is derived by frequency domain methods that provide insight into its distributional properties, but may be expressed as a quadratic form in the residual autocorrelations from maximum likelihood estimation.

Keywords: DIAGNOSTIC STATISTIC, EWMA, FREQUENCY DOMAIN, MODEL ROBUSTNESS, MULTI-STEP ERRORS

1 Introduction

The use of ARIMA models for time series forecasting generally involves three stages of modelling: model selection, estimation and checking. The estimation of the model parameters is usually by maximum likelihood, or some close approximation to this, that is asymptotically equivalent to finding the parameters that minimise the sum of squares of the in-sample *one step ahead* forecast errors. The guarantee that the model will also produce good multi-step forecasts, at lead times $L > 1$, lies with the stages of model selection and checking. Model selection ensures that the model is capable of representing the observed statistical features of the series. Model checking ensures that it has satisfactorily achieved this aim, following parameter estimation. The standard model check is the Box-Pierce test (Box and Pierce 1970) for lack of residual correlation, or its modified version,

*School of Mathematical and Computing Sciences, Victoria University, PO Box 600, Wellington, NZ.
†Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK.

the Ljung-Box test (Ljung and Box 1978). These tests encompass a wide range of alternatives, so are not as powerful as a test against a specific alternative.

In this paper we consider a specific test to be used as a model check. The test is whether the fitted model is robust with respect to multi-step prediction: by varying the parameters of the fitted model, can significantly better forecasts be obtained for some lead time $L > 1$? Tiao and Xu (1993) investigated in depth a particular example in which this test is appropriate. They proposed two statistics D_J^2 and T_J^2, where $J = L - 1$, based on the change in the estimate of the single parameter η of the IMA(1,1) model, when it was re-estimated to minimise the L-step, rather than the 1-step, prediction error sum of squares. They established the asymptotic distributions of D_J^2 and T_J^2 as χ_1^2 and χ_J^2 respectively under the null hypothesis, and investigated in finite samples the size and power properties of the chi-squared tests based upon D_J^2 and T_J^2. For the power study they took an ARMA(1,1) process as the alternative model.

In the next section we review relevant frequency domain estimation procedures. Then in section 3 we propose a score statistic for testing the same problem considered by Tiao and Xu (1993), the score being the gradient of the multi step prediction error sum of squares with respect to the model parameters, evaluated at the maximum likelihood estimate of the parameters. The advantage of this statistic is that it can be readily evaluated without re-estimation of the model, and its distribution can be more reliably determined, because it depends only on the residuals obtained following maximum likelihood estimation. However, in common with all score tests, it may lack power because it depends on local rather than global properties of the model. In section 4 we compare the performance of this score test on the same example investigated by Tiao and Xu. They suggested that further work was needed before their test could be used for lead times greater than two, whereas we will show that our test may be used for these higher lead times. Finally in section 5 we apply our test to an autoregressive model of order 6. We demonstrate good size properties under the null hypothesis and substantial power for rejection, when the AR(6) model is fitted to a series generated as an ARMA(3,3) process.

2 Frequency domain estimation

Haywood and Tunnicliffe Wilson (1997), hereafter referred to as HTW, presented a frequency domain approach to estimating time series model parameters by minimising the multi-step ahead prediction error sum of squares. The score statistic that we now investigate is proposed in the discussion of that paper. In recent years there has been renewed appreciation of frequency domain methods for the estimation and testing of parametric time series models (e.g. Harvey 1989, section 4.3), using what is often termed the Whittle likelihood. The (asymptotic) independence of the ordinates of the sample spectrum help to simplify theoretical analysis, and the numerical computations are also straightforward. For estimation and testing using a multi-step error criterion, the computations require finite Fourier transformations between time domain and frequency domain coefficients, so we begin by reviewing the computations required for construction of the test statistic, as presented in HTW.

We assume that the time series x_t under consideration, may be modelled as an integrated process that requires differencing of order d to yield a stationary process w_t. Suppose that the model for w_t has a spectrum $S(f)$, for $0 \le f \le \frac{1}{2}$, which has a linear form with coefficients $\beta_1, \beta_2, \ldots \beta_k$:

$$S(f) = \sum_{i=1}^{k} \beta_i S_i(f) = S_\beta \beta, \tag{1}$$

where S_β is the row vector of component functions $S_i(f)$, and β is the column vector of coefficients. We shall later extend (1) to the assumption that the spectrum has a linear *approximation* of this form, local to some set of coefficients, so that S_β is the vector of derivatives of $S(f)$ with respect to β. Let the infinite moving average representation of the model for w_t be

