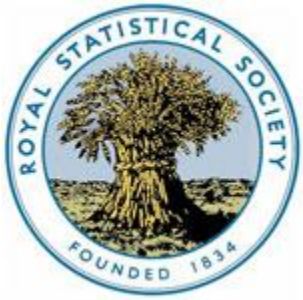

Leave-k-Out Diagnostics for Time Series
Author(s): Andrew G. Bruce and R. Douglas Martin
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 51, No. 3 (1989), pp. 363-424
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2345449
Accessed: 24-11-2015 09:52 UTC

J. R. Statist. Soc. B (1989)
51, *No.* 3, *pp.* 363–424

Leave-*k*-out Diagnostics for Time Series

By ANDREW G. BRUCE and R. DOUGLAS MARTIN†

University of Washington, Seattle, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 8th, 1989*, Professor A. P. Dawid *in the Chair*]

SUMMARY

We propose diagnostics for autoregressive integrated moving average (ARIMA) model fitting for time series formed by deleting observations from the data and measuring the change in the estimates of the parameters. The use of leave-one-out diagnostics is a well-established tool in regression analysis. We demonstrate the efficacy of observation-deletion-based diagnostics for ARIMA models, addressing issues special to the time series setting. It is shown that the dependency aspect of time series data gives rise to a 'smearing' effect, which confounds the diagnostics for the coefficients. It is also shown that the diagnostics based on the innovations variance are much clearer and more sensitive than those for the coefficients. A 'leave-*k*-out' diagnostics approach is proposed to deal with patches of outliers, and problems caused by 'masking' are handled by use of iterative deletion. An overall strategy for ARIMA model fitting is given and applied to two data sets.

Keywords: AUTOREGRESSIVE MOVING AVERAGE MODELS; DIAGNOSTICS; INFLUENCE; MISSING DATA; OUTLIERS; TIME SERIES

1. INTRODUCTION

Regression diagnostics are becoming a well-accepted tool in the practice of statistics. This is evidenced not only by books devoted to the subject (e.g. Belsley *et al.* (1980), Cook and Weisberg (1982) and Atkinson (1985) but also by the penetration of the concepts into standard texts on regression (e.g. Weisberg (1980)) and the increasingly widespread availability of software for computing the diagnostics. We also see the basic leave-one-out diagnostic idea for linear regression being carried over to somewhat more complicated settings such as logistic regression (Pregibon, 1981) and Cox regression (Storer and Crowley, 1985).

However, the literature appears to be relatively devoid of analogous results in the time series setting, in spite of a rather obvious way to obtain leave-one-out diagnostics in the context of autoregressive integrated moving average (ARIMA) model fitting for time series: one deletes a single observation at a time, and for each deletion computes a Gaussian maximum likelihood estimate (MLE) for missing data (see, for example, Jones (1980), Harvey and Pierse (1984) and Kohn and Ansley (1986)). The use of Gaussian MLEs for missing data entails intuitively appealing use of predictions in place of missing data. A diagnostic display is obtained by comparing the leave-one-out MLEs with the Gaussian MLEs for the full data set versus time, on an appropriate comparison scale. This idea was articulated by Brillinger (1966), but only the advent of powerful computers and algorithms for fitting autoregressive moving

†*Address for correspondence*: Department of Statistics, GN-22, University of Washington, Seattle, WA 98195, USA.

 0035–9246/89/51363 $2.00

average (ARMA) and ARIMA models with missing data has placed actual use of the procedure within reach.

In this paper we demonstrate the efficacy of observation deletion diagnostics for time series, addressing in the process some issues which are special to the time series setting. In particular, we consider not only diagnostics based on ARIMA model coefficients but also diagnostics based on the innovations variance. We show that the time series problem gives rise to a 'smearing' effect which is not encountered in the usual independent observation setting. For diagnostics based on coefficients, this smearing can result in considerable ambiguity concerning the numbers and locations of outliers. By both examples and theoretical calculations, we show that diagnostics based on the innovations variance are far superior to coefficient-based diagnostics in this regard.

Furthermore, outliers and other influential observations frequently occur in patches in the time series setting. Thus we propose a 'leave-k-out' diagnostic approach which is both effective and within computational reach.

The paper is organized as follows. Section 2 presents the basic leave-k-out diagnostic, based on the coefficients and innovations variance, including a proposal for scaling. Some artificial examples are given in Section 3 which illustrate that the innovations variance is a better diagnostic tool. Analytical results on smearing effects associated with leave-k-out diagnostics are also presented. Section 4 presents a strategy for detecting influential patches. An iterative deletion procedure is given to overcome problems caused by 'masking'. Techniques are also discussed for handling other types of disturbance, such as level shifts and variance changes. Finally, we give an overall strategy for ARIMA model identification and fitting using the leave-k-out diagnostics. This strategy is applied in Section 5 to two real data sets. Section 6 yields some insight into the diagnostic based on the innovations variance by decomposing the influence. The relationship is explored in Section 7 between maximum likelihood estimation with missing data and maximum likelihood estimation using a special additive outliers (AO) model, revealing an important connection between the diagnostic and Fox (1972) tests for the presence of AO. Finally, some comments about scaling and other related work are given in Section 8.

2. DIAGNOSTICS FOR AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS

We define two different diagnostics for ARIMA models based on deleting observations, and measuring the change in the estimated parameters. The diagnostic DV measures the change in the estimated innovations variance, and the diagnostic DC measures the change in the estimated ARIMA coefficients.

2.1. *The Model*

Consider a non-stationary process x_i, $t = 1, \ldots, n$, which can be represented by an ARIMA$(p, d, q) \times (P, D, Q)$ model

$$\mathbf{\Phi}(B^s)\mathbf{\phi}(B)\nabla^d\nabla_s^D x_t = \gamma + \mathbf{\Theta}(B^s)\mathbf{\theta}(B)\varepsilon_t \tag{2.1}$$

where the ε_t are the innovations. These are assumed to be independent Gaussian random variables with zero mean and variance σ^2. B is the backshift operator, and the regular and seasonal difference operators are $\nabla = 1 - B$ and $\nabla_s = 1 - B^s$

respectively. The intercept term is γ, the ordinary autoregressive and moving average operators are

$$\begin{aligned} \boldsymbol{\phi}(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p, \\ \boldsymbol{\theta}(B) &= 1 - \theta_1 B - \cdots - \theta_q B^q \end{aligned} \tag{2.2a}$$

and the corresponding 'seasonal' operators are

$$\begin{aligned} \boldsymbol{\Phi}(B^s) &= 1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP}, \\ \boldsymbol{\Theta}(B^s) &= 1 - \Theta_1 B^s - \cdots - \Theta_Q B^{sQ}. \end{aligned} \tag{2.2b}$$

Let $\boldsymbol{\alpha}$ denote the $r \times 1$ vector of parameters,

$$\boldsymbol{\alpha}^{\mathrm{T}} = (\phi_1, \ldots, \phi_p, \Phi_1, \ldots, \Phi_p, \theta_1, \ldots, \theta_q, \Theta_1, \ldots, \Theta_q)^{\mathrm{T}} \tag{2.3}$$

where $r = p + P + q + Q$. Assume that the polynomials in equations (2.2) have their roots outside the unit circle, so that the process $w_t \equiv \nabla^d \nabla_s^D x_t$ is stationary and invertible.

2.2. *Estimation of Autoregressive Integrated Moving Average Models with Missing Data*

Exact MLEs with missing data can be obtained using the state space representation of an ARIMA model. Various formulations have been given by Jones (1980), Harvey and Pierse (1984), Kohn and Ansley (1986) and Bell and Hillmer (1987). We have implemented the Harvey and Pierse approach, which has the advantage of simplicity, but the disadvantage that missing values are not allowed in the first $d + sD$ observations. If a missing value does occur at the beginning of a series, we deal with this problem in the current version of our diagnostic by computing the likelihood for the reversed series.

The Kohn and Ansley approach has the attractive feature that missing values are allowed anywhere in the series, and we hope to implement their version in the near future.

2.3. *Diagnostics for Coefficients DC*

Denote the MLE of $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$. Let $A = \{t_1, t_2, \ldots, t_k\}$ be an arbitrary subset of $\{1, 2, \ldots, n\}$, and let $\hat{\boldsymbol{\alpha}}_A$ denote the MLE with observations $y_{t_1}, \ldots, y_{t_k}$ treated as missing. If some of the observations in A have an undue influence on the estimate $\hat{\boldsymbol{\alpha}}_A$, then this will often reveal itself in the form of a substantial difference between $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\alpha}}_A$. We define the *empirical influence on the coefficients of the subset* A by

$$\mathbf{EIC}(A) = -n(\hat{\boldsymbol{\alpha}}_A - \hat{\boldsymbol{\alpha}}). \tag{2.4}$$

Standardizing by the factor n leads to a non-degenerate asymptotic form for equation (2.4).

The empirical influence $\mathbf{EIC}(A)$ is an r-dimensional vector, and as such is difficult to interpret. Further, the empirical influence is relative and comparable only within a data set. Following the ordinary regression approach, we would use a diagnostic on a quadratic form of the empirical influence function, namely

$$\mathrm{DC}(A) = \frac{1}{n} \mathbf{EIC}^{\mathrm{T}}(A) \hat{\mathbf{C}}^{-1} \mathbf{EIC}(A) \tag{2.5}$$

where $\hat{\mathbf{C}}$ is an estimate of the covariance matrix $\mathbf{C}$ of $\hat{\boldsymbol{\alpha}}$. This estimate is easily constructed via the following considerations.

Under regularity conditions that $\hat{\boldsymbol{\alpha}}$ is asymptotically normal (see, for example, Fuller (1976)),

$$(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\sqrt{n} \rightarrow N_r(\mathbf{0}, \mathbf{C}(\boldsymbol{\alpha}))$$

with asymptotic covariance matrix $\mathbf{C}(\boldsymbol{\alpha})$ which is related to $\mathbf{I}(\boldsymbol{\alpha})$, the asymptotic information matrix, by

$$\mathbf{C}(\boldsymbol{\alpha})^{-1} = \mathbf{I}(\boldsymbol{\alpha}). \tag{2.6}$$

If $\hat{\mathbf{I}}(\boldsymbol{\alpha})$ is a consistent estimator of $\mathbf{I}(\boldsymbol{\alpha})$, then the Mann–Wald theorem implies that

$$n(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^{\mathrm{T}}\hat{\mathbf{I}}(\boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \rightarrow \chi_r^2 \tag{2.7}$$

where χ_r^2 denotes a chi-squared random variable with r degrees of freedom. Thus, it is natural to choose $\hat{\mathbf{C}}^{-1}$ to be $\hat{\mathbf{I}}(\boldsymbol{\alpha})$.

One estimator of $\mathbf{I}(\boldsymbol{\alpha})$ is $\mathbf{I}(\hat{\boldsymbol{\alpha}})$, the expected information evaluated at the MLE. Although not commonly available in the literature, a closed form expression for $\mathbf{I}(\boldsymbol{\alpha})$ in terms of $\boldsymbol{\alpha}$ exists (see Appendix A). Using this expression, we take as our leave-k-out diagnostic for coefficients

$$\begin{aligned} \mathrm{DC}(A) &= \frac{1}{n}\mathbf{EIC}^{\mathrm{T}}(A)\mathbf{I}(\hat{\boldsymbol{\alpha}})\mathbf{EIC}(A) \qquad (2.8) \\ &= n(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_A)^{\mathrm{T}}\mathbf{I}(\hat{\boldsymbol{\alpha}})(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_A). \end{aligned}$$

Although the distribution of $\mathrm{DC}(A)$ is not known, the use of the χ_r^2 distribution allows us to view $\mathrm{DC}(A)$ on a familiar scale. Namely, this measures the influence of a subset A with respect to the approximate confidence intervals about $\hat{\boldsymbol{\alpha}}$. It corresponds to using the F distribution as a reference for Cook's distance (Cook and Weisberg, 1982) and DFFITS (Belsley *et al.*, 1980). The χ_r^2 distribution is used in the time series case, rather than an F distribution, since $\mathbf{I}(\boldsymbol{\alpha})$ does not involve the nuisance parameter σ^2.

