RESEARCH REPORT SERIES
(Statistics #2007-16)

Goodness-of-Fit and Badness-of-Fit Diagnostic Tests for Time Series Models

Tucker McElroy
Scott Holan*

University of Missouri-Columbia*

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: November 13, 2007

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

Goodness-of-Fit and Badness-of-Fit Diagnostic Tests for Time Series Models

Tucker McElroy*and Scott Holan[†]
U.S. Census Bureau and University of Missouri-Columbia

Abstract

Diagnostics for testing model goodness-of-fit and badness-of-fit for time series data are formulated by considering a convenient metrization of a spectral density's departure from constancy. The method is illustrated through numerical experiments and several case studies.

Keywords. ARMA, EXP Models, Frequency Domain, Nonstationary Time Series.

Disclaimer This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

This paper presents diagnostics for testing model goodness-of-fit (gof) and badness-of-fit (bof) for time series data, by considering a convenient metrization of a spectral density's departure from constancy. In the model-based approach to time series analysis, estimated residuals are computed once a fitted model has been obtained from the data, and these are then tested for "whiteness" i.e., it is determined whether they behave like white noise (Brockwell and Davis, 1996). Tests for residual whiteness include Portmanteau tests, such as Ljung and Box (1978), Li(2004), and Peña and Rodríguez (2002), and frequency domain tests – Beran (1994), Paparoditis (2000), Chen and Deo (2004), McElroy and Holan (2006), and Drouiche (2007); these procedures generally postulate whiteness of the residuals as the Null Hypothesis, so that significant rejections indicate model *inadequacy*. Since the classical statistical paradigm dictates that the practitioner seeks to reject

[*] Statistical Research Division, U.S. Census Bureau, 4700 Silver Hill Road, Washington, D.C. 20233-9100, tucker.s.mcelroy@census.gov

[†] Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100, holans@missouri.edu

Null Hypotheses by obtaining low p-values, we have the paradoxical situation that gof tests are actually designed to identify bad models. Nevertheless, what is wanted is an indication that the fitted model is good, or at least adequate. In order to obtain a statistically significant indication of model *adequacy*, one must reject a Null Hypothesis of non-whiteness; in other words, a badness-of-fit (bof) diagnostic is needed in order to find good models.

Here we present a convenient metrization of whiteness of a time series, which is quite similar to the approach of Drouiche (2007) and was developed independently. We demonstrate that our metrization intuitively captures gof/bof for time series, and can be used to test for model inadequacy/adequacy in a flexible and rigorous fashion. The work of Drouiche (2007), while similar in spirit, does not consider the bof applications. Section 2 defines our spectral measure – our metrization of whiteness – and motivates this choice by making connections to the work of Peña and Rodríguez (2003, 2006). We also derive some basic properties of our measure, which include those of Drouiche (2007), but with an additional facet that is arguably advantageous in the gof/bof context. Our formulation is at first very general, allowing for the application of the spectral measure in local frequency bands to nonstationary time series. Some examples are provided on familiar time series models in Section 3. Section 4 considers the statistical estimate of the spectral measure and its distributional properties. The asymptotics involve known techniques, but the computation of asymptotic variances is delicate and is included in an appendix, along with all proofs. Section 5 provides an explicit discussion of gof/bof testing using our diagnostics, and Section 6 demonstrates their empirical properties through some simulation experiments and case studies. We compare the gof test to the Ljung-Box diagnostic and also provide some power results for the bof procedure.

2 Metrization of Whiteness

We make use of some basic notations in this paper. Suppose that, after suitable transformations and differencing if necessary, we have a mean zero stationary time series $X_1, X_2, \cdots, X_n$, which will sometimes be denoted by the vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)^{'}$. At the end of this subsection, we consider the case of nonstationary time series in detail. When the autocovariance function $\gamma_f(h)$ is absolutely summable, the spectral density f can be defined by

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_f(h) e^{-ih\lambda} \qquad (1)$$

with $i = \sqrt{-1}$ and $\lambda \in [-\pi, \pi]$. For a general function g that is integrable on $[-\pi, \pi]$, we define its inverse Fourier transform via

$$\gamma_g(h) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) e^{ih\lambda} \, d\lambda,$$

a relation that we will use repeatedly in the sequel. That is, γ_g and g are Fourier transform pairs. Furthermore, denote the $n \times n$ Toeplitz matrix associated with g by $\Sigma(g)$, which is defined by

$$\Sigma_{jk}(g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) e^{i(j-k)\lambda} d\lambda.$$

So if f is the spectral density of a stationary process, $\Sigma(f)$ is the associated $n \times n$ autocovariance matrix. Finally, let $\widehat{f}(\lambda)$ denote the periodogram defined on a continuum of frequencies:

$$\widehat{f}(\lambda) = \frac{1}{n} \left| \sum_{t=1}^{n} X_t e^{-it\lambda} \right|^2 = \sum_{h=1-n}^{n-1} R(h) e^{-ih\lambda} \qquad \lambda \in [-\pi, \pi],$$

with $R(h)$ equal to the sample (uncentered) autocovariance function. We adopt the notation

$$\theta_A(g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} A(\lambda) g(\lambda) \, d\lambda. \tag{2}$$

Here the function A is a fixed-bandwidth kernel that centers attention on a range of frequencies, which roughly speaking are given by the support region of A. The main gof measure that we consider in this paper is a local spectral variance of the logged spectral density, given by

$$\psi_A(f) = \theta_A(\log^2 f) - \theta_A^2(\log f), \tag{3}$$

where f is a spectral density for an invertible model (f nonzero everywhere). The empirical version of this measure is obtained by replacing f by the periodogram $\widehat{f}$, and replacing the integrals by a Riemann sum over grid points located at the Fourier frequencies. This is discussed more in Section 3; here we discuss some of the properties of (3) and its motivation.

In Peña and Rodríguez (2006) a gof measure for time series data is introduced and developed, which is based on the logarithm of the determinant of the (empirical) autocorrelation matrix. The use of this quantity for gof tests is justified in several ways in Peña and Rodríguez (2006); for one, this quantity appears in the logarithm of the Gaussian likelihood function. Note that the above authors apply this statistic to the residual autocorrelations obtained from fitting a time series model to the data, and thus their method is similar in spirit to the use of Ljung-Box statistics (Ljung and Box, 1978). Essentially, the statistic of Peña and Rodríguez (2006) can be written as

$$\widehat{D} = \frac{1}{n} \log \det \Sigma(\widehat{f}).$$

The above authors consider the case that Σ is m-dimensional, where m is a fixed integer (i.e., it does not expand with sample size in their asymptotics). Also they consider the autocovariance matrix associated with estimated residuals normalized to have unit sample variance, and thus $\widehat{f}$ would be the periodogram of such residuals. Now following the treatment in Taniguchi and Kakizawa (2000), under some conditions there exist approximations for the Toeplitz matrices $\Sigma(g)$ of the form

$$\Sigma(g) \doteq Q \, D(g) \, Q^* \tag{4}$$

with $Q_{jk} = n^{-1/2} \exp\{i2\pi jk/n\}$, and $*$ denoting the conjugate transpose. Here $D(g)$ is a diagonal matrix with entries given by $g(2\pi k/n)$ for $k = 1, \cdots, n$. Substituting $\widehat{f}$ for g in (4), we obtain

$$\frac{1}{n} \sum_{k=1}^{n} \log \widehat{f}(2\pi k/n),$$

which is the Riemann-sum approximation to the integral of the log periodogram. Generalizing this measure by introducing a local spectral weighting kernel A, we obtain

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} A(\lambda) \log \widehat{f}(\lambda)\, d\lambda. \tag{5}$$

Although such a measure looks quite a bit different from the $\widehat{D}$ of Peña and Rodríguez (2006), the above heuristics show that it is similar in spirit to their statistic.

Now, the theoretical measure associated with (5) is $\theta_A(\log f)$ as in (2), a weighted log moment of the spectrum. This is similar to the spectral moment approach of Miller and Rochwarger (1970), although that work considers integrals of polynomials multiplying the spectrum. The idea is that moments of the spectrum (or log spectrum in our case) can reveal important properties of the frequency domain representation of a time series, and thus a statistic based on this measure will be useful for assessing the gof of a particular time series model.

In Miller and Rochwarger (1970) some of the interest focuses on a spectral variance, which can be used to assess the spread (or entropy, loosely defined) in a spectrum. In a similar fashion, we will focus on the spectral variation measure $\psi_A(f)$ rather than $\theta_A(\log f)$. It is easy to see that

$$\psi_A(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\log f(\lambda) - \theta_A(\log f))^2 A(\lambda)\, d\lambda \tag{6}$$

so long as $\gamma_A(0) = 1$, i.e., the kernel A integrates to unity. In contrast, the spectral measure of Drouiche (2007) is $D(f) = \log \theta_{1_{[-\pi,\pi]}}(f) - \theta_{1_{[-\pi,\pi]}}(\log f)$. Now so long as A is non-negative, we have from (6) the following properties of $\psi_A(f)$:

1. ψ_A is non-negative.

2. $\psi_A(c\, f) = \psi_A(f)$ for any $c > 0$.

3. $\psi_A(f) = 0$ iff $f(\lambda) \propto 1$ for all $\lambda \in supp(A)$.

4. $\psi_A(f) = \psi_A(1/f)$.

So ψ_A generalizes D to local frequency bands given by $supp(A)$, the support of the kernel A; D possesses (1), (2), and (3), but not (4) in general. This latter property has the following benefit. If f is the spectrum of theoretical model residuals and $A = 1_{[-\pi,\pi]}$, we may consider that peaks and troughs in f are equally persuasive in indicating departures from whiteness; in particular, $\psi_{1_{[-\pi,\pi]}}$ gives equal measure to both f and its reciprocal, so that peaks and troughs contribute equally to the assessment of non-whiteness.

Finally, we note that the natural domain of ψ_A extends beyond the continuous functions on $[-\pi, \pi]$, since f is allowed to have a singularity or zero at frequencies outside the support of A. Thus we can extend our discussion to nonstationary time series. We say the data is homogeneously nonstationary when Y_t is nonstationary and there exists some differencing operator $\delta(B)$ such that $\delta(B)Y_t = X_t$ is stationary, where Y_t now denotes the observed data. It may be of interest to consider the spectral variation measure on an estimate of the pseudo-spectrum of Y_t rather than of the spectrum of X_t. The pseudo-spectrum is given by

$$f_\delta(\lambda) = \frac{f(\lambda)}{|\delta(e^{-i\lambda})|^2},$$

and we define the spectral variation measure by

$$\psi_A(f_\delta) = \theta_A(\log^2 f_\delta) - \theta_A^2(\log f_\delta),$$

when the poles and zeroes of f_δ lie outside *supp*(A).