$$w_t = (1 + \psi_1 B + \psi_2 B^2 + \cdots)e_t = \psi(B)e_t, \tag{2}$$

where e_t, with variance σ^2, is the white noise linear innovation of w_t. Then the model spectrum for w_t is given by

$$S(f) = \sigma^2 |\psi\{\exp(i\omega)\}|^2 = |\nu\{\exp(i\omega)\}|^2, \tag{3}$$

where $\nu(B) = \sigma\psi(B)$ and $\omega = 2\pi f$. The infinite moving average representation of the integrated process $x_t = (1 - B)^{-d} w_t$ is then formally given as

$$x_t = (1 - B)^{-d}\psi(B)e_t = \Psi(B)e_t. \tag{4}$$

Now define the truncated operator:

$$T(B) = \sigma\left(\Psi_0 + \Psi_1 B + \cdots + \Psi_{L-1}B^{L-1}\right). \tag{5}$$

Then the error $e_t(L)$ in the multi-step prediction of x_{t+L} made at time t, that results from applying this model, may be expressed:

$$e_t(L) = \frac{T(B)}{\nu(B)} w_t. \tag{6}$$

The mean sum of squares of L-step ahead prediction errors then has the frequency domain approximation as

$$F = \int_{-\frac{1}{2}}^{\frac{1}{2}} G(f)\frac{S^*(f)}{S(f)}df = \left[G(f)\frac{S^*(f)}{S(f)}\right]_0 \tag{7}$$

where

$$G(f) = |T\{\exp(i\omega)\}|^2 \tag{8}$$

and

$$S^*(f) = \frac{1}{n}\left|\sum_{t=1}^{n} w_t \exp(i\omega t)\right|^2 \tag{9}$$

is the sample spectrum of w_t. We have also introduced here the notation $[A]_k$ for the kth Fourier coefficient of the function $A(f)$ of frequency f over the range $-\frac{1}{2} < f \leq \frac{1}{2}$. An important point, both practical and theoretical, is that all such functions of frequency will be evaluated at the grid of frequencies $f_j = j/n$, where n is the length of the series w_t and $j = 0, 1, \ldots, [\frac{n}{2}]$. The finite Fourier transform (and its inverse) will be applied to form Fourier series at these frequency points, and to extract their coefficients. We shall also use these transforms to construct, from $A(f)$, further functions of frequency defined as the truncated Fourier series:

$$[A]_l^m \sim \sum_{l}^{m} [A]_k \exp(ik\omega) \tag{10}$$

and

$$[A]_{\frac{1}{2}}^m \sim \tfrac{1}{2}[A]_0 + \sum_{1}^{m} [A]_k \exp(ik\omega). \tag{11}$$

The minimisation of F with respect to β requires the iterative solution of equations

$$H\delta = g \tag{12}$$

where $g = -F_\beta$ is the negative transpose of the score F_β, the derivative of F with respect to β, and H is an estimate of the Hessian of F. These quantities are evaluated at a given parameter vector $\beta = \beta_{\text{old}}$, and the equations solved to obtain the parameter correction δ, which is used to construct new parameters $\beta_{\text{new}} = \beta_{\text{old}} + \delta$.

The expressions for g and H are, using the notation we have introduced above,

$$g = \left[G \frac{S'_\beta}{S} \frac{S^*}{S} \right]_0 - 2 \left[\bar{T} \left[T \left[\frac{S'_\beta}{S} \right]_{\frac{1}{2}}^{\infty} \right]_0^{L-1} \frac{S^*}{S} \right]_0, \tag{13}$$

and

$$H = \left[G \frac{S'_\beta}{S} \frac{S_\beta}{S} \right]_0 - 2 \left[\bar{T} \left[T \left[\frac{S'_\beta}{S} \right]_{\frac{1}{2}}^{\infty} \right]_0^{L-1} \frac{S_\beta}{S} \right]_0. \tag{14}$$

To simplify the notation, and to specify the numerical procedure more precisely, we introduce first the n-vector $Y = S^*/S$ and $n \times k$ matrix $X = S_\beta/S$. The rows of these correspond to the frequency grid f_j, and the columns of X to the coefficients β. The ratios in these expressions are computed row-wise. We further introduce the $n \times k$ matrix

$$Z = GX - 2\bar{T} \left[T \left[X \right]_{\frac{1}{2}}^{\infty} \right]_0^{L-1}, \tag{15}$$