While the exact definition of what is influential depends on the problem, a rough guide is to judge a subset A of points to be influential if the 'p value' of $\mathrm{DC}(A)$ based on the χ_r^2 reference distribution is smaller than 0.5 (not 0.05; see Cook and Weisberg (1982)). Empirical evidence shows that this guideline is quite useful, except near the region of non-invertibility or non-stationarity. This guideline is not a significance test and merely serves as a general purpose indication of influence. See Section 8 for more about scaling and reference distributions.

2.4. *Diagnostics for the Innovations Variance DV*

The influence of a subset A can also be measured by evaluating the effect of its removal on the MLE of the innovations variance $\hat{\sigma}^2$. We define the *empirical influence on the innovations variance of a subset A* by

$$\mathrm{EIV}(A) = -n(\hat{\sigma}_A^2 - \hat{\sigma}^2) \tag{2.9}$$

where $\hat{\sigma}_A^2$ is the MLE of σ^2 with observations at times $t \in A$ treated as missing. The diagnostic is formed in the same manner as earlier: a standardized version of $\mathrm{EIV}(A)$

is computed, based on asymptotic theory. Under regularity conditions, $\hat{\sigma}^2$ is asymptotically normal (and independent of $\hat{\boldsymbol{\alpha}}$):

$$(\hat{\sigma}^2 - \sigma^2)\sqrt{n} \to N(0, 2\sigma^4). \tag{2.10}$$

Then by the Mann–Wald theorem

$$\frac{n}{2}\left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right)^2 \to \chi_1^2.$$

Thus, we propose to use as leave-k-out diagnostic for innovations variance

$$\mathrm{DV}(A) = \frac{n}{2}\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_A^2} - 1\right)^2 \tag{2.11}$$

with the reference distribution being a chi-squared distribution with one degree of freedom (χ_1^2). Again, we suspect a subset A of observations to be influential if the p value for $\mathrm{DV}(A)$ is less than 0.5 using a χ_1^2 distribution.

2.5. *Diagnostics for Patches*

For independent observations, computational considerations usually result in deletion of a single observation at a time. The corresponding hope is that masking will not prevent us from discovering all influential data by iterative deletion. However, the time series situation differs from the case of independent observations in at least two important ways:

(a) structure is imposed by time ordering and
(b) influential observations often come in the form of an 'outlier patch' or other local 'structural' change extending over several observations.

Leave-one-out diagnostics can fail to give clear evidence of influence in the case of patchy disturbances such as outliers, as we show later in a concrete example. Such behaviour might be regarded as a form of masking since the effect of any single outlier in such a patch can be overwhelmed by the effect of the other outliers. Fortunately, this kind of situation is easily dealt with in time series (unlike as in unstructured independent observation problems) by leaving out k consecutive observations, i.e. by taking $A = A_{k,t}$ to consist of the k time points centred at t: $(t - [(k-1)/2], \ldots, t + [k/2])$, where $[x]$ denotes the largest integer less than or equal to x. For even k, t is the point closest to the left of the centre of the patch $A_{k,t}$. To simplify notation, we denote $\mathrm{DC}(A_{k,t})$, $\mathrm{DV}(A_{k,t})$, $\hat{\boldsymbol{\alpha}}_{A_{k,t}}$ and $\hat{\sigma}^2_{A_{k,t}}$ by $\mathrm{DC}(k, t)$, $\mathrm{DV}(k, t)$, $\hat{\boldsymbol{\alpha}}_{k,t}$ and $\hat{\sigma}^2_{k,t}$ respectively. When $k = 1$, we shall often simply write $\mathrm{DC}(t)$, $\mathrm{DV}(t)$, $\hat{\boldsymbol{\alpha}}_t$ and $\hat{\sigma}^2_t$.

For patches at the ends of the series, where $t \leqslant [(k-1)/2]$ or $t > n - [k/2]$, $\mathrm{DC}(k, t)$ and $\mathrm{DV}(k, t)$ are computed with the patch truncated in the obvious manner. For non-stationary models, the series will be reversed to obtain $\mathrm{DC}(k, t)$ and $\mathrm{DV}(k, t)$ for $t = 1, \ldots, d_0 + [(k-1)/2]$ where d_0 is the order of the differencing.

3. SUPERIORITY OF DV OVER DC

At first thought, DC might appear to be the diagnostic of choice since it is the ARIMA model analogue to Cook's distance (Cook and Weisberg, 1982). In the

regression case, the error variance is a nuisance parameter and therefore has less intuitive appeal as the basis for a diagnostic. However, in time series models, the innovations variance often plays a fundamental role and, for some problems, the innovations variance is the parameter of interest. Furthermore, as we show in this section, DV has better properties than DC.

The major difference between leave-k-out coefficients DC for time series (including $k = 1$) and the usual regression coefficients diagnostics for independent data is a *smearing* of the effect of an isolated outlier or patch of outliers to adjacent points. A given point may be judged influential using the diagnostics for the coefficients because of an outlier at an adjacent point. For example, in the AR(p) case, an isolated outlier can result in significant values for DC at p times before and after the occurrence of the outlier. Hence, interpretation of leave-k-out diagnostics for DC is not so clear as in the usual regression case. In contrast, diagnostics for the innovations variance display much smaller, and often negligible, smearing effects. In this section, we first demonstrate this using some artificial examples. Then an asymptotic approximation is given to establish an analytical rationale for these different smearing effects.

3.1. *Outlier Models and Examples*

In the following examples, we focus on influential points caused by *outliers*. Influential observations may also be the result of structural changes, such as level shifts or variance changes. We discuss the application of leave-k-out diagnostics to such problems in Sections 4 and 5.

We examine the performance of DC and DV under two types of contamination commonly used in other studies (see, for example, Fox (1972), Denby and Martin (1979), Martin and Yohai (1986) and Tsay (1986)): the AO model and the innovations outliers (IO) model.

Let x_t be a Gaussian ARIMA process specified by model (2.1). Then y_t behaves according to a *fixed magnitude* AO model if

$$y_t = x_t + \zeta_t z_t \tag{3.1}$$

where ζ_t is fixed (but may depend on t) and z_t is a 0–1 process. The magnitude of the outliers is ζ_t; isolated outliers and patches are created by appropriate choice of zeros and ones for z_t.

A *fixed magnitude* IO model is formed through contamination in the innovations process ε_t. Let ε_t be a contaminated white noise process, with

$$\varepsilon_t = \tilde{\varepsilon}_t + \zeta_t z_t \tag{3.2}$$

where $\tilde{\varepsilon}_t$ are independent Gaussian random variables with zero mean and variance σ^2, and ζ_t, z_t are as before. Then y_t follows an ARIMA IO model if it is generated by model (2.1) with the ε_t given by model (3.2).

The models (3.1) and (3.2) are rather special AO and IO forms. More general AO and IO (and other) outlier models for time series are possible (see Martin and Yohai (1986) for a very flexible 'general replacement' model).

3.1.1. *Example 1: AR(1), $\phi = 0.4$, $\sigma^2 = 1$, additive outliers model with one isolated outlier*

Starting with a simple case, we examine a simulated AR(1) series of 100 points from Gaussian white noise with $\phi = 0.4$ and a single AO of $+4$ at point 28. The MLE fit

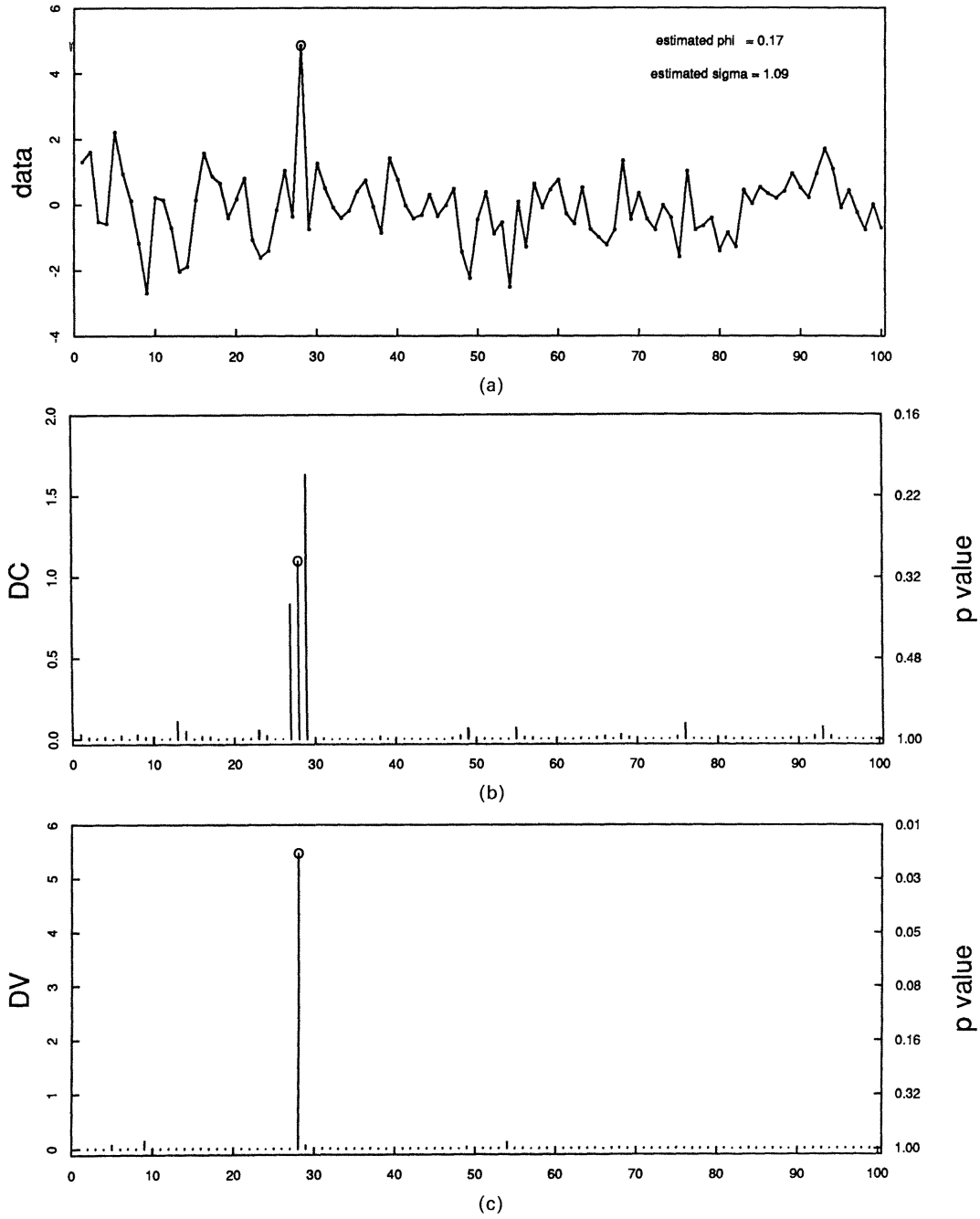

Fig. 1. Example 1, simulated AR(0.4) model with one isolated AO: (a) plot of data; (b) scaled leave-one-out diagnostics—ϕ; (c) scaled leave-one-out diagnostics—innovations variance

of an AR(1) model with the entire series yields $\hat{\phi} = 0.17$ and $\hat{\sigma}^2 = 1.09$. The data are plotted in Fig. 1(a) and the outlier is marked by 'o'. The leave-one-out diagnostics, DC($\cdot$) and DV($\cdot$), for $\hat{\phi}$ and $\hat{\sigma}^2$ are displayed in Figs 1(b) and 1(c). The p values corresponding to a χ^2_1 distribution are displayed on the right-hand axis. The p values

for DC(t) at t values of 27, 28 and 29 are all smaller than 0.5, while the p value for DC(t) is considerably greater than 0.5 for all other times. Thus y_{27}, y_{28} and y_{29} are judged to be influential. By contrast, among all DV(t) values, only DV(28) is significant, and its p value of about 0.02 is much smaller than DC(28). This example is indicative of a general pattern which we establish analytically later: an outlier is smeared across several values of DC($\cdot$) but is identified exactly by DV($\cdot$) and with greater power! In particular, for AR(1) processes, the smearing for DC($\cdot$) extends by one time unit in each direction from $t = 28$.