3 Examples - $\psi_A(f)$ for Some Familiar Models

Recall that the basic EXP model (Bloomfield, 1973) can be defined by a Fourier expansion of the log spectrum:

$$\log f(\lambda) = \sum_j \xi_j e^{-i\lambda j}.$$

By the evenness of the spectrum, $\xi_j = \xi_{-j}$. An EXP(m) model has $\xi_j = 0$ for $j > m$; we will also consider EXP(∞) models in this section, where $\xi_j \neq 0$ for all $j.m.$ Now for such models the spectral variation measure is given by

$$\psi_A(f) = \sum_{j,l} \xi_j \xi_l \left[\gamma_A(j-l) - \gamma_A(j)\gamma_A(l)\right].$$

In the special case that $A = 1_{[-\pi,\pi]}$, we have

$$\psi_A(f) = \sum_{j \neq 0} \xi_j^2.$$

Since the ξ_0 parameter essentially corresponds to the scale of the spectral density, it makes sense (given the comments above) that ξ_0 is omitted from the summation. Below we give some simple examples of this formula.

Example 1: $EXP(1)$ Let $\log f(\lambda) = \xi_0 + \xi_1(e^{-i\lambda} + e^{i\lambda})$. Then assuming $\gamma_A(0) = 1$ we obtain

$$\psi_A(f) = 2\xi_1^2 \left(1 + \gamma_A(2) - 2\gamma_A^2(1)\right).$$

Example 2: $MA(1)$ Let $f(\lambda) = |1 - \theta e^{-i\lambda}|^2 \sigma^2$, so that

$$\log f(\lambda) = \log \sigma^2 - \sum_{j \geq 1} \frac{\theta^j}{j} (e^{i\lambda j} + e^{-i\lambda j}).$$

Hence $\xi_j = -\theta^{|j|}/|j|$ and

$$\psi_{1_{[-\pi,\pi]}}(f) = \sum_{j \neq 0} \frac{\theta^{2|j|}}{j^2} = 2 \int_0^{\theta^2} - \frac{\log(1-u)}{u} \, du.$$

It is clear that the overall spread of f increases with $|\theta|$, and $\psi_{1_{[-\pi,\pi]}}(f)$ picks this up in a fairly direct fashion, giving a measure bounded between zero and $\pi^2/3$; Figure 2 plots θ against $\psi_{1_{[-\pi,\pi]}}(f)$. Let $h(x) = 2 \int_0^x - \frac{\log(1-u)}{u} \, du$ (when $x < 0$, we have $h(x) = 2 \int_x^0 \frac{\log(1-u)}{u} \, du$). Now by property (4) of Section 2, the preceding analysis also holds for $AR(1)$ models.

Example 3: $MA(2)$ First considering the case that the MA polynomial has two real roots, we can write

$$f(\lambda) = |1 - \theta_1 e^{-i\lambda}|^2 |1 - \theta_2 e^{-i\lambda}|^2 \sigma^2$$
$$\log f(\lambda) = \log \sigma^2 - \sum_{j \geq 1} \frac{\theta_1^j + \theta_2^j}{j} (e^{i\lambda j} + e^{-i\lambda j}).$$

This defines ξ_j, and we obtain

$$\psi_{1_{[-\pi,\pi]}}(f) = \sum_{j \neq 0} \frac{(\theta_1^{|j|} + \theta_2^{|j|})^2}{j^2} = h(\theta_1^2) + 2h(\theta_1 \theta_2) + h(\theta_2^2).$$

If there are complex conjugate roots, we can write

$$f(\lambda) = |1 - \rho e^{-i(\lambda - \omega)}|^2 |1 - \rho e^{-i(\lambda + \omega)}|^2 \sigma^2.$$

By the same Taylor series techniques, we find that

$$\xi_j = -\frac{\rho^{|j|}}{|j|} 2 \cos \omega j.$$

Hence the spectral variance measure is

$$\psi_{1_{[-\pi,\pi]}}(f) = \sum_{j \neq 0} \frac{\rho^{2|j|}}{j^2} \, 4 \cos^2 \omega j.$$

It is interesting that the spectral peak/trough location parameter ω affects the variation. Since the integral is computed over all frequencies in $[-\pi, \pi]$, there is more variability when the peak/trough is in the center ($\omega = 0$) or at the end ($\omega = \pm\pi$). Also, the variation increases with ρ, which parametrizes the strength of the peak/trough.

Example 4: *ARMA* More generally, when f is the spectral density of an $ARMA(p,q)$ process we have

$$f(\lambda) = \frac{\Pi_{j=1}^{q} |1 - \theta_j e^{-i\lambda}|^2}{\Pi_{j=1}^{p} |1 - \phi_j e^{-i\lambda}|^2} \sigma^2$$

for (possibly complex) roots $1/\theta_j$ and $1/\phi_j$. Then the log spectrum is

$$\log f(\lambda) = \log \sigma^2 + \sum_{j=1}^{q} \log |1 - \theta_j e^{-i\lambda}|^2 - \sum_{j=1}^{p} \log |1 - \phi_j e^{-i\lambda}|^2.$$

At this point, we refer to the previous examples to expand the case of real roots or pairs of complex conjugate roots.

4 Statistical Properties

Although $\psi_A(\widehat{f})$ is our statistic of interest, there is some question of how to compute it, given that it involves an integral of the periodogram. The most straightforward approach – following Chiu (1988) and Taniguchi and Kakizawa (2000) – is to use a Riemann sum approximation with mesh points given by the Fourier frequencies. A delicate (and non-obvious) issue is that the asymptotic variance of any integral approximation to $\psi_A(\widehat{f})$ depends on the mesh size – see Deo and Chen (2000) for a related discussion. For coherency of treatment with the literature, we use the approximation described Section 6.4 of Taniguchi and Kakizawa (2000):

$$\theta_A(\widehat{f}) \approx \frac{1}{n} \sum_{j=-n/2}^{n/2} A(\lambda_j) \widehat{f}(\lambda_j).$$

Here we suppose that n is even (else replace $n/2$ by its greatest integer), and $\lambda_j = 2\pi j/n$ (the Fourier frequencies). The generalizations to $\theta_A(\log \widehat{f})$, etc. are obvious. We denote these Riemann sum approximation to θ_A and ψ_A via $\tilde{\theta}_A$ and $\tilde{\psi}_A$ (although these approximations depend on n, this will be suppressed in the notation).

We now present the asymptotic theory for the measure $\tilde{\psi}_A(\widehat{f})$, which is an estimate of $\psi_A(\tilde{f})$. Here $\tilde{f}$ is the true spectral density of the data, and may differ from a specified model f. Let $b(\lambda) = |\delta(e^{-i\lambda})|^2$, so that our measure is $\tilde{\psi}_A(\widehat{f}/b)$; when the data is stationary, $b = 1$. So $\tilde{f}$ denotes the true spectral density for the (stationary) differenced data $X_t = \delta(B)Y_t$. Theorem 1 gives the joint asymptotic normality result for the first and second log moments, i.e., $\tilde{\theta}_A(\log \widehat{f}/b)$ and $\tilde{\theta}_A(\log^2 \widehat{f}/b)$. The basic assumption that we use is that the data are Gaussian. Let Γ denote the gamma function, and let $\dot{\Gamma}(x)$ denote the first derivative of the gamma function at x (and so on for higher derivatives).

Theorem 1 *Suppose that $\{X_t\}$ is a stationary zero mean Gaussian time series with $\sum_k |k\gamma_X(k)| < \infty$ and spectral density bounded away from zero. Also suppose that the kernel A is bounded and*

Lipschitz continuous such that $A \log b$ is also bounded and Lipschitz. Then

$$\begin{bmatrix} \sqrt{n}\left(\tilde{\theta}_A(\log \widehat{f}/b) - \theta_A(\log \tilde{f}/b) - \dot{\Gamma}(1)\gamma_A(0)\right) \\ \sqrt{n}\left(\tilde{\theta}_A(\log^2 \widehat{f}/b) - \theta_A(\log^2 \tilde{f}/b) - 2\dot{\Gamma}(1)\theta_A(\log \tilde{f}/b) - \ddot{\Gamma}(1)\gamma_A(0)\right) \end{bmatrix}$$

is asymptotically bivariate normal with zero mean vector and variance matrix V. The entries of V are given by:

$$\begin{aligned} V_{11} &= \left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\gamma_{A^2+AA^-}(0) \\ V_{12} &= 2\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\theta_{A^2+AA^-}(\log \tilde{f}/b) + \left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\gamma_{A^2+AA^-}(0) \\ V_{22} &= 4\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\theta_{A^2+AA^-}(\log^2 \tilde{f}/b) + 4\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\theta_{A^2+AA^-}(\log \tilde{f}/b) \\ &\quad + \left(\ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)\right)\gamma_{A^2+AA^-}(0), \end{aligned}$$

where $A^-(\lambda) = A(-\lambda)$.

Corollary 1 *Under the same assumptions and notation of Theorem 1 and with $\gamma_A(0) = 1$,*

$$\begin{aligned} &\sqrt{n}\left(\tilde{\psi}_A(\widehat{f}/b) - \psi_A(\tilde{f}/b) - \ddot{\Gamma}(1) + \dot{\Gamma}^2(1)\right) \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, W) \\ &W = V_{22} - 4\left(\dot{\Gamma}(1) + \theta_A(\log \tilde{f}/b)\right)V_{12} + 4\left(\dot{\Gamma}(1) + \theta_A(\log \tilde{f}/b)\right)^2 V_{11}, \end{aligned}$$

with V_{11}, V_{12}, V_{22} as stated in Theorem 1.