where again, the multiplications by G, $\bar{T}$ and T are row-wise. We may now express $g = hZ'Y$ and $H = hZ'X$, where $h = 1/n$ is the interval of the frequency grid. Because all of g, H, Y and X are real, we can make the following modifications to Y, X and Z. The symmetry of spectra allows us to truncate Y, X and Z, retaining only the first m rows corresponding to the frequencies $0 \le f \le \frac{1}{2}$ and discarding the remainder. We also discard the imaginary part of Z. Consequently, in computing $Z'Y$ and $Z'X$, a weighting of $\frac{1}{2}$ should be applied to the first of the remaining rows, corresponding to frequency $f_0 = 0$ and, if n is even, also to the last of the remaining rows, corresponding to $f_{n/2} = \frac{1}{2}$. The value of h should also become $2/n$. The weighting of $\frac{1}{2}$ is conveniently implemented by multiplying the rows that correspond to frequencies $f = 0$ and $f = \frac{1}{2}$ in Y, X and Z, by $\sqrt{\frac{1}{2}}$. We shall hereafter assume that these modifications have all been made, so that the expressions of $g = hZ'Y$ and $H = hZ'X$ remain valid. Besides leading to some numerical efficiency, these modifications will also assist the later distributional analysis. We shall also later use a vector of ones, modified in the same way.

3 The score test statistic

We propose the following test. First, estimate the parameters β by approximate maximum likelihood, which is equivalent to minimising the sum of squares of one-step ahead prediction errors. This is done in the frequency domain by maximising the Whittle likelihood, using iteratively re-weighted least squares. We repeatedly solve

$$X'Y = X'X\beta \tag{16}$$

with the new value of β being used to redefine the fitted model spectrum S, and thence the vector $Y = S^*/S$ and matrix $X = S_\beta/S$ for the next iteration. Let the final (converged) solution be $\hat{\beta}$.

Starting with this value of $\beta = \hat{\beta}$, form Y, X, Z and the gradient g and Hessian H, using the lead time L that has been specified for the test. Solve

$$g = H\delta \tag{17}$$

for the step δ in the parameter β towards the estimate $\tilde{\beta}$ which minimises the frequency domain approximation F of the sum of squares of L-step ahead prediction errors. Finally, evaluate the approximation to the reduction in F achieved by this parameter step, assuming F to be quadratic in a region containing $\beta = \hat{\beta}$ and $\tilde{\beta}$. This assumption is reasonable under the hypothesis that S is the true model for w_t, because $\tilde{\beta}$ should then be close to $\hat{\beta}$. The approximate reduction is $\frac{1}{2}q$ where

$$q = g'\delta = g'H^{-1}g. \tag{18}$$

This is our basic test statistic; a quadratic form in the score g. Note that it is not a direct measure of the size of the parameter step, such as $\delta'\delta$, or an indirect, but commonly used, measure of the size of this step, such as $\delta'V^{-1}\delta$, where V is a consistent estimate of $\mathrm{Var}\,\delta$. Rather, our statistic is a measure of the reduction in the L-step ahead prediction error variance that may be achieved by re-estimation of the parameters. In practise, iteration of this process would be required to achieve the exact minimum of F. Asymptotically however, one step will be sufficient, and we base our finite sample score statistic q on this single iterate reduction.

To employ the statistic q, we shall need a consistent estimate of its distribution. First, however, we must draw attention to the fact that H is singular, and explain how we overcome this. In HTW we noted that $\beta'g = 0$ and $\beta'H = 0$ for the values $\beta = \beta_{\mathrm{old}}$ used to construct g and H. The solution of $H\delta = g$ for the step δ therefore contains an arbitrary multiple of β_{old}. The source of this singularity is that F is invariant to any scaling of S, and hence of β. In HTW we proposed a normalisation of β to obtain a unique estimate. However, because $q = \frac{1}{2}g'\delta$, and $g'\beta_{\mathrm{old}} = 0$, the value of the score statistic is not changed by the addition of any arbitrary multiple of β_{old} to δ. We therefore employ a generalised inverse of H in (18).

In the remainder of this section we investigate the sampling properties of q, and show that it may be approximated as a weighted sum of $(k-1)$ independent chi-squared variables C_i on 1 degree of freedom

$$q \approx \sum_{i=1}^{k-1} D_{i,i}\, C_i. \tag{19}$$

Our development begins by expressing, from (18),

$$q = h^2 \hat{Y}' Z H^{-1} Z' \hat{Y} \tag{20}$$

where we have written $\hat{Y}$ in place of Y, to emphasise that q is formed at $\beta = \hat{\beta}$, so that $Y = \hat{Y} = S^*/\hat{S}$, the sample spectrum of the residuals from (approximate) maximum likelihood estimation of the model. Thus q is a quadratic form in the residual sample spectrum (and hence also a quadratic form in the residual sample autocorrelations). We therefore require an approximate formula for this residual sample spectrum, which we develop as follows. In large samples, for which the likelihood may be accurately approximated by a quadratic about the true value of β, a single iteration only is required of the reweighted least squares regression, to obtain the approximate MLE