3.1.2. *Example 2: AR(1), $\phi = 0.4$, $\sigma^2 = 1$, innovations outliers model with one isolated outlier*

Example 2 is the same series as in example 1, except that the outlier at point 28 is of the IO type. The MLEs of the parameters are $\hat{\phi} = 0.27$ and $\hat{\sigma}^2 = 1.06$. Fig. 2 displays the data and leave-one-out diagnostics DC($\cdot$) and DV($\cdot$). The problem of smearing for DC is considerably worse than in example 1. The diagnostic DC is significant only at time $t = 27$, while the p value of about 0.7 at $t = 28$ is quite insignificant. However, there is no smearing with DV($\cdot$): DV(28) is several magnitudes larger than DV(t) for any other t, and hence DV identifies the outlier quite clearly. Again, this is general behaviour, established later. For the AR(1) case, DC is large at the time point just before the occurrence of an isolated IO-type outlier but is small at the time of occurrence of the outlier, while DV is large only at the time of occurrence of the outlier.

3.2. *Smearing and Expected Asymptotic Diagnostic*

To understand the smearing behaviour of the diagnostic (2.8) for coefficients, it is helpful to use an asymptotic representation of DC(A) for general subset deletions A. For subsets $A_{k,t}$ of *fixed size* k, the 'usual' $\sqrt{n}$ asymptotics do not apply, and typically $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_{k,t} = O(n^{-1})$. Correspondingly, we are interested in the asymptotic behaviour of nDC(A). Since the asymptotic distribution of this quantity is quite complicated, we work with an *expected asymptotic diagnostic for the coefficients*, defined by

$$\begin{aligned} \mathrm{EDC}(A) &= E\left[\lim_{n\to\infty} n\mathrm{DC}(A)\right] \\ &= E\left[\lim_{n\to\infty} n^2(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_A)^{\mathrm{T}}\mathbf{I}(\hat{\boldsymbol{\alpha}})(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_A)\right]. \end{aligned} \tag{3.3}$$

In the same spirit as in equation (3.3), we shall use an *expected asymptotic diagnostic for the innovations variance*:

$$\mathrm{EDV}(A) = E\left[\lim_{n\to\infty} n\mathrm{DV}(A)\right]. \tag{3.4}$$

Calculation of EDC and EDV is very tedious, so we restrict our attention to the AR(1) case. The computations are based on a Taylor series expansion of the efficient score functions for the full likelihood and for the likelihood with subset A treated as missing. Details of the calculations made for the discussion to follow may be found in Bruce and Martin (1987).

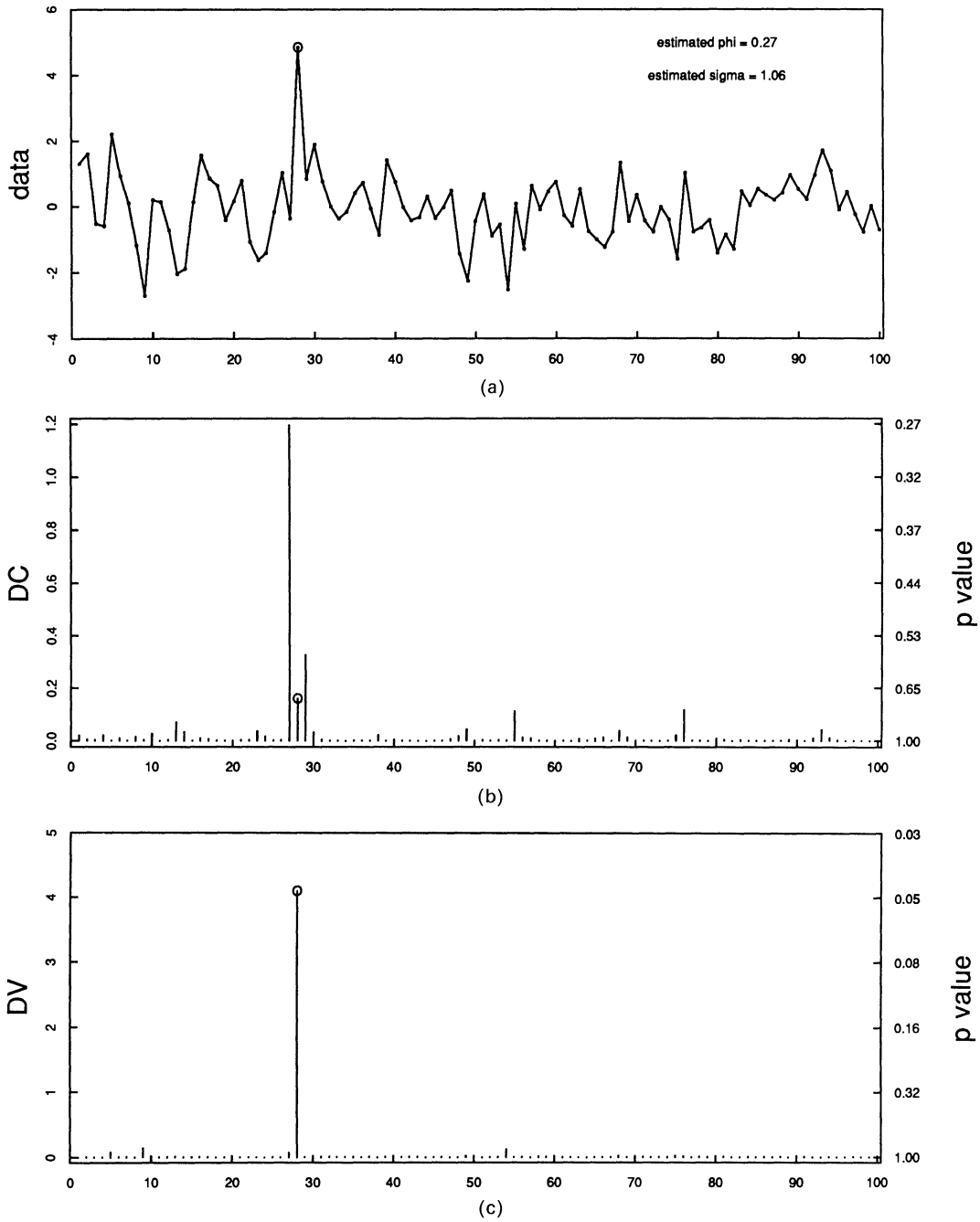

Fig. 2. Example 2, simulated AR(0.4) model with one isolated IO: (a) plot of data; (b) scaled leave-one-out diagnostics—ϕ; (c) scaled leave-one-out diagnostics—innovations variance

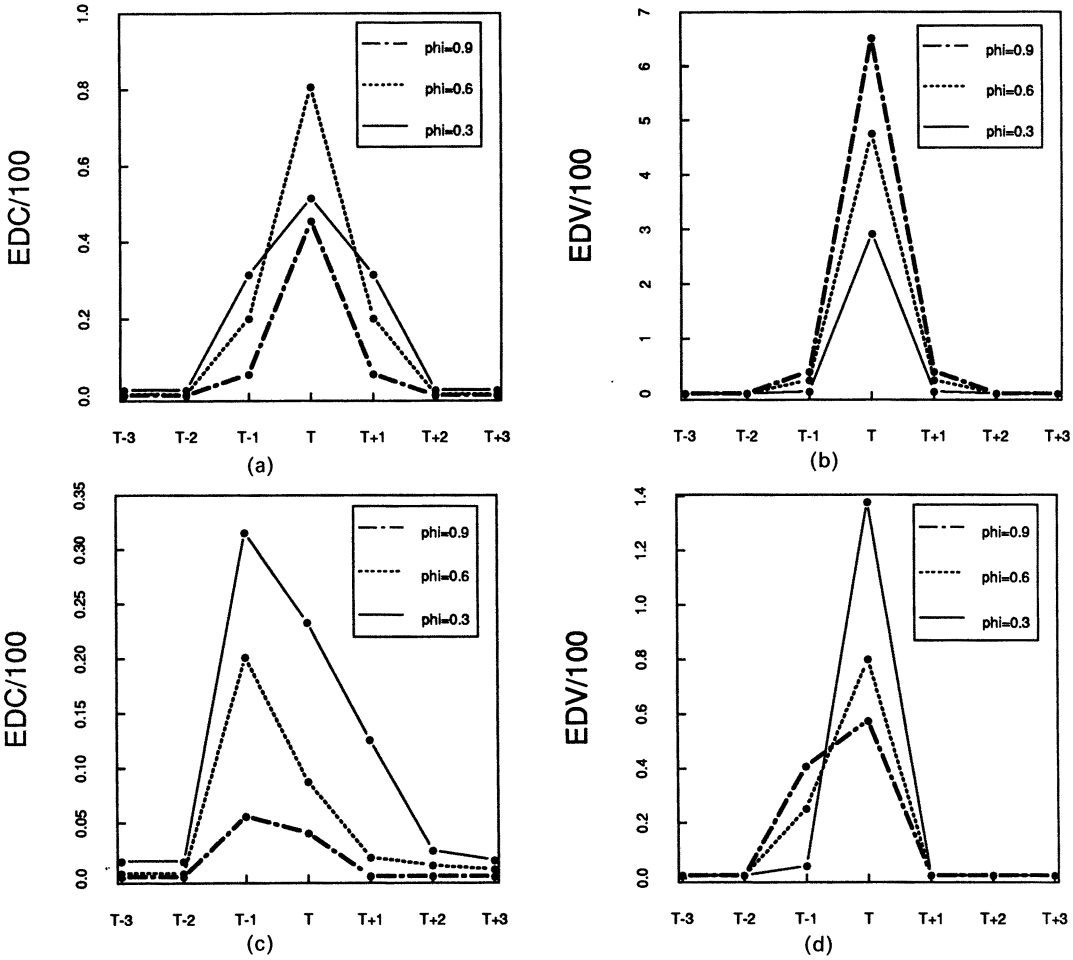

Fig. 3. Expected asymptotic diagnostic curves: (a) EDC for AO-type outlier of size +4 at time T; (b) EDV for AO-type outlier of size +4 at time T; (c) EDC for IO-type outlier of size +4 at time T; (d) EDV for IO-type outlier of size +4 at time T

3.2.1. *Smearing for AR(1) model: additive outliers case*

We can understand the problem of smearing by computing the expected asymptotic diagnostic for various outlier models. For the AO case, the expectation in equations (3.3) and (3.4) is with respect to the process y_t given by model (3.1). In particular, if x_t is a Gaussian ARIMA process and A is a fixed subset of time indices, then for the process

$$y_t = \begin{cases} x_t & t \notin A \\ x_t + \zeta & t \in A \end{cases}$$

we shall denote the expected asymptotic diagnostics by $\mathrm{EDC}^{\mathrm{AO}}_{(\zeta;A)}$ and $\mathrm{EDV}^{\mathrm{AO}}_{(\zeta;A)}$.

Fig. 3(a) displays $\mathrm{EDC}^{\mathrm{AO}}_{(+4;T)}(t)/100$, i.e. the expected asymptotic diagnostic for the coefficients assuming a single outlier of size 4 at time T, for $t = T-3, T-2, \ldots, T+3$ with ϕ values of 0.3, 0.6 and 0.9. Fig. 3(b) gives the corresponding

plot for $\mathrm{EDV}^{\mathrm{AO}}_{(+4;T)}(t)/100$. The scaling factor of 1/100 approximates the expected value of the diagnostics for a sample size of 100. The asymptotic approximations verify what was observed in example 1 for AO models: the smearing is worse for DC, and DV tends to be more sensitive.

Because of sampling fluctuation, the patterns of diagnostics observed in example 1 differ from the expected diagnostics in two regards: the magnitude of DC and DV is larger than EDC and EDV, and the pattern over time for DC is not the same in that the largest diagnostic occurs after the outlier ($t = 28$).

In Fig. 4(a), we graphically compare the amount of smearing for DC and DV as a function of ϕ. The ratios

$$\frac{\mathrm{EDC}^{\mathrm{AO}}_{(+4;T)}(T-1)}{\mathrm{EDC}^{\mathrm{AO}}_{(+4;T)}(T)}$$

(full line) and

$$\frac{\mathrm{EDV}^{\mathrm{AO}}_{(+4;T)}(T-1)}{\mathrm{EDV}^{\mathrm{AO}}_{(+4;T)}(T)}$$

(broken line) represent the proportional amount of smearing for an outlier of size 4 at T. These ratios are always less than unity. However, the expected asymptotic smearing for DV is small in absolute terms for all ϕ and also substantially smaller than that of DC for all but quite large values of ϕ. The smearing for DC is greater than 0.5 for a large range of ϕ values, and in such situations smearing may lead to some confusion when examining DC.