5 Goodness-of-Fit and Badness-of-Fit Testing

We now focus on testing model residuals for whiteness; these residuals will be assumed to be stationary, so that $b = 1$. Also $A = 1_{[-\pi,\pi]}$, and we view f as the spectrum of the model residuals. We write ψ_1 for $\psi_{1_{[-\pi,\pi]}}$ throughout. In this case, Corollary 1 simplifies to the following:

Corollary 2 *Under the same assumptions and notation of Theorem 1 and with $A = 1_{[-\pi,\pi]}$,*

$$\begin{aligned} &\sqrt{n}\left(\tilde{\psi}_1(\widehat{f}) - \psi_1(\tilde{f}) - \ddot{\Gamma}(1) + \dot{\Gamma}^2(1)\right) \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, W) \\ &W = 8(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1))(\psi_1(\tilde{f}) + \dot{\Gamma}^2(1)) + 2(\ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)) - 8(\dot{\Gamma}(1)\dddot{\Gamma}(1) - \ddot{\Gamma}(1)\dot{\Gamma}^2(1)). \end{aligned}$$

As discussed in Section 1, gof testing seeks to reject whiteness of model residuals, which is equivalent to the Null Hypothesis that $\psi_1(\tilde{f}) = 0$ by property (3) of Section 2. So for the gof procedure, we have

$$\begin{aligned} H_0 &: \psi_1(\tilde{f}) = 0 \\ H_a &: \psi_1(\tilde{f}) > 0. \end{aligned}$$

Note that by Corollary 2, we can determine asymptotic power with only a knowledge of the value of $\psi_1(f)$; it is not necessary to know the full spectrum $\tilde{f}$, only its metrization through ψ_1. For bof testing, we instead specify a given level of non-whiteness $\mu_0 > 0$, and test the hypotheses

$$H_0 : \psi_1(\tilde{f}) = \mu_0$$
$$H_a : \psi_1(\tilde{f}) < \mu_0.$$

Here we note that the alternative is lower one-sided, in contrast to the upper one-sided gof procedure. Since by Corollary 2 the asymptotic distribution of $\tilde{\psi}_1(\hat{f})$ only depends on $\tilde{f}$ through the metrization $\psi_1(\tilde{f})$, we can compute the mean and variance under H_0. It is not clear how to adapt Ljung-Box statistics or D to bof testing, since the asymptotic distribution of these statistics under non-whiteness of the underlying process is either unknown, or is a complex function of the entire spectral density.

What is the advantage of bof over gof? As mentioned in Section 1, gof tests are used to find bad models, whereas bof tests can be used to find good models. For a gof test, we seek to reject the Null Hypothesis (that the model fit is perfect) with significant p-values. Failing to reject this Null Hypothesis corresponds to having a good model, but how high should the p-values be? 20% ? 50%? 75%? This part of the analysis becomes vague. The bof test starts with a certain assumed level of badness of model fit in the Null Hypothesis, and seeks to reject this with significance – in the direction of better model fit. Thus one can say – with an associated significant p-value – that a given model is adequate. Of course such a procedure requires that the asymptotic distribution of the test statistic only depends on f through the chosen metrization of whiteness. Moreover, the power of the procedure will understandably depend on the choice of μ_0; in fact, the power will be an increasing function of μ_0. Figure 4 gives an indication of these relationships, using the formula for bof asymptotic power given below. Let $c_1 = \ddot{\Gamma}(1) - \dot{\Gamma}^2(1)$ and

$$c_2 = 8\left(c_1\dot{\Gamma}^2(1) - \dot{\Gamma}(1)\dddot{\Gamma}(1) + \ddot{\Gamma}(1)\dot{\Gamma}^2(1)\right) + 2(\ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)).$$

Note that $c_1 = \pi^2/6$ and $c_2 \approx 23.811$. The bof asymptotic power for a δ-level test is given by

$$\Phi\left(\frac{\sqrt{n}(\mu_0 - \mu_a) + z_\delta\sqrt{8c_1\mu_0 + c_2}}{\sqrt{8c_1\mu_a + c_2}}\right), \tag{7}$$

where Φ is the standard normal cdf and $z_\alpha = \Phi^{-1}(\alpha)$. Here μ_0 and μ_a are Null and Alternative specifications of $\psi_1(f)$ (and μ_a is truth).

There is an exact relationship between the gof and bof test statistics. For a specified μ_0 (which equals zero in the gof context), we define $\hat{\psi}(\mu_0)$ to be the corresponding test statistic, which is given by

$$\hat{\psi}(\mu_0) = \sqrt{n}\,\frac{\tilde{\psi}_1(\widehat{f}) - \mu_0 - c_1}{\sqrt{8c_1\mu_0 + c_2}}.$$

Now the bof test statistic is given by the above formula when $\mu_0 > 0$, but the gof test statistic corresponds to $\hat{\psi}(0)$. These test statistics are related by the formula

$$\hat{\psi}(\mu_0) = \frac{\sqrt{c_2}\hat{\psi}(0) - \mu_0\sqrt{n}}{\sqrt{8c_1\mu_0 + c_2}}. \tag{8}$$

Thus, there is an equivalence between gof and bof, and the link is μ_0. We see from (8) that small values of $\hat{\psi}(0)$ – associated with failure to reject gof – result in negative values of $\hat{\psi}(\mu_0)$, so long as $\mu_0\sqrt{n/c_2}$ is larger than the gof test statistic. Thus, $\hat{\psi}(\mu_0)$ is significant if μ_0 is large enough, or if the sample size is large enough. In general, small gof p-values result in large bof p-values, and vice versa, so long as a μ_0 is suitably large with respect to the sample size. In fact, (8) can be used to determine the choice of the threshold μ_0 in bof testing. Letting α and δ be the significance levels of the gof and bof procedures respectively, we obtain the relation

$$z_\delta = \frac{\sqrt{c_2}z_{1-\alpha} - \mu_0\sqrt{n}}{\sqrt{8c_1\mu_0 + c_2}}. \tag{9}$$

If the bof procedure were significant at the 5% level (i.e., $\delta = .05$), then we would like the gof procedure to fail to reject with at least α probability, where now α is larger than .05 and can be selected by the user. For example, one might choose $\alpha = .50$ or $\alpha = .20$. One can easily show via (7), that for μ_0 satisfying (9), the bof power is approximately $1 - \alpha$ when the alternative is white noise (i.e., $\mu_a = 0$). In other words, given a δ-level bof test, we can choose μ_0 according to (9) to ensure approximate $1 - \alpha$ power against a white noise alternative, where α is chosen by the user. Solving (9) for μ_0 when $\delta = .05$, we obtain two roots by the quadratic formula. So long as $.05 < \alpha < .95$, the smaller root is negative, and so we let μ_0 be given by the larger root:

$$\mu_0 = \frac{\sqrt{c_2}z_{1-\alpha}}{\sqrt{n}} + \frac{4c_1z_{.05}^2}{n} + \sqrt{\frac{16c_1^2z_{.05}^4}{n^2} + \frac{8c_1\sqrt{c_2}z_{1-\alpha}z_{.05}^2}{n^{3/2}} + \frac{c_2z_{.05}^2}{n}}. \tag{10}$$

We denote this choice of μ_0 by $\mu_0(n)$ (since it depends on sample size), when we use (10) to determine the parameter. To re-iterate the property of this choice of μ_0: when the bof test statistic has p-value less than .05, the gof will have p-value greater than α percent, and the bof will have approximate power $1-\alpha$ against the white noise alternative. Since this represents a fairly intuitive relationship between bof and gof testing, we recommend using (10) for determining μ_0. An example of this relation is depicted in Figure 2 for $\alpha = .2$ and $\delta = .05$.

6 Empirical Studies

In this section we evaluate both the gof and bof measures using Monte Carlo simulation and real time series data. Although our measure is similar to the diagnostics of Peña and Rodríguez (2006), we do not make direct comparisons here. Instead we make comparisons to the Ljung-Box statistic. The reason we do not make a direct comparison to Peña and Rodríguez (2006) is four-fold.

First, we were unable to duplicate the results presented in their simulation studies. Second, even assuming the results of the simulation study they conduct for their gof diagnostic are correct, they claim to beat the performance of the Ljung-Box statistic. We acknowledge that the Ljung-Box gof diagnostic out-performs our gof diagnostic, and we present the results of a simulation study that quantifies to what extent its performance is superior. Thirdly, our diagnostic is capable of bof testing, which neither the Peña-Rodríguez nor the Ljung-Box procedures can accomplish. Finally, in Peña and Rodríguez (2006) the simulation is conducted without postulating a specific model under the Null Hypothesis, but rather their hypothesis is that the true model belongs to a certain class of models (i.e., an $ARMA(p, q)$ with fixed p and q). This formulation is incompatible with the testing paradigm that we establish.

6.1 Simulations

We determine the size and power of both our gof/bof diagnostics under several different departures from white noise residuals. First we consider the size of our gof diagnostic under the null hypothesis of white noise. Additionally, we evaluate the distribution of our test statistic in finite samples. In order to do this we performed 10,000 Gaussian simulations of various samples sizes ($n = 150, 250, 350, 500$) and calculated the mean, standard deviation and α-level of our diagnostic under a nominal α-level of $\alpha = .05$. Further, we investigated the size of the Ljung-Box under identical conditions using $m = 5$, 10, and 20 autocorrelations in the calculation of the statistic (see Table 1). Although the size of the Ljung-Box statistic is moderately better than ours, both diagnostics are fairly close to the nominal level and approach .05 as the sample size increases, as expected. Similarly the mean and standard deviation of our test statistic under the null hypothesis approach the correct mean and standard deviation as the sample size increases (see Table 1). Although it is crucial that the mean and standard deviation approach 0 and 1 respectively, it is equally important that the distribution be normal. As can be seen from Figure 1, the distribution of the gof statistic is well-approximated by the normal for sample size 500.

Next, we compare the power of our gof diagnostic with the power of the Ljung-Box statistic (at $m = 5, 10, 20$) and the turning point diagnostic for independence (see Brockwell and Davis, 1991, Pages 312-313), defined by the asymptotic distribution of the number of turning points in the series of model residuals. To assess the the performance we simulated from an $MA(1)$ data generating process with $\theta = .9$. We then computed model residuals obtained from an $MA(1)$ with θ ranging between .1 and .8. The residuals were then tested for whiteness, as discussed in Section 5. In our simulation as θ decreases from .9, the departure from whiteness in the estimated residuals increases, and it should be easier to reject the H_0. This simulation was conducted at the nominal α-level of .05 with 1000 simulations of various sample sizes (see Table 2). In general our power does not perform as well as the Ljung-Box or turning point statistics. However, one advantage of our gof diagnostic (incidentally, shared by Drouiche (2007)) over Ljung-Box is that only one test statistic is formed and only one p-value is produced. So the practitioner is freed from having to choose m when testing for model adequacy.

We now turn our attention to the bof diagnostic. For this diagnostic we evaluated the power under several formulations. In the first simulation we chose the threshold value, $\mu_0(n)$, adaptively based on the sample size using (10) with $\alpha = .2$ and $\delta = .05$. This method of choosing μ_0 induces a Null hypothesis $\psi_1(f) = \mu_0$ with an Alternative hypothesis $\psi_1(f) < \mu_0$, while simultaneously controlling the power. Further in this first power study we suppose that the model residual process follows an $MA(1)$ with parameter θ between 0 and approximately .42. We simulated Gaussian $MA(1)$ processes with $\theta = .4, .3, .2, .1, 0$ for various sample sizes between $n = 100$ and $n = 500$ and determined power using 10,000 Monte Carlo replications and nominal level equal to .05; see Table 3. As a result of choosing μ_0 via (10) we find that the power is rather good even for sample sizes as small as $n = 100$. In fact, we see that the power of our test remains fairly constant across sample size and is approximately equal to .8 for a white noise alternative. Here it is $\mu_0(n)$ that changes with sample size. Specifically, $\mu_0(n)$ is a decreasing function of n, and as such when the sample size increases so does our assurance that rejecting the Null hypothesis of a "bad" model in the direction of a better fitting model actually results "good" model and not just a model superior to that postulated under the Null. Additionally, our simulations confirm that the power under a white noise alternative turns out to be approximately $1 - \alpha$. Simulations for several other values of α confirm this result.