$\hat{\beta}$. Thus $\hat{\beta} = (X'X)^{-1}X'Y$, where Y is the sample spectrum S^*/S of the *true* model innovations e_t. Consequently we can, in large samples, set

$$\frac{\hat{S}}{S} = X(X'X)^{-1}X'Y = UY. \tag{21}$$

We further approximate

$$\begin{aligned} \hat{Y} = \frac{S^*}{\hat{S}} &= 1 + \frac{S^* - \hat{S}}{\hat{S}} \\ &= 1 + \frac{S^* - \hat{S}}{S}\frac{S}{\hat{S}} \\ &= 1 + (I - U)Y\frac{S}{\hat{S}} \\ &\approx 1 + (I - U)Y \\ &= 1 + PY. \end{aligned} \tag{22}$$

In (22) we have first substituted for $\hat{S}/S$ from (21), then approximated $S/\hat{S}$ by its limit in probability, which is 1. The error in this approximation is neglected as being of second order. We obtain our final expression for q by noting, from HTW, that $Z'1 = 0$, where 1 is the (modified) vector of ones, and also $P\,1 = 0$, giving

$$q \approx h^2(Y-1)'PZH^{-1}Z'P(Y-1) = (Y-1)'ADA'(Y-1), \tag{23}$$

where D is a $(k-1)\times(k-1)$ diagonal matrix, and A is an orthonormal $m\times(k-1)$ matrix. Now D and A are functions of $\hat{\beta}$. The diagonals of D will be consistently estimated constants, and the columns of A are consistently estimated functions of frequency. Each element of $A'(Y-1)$ is an orthonormal linear function of the mean corrected spectral ordinates of unit variance white noise, and by the central limit theorem is asymptotically standard normal. The squares of these elements, C_i, are therefore each asymptotically chi-squared on 1 degree of freedom.

A further statistic, mentioned above, is the invariant measure of the magnitude of the parameter step δ. This is equivalently, and more conveniently, considered as the invariant measure of the magnitude of the score $g = Z'P(Y-1)$, which is given by $r = g'W^{-1}g$, where $W = \operatorname{Var} g$. By a similar argument to that used for q, we can demonstrate that r has, asymptotically, a chi-squared distribution on $(k-1)$ degrees of freedom.

4 Assessment of the statistic for the IMA(1,1) model

The IMA(1,1) model was investigated by Tiao and Xu (1993). We use the parameterisation of the spectrum of $w_t = (1-B)x_t$ as

$$S = \beta_1 + \beta_2(2 - 2\cos w) \tag{24}$$

which corresponds to taking x_t to be the sum of a random walk with innovation variance β_1 and independent white noise with variance β_2. In this case H is of rank 1, so the test statistic q is simply of the form $D_{1,1}C_1$ with $D_{1,1} = \frac{1}{2}h^2\mathrm{tr}(PZH^{-1}Z'P)$. We can therefore take $q/D_{1,1}$ to be asymptotically chi-squared on 1 degree of freedom. For the rank 1 case, $q/D_{1,1}$ is the same as the r statistic. It may be shown that for this example, the statistic is asymptotically the square of a linear combination of the sample residual autocorrelations, with the coefficient of the autocorrelation at lag j being proportional to the coefficient of z^j in

$$(1-\eta z)^{-1}\left[(1-\eta)(z + z^2 + \cdots + z^{L-1}) + z^L\right], \tag{25}$$

where η is the moving average parameter in the IMA(1,1) representation of the model. If η is close to 1, most of the weight will be placed on residual sample autocorrelations around lag L.

We carried out a simulation study to investigate the accuracy, under the null hypothesis, of the approximation we have presented for this statistic. For each combination of nominal size, lead time, value of η and series length given in Table 1, 10,000 replications were performed.

Lead time	Nominal size	Length of series	η = 0	0.2	0.4	0.6	0.8
2	10%	50	0.094	0.096	0.091	0.091	0.099
		100	0.094	0.096	0.093	0.094	0.093
		200	0.099	0.097	0.092	0.093	0.097
	5 %	50	0.046	0.046	0.043	0.040	0.054
		100	0.046	0.046	0.045	0.045	0.047
		200	0.049	0.048	0.046	0.043	0.047
4	10%	50	0.088	0.083	0.079	0.080	0.076
		100	0.091	0.091	0.089	0.089	0.085
		200	0.095	0.094	0.094	0.094	0.094
	5 %	50	0.040	0.038	0.039	0.036	0.039
		100	0.044	0.041	0.041	0.042	0.043
		200	0.047	0.047	0.045	0.046	0.047
10	10%	50	0.075	0.069	0.066	0.063	0.053
		100	0.082	0.082	0.083	0.081	0.071
		200	0.090	0.093	0.087	0.088	0.085
	5 %	50	0.041	0.036	0.032	0.027	0.021
		100	0.043	0.041	0.041	0.037	0.032
		200	0.045	0.043	0.044	0.041	0.041

Table 1: Empirical significance levels for proposed test statistic at various lead times. 10,000 replications for each tabulated entry.