The potential for confusion becomes unquestionably serious in situations where there is more than one outlier. We demonstrate this in the very simplest context. Suppose y_t is observed with two isolated AO-type outliers of size 4 at times $T - 1$ and $T + 1$. Fig. 4(b) exhibits

$$\frac{\mathrm{EDC}^{\mathrm{AO}}_{(+4;T-1,T+1)}(T)}{\mathrm{EDC}^{\mathrm{AO}}_{(+4;T-1,T+1)}(T-1)}$$

(full line) and

$$\frac{\mathrm{EDV}^{\mathrm{AO}}_{(+4;T-1,T+1)}(T)}{\mathrm{EDV}^{\mathrm{AO}}_{(+4;T-1,T+1)}(T-1)}$$

(broken line) as a function of ϕ. The expected asymptotic value of $\mathrm{DC}(T)$ with outliers at $T - 1$ and $T + 1$ is larger than the expected asymptotic diagnostic at the $T - 1$ (and, by symmetry, $T + 1$) outlier position for all ϕ and has a maximum value almost six times larger! Thus, DC will be totally ineffective in revealing such a configuration of outliers. By contrast, the ratio for EDV stays below unity for all ϕ and is substantially smaller than unity except for $|\phi|$ greater than about 0.6. We therefore expect DV to be far superior to DC in revealing such outlier configurations.

3.2.2. *Smearing for AR(1) model: innovations outliers case*

The analysis of smearing for IO models parallels that for AO models. However, since the outliers occur in the innovations of the process, the difference in the score functions for ϕ is not symmetric. Here, the expected asymptotic diagnostics are not symmetric, as was the case for AO models. Suppose y_t is observed with an IO-type

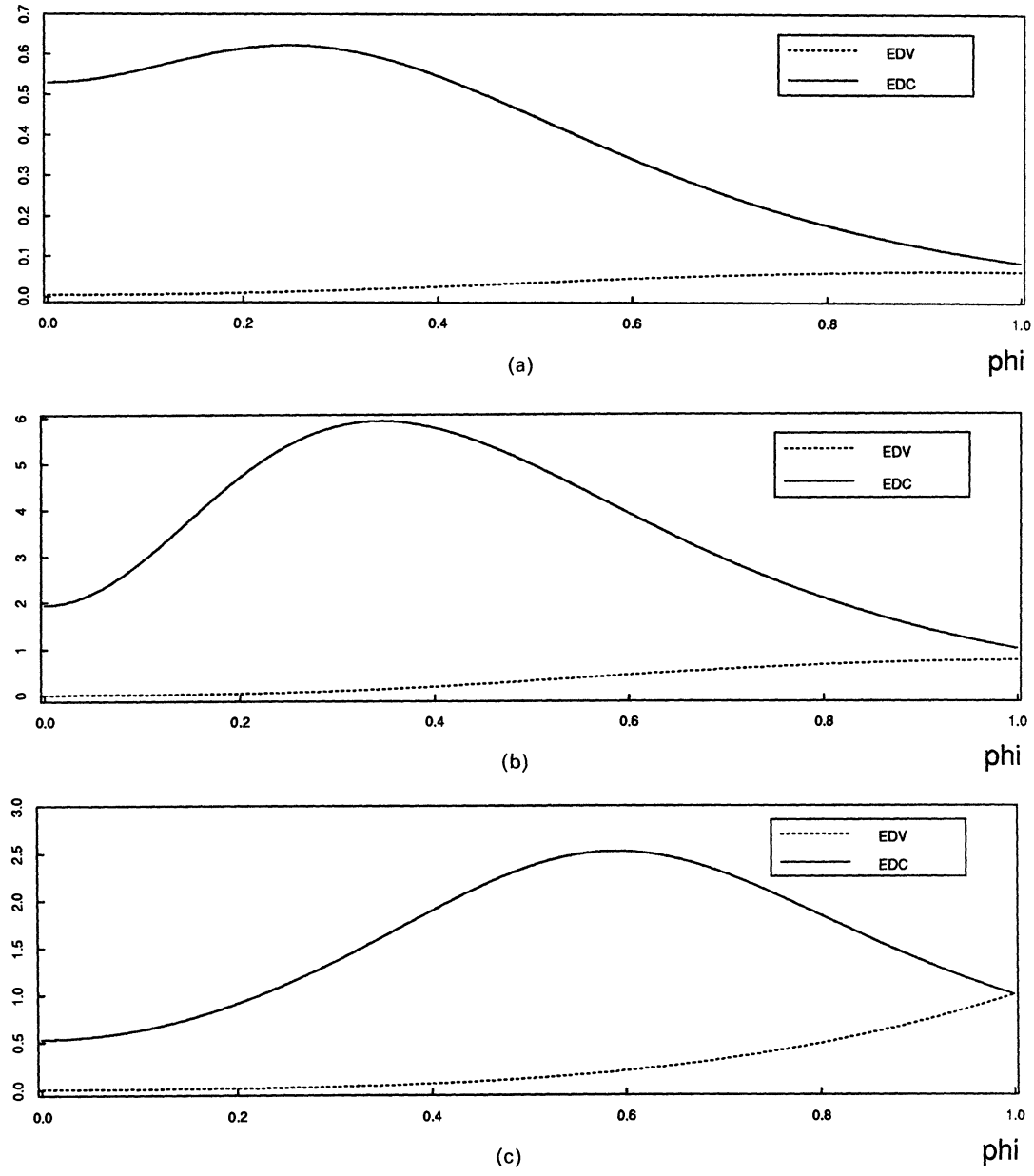

Fig. 4. Expected asymptotic diagnostics—smearing plots: (a) proportion of smearing due to an isolated AO; (b) proportion of smearing due to two adjacent isolated AOs; (c) proportion of smearing due to an isolated IO

outlier of magnitude ζ at T, so ε_t is given by model (3.2) with $z_t = 0$ except at $t = T$ where $z_t = 1$. If x_t represents the series *without* the innovations outlier, then it is easy to check that

$$y_t = \begin{cases} x_t & t < T \\ x_t + \zeta\phi^{t-T} & t \geqslant T. \end{cases}$$

We use the notation $\mathrm{EDC}^{\mathrm{IO}}_{(\zeta,T)}$ and $\mathrm{EDV}^{\mathrm{IO}}_{(\zeta,T)}$ to denote the corresponding expected asymptotic diagnostics.

With IO models, the behaviour of the smearing effects of an outlier differ even more dramatically for EDC and EDV. For EDC, the effect of smearing is not restricted to points immediately adjacent to the time of occurrence T of an outlier. Specifically, an outlier affects $\mathrm{EDC}^{\mathrm{IO}}_{(\zeta,T)}(t)$ for all $t \geqslant T - 1$ and has the maximum effect at $t = T - 1$! This is quite misleading. By way of contrast, the effects of an outlier at T are seen only at $t = T - 1$ and $t = T$ for $\mathrm{EDV}^{\mathrm{IO}}_{(\zeta,T)}(t)$ and the maximum value occurs at $t = T$, as we wish. Figs 3(c) and 3(d) display $\mathrm{EDC}^{\mathrm{IO}}_{(+4;T)}(t)/100$ and $\mathrm{EDV}^{\mathrm{IO}}_{(+4;T)}(t)/100$ for $t = T - 3, T - 2, \ldots, T + 3$ with ϕ values of 0.3, 0.6 and 0.9. The severe smearing of DC at $t = T - 1$ is reflected in Fig. 3(c), where $\mathrm{EDC}^{\mathrm{IO}}_{(+4,T)}(T - 1)$ dominates $\mathrm{EDC}^{\mathrm{IO}}_{(+4;T)}(T)$. This is also demonstrated by Fig. 4(c), which shows

$$\frac{\mathrm{EDC}^{\mathrm{IO}}_{(+4;T)}(T - 1)}{\mathrm{EDC}^{\mathrm{IO}}_{(+4;T)}(T)}$$

(full line) and

$$\frac{\mathrm{EDV}^{\mathrm{IO}}_{(+4;T)}(T - 1)}{\mathrm{EDV}^{\mathrm{IO}}_{(+4;T)}(T)}$$

(broken line). For most values of ϕ, the ratio for $\mathrm{EDC}^{\mathrm{IO}}$ is greater than unity, while the ratio for EDV stays well below unity. For $|\phi|$ close to unity, we cannot expect such good results from even EDV in the presence of an IO.

These results extend to AR(p) models: dominant values of EDC can occur at the p consecutive times *preceding* an isolated IO. The use of EDV(t) is obviously preferred for IO as well as AO situations.

4. OVERALL STRATEGY

In this section we present an overall strategy for ARIMA model fitting using leave-k-out diagnostics. The diagnostics defined in Section 2 are used in a simple recipe to determine the length of a patch of influential points. This strategy is embedded in an iterative deletion procedure, which often overcomes problems caused by masking. Since in some cases the iterative deletion procedure fails, more flexible subset deletion techniques are introduced. The problem of ARIMA model identification is also discussed.

4.1. *Patch Length Determination Strategy*

Outliers and other influential observations typically come in patches: there are often several aberrant values adjacent in time. The length of the patch of influential observations often depends solely on the sampling interval. It is thus imperative to search for influential *patches* and not just isolated observations. Motivated by an example, a recipe is given for determining the length of a patch of influential observations using the diagnostic DV.

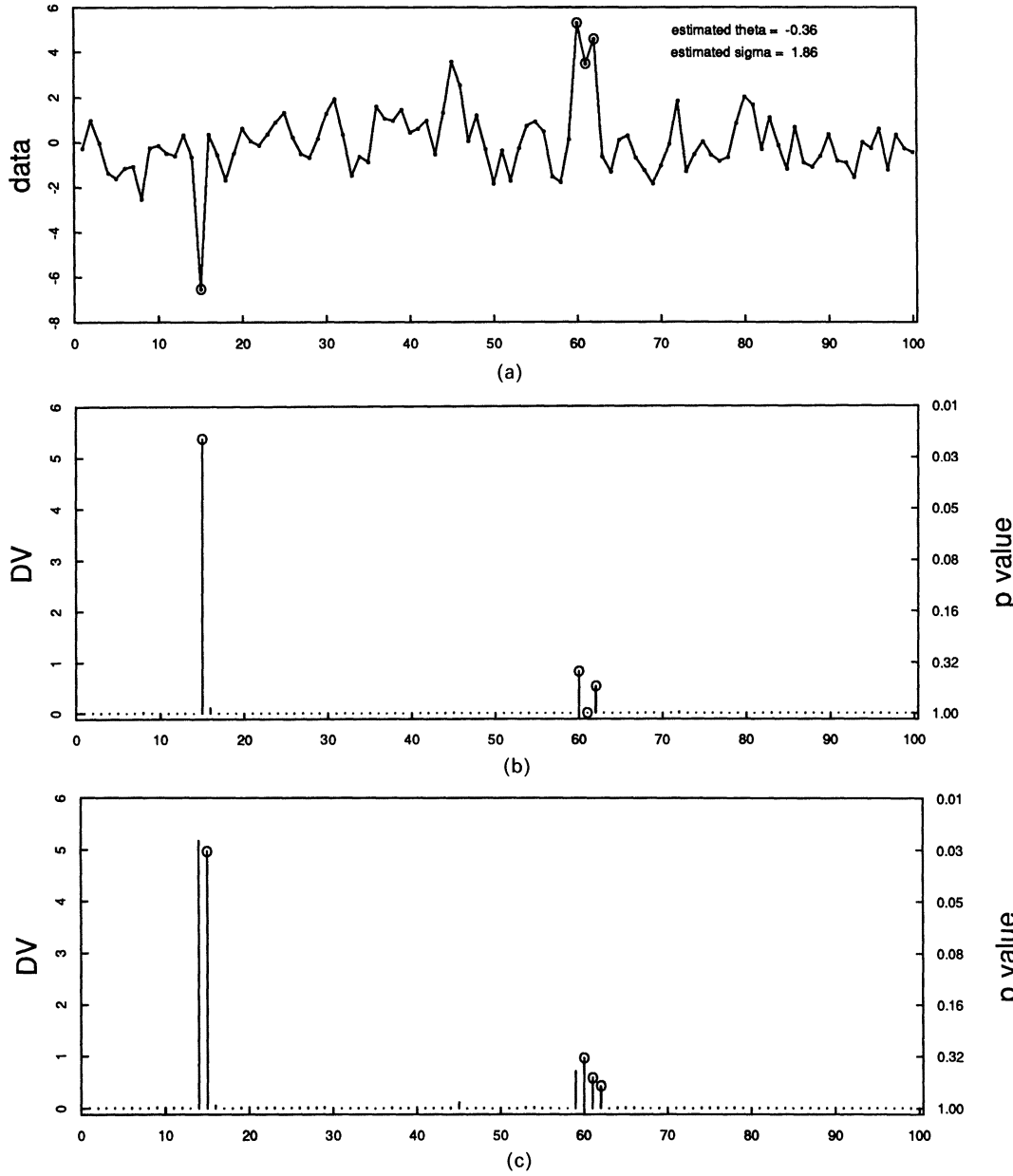

Fig. 5. Example 3, simulated MA(−0.5) model with one patch and one isolated AO: (a) plot of data; (b) scaled leave-one-out diagnostics—innovations variance; (c) scaled leave-two-out diagnostics—innovations variance; (d) scaled leave-three-out diagnostics—innovations variance; (e) scaled leave-four-out diagnostics—innovations variance