In our second power study we mapped the different values of μ_0 into equivalent $MA(1)$ processes; if θ is the MA parameter, then $\mu_0 = 2\sum_{j=1}^{\infty} \theta^{2j}/j^2$ as shown in Example 2 of Section 3. For a graph of this mapping see Figure 2. Thus here we are keeping μ_0 fixed across different sample sizes. In the first power study, under μ_0 "fixed", we suppose that the model residual process follows an $MA(1)$ with parameter θ between 0 and .66, which corresponds to $\psi_1(f)$ values between 0 and 1. Our Null Hypothesis states that $\psi_1(f) = 1$, and the Alternative Hypothesis states that $\psi_1(f) < 1$, or equivalently that $\theta < .66$. Again, we simulated Gaussian $MA(1)$ processes with $\theta = .4, .3, .2, .1, 0$, with three different sample sizes, and determined the power using $10,000$ Monte Carlo replications and an α level of .05. In the second study, for "fixed" μ_0, we now let $\mu_0 = .5$, which for an $MA(1)$ corresponds to $\theta = .48$. So we simulated Gaussian $MA(1)$ processes with $\theta = .4, .3, .2, .1, 0$ with three different sample sizes, and determined the power. The results are reported in Table 4. Even though, in this case, the power depends on what size departures one is willing to accept under the Null Hypothesis, this is still a very sensible way to test for model inadequacy. This flexibility allows the practitioner the control of deciding what degree of "badness" is acceptable for a given application. If μ_0 is chosen equal to .5 a priori, large samples are needed for high power if one only wishes to consider slight departures from whiteness. On the other hand, for $\mu_0 = 1$ a priori fair power is achieved for relatively moderate sample sizes.

6.2 Case Studies

Next, we consider the diagnostics on several time series: *m00110*, *m00100*, *France*, and *Shoe*. The first two time series are from the Foreign Trade Division of the U.S. Census Bureau; the first series is Imports of Meat Products, and the second series is Imports of Dairy Products and Eggs. Both of these series are for the time period from January 1989 to December 2003. The *France* series refers to the sales volume for Grands Magasins produced by the Chamber of Commerce and of Industry of Paris (CCIP), from January 1990 through March 2004. The *Shoe* series is U.S. Retail Sales of Shoe Stores data from the monthly Retail Trade Survey of the Census Bureau, from 1984 to 1998.

In order to illustrate our diagnostics usefulness in practice we fit models to the data using the "automodl" command of the 2007 update (version 0.3) of X-12-ARIMA, which follows closely the automatic modeling procedure of Gómez and Maravall (2001). We adjusted for regression effects (such as outliers and trading day) when applicable. Next we obtained the estimated residuals from the fitted model and calculated our gof and bof diagnostic tests using $\mu_0(n)$ with $\alpha = .2$ and $\delta = .05$. Below, n is the effective number of observations (the sample size of the differenced series). For comparison we constructed a time series plot and acf plot of the residuals along with the p-values through lag 20 for the Ljung-Box statistic. This was done using the "tsdiag" command in R (R Development Core Team, 2005).

The first series we consider is the *Shoe* series ($n = 157$). The automodl procedure of X-12-ARIMA provided the following model:

$$(1-B)(1-B^{12})X_t = (1-.572B)(1-.336B^{12})\varepsilon_t.$$

The p-value from our gof diagnostic was .489, while the p-values for our bof diagnostic with $\mu_0(n) = 1.146$ was .011. This can be contrasted with various other measures constructed from the residuals (see Figure 5). It appears that the acf plot of the residuals and Ljung-Box statistics seem to indicate an adequate fit. In this case our results agree with this assessment.

The next series we consider is the *France* series ($n = 158$). The automodl procedure of X-12-ARIMA obtained the following model for the log transformed data:

$$(1-B)(1-B^{12})X_t = (1-1.0382B+.3381B^2)\varepsilon_t.$$

The p-value from our goodness-of-fit diagnostic was .927, while the p-values for bof with $\mu_0(n) = 1.142$ was $< .001$. Next we compared our analysis with the acf plot of the residuals and the Ljung-Box statistic out to lag 20 (see Figure 6). In this case our results corroborate the results of the acf plot and the Ljung-Box test, namely that the model is very good.

We next focus our attention on the *m00100* series ($n = 167$). The automodl feature of X-12-ARIMA fitted the following model to the log transformed data:

$$(1-B)(1-B^{12})X_t = (1-.8390B^{12})\varepsilon_t.$$

The p-value from our gof diagnostic was .134, while the p-values for bof with $\mu_0(n) = 1.106$ was .075. We can compare this with the time series plot and acf of the residuals along with the Ljung-Box out to lag 20 (see Figure 7). This example illustrates an instance when using Ljung-Box and acf plots is somewhat ambiguous. When looking at the Ljung-Box p-values, the decision rule is dependent on which lag is being considered. Moreover, it requires the practitioner to make a choice as to what constitutes a high p-value.

Finally, we examine the *m00110* series ($n = 167$). The automodl procedure of X-12-ARIMA obtained the following model for the log transformed data:

$$(1 - .5907B)(1 - B)(1 - B^{12})X_t = (1 - .9380B)(1 - .9377B^{12})\varepsilon_t.$$

The p-value from our gof diagnostic was .541, while the p-values for our bof diagnostics with $\mu_0(n) = 1.106$ was .008. This series illustrates an instance where the determination of model adequacy is less definitive. Specifically, the decision to accept a model is markedly different, depending on whether the number of correlations used in the Ljung-Box test are less than or greater then 14 (see Figure 8). However, our bof statistic avoids this complication. The practitioner can use these diagnostics to assess what type of departure from whiteness is deemed acceptable, and take action accordingly.

6.3 Discussion

This paper introduces the concept of badness-of-fit testing for time series modeling, using a convenient metrization of whiteness. We explicitly demonstrate how gof and bof testing can be implemented, and the relationship between them; we also describe how the crucial μ_0 parameter can be chosen in an intuitive fashion. Our method is illustrated on four economic time series, and the results are compatible with the information from acf and pacf plots.

A potential criticism of bof testing raises the question of the choice of μ_0. Given a significant rejection of H_0 in the bof test, can we be assured that the residuals are truly white? How close to white are they then, and are they "close enough"? However, similar questions could be leveled at gof testing: given that whiteness is rejected with significance, can it be that model residuals are still close enough to being white noise, such that the model fit might be deemed decent – perhaps through some other assessment such as out-of-sample forecasting performance? We have attempted to resolve the issue of the choice of μ_0 as follows: one chooses an α such that the asymptotic power of the bof test is $1 - \alpha$ for a white noise alternative, so long as μ_0 is chosen according to (10). In our simulations and case studies we have taken $\alpha = .2$, since this gives high power (about 80%) against the white noise alternative. Clearly, other choices are available.

Graphs such as Figure 3 give an indication of the relationship between $\psi_1(f)$ and flatness of the spectrum for $MA(1)$ models. Furthermore, the four case studies presented illustrate that low p-values (the *France* series) indeed correspond to white residuals, whereas moderate p-values (the *Shoe* and *m00110* series) still indicate residuals that are close to being uncorrelated – witness the plots of standardized residuals in Figures 5 and 8. The high p-value for the *m00100* series seems to indicate that the residuals are too far from whiteness; now one should flip things around, and do gof testing to reject the given model (this is a borderline case, because the gof p-value is .134 and only one Ljung-Box statistic is significant at the 5% level).

In practice, plots of residual ACFs and standardized residuals will be helpful in determining model adequacy; our bof testing procedure allows for quantization of this concept via an appropriate metrization of whiteness, allowing one to "accept" fitted models with statistical significance.

Appendix

Proof of Theorem 1. We use the notation for Riemann sums introduced in the beginning of Section 2.3. We wish to consider the convergence, for any real numbers a and c, of

$$\begin{aligned} &\frac{1}{n}\sum_{j=-n/2}^{n/2} A(\lambda_j)\left(a\log\widehat{f}(\lambda_j)/b(\lambda_j) + c\log^2\widehat{f}(\lambda_j)/b(\lambda_j)\right) \qquad \text{(A.1)} \\ &= \frac{1}{n}\sum_{j=-n/2}^{n/2} A(\lambda_j)\left((a-2c\log b(\lambda_j))\log\widehat{f}(\lambda_j) + c\log^2\widehat{f}(\lambda_j)\right) \\ &- \frac{1}{n}\sum_{j=-n/2}^{n/2} A(\lambda_j)\left(a\log b(\lambda_j) - c\log^2 b(\lambda_j)\right). \end{aligned}$$

The second term is deterministic, and thus contributes to the mean; it is asymptotic to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} A(\lambda)\left(a\log b(\lambda) - c\log^2 b(\lambda)\right)\, d\lambda.$$

The first term can be written as

$$\frac{1}{n}\sum_{j=-n/2}^{n/2}\left(A(\lambda_j)\zeta(\widehat{f}(\lambda_j)) + A(\lambda_j)\log b(\lambda_j)\xi(\widehat{f}(\lambda_j))\right),$$

where $\zeta(x) = a\log x + c\log^2 x$ and $\xi(x) = -2c\log x$. The convergence of this type of functional is obtained by a straightforward generalization of Theorem 6.4.3 of Taniguchi and Kakizawa (2000), under the conditions that both A and $A\log b$ are bounded and Lipschitz (the generalization involves extending the limit theorem to functionals of type $A(\lambda)\zeta(\hat{f}) + B(\lambda)\xi(\hat{f})$; Theorem 6.4.3 applies to each summand separately). Then the asymptotic mean is

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} A(\lambda)\left(\int_0^\infty \zeta(\tilde{f}(\lambda)r)e^{-r}\, dr + \log b(\lambda)\int_0^\infty \xi(\tilde{f}(\lambda)r)e^{-r}\, dr\right) d\lambda.$$

Applying this, we see that the asymptotic mean of (A.1) is given by

$$a\left(\theta_A(\log\tilde{f}/b) + \gamma_A(0)\dot{\Gamma}(1)\right) + c\left(\theta_A(\log^2\tilde{f}/b) + 2\theta_A(\log\tilde{f}/b)\dot{\Gamma}(1) + \gamma_A(0)\ddot{\Gamma}(1)\right).$$