The empirical size of the test statistic is accurate, or conservative, at all combinations of parameter values and lead times reported in Table 1, including $L = 10$. Thus there is no evidence of the excessive skewness (compared to the relevant asymptotic chi-squared distribution) reported by Tiao and Xu (1993) for all $L > 2$, which gave infeasible large positive size distortions for their proposed statistics, T_J^2 and D_J^2 ($J = L - 1$). While our test size does improve with series length, differences are often quite small and the test appears reasonably sized at moderate lead times, certainly for series of length 100 or greater.

We also investigated the power of the test when it was applied to the ARMA(1,1) process as an alternative data generating model, to which the IMA(1,1) was fitted. In this case, 1,000 replications were performed for series generated using each of six combinations of ARMA parameters ($\phi > \theta \geq 0$) at five lead times and two nominal sizes, with all series of length 200. Our six combinations of ARMA parameters form a subset of those considered by Tiao and Xu (1993) in their Table 3, where comparable (maximum) powers for the Ljung-Box statistic are also given. (Note Tiao and Xu gave empirical powers only for 10% nominal size.) Table 2 presents our results.

Some general patterns are suggested for our test statistic: power decreases with lead time for 'moderate' values of the AR parameter, $\phi < 0.5$ say, while power is maximised at medium lead

Nominal size	Lead time	(ϕ, θ)					
		$(0.1, 0.0)$	$(0.4, 0.1)$	$(0.7, 0.4)$	$(0.9, 0.0)$	$(0.9, 0.4)$	$(0.95, 0.3)$
10%	2	0.442	0.890	0.564	0.181	0.163	0.110
	4	0.271	0.788	0.694	0.265	0.268	0.135
	6	0.204	0.638	0.695	0.331	0.342	0.172
	8	0.165	0.542	0.667	0.356	0.353	0.173
	10	0.130	0.469	0.611	0.372	0.347	0.156
5 %	2	0.336	0.808	0.444	0.103	0.091	0.057
	4	0.200	0.698	0.578	0.163	0.165	0.063
	6	0.128	0.529	0.577	0.206	0.215	0.102
	8	0.095	0.432	0.542	0.212	0.210	0.085
	10	0.076	0.357	0.457	0.189	0.208	0.072

Table 2: Empirical powers for proposed test statistic at various lead times. 1,000 replications with series of length 200. Testing the $(1, \eta)$ estimated model against the (ϕ, θ) data generating process.

times ($L > 2, L < 10$) for higher values of ϕ. When compared to Tiao and Xu's (1993, Table 3) statistic T_1^2, our test has comparable empirical power for some cases (e.g. $\phi = 0.9$) but appears less powerful in others. However, our accurate (or conservative) empirical size may explain that difference; for example, Tiao and Xu report empirical sizes for T_1^2 of 16.5% and 11.7% (at nominal sizes of 10% and 5%) with series of length 100 and $\eta = 0.8$.

To interpret the observed patterns in the power of our test statistic, it is necessary to consider when the IMA(1,1) model could reasonably be expected to fit well a series generated by an ARMA(1,1) process. With η near unity the IMA(1,1) model can fit well a process that is close to white noise. Hence combinations of ARMA parameters that impose little structure can be well modelled by the IMA(1,1) (e.g. $\phi \approx \theta$). Similarly, with η near zero the IMA(1,1) model can fit well an ARMA process with ϕ close to one. Generated processes that are more clearly stationary within the sample period (with moderate values of ϕ) display reasonably rapid reversion to the mean; in such cases power will be greatest at low lead times, where the divergence between the autocorrelation structures beyond lag 1 is marked but forecast uncertainty is still moderate. Conversely, for generated processes that display more persistence (ϕ close to one), power will increase with lead time, since a greater horizon is required to differentiate between stationary and integrated observed behaviour. However, at long lead times (e.g. $L = 10$) power may reduce even for ϕ close to one, since with the IMA(1,1) model there is the additional uncertainty in predictions that accumulates rapidly with lead time, as is the case for integrated processes.