4.1.1. *Example 3: MA(1), $\theta = -0.5$, $\sigma^2 = 1$, additive outliers model with one patch and one isolated outlier*

This example is a simulated MA(1) series with $\theta = -0.5$ and both a patch of three outliers of size $+5$ at points 60–62 and an isolated outlier of size -4 at time 15. The outliers are all of the AO type. Fig. 5(a) shows the data; the MLEs are $\hat{\theta} = -0.36$ and $\hat{\sigma}^2 = 1.86$. Leave-one-out to leave-four-out diagnostics for DV are displayed in Figs 5(b)–5(e).

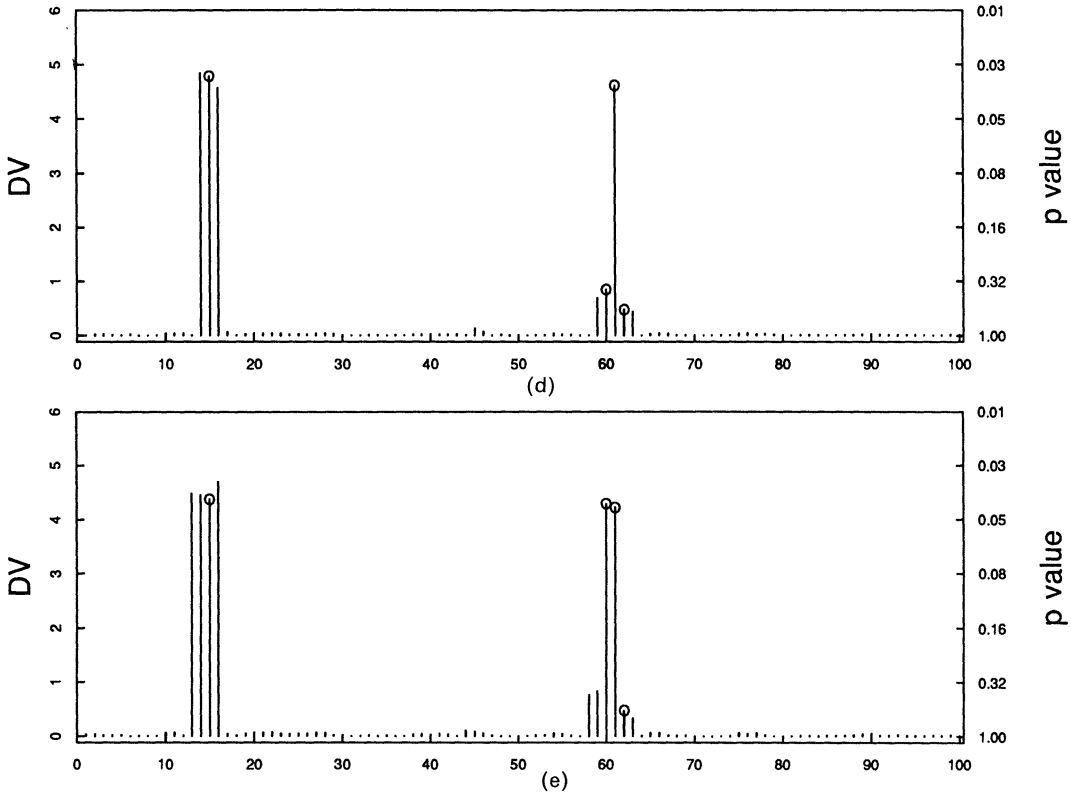

Fig. 5 (*continued*)

Recall that, for $k \geqslant 2$, DV(k, t) represents the influence of a patch of k observations centred at t and that, for even k, t is the closest point to the left of the 'centre' of the patch. For example, with $k = 2$, DV(k, t) corresponds to the diagnostic computed when y_t and y_{t+1} are left out.

Although leave-one-out diagnostics clearly identify the isolated outlier, there is barely an indication (using a p value of 0.5 as a guideline) of something happening at $t = 60$ and $t = 62$. The leave-one-out diagnostics are not adequate for detecting the patch of outliers: leaving a single point out in the patch is insufficient because the remaining outliers in the patch comprise the bulk of the influence of the patch. Leave-two-out and leave-three-out diagnostics provide progressively stronger evidence of the patch of outliers. The value DV(3, 61) is over five times larger than other neighbouring diagnostic values.

4.1.2. *Patch length determination strategy*

The isolated outlier in Fig. 5 is smeared in the leave-k-out diagnostics for $k = 2$, 3, 4. For $k = 2$, both DV(2, 14) and DV(2, 15) are highly significant and have nearly the same value as DV(1, 15). Similar behaviour is observed for $k = 3$ and $k = 4$. The general pattern is as follows: $k - 1$ values of DV(k, $\cdot$) surrounding the location of an isolated outlier at T are significant and have nearly the same value as DV(k, T)! This corresponds to what we might intuitively expect for an isolated outlier: deletion of a patch which includes an isolated outlier has nearly the same effect as deleting only the isolated outlier.

Similar behaviour occurs for a patch of outliers. For example, DV(4, t) yields values at $t = 60$ and $t = 61$ which are nearly equal to DV(3, 61). In general, for a patch of k_0 outliers centred at T, the following patch property (PP) holds.

> For $k \geqslant k_0$, there are $k - k_0 + 1$ subsets $A_{k,t}$ which completely overlap the patch, and for deletion of these subsets, the magnitude of DV(k, t) is roughly the same and significant (i.e. the associated p value is less that 0.5).

Thus, we judge an influential patch to be of length $k_0 \geqslant 1$ centred at T if DV(k_0, T) is significant, and the PP holds. If DV(k_0, T) is significant and the PP fails to hold, then this is an indication that a broader patch of outliers is present.

This provides us with an initial strategy for identifying the length of patches of influential points.

> Compute leave-k-out diagnostics for increasing $k = 1, 2, \ldots$, until the magnitude of DV(k, t) does not 'significantly' increase for any t. The length of a patch will be estimated as one less than the first value of k for which nearly uniform smearing is in evidence.

4.2. *Iterative Deletion Strategy*

The masking of influential points (e.g. outliers) by other influential points is a problem encountered in all types of diagnostics. As we have already seen, masking caused by a single patch of outliers can be handled adequately by leave-k-out diagnostics. However, sometimes the presence of a gross outlier will have sufficient influence that deletion of aberrant values elsewhere in the series has little effect on the estimate. More subtle types of masking occur when moderate outliers occur close to one another. These types of masking can often be effectively uncovered by an iterative deletion process which consists of removing suspected outliers from the data and recomputing the diagnostics.

To deal with problems caused by masking, we build on the initial patch length determination strategy as follows.

(a) Run leave-k-out diagnostics on the data, for $k = 1, 2, \ldots$, until either the length of the most influential (significant) patch is determined using the guidelines of Section 4.1 or $k = K_{\max}$, where $K_{\max}$ is determined by the user. In principle, $K_{\max}$ is the length of the longest patch of outliers thought to be present in the data. However, computational costs may require that $K_{\max}$ be reasonably small. For 'short' time series, i.e. $n < 250$, setting $K_{\max} = 5$ will often reveal most if not all problems with the data. The second case can result from two possibilities: either no influential observations were detected or the length of an influential patch is ill determined (according to the guidelines of Section 4.1). The latter case may be due to a patch of length greater than $K_{\max}$, or perhaps the patch length simply cannot be determined from the data. In this case, we determine the 'most influential' patch as that corresponding to the most significant diagnostic.

(b) If no influential points are found, then conclude the analysis. If influential points are found, then delete the most influential points as identified in step (a) and go back to step (a). The new leave-k-out coefficients should be scaled

according to the MLE computed with the outliers removed, to gauge additional influence of the remaining points.

The next artificial example illustrates the efficacy of the iterative deletion procedure in handling problems caused by masking.

4.2.1. *Example 4: simulated MA(1), $\theta = -0.5$, additive outliers model with one patch and one adjacent isolated outlier*

The data are the same as those used in example 3, except that the isolated outlier is moved from point 15 to point 58, adjacent to the patch of outliers at points 60–62. The data are plotted in Fig. 6(a). Leave-one-out to leave-four-out diagnostics for DV are given in Figs 6(b)–6(e).

The masking is much more severe than in example 3: the isolated outlier is now completely masked for the leave-one-out case (cf. example 3), though the patch still shows up prominently in the leave-three-out diagnostics. However, leave-four-out diagnostics are only slightly more significant than the leave-three-out diagnostics, and the pattern of smearing is reasonably consistent with a patch of three outliers. So, following our strategy, we delete points 60–62 and recompute the diagnostics. The isolated outlier is now easily identified by the recomputed leave-one-out diagnostics of Fig. 6(f): removal of the patch eliminates the masking problem.

4.3. *Structural Changes and Flexible Subset Deletion Techniques*

Until now, we have concentrated on influential points in the form of outliers. However, influential points may also be due to other types of disturbance, such as level shifts or variance changes. The iterative deletion procedure of Section 4.2 is often effective for uncovering these types of problem. However, the procedure will sometimes fail in the presence of an influential patch longer than $K_{\max}$; an example of this is provided in Section 5. Masking may prevent a long patch, or any points in the patch, from being detected. Even when the patch is detected, if the disturbance spans a time period considerably greater than $K_{\max}$, the iterative deletion may require an intolerable number of iterations. To handle these failures, one should be prepared to adopt a flexible approach to subset deletion guided by the data at hand.

We consider abandoning the iterative deletion procedure and using the flexible deletion approach primarily in two kinds of situation. First, when examining the data and the residuals, the analyst may suspect a structural change in the data. Secondly, the leave-k-out diagnostics may indicate a local disturbance of duration greater than $K_{\max}$ (e.g. when the patch length is ill determined; see step (a) of Section 4.2). In either case, flexible subset deletion techniques can help to identify the structure more precisely.

An attractive way of carrying out flexible subset deletions is through the use of interactive graphics on a computer workstation. Candidate subsets are identified on computer graphics plots of the data and/or the residuals, and DV is computed for such subsets. For example, if the analyst believes a local level shift is present somewhere between the times t_0 and t_1, then DV could be computed for a judicious selection of patches between t_0 and t_1 to clarify the jump points. This procedure may easily be carried out with the aid of a 'mouse' and appropriate software.