Taking $a = 1, c = 0$ and $a = 0, c = 1$ respectively gives the means stated in the theorem. Letting

$$\begin{aligned}
v_1(\lambda) &= \int_0^\infty \zeta^2(\tilde{f}(\lambda)r)e^{-r}\,dr - \left(\int_0^\infty \zeta(\tilde{f}(\lambda)r)e^{-r}\,dr\right)^2 \\
v_2(\lambda) &= \int_0^\infty \zeta(\tilde{f}(\lambda)r)\xi(\tilde{f}(\lambda)r)e^{-r}\,dr - \left(\int_0^\infty \zeta(\tilde{f}(\lambda)r)e^{-r}\,dr\right)\left(\int_0^\infty \xi(\tilde{f}(\lambda)r)e^{-r}\,dr\right) \\
v_3(\lambda) &= \int_0^\infty \xi^2(\tilde{f}(\lambda)r)e^{-r}\,dr - \left(\int_0^\infty \xi(\tilde{f}(\lambda)r)e^{-r}\,dr\right)^2,
\end{aligned}$$

the asymptotic variance is given by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[A^2(\lambda) + A(\lambda)A(-\lambda)\right]\left(v_1(\lambda) + 2\log b(\lambda)v_2(\lambda) + \log^2 b(\lambda)v_3(\lambda)\right)\,d\lambda.$$

We next compute these variance functions:

$$\begin{aligned}
v_1(\lambda) &= a^2\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right) \\
&\quad + 2ac\left(2\log\tilde{f}(\lambda)\ddot{\Gamma}(1) - 2\log\tilde{f}(\lambda)\dot{\Gamma}^2(1) + \dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right) \\
&\quad + c^2\left(4\log^2\tilde{f}(\lambda)\ddot{\Gamma}(1) - 4\log^2\tilde{f}(\lambda)\dot{\Gamma}^2(1) + 4\log\tilde{f}(\lambda)\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right) + \ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)\right) \\
v_2(\lambda) &= 2ac\left(\dot{\Gamma}^2(1) - \ddot{\Gamma}(1)\right) + 2c^2\left(2\log\tilde{f}(\lambda)\left(\dot{\Gamma}^2(1) - \ddot{\Gamma}(1)\right) + \dot{\Gamma}(1)\ddot{\Gamma}(1) - \dddot{\Gamma}(1)\right) \\
v_3(\lambda) &= 4c^2\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right).
\end{aligned}$$

It follows that the asymptotic variance of (A.1) is

$$\begin{aligned}
&a^2\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\gamma_{A^2+AA^-}(0) \\
&+ 2ac\left(2\theta_{A^2+AA^-}(\log\tilde{f})\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right) + \left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\gamma_{A^2+AA^-}(0)\right) \\
&+ c^2(4\theta_{A^2+AA^-}(\log^2\tilde{f})\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right) + 4\theta_{A^2+AA^-}(\log\tilde{f})\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right) \\
&+ \left(\ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)\right)\gamma_{A^2+AA^-}(0)) \\
&- 4ac\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\gamma_{(A^2+AA^-)\log b}(0) \\
&- 4c^2\left(\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\gamma_{(A^2+AA^-)\log b}(0) + 2\theta_{(A^2+AA^-)\log b}(\log\tilde{f})\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\right) \\
&+ 4c^2\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\gamma_{(A^2+AA^-)\log^2 b}(0).
\end{aligned}$$

Finally, setting $a = 1$ and $c = 0$ yields the asymptotic variance V_{11} of the first log moment, while $a = 0$ and $c = 1$ yields V_{22}, the second log moment. If we set $a = 1 = c$, then we should subtract off $V_{11} + V_{22}$ from the resulting quantity, which yields $2V_{12}$. In this way the asymptotic covariance matrix V is obtained, and the log moments are asymptotically normal with the indicated mean and covariance matrix V. □

Proof of Corollary 1. We first observe that

$$\begin{aligned}
&\tilde{\theta}_A(\log^2 \widehat{f}/b) - \tilde{\theta}_A^2(\log \widehat{f}/b) - \theta_A(\log^2 \tilde{f}/b) + \theta_A^2(\log \tilde{f}/b) - \ddot{\Gamma}(1) + \dot{\Gamma}^2(1) \\
&= \left(\tilde{\theta}_A(\log^2 \widehat{f}/b) - \theta_A(\log^2 \tilde{f}/b) - 2\dot{\Gamma}(1)\theta_A(\log \tilde{f}/b) - \ddot{\Gamma}(1)\right) \\
&- \left(\tilde{\theta}_A(\log \widehat{f}/b) - \theta_A(\log \tilde{f}/b) - \dot{\Gamma}(1)\right) \left(\tilde{\theta}_A(\log \widehat{f}/b) + \theta_A(\log \tilde{f}/b) + \dot{\Gamma}(1)\right) .
\end{aligned}$$

Now since $\tilde{\theta}_A(\log \widehat{f}/b) \xrightarrow{P} \theta_A(\log \tilde{f}/b) + \dot{\Gamma}(1)$, we use Theorem 1 to deduce that

$$\sqrt{n} \left(\tilde{\psi}_A(\log \widehat{f}/b) - \psi_A(\log \tilde{f}/b) - \ddot{\Gamma}(1) + \dot{\Gamma}^2(1)\right) \overset{\mathcal{L}}{\Longrightarrow} G_2 - 2 \left(\theta_A(\log \tilde{f}/b) + \dot{\Gamma}(1)\right) G_1,$$

where $[G_1, G_2]'$ is bivariate normal with covariance matrix V from Theorem 1. Hence the limiting variance is W, as given in the statement of the Corollary 1. □

Proof of Corollary 2. Using the abbreviation $B = A^2 + A\,A^-$, we can simplify W:

$$\begin{aligned}
W &= 4\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right) \theta_B(\log^2 \widetilde{f}/b) + 4\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right) \theta_B(\log \widetilde{f}/b) + \left(\ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)\right) \gamma_B(0) \\
&- 4\left(\dot{\Gamma}(1) + \theta_A(\log \widetilde{f}/b)\right) \left(2(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1))\theta_B(\log \widetilde{f}/b) + (\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1))\gamma_B(0)\right) \\
&+ 4\left(\dot{\Gamma}(1) + \theta_A(\log \widetilde{f}/b)\right)^2 (\ddot{\Gamma}(1) - \dot{\Gamma}^2(1))\gamma_B(0).
\end{aligned}$$

Now letting $A = 1_{[-\pi,\pi]}$ yields the stated value for W after some simplifications. □

References

[1] Beran, J. (1994) *Statistics for Long Memory Processes.* New York: Chapman and Hall.

[2] Bloomfield, P. (1973) An exponential model for the spectrum of a scalar time series. *Biometrika* **60**, 217-226.

[3] Brockwell, P. and Davis, R. (1991) *Time Series: Theory and Methods* New York: Springer-Verlag.

[4] Chen, W. and Deo, R. (2000) On the integral of the squared periodogram. *Stochastic Processes and Their Applications* **85**, 159–176.

[5] Chen, W. and Deo, R. (2004) A generalized portmanteau goodness-of-fit test for time series models. *Econometric Theory* **20**, 382 – 416. 5

[6] Drouiche, K. (2007) A test for spectrum flatness. *Journal of Time Series Analysis*, forthcoming.

[7] Gómez, V. and Maravall, A. (2001) “Automatic modeling methods for univariate time series,” in *A Course in Time Series*, eds. Pẽna, D., Tiao, G., and Tsay, R., 171–201, New York: John Wiley & Sons.

[8] Li, W. (2004) *Diagnostic Checks in Time Series.* CRC Press.

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

[10] McElroy, T. and Holan, S. (2006) A spectral approach for locally assessing time series model misspecification. SRD Research Report No. 2006/12, U.S. Census Bureau.
http://www.census.gov/srd/papers/pdf/rrs2006-12.pdf

[11] Miller, K. and Rochwarger, M. (1970) Estimation of spectral moments of time series. *Biometrika* **57**, 513–517.

[12] Paparoditis, E. (2000) Spectral density based goodness-of-fit tests for time series models. *Scandinavian Journal of Statistics* **27**, 143 – 176.

[13] Peña, D. and Rodríguez, J. (2002) A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association* **97**, 601–610.

[14] Peña, D. and Rodríguez, J. (2003) Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis* **85**, 361–374.

[15] Peña, D. and Rodríguez, J. (2006) The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series. *Journal of Statistical Planning and Inference* **136**, 2706–2718.

[16] R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

	ψ_1			**Ljung-Box: α-level**		
n	Mean	Std.	α-level	$m = 10$	$m = 15$	$m = 20$
150	.081	1.091	.0775	.0541	.0605	.0646
250	.060	1.056	.0725	.0555	.0597	.0652
350	.042	1.033	.0697	.0528	.0559	.0592
500	.026	1.032	.0666	.0493	.0554	.0588

Table 1: Comparison of level for goodness-of-fit diagnostic versus Ljung-Box. Note the number of Monte Carlo simulations was 10,000, at a nominal α-level of .05. Distributional results are presented in the form of mean and standard deviation.

$n = 150$	ψ_1	LB $m = 5$	LB $m = 10$	LB $m = 20$	Turning Point
.8	.102	.106	.112	.117	.106
.7	.193	.241	.203	.206	.235
.6	.306	.485	.417	.377	.467
.5	.470	.705	.598	.554	.662
.4	.675	.899	.793	.735	.834
.3	.820	.981	.915	.884	.940
.2	.921	.999	.987	.968	.981
.1	.961	1	.997	.990	.996
$n = 250$	ψ_1	LB $m = 5$	LB $m = 10$	LB $m = 20$	Turning Point
.8	.115	.173	.169	.156	.154
.7	.228	.494	.398	.373	.391
.6	.413	.828	.731	.665	.678
.5	.607	.978	.925	.874	.913
.4	.849	.998	.993	.977	.972
.3	.948	1	.999	.996	.999
.2	.987	1	1	1	1
.1	.998	1	1	1	1
$n = 350$	ψ_1	LB $m = 5$	LB $m = 10$	LB $m = 20$	Turning Point
.8	.116	.260	.236	.204	.195
.7	.275	.731	.593	.526	.518
.6	.469	.986	.944	.849	.843
.5	.779	1	1	.991	.982
.4	.935	1	1	1	.998
.3	.997	1	1	1	1
.2	.999	1	1	1	1
.1	1	1	1	1	1

Table 2: This table compares the power of our gof diagnostic (ψ_1) with the Ljung-Box (LB) and Turning Point diagnostics. Note the number of Monte Carlo simulations was 1000, at a nominal α-level of .05.