5 Assessment of the statistic for an AR(6) model

Autoregressive models are often used to approximate the structure of stationary time series for the purpose of prediction. We have therefore investigated the power of our test to detect when such an approximation might be significantly sub-optimal from the point of view of multi-step prediction. We chose as an alternative model, a series composed of the sum of three independent components; a near unit root AR(1) process, an AR(2) process with a stochastic cycle of period 25, and white noise. Thus $x_t = u_t + v_t + z_t$ where $u_t = \phi u_{t-1} + a_t$, with $\phi = 0.99$ and $\mathrm{Var}\, a_t = 1.0$; $v_t = 2r \cos \lambda v_{t-1} + b_t$ with $r = 0.98$, $\lambda = 2\pi/25$ and $\mathrm{Var}\, b_t = 0.3$; and z_t, uncorrelated with variance

1.0. To check the size of the test, we required an approximating autoregressive model of suitable order. From a simulation of the process x_t, use of the AIC selected an autoregressive approximation of order 6. We therefore determined the AR(6) model $\phi(B)x_t = e_t$ that provided the minimum variance one-step ahead predictor of x_t.

We simulated 1000 samples of time series of length 200 from this AR(6) model, fitted an AR(6) model to each sample in the frequency domain, and evaluated the test statistic for lead times $L = 8$ and $L = 16$. These were chosen as substantial fractions of the period of the stochastic cycle, at which it would be important to detect inadequacy of predictions. The frequency domain form of the AR(6) model spectrum S is given by the reciprocal linear expression $1/S = R = \sum_{j=0}^{6} \beta_j \cos j\omega = R_\beta \beta$. The approximate maximum likelihood estimation procedure in the frequency domain is the iterative solution of the weighted least squares equations $X'Y = X'X\beta$, where $X = R_\beta/R$ and $Y = 2 - RS^*$ is the adjusted response (see McCullagh and Nelder 1983, section 2.5). With these definitions of X and Y the evaluation of our multi-step test statistics proceeds exactly as previously described.

The results of these simulations are shown in Table 3. The empirical sizes are shown for the nominal sizes of 10% and 5%, for the tests based on both our statistics q and r, and for the Box-Pierce and Ljung-Box test statistics. For the q statistic the empirical sizes are not significantly different from the nominal sizes, though they are slightly over-sized for lead time 8 and slightly undersized for lead time 16. The r statistic at lead time 8 has empirical size close to the nominal, but this statistic is somewhat undersized for lead time 16. The Box-Pierce test is even more conservative and the Ljung-Box test is the most over-sized, particularly so at the 10% level.

Lead time	Nominal size	q Statistic	r Statistic	Box-Pierce	Ljung-Box
8	10%	0.112	0.108	0.067	0.128
	5 %	0.053	0.051	0.023	0.062
16	10%	0.098	0.079	0.067	0.128
	5 %	0.040	0.028	0.023	0.062

Table 3: Empirical size of tests of model adequacy for given lead time and nominal size, for an AR(6) null model.

We proceeded to estimate the power of the tests for rejecting the AR(6) model, when applying it to the alternative model described above, using the same series length and number of replications. Table 4 shows that the highest powers occurred for the test based on the q statistic, applied for lead time 16. The test of nominal size 5% then demonstrated an empirical power of 60% against the alternative. Of course if the selected model, in this case the AR(6), has been sensibly chosen, and if a check was made on the portmanteau diagnostic statistic, one would not expect to find a test that is capable of rejecting the null with exceptionally high power.

Lead time	Nominal size	q Statistic	r Statistic	Box-Pierce	Ljung-Box
8	10%	0.369	0.153	0.153	0.254
	5 %	0.272	0.256	0.086	0.146
16	10%	0.708	0.478	0.153	0.254
	5 %	0.599	0.351	0.086	0.146

Table 4: Empirical powers of tests of model adequacy for given lead time and nominal size for an AR(6) model fitted to the alternative process.

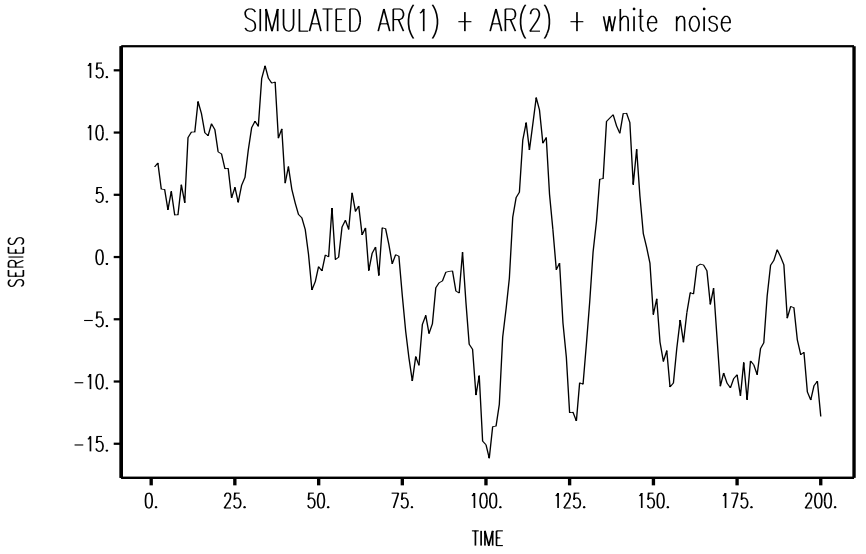

Figure 1: A realisation of the alternative model used for the power study.