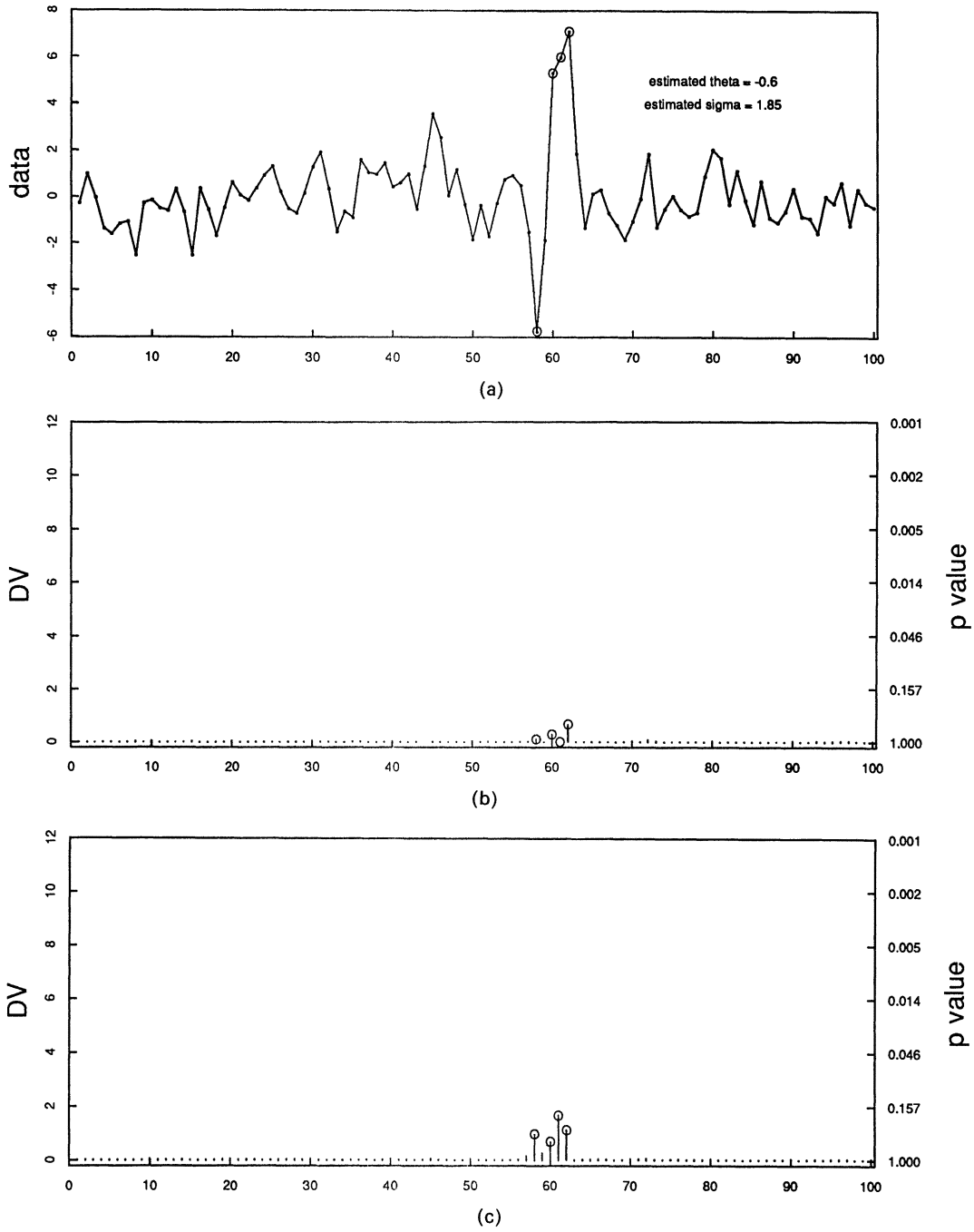

Fig. 6. Example 4, simulated MA(−0.5) model with one patch and one adjacent isolated IO: (a) plot of data; (b) scaled leave-one-out diagnostics—innovations variance; (c) scaled leave-two-out diagnostics—innovations variance; (d) scaled leave-three-out diagnostics—innovations variance; (e) scaled leave-four-out diagnostics—innovations variance; (f) scaled leave-one-out diagnostics after iterative deletion

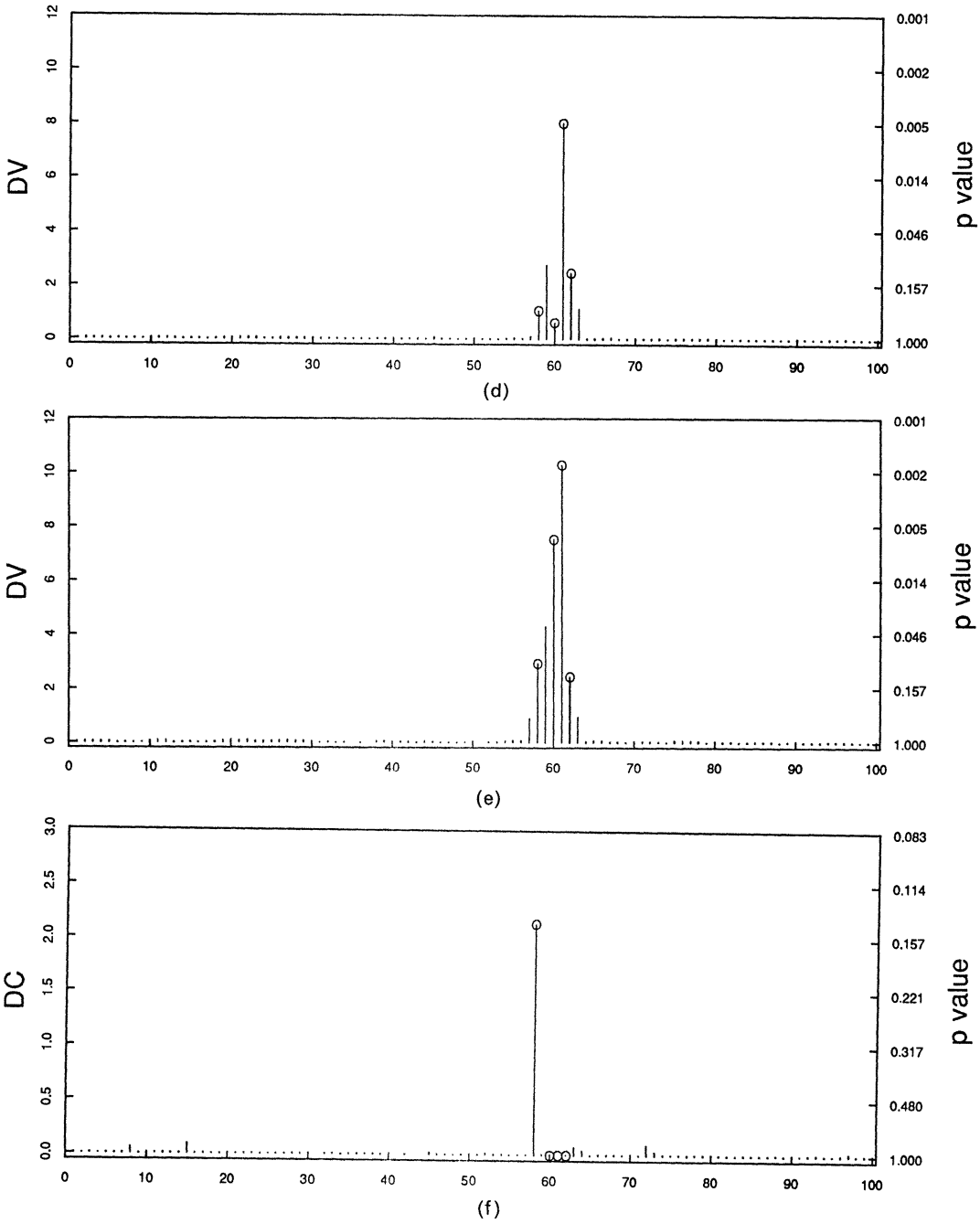

Fig. 6 (*continued*)

A non-interactive and computationally expensive approach is to run leave-k-out diagnostics on the data for selected values of k between $K_{\max}$ and $n/2$. We might for example choose $k = [n/2], [n/4], \ldots, [n/2^r]$ where r is the largest integer such that $n/2^r > K_{\max}$. An alternative would be to choose $k = 2^s, 2^{s+1}, \ldots, 2^t$ where s is the

smallest integer such that $2^s \geqslant K_{\max}$ and t is the largest integer such that $2^t \leqslant n/2$. From these diagnostics, the disturbances can often be clarified.

Another application of the 'bottom-up' or 'top-down' diagnostics would be to provide a final check on the model after the analysis is completed. This ensures detection of long patches of influential points and structural changes.

4.4. *Model Identification and Analysis of Residuals*

4.4.1. *Model identification*

The foregoing analysis presumes that the degree of differencing and the order for the model was known. In practice, this is rarely the case, and the model must be determined by some criteria such as the Box–Jenkins identification procedure. However, outliers may cause improper model specification. To handle order selection in the presence of outliers and structural changes, we can embed the iterative deletion strategy in an iterative procedure similar to that used by Tsay (1986). The initial model order is selected, and the iterative deletion strategy performed on the initial model. After removing all influential points, the model is identified again. If the same model is selected, then the analysis is concluded. Otherwise, iterative deletion is performed again, and the cycle is repeated until the same model is identified in successive rounds.

While this procedure is usually adequate, it may fail in some situations where a poor initial model is selected. If the wrong model is identified, then removal of the influential points *under that model* may lead to selection of the same model (see Bruce (1988) for an example).

4.4.2. *Analysis of residuals and influential points*

In the presence of outliers and structural changes, the usual prediction residuals are often misleading for identifying the influential observations. Instead, we recommend the examination of the residuals based on the predictions formed when the observations identified as influential are treated as missing. Since the predictions are not distorted by influential observations, this procedure reveals outliers and structural changes more clearly.

After selecting the final model, a careful analysis of the influential data points should be carried out. Of particular interest is the determination of any physical causes or events related to such points. Also, one may be able to categorize influential points as isolated or patches of outliers, or perhaps associate them with a level shift or variance change.

Points diagnosed as outliers can be further classified by type (AO versus IO). A formal way of determining whether an outlier identified by DV is IO or AO is to use a robust version of Fox's (1972) test, as described in Martin and Zeh (1977). See also the non-robust use of Fox-type tests in the outlier identification and model fitting scheme proposed by Chang and Tiao (1983), Hillmer *et al.* (1983) and Tsay (1986).

A less formal way of determining whether an outlier is IO or AO is to examine a lag-1 scatter plot of the residuals. As was pointed out by Martin and Zeh (1977), IOs tend to fall near the abscissa and ordinate of such a plot, whereas AOs tend to appear away from the abscissa and ordinate, assuming that robust parameter estimates have

been used to form the residuals. In the present context, we recommend the use of the parameter estimates obtained with the outliers identified by DV deleted.

4.4.3. *Use of intervention analysis*

A variety of structural changes, such as outlier patches, level shifts and even variance shifts, which may be detected by the leave-k-out strategy, can be handled by intervention analysis, as in Box and Tiao (1975). The prediction residuals for local structural changes provide information which may suggest a small palette of intervention 'shapes'. We note that the diagnostics may suggest intervention analysis which might be otherwise overlooked because the investigator was unaware of any particular cause (e.g. policy change).

4.4.4. *Need for flexible approach*

The analyst will often want to deviate from the procedures outlined as dictated by the data. Our initial attempts at specifying a fixed 'overall strategy' failed because the data that we wanted to analyse often forced us to take a very flexible and improvisational approach. Our feeling now is that different data sets have their own special problems, and establishing a rigid overall framework for analysis detracts considerably from the potential benefits of subset deletion diagnostics.

5. APPLICATIONS TO REAL DATA

In this section, we analyse two economic time series using the strategy articulated in Section 4. The first series is relatively well behaved, except for several patches of outliers. The second is more difficult to model, since it contains several local non-stationarities and disturbances, including level shifts and a variance change. For brevity, we omit the details of model selection in the examples to follow and concentrate instead on the diagnostics.

5.1. *Example 5: Exports to Latin-American Republics—1966–83*

In this example, we study monthly unadjusted data on exports from the USA to Latin-American republics. This series was examined by Burman (1985), who focused on outliers and forecasting in US Census Bureau data. A plot of the logarithm of the data is given in Fig. 7(a); the circled values represent points eventually deleted from the series. We fit an ARIMA(0, 1, 2) model to this series, and the residuals from the MLE fit are plotted in Fig. 7(b). The residuals based on one-step predictions computed from the data with the outliers removed (i.e. treated as missing data) are given in Fig. 7(c).