Power with μ_0 as a function of sample size - $n(\mu_0)$								
H_a: θ	100 (1.495)	150 (1.177)	200 (.997)	250 (.878)	300 (.793)	350 (.728)	400 (.676)	500 (.598)
.4	.5733	.5160	.4705	.4286	.3837	.3571	.3111	.2636
.3	.6826	.6610	.6223	.6100	.5888	.5550	.5395	.5070
.2	.7496	.7283	.7332	.7192	.7160	.6980	.6971	.6854
.1	.7878	.7877	.7799	.7736	.7801	.7767	.7814	.7768
WN	.8025	.8016	.7919	.7987	.7974	.7944	.7992	.7951

Table 3: This table examines the power of the bof diagnostic under several different departures from white noise. Note the number of Monte Carlo simulations was 10,000, at a nominal α-level of .05. Note that μ_0 is determined as a function of sample size with $\alpha = .2$ and $\delta = .05$.

	H_0: $\theta \approx .66$ or $\mu_0 = 1$			H_0: $\theta \approx .48$ or $\mu_0 = .5$		
H_a: θ	$n = 150$	$n = 250$	$n = 350$	$n = 150$	$n = 250$	$n = 350$
.4	.3803	.5475	.6724	.0648	.0945	.1167
.3	.5200	.7122	.8359	.1271	.2081	.2694
.2	.6231	.8119	.9025	.2066	.3178	.4090
.1	.6684	.8524	.9318	.2653	.4002	.5022
WN	.7011	.8654	.9400	.2836	.4297	.5483

Table 4: This table examines the power of the bof diagnostic under several different departures from white noise. Note the number of Monte Carlo simulations was 10,000, at a nominal α-level of .05.

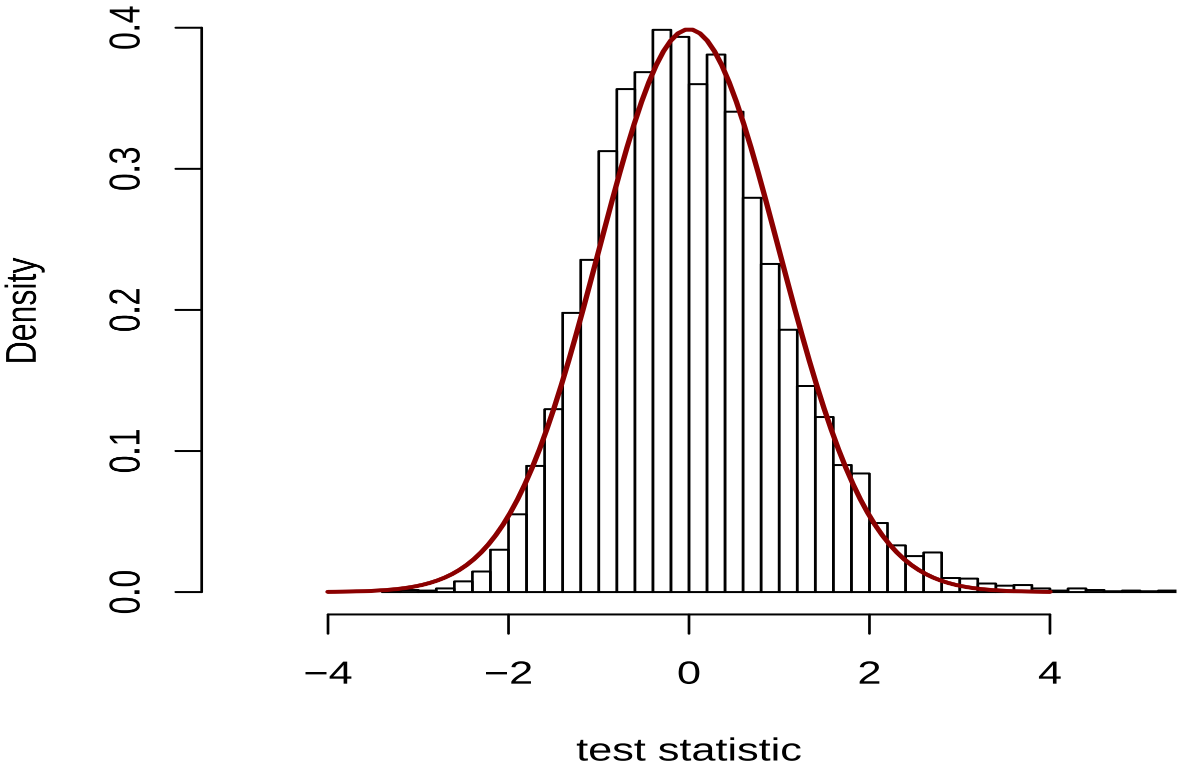

Figure 1: This figure contains a histogram for the distribution of the gof statistic from a simulation with 10,000 repetitions of sample size 500, under a white noise null hypothesis. Note the theoretical Normal(0,1) pdf is superimposed for convenience.

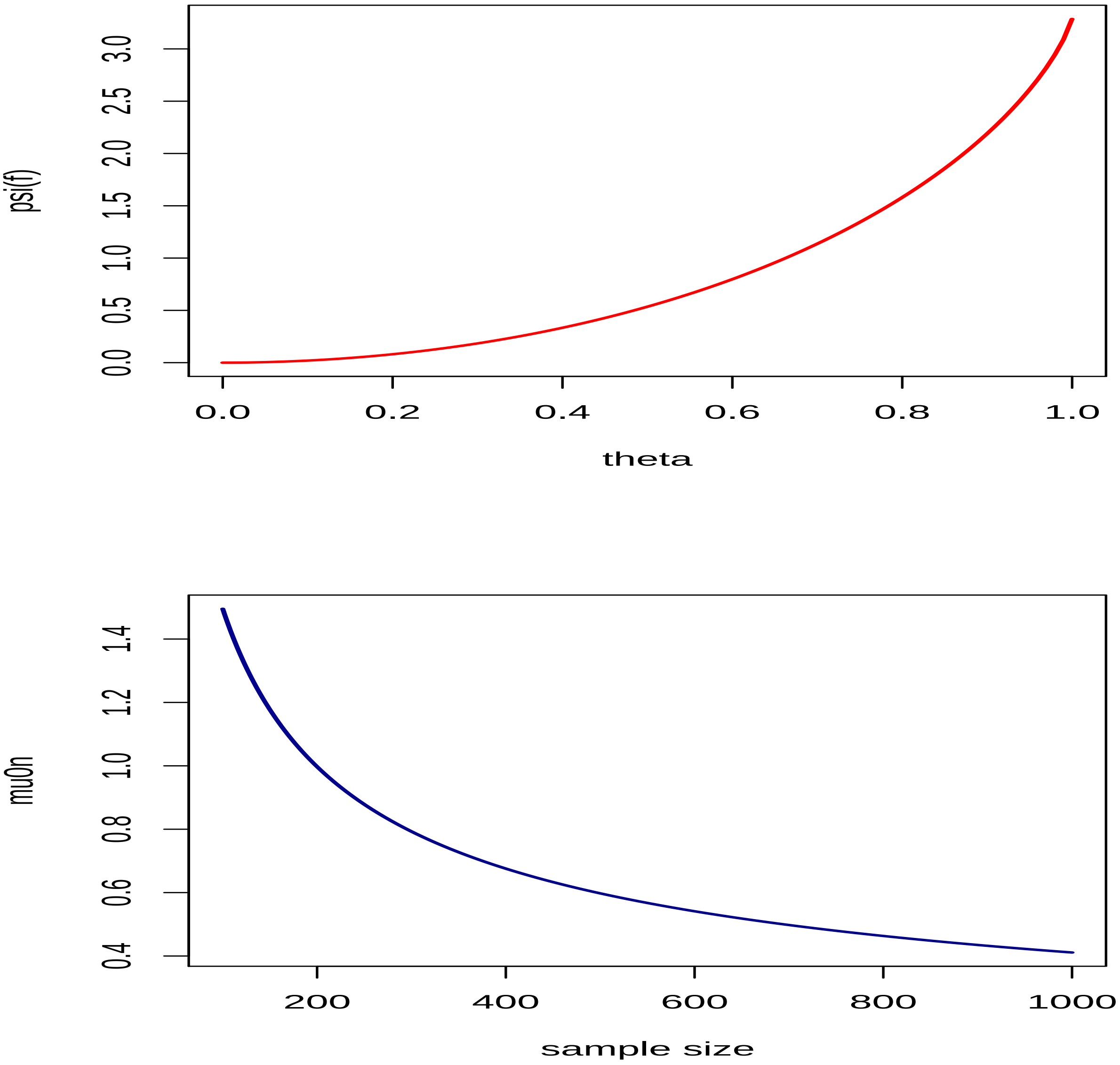

Figure 2: This top panel contains a graph of θ vs. $\psi_1(f)$. Using this graph one can find equivalent $MA(1)$ processes associated with different values of $\psi_1(f)$. The bottom panel contains a plot of μ_0 as a function of sample size such that $\alpha = .2$ and $\delta = .05$.

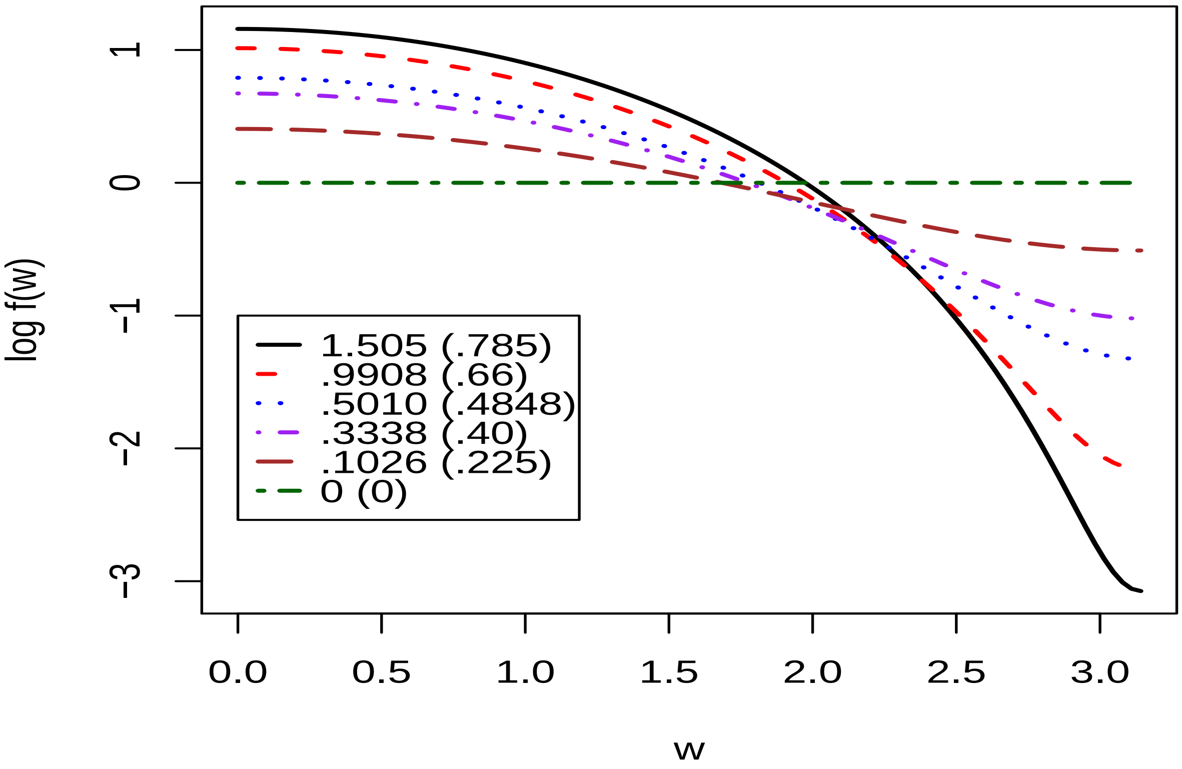

Figure 3: This figure contains a graph of the theoretical log spectral density for several values of μ_0. Further we display the parameters μ_0, along with their associated $MA(1)$ parameters in parenthesis.