We have determined that for the alternative model, the minimum prediction error variances that can be achieved for lead times 1, 8 and 16, using the true model, are respectively 3.28, 24.51 and 31.82. The error variances that are achieved at the same lead times, using the AR(6) model that minimises the one-step prediction error variance, are respectively 3.65, 46.15 and 72.61, and are clearly much greater for the lead times greater than 1. The error variances that can be achieved, using, for each respective lead time, the minimum variance linear predictors based upon the 6 most recent observations, are 3.65, 40.93 and 38.14. This shows that the potential reduction in prediction error variance that may be achieved is relatively small at lead time 8, from 46.15 to 40.93. But it is very substantial at lead time 16, from 72.61 to 38.14, which is not much greater than the value of 31.82 achieved by the true model. This accords with the results of the power investigation, that suggests it is difficult to detect model inadequacy using a lead time of 8, but quite possible at a lead time of 16.

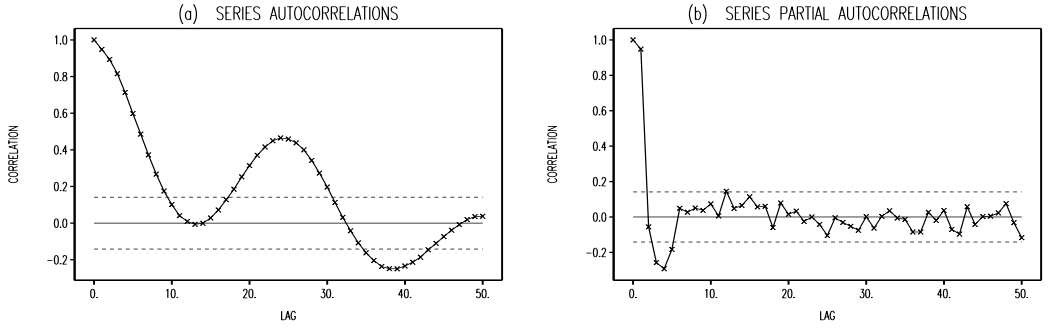

Figure 2: The sample autocorrelations and partial autocorrelations of the realised series.

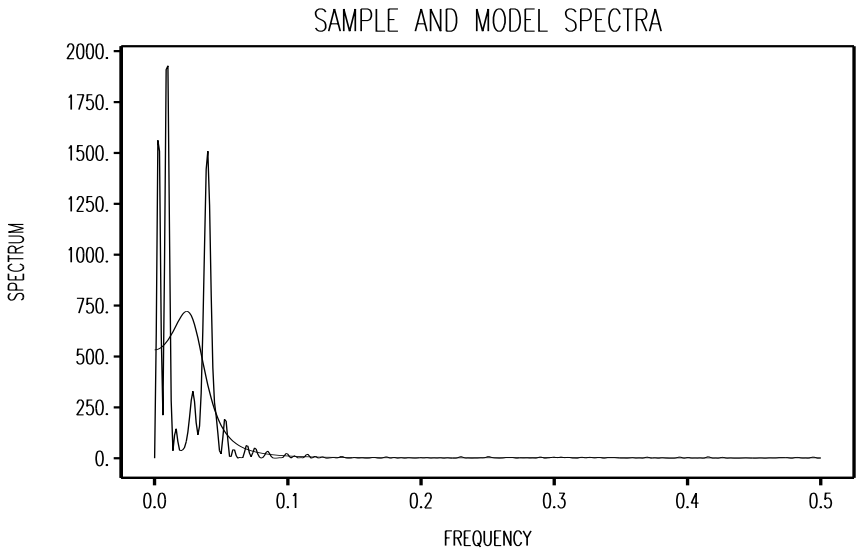

Figure 3: The sample spectrum of the realised series.