Leave-k-out diagnostics for $k = 1, 2, 3$ are displayed in Figs 7(d)–7(f) for DV. Leaving out longer patches reveals nothing new, since the series is dominated by the outliers at 1/69 and 2/69 (i.e. January 1969 and February 1969). The effect on the innovations variance of leaving these two points out is dramatic ($p < 0.0001$): see Fig. 7(e). It is unlikely that a broader patch of outliers is present in this time period, since leave-three-out diagnostics yield no increased significance, and the smearing is consistent with a patch of two outliers. The plots also hint at outliers in the last quarter of 1971. DV($\cdot$, 2) is clearly significant for other points (e.g. 10/71) and suffers from masking since the scale of the diagnostic at 1/69 and 2/69 is so large.

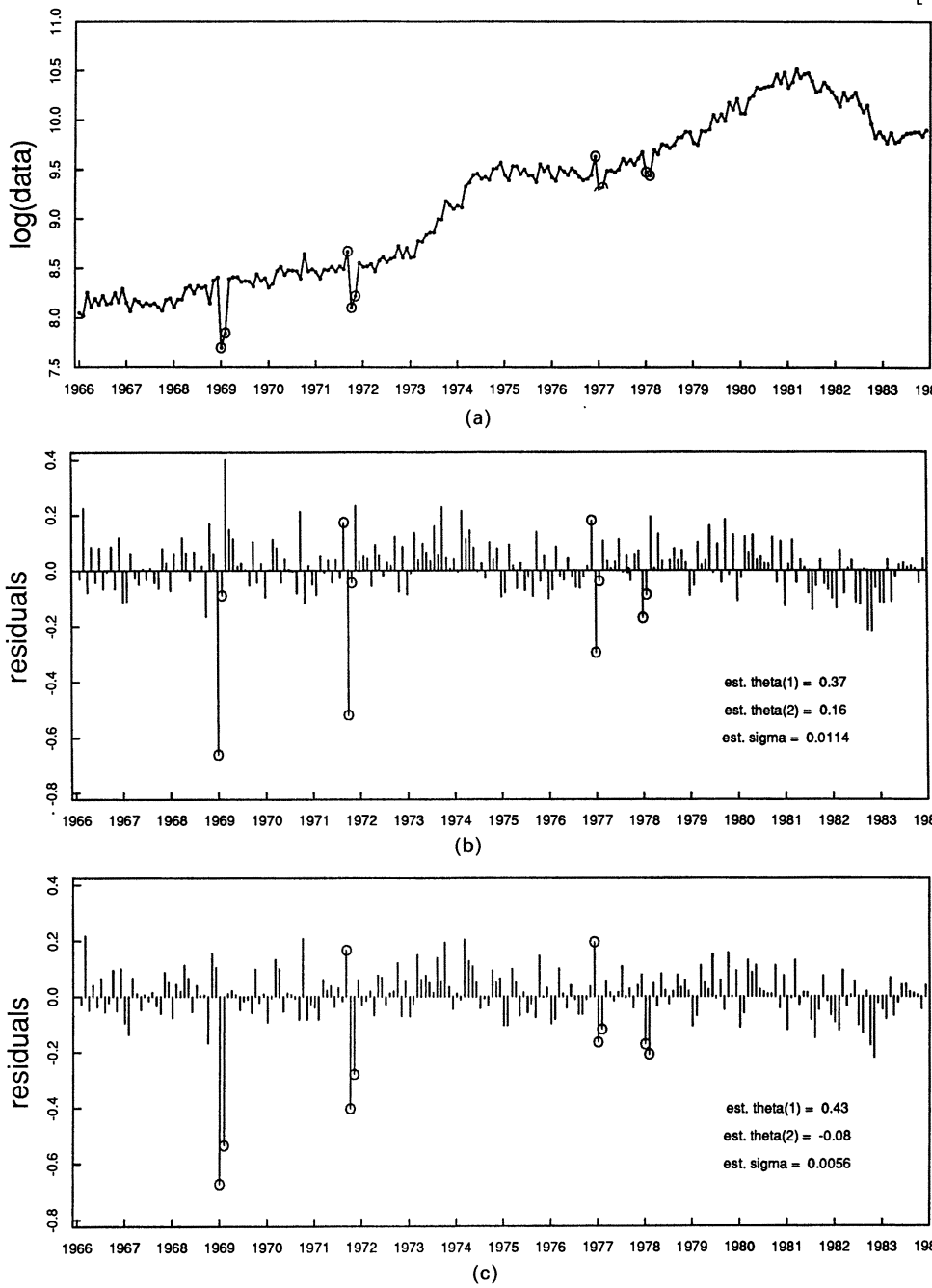

Fig. 7. Example 5, logarithm of exports to Latin-America: (a) plot of data; (b) residuals from ARIMA(0, 1, 2) fit; (c) residuals from ARIMA(0, 1, 2) fit—outliers removed; (d) scaled leave-one-out diagnostics—innovations variance; (e) scaled leave-two-out diagnostics—innovations variance; (f) scaled leave-three-out diagnostics—innovations variance; (g) DV—leave-one-out diagnostics after one round of iterative deletion; (h) DV—leave-two-out diagnostics after one round of iterative deletion; (i) DV—leave-three-out diagnostics after one round of iterative deletion; (j) DV—leave-four-out diagnostics after one round of iterative deletion; (k) DV—leave-one-out diagnostics after two rounds of iterative deletion; (l) DV—leave-two-out diagnostics after two rounds of iterative deletion; (m) DV—leave-three-out diagnostics after two rounds of iterative deletion; (n) DV—leave-four-out diagnostics after two rounds of iterative deletion

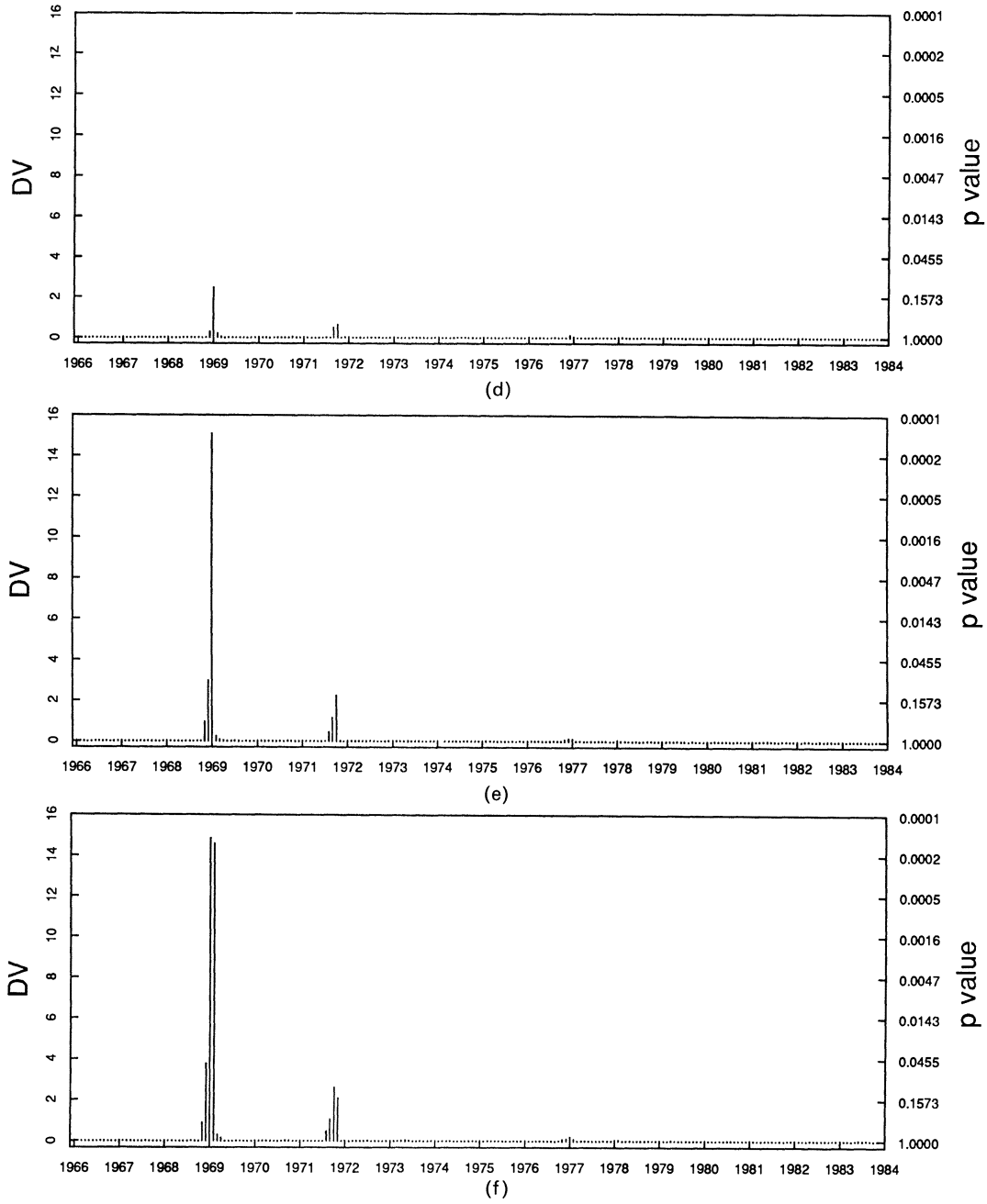

Fig. 7 (*continued*)

Following the strategy of Section 4, we remove the points at 1/69 and 2/69 and recompute the diagnostics for $k = 1, 2, 3, 4$. The results of the first round of iterative deletion are displayed in Figs 7(g)–7(j). Using the guidelines of Section 3, DV identifies the patch 9/71, 10/71 and 11/71 as outliers, with $p < 0.01$. Evidence for including 9/71 as part of the patch is weaker than for 10/71 and 11/71: the increase of DV(3, t) over DV(2, t) for $t = 10/71$ is relatively small. However, the pattern of smearing in

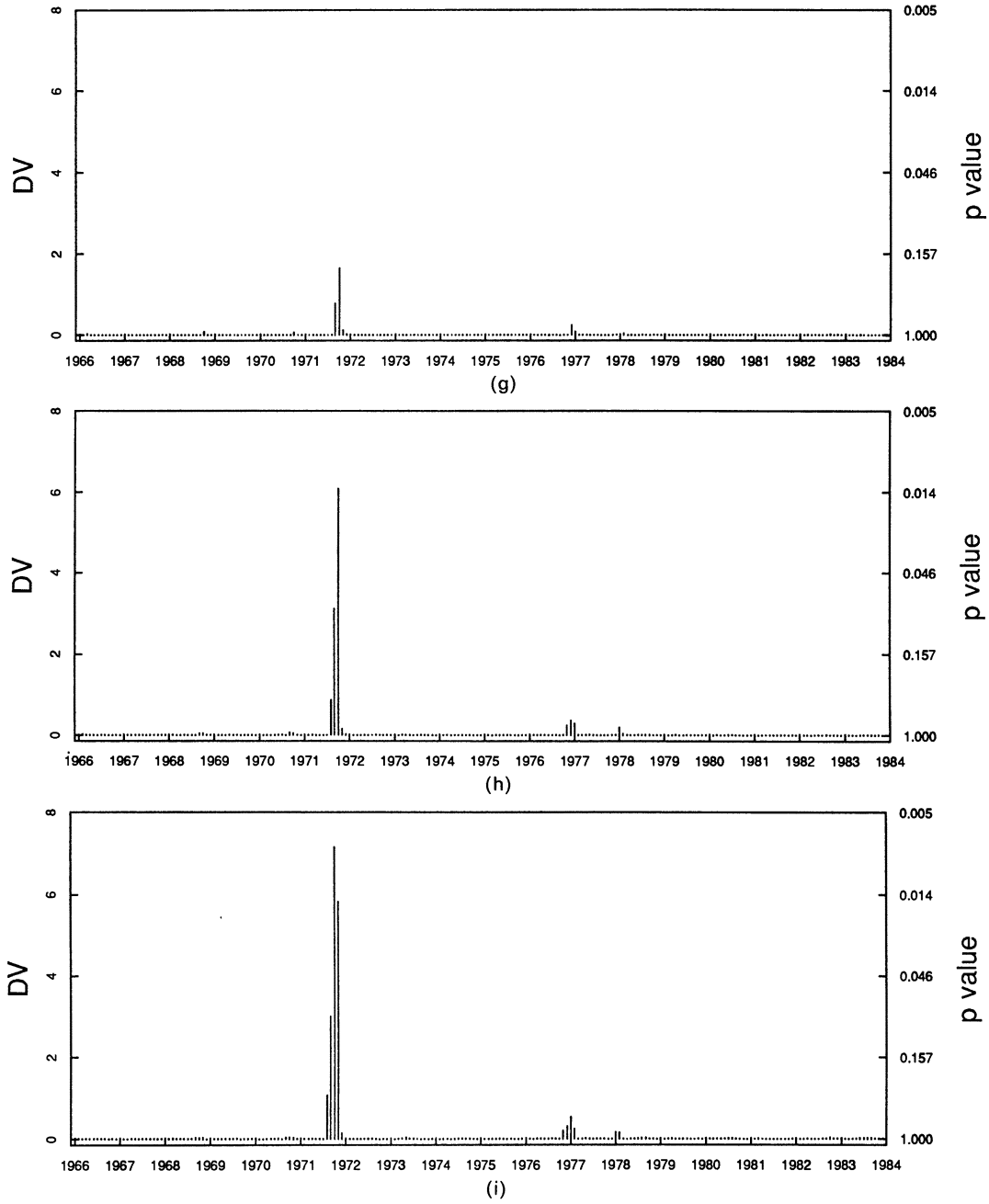

Fig. 7 (*continued*)

Figs 7(h)–7(j) is more consistent with a patch of three outliers than with a patch of two. Hence, these points are removed, and the diagnostics are recomputed.