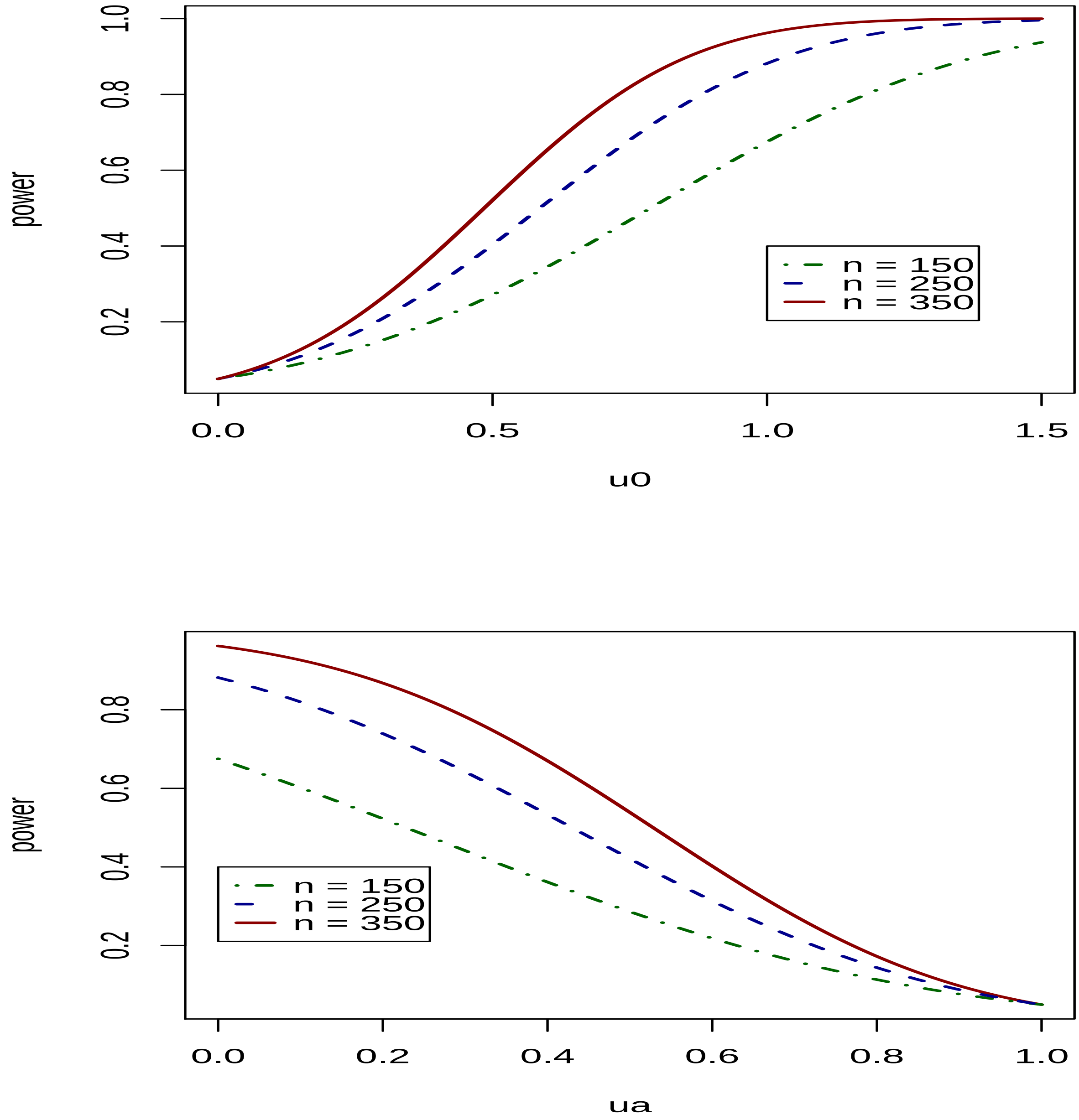

Figure 4: The top panel of this figure contains a graph of the maximum theoretical power for $\mu_0 \in [0, 1.5]$ (i.e., $\mu_a = 0$). The bottom panel displays the theoretical power for different values of μ_a when μ_0 is held fixed and equal to 1.

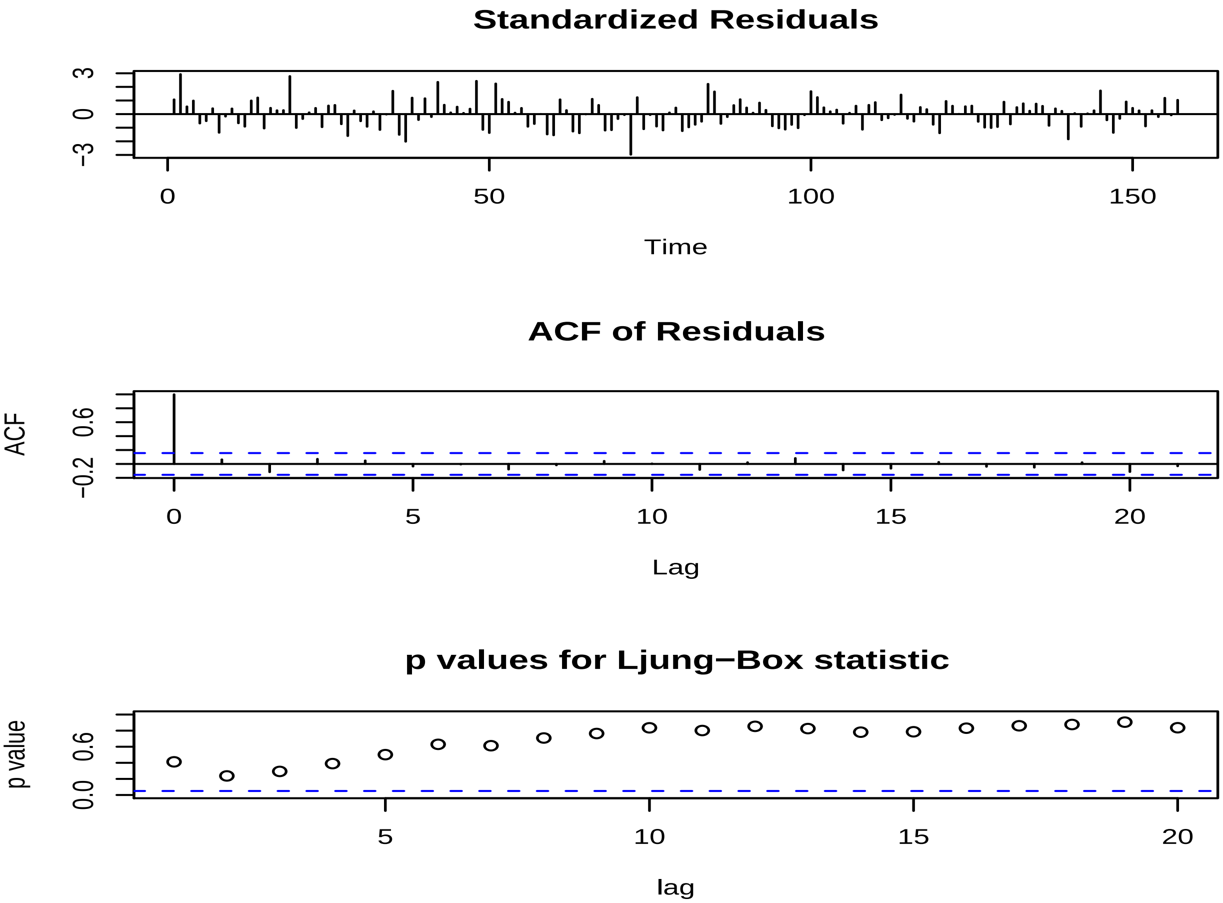

Figure 5: This figure contains a time series plot of the residuals from the model fit to the *Shoe* data using the auto-model feature in X-12-ARIMA, along with a plot of the acf of the residuals. Finally, the p-values of the Ljung-Box statistic are plotted for lags up to 20.