For this example it is worth looking more closely at one realisation of the alternative process. Figure 1 shows the realised series, Figure 2 its sample autocorrelation and sample partial autocorrelation, and Figure 3 its sample spectrum (plotted on a finer grid of frequencies than the harmonics). Superimposed on the sample spectrum is the spectrum of the fitted model. Figure 4 shows the residual series from fitting the AR(6) model, with its sample autocorrelations. Figure 5 shows the residual sample spectrum. Figure 2 suggests in fact that an AR(5) model would have been sufficient for this series, and Figures 4 and 5 suggest no obvious inadequacy in the residuals. However, Figure 3 shows that the fitted AR(6) fails to resolve the two peaks at frequency zero and 0.04, in the sample spectrum of the series. The residual spectrum in Figure 5 is the ratio of the sample spectrum to the model spectrum in Figure 3, and we might now notice the somewhat higher values in Figure 5, around these 'peak frequencies', with a tendency for lower values between them. But the natural variability of the white noise spectrum makes it difficult to assess the significance of these departures from uniformity of the spectrum. The test statistic q for the lead time 16 in this case has a p value of 2.14%, which indicates model inadequacy. The Box-Pierce and Ljung-Box statistics have values of 7.254 and 8.108, calculated from residual sample autocorrelations up to lag 20. These are not at all significant when referred to the chi-squared distribution on 14 degrees of freedom.

We believe that this example is typical of some of the real series that occur in econometric modelling. It may explain why some models prove to be poor at predicting turning points of cyclical components, because of their failure to properly resolve the low frequency structure. Further, it supports the application of the diagnostic statistic proposed in this paper.

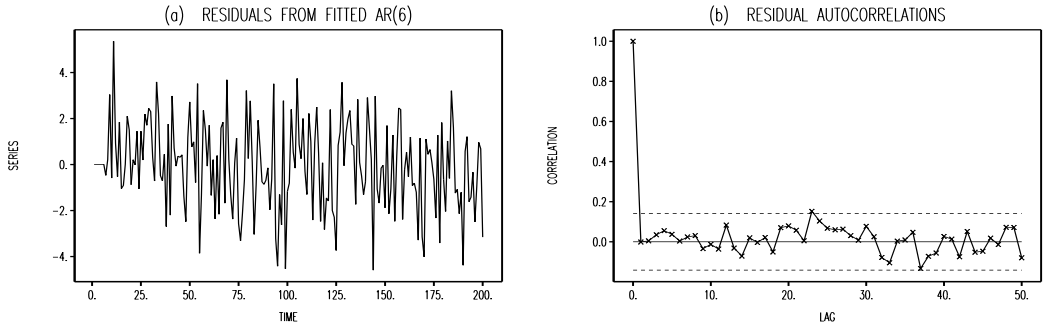

Figure 4: The residuals from fitting an AR(6) model to the realised series, and their sample autocorrelations.

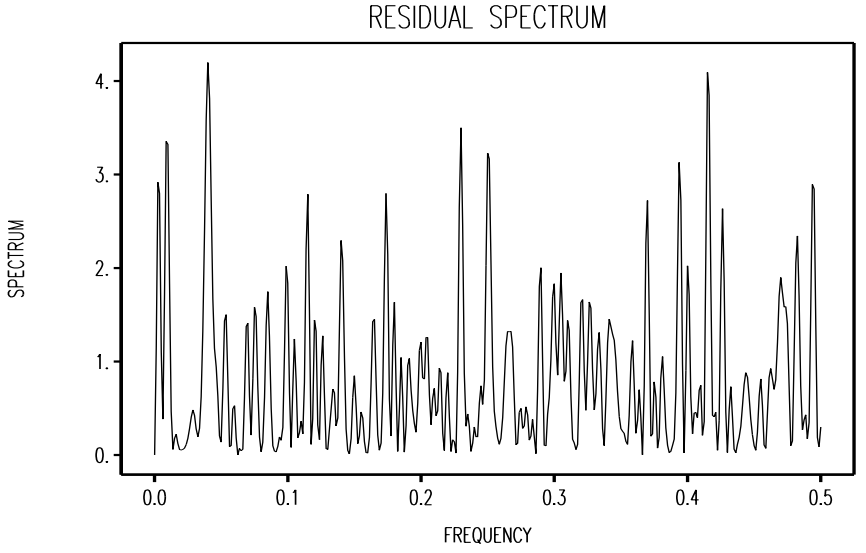

Figure 5: The residual spectrum after fitting an AR(6) model to the realised series.

References

Box, G.E.P. and Pierce, D.A. (1970) Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association*, 65, 1509–1526.

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

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

Ljung, G.M. and Box, G.E.P. (1978) On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303.

McCullagh, P. and Nelder, J.A. (1983) *Generalized Linear Models.* London: Chapman and Hall.

Tiao, G.C. and Xu, D. (1993) Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika*, 80, 623–641.