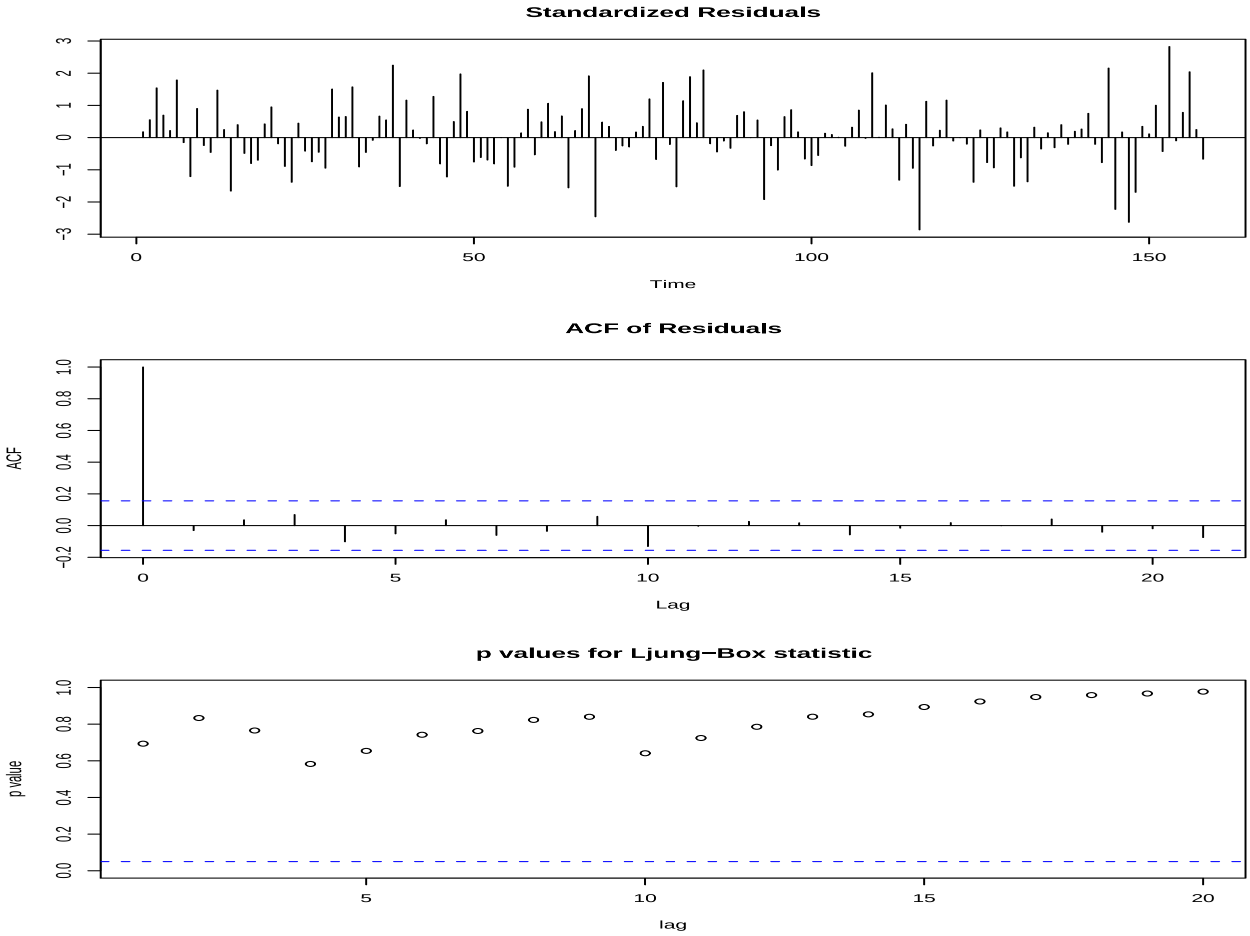

Figure 6: This figure contains a time series plot of the residuals from the model fit to the *France* data using the auto-model feature in X-12-ARIMA, along with a plot of the acf of the residuals. Finally, the p-values of the Ljung-Box statistic are plotted for lags up to 20.

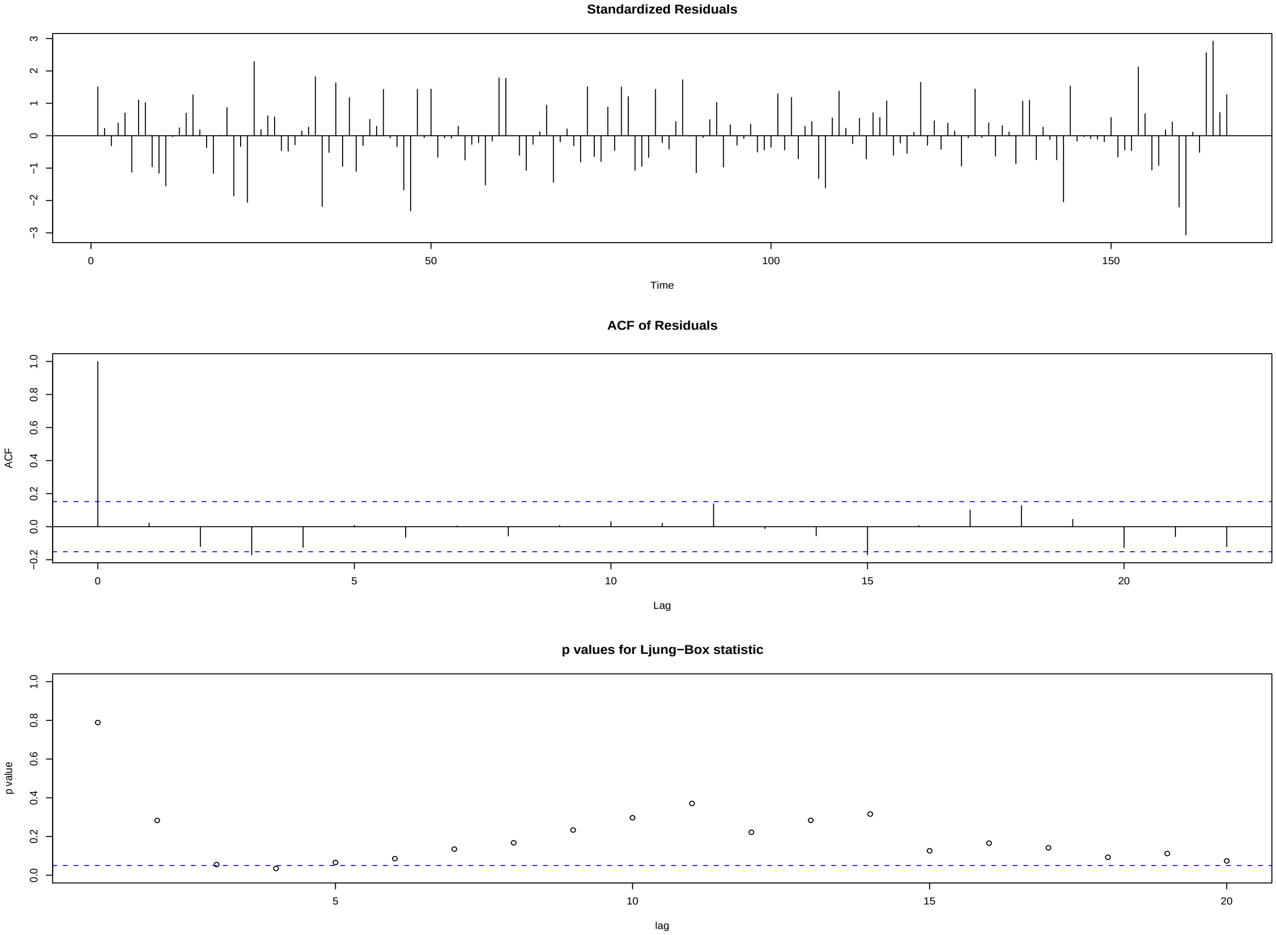

Figure 7: This figure contains a time series plot of the residuals from the model fit to the *m00100* data using the auto-model feature in X-12-ARIMA, along with a plot of the acf of the residuals. Finally, the p-values of the Ljung-Box statistic are plotted for lags up to 20.

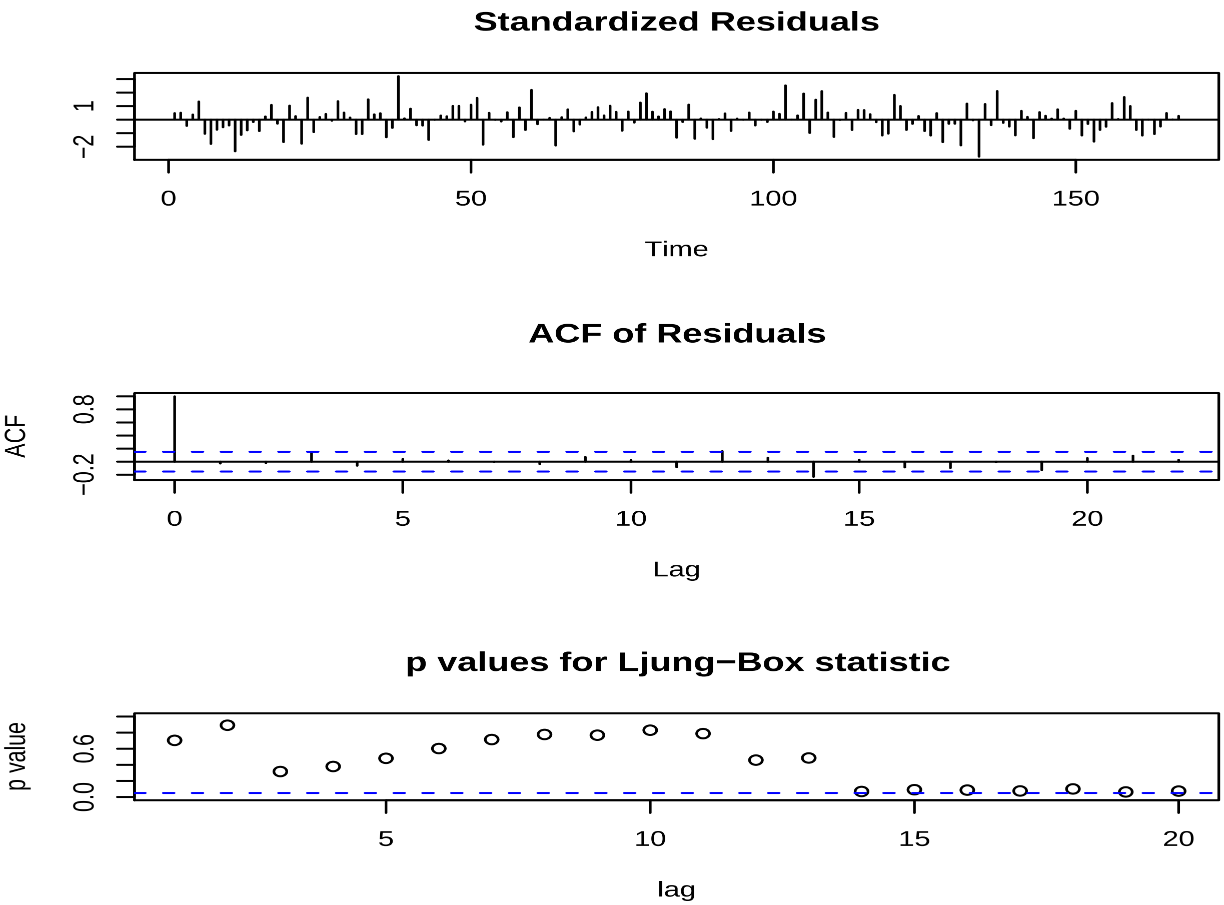

Figure 8: This figure contains a time series plot of the residuals from the model fit to the *m00110* data using the auto-model feature in X-12-ARIMA along with a plot of the acf of the residuals. Finally, the p-values of the Ljung-Box statistic are plotted for lags up to 20.