Leave-k-out diagnostics for a second round of iterative deletion are performed, and an influential patch is identified at 12/76–2/77: see the leave-one-out to leave-four-out diagnostics plotted in Figs 7(k)–7(n). The significance of this patch is much smaller than in previous rounds ($p = 0.29$). Another round of diagnostics is run and an

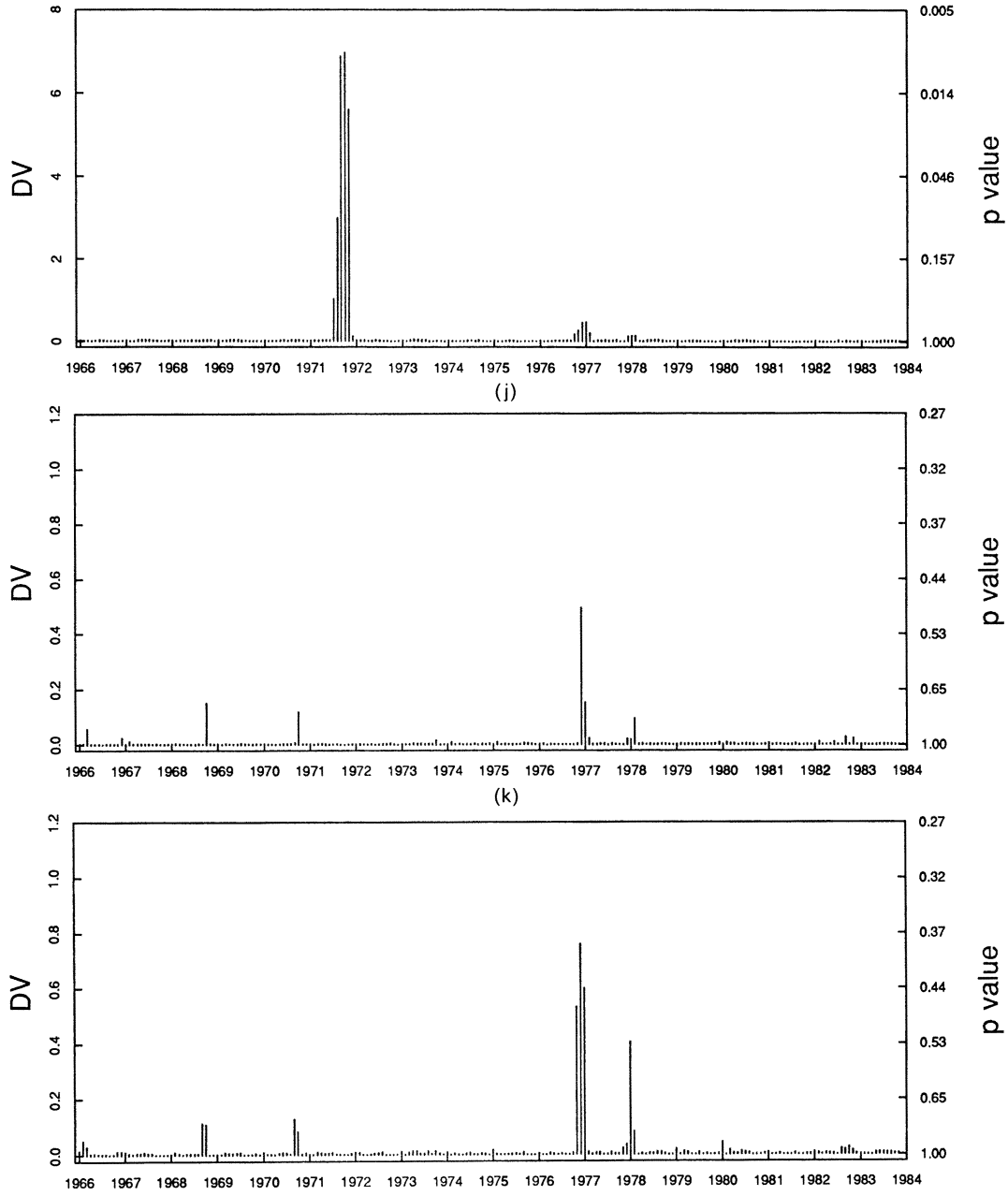

Fig. 7 (*continued*)

influential patch at 1/78–2/78 is detected. The plots are omitted for this round, since the pattern of diagnostics is similar to the previous round: see Figs 7(k)–7(n). One interesting feature is that leave-one-out diagnostics do not pick up the patch (1/78–2/78): we need leave-two-out diagnostics to identify these points as influential.

Other potential outliers are weakly indicated at 10/68 and 10/70 (see Fig. 7(k)); they are associated with fairly high p values of 0.74 and 0.71. These points correspond to

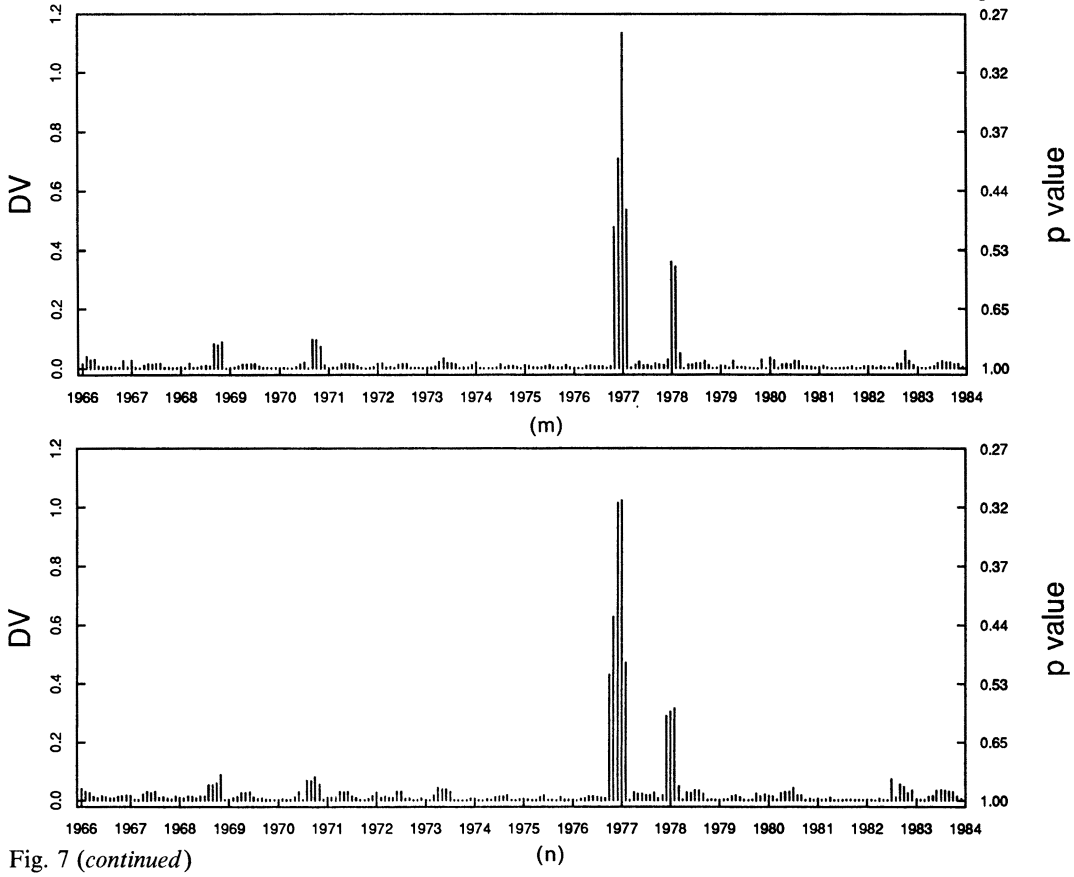

Fig. 7 (*continued*)

moderately large residuals in Figs 7(b) and 7(c), but evidently do not significantly influence the estimate of the innovation variance.

In the final analysis, four groups of outliers were identified and removed using three rounds of iterative deletion. The points which were deleted at each stage, and the corresponding MLEs, are given in Table 1. Removal of the influential points results in a drop in the estimated innovations variance by a factor of 2. The first two groups of outliers at 1/69–2/69 and 9/71–11/71 correspond to dock strikes and forestalling, yielding large negative and positive outliers respectively. The other groups (12/76–2/77 and 1/78–2/78) have no known cause and exert considerably less influence on the estimated parameters. Burman (1985) identified the first two groups as outliers, along with 10/68 and 10/70, using the model-based methodology of Hillmer *et al.* (1983). The points 10/68 and 10/70 show up in the leave-k-out diagnostics, but not so prominently as the other patches at 12/76–2/77 and 1/78–2/78, which were not identified by Burman. Once again, we see the importance of searching for influential patches as well as isolated outliers.

5.1.1. *Analysis of residuals*

Fig. 7(c) gives the residuals based on one-step predictions from the data with the outliers removed, where the predicted values are computed from the MLE estimated with the outliers removed. The general pattern is similar to the original set of residuals

TABLE 1
Parameter fit to export data

Iteration step	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}^2$	*Points deleted*	*p value*
0	0.367	0.160	0.0114	—	—
1	0.460	0.003	0.0083	1/69, 2/69	0.00010
2	0.448	−0.041	0.0066	9/71, 10/71, 11/71	0.0075
3	0.431	−0.058	0.0060	12/76, 1/77, 2/77	0.29
4	0.43	−0.08	0.0056	1/78, 2/78	0.45

(see Fig. 7(b)), but with an important difference: the large residuals in Fig. 7(c) correspond to the points identified as outliers in this analysis. Specifically, that last outlier in each patch, masked in Fig. 7(b), shows up prominently in the residual plot of Fig. 7(c). Correspondingly, the residuals following the patch of outliers, which are large in Fig. 7(b), reveal nothing unusual in Fig. 7(c). Thus, the plot of residuals with the influential data points treated as missing provides a useful graphical display to be compared with the raw residual plot.

5.2. *Example 6: Value of Unfilled Orders, Radio and Television*

Fig. 8(a) displays the monthly value UNFTV (in millions of dollars) of unfilled orders for radios and televisions from 1958 to 1981. This series was previously studied by Martin *et al.* (1983), who used a robust filter to fit an ARIMA$(0, 1, 1) \times (0, 1, 1,)_{12}$ model. The series was also analysed by Engle and Kraft (1983), who fit an autoregressive conditionally heteroscedastic (ARCH) model to the data. Our initial fit is an ARIMA$(0, 1, 1) \times (0, 1, 1)_{12}$ model; the MLEs are given in Table 2 and the residuals from the MLE fit are given in Fig. 8(b).

Examination of Fig. 8(b) shows that the end of series has many more large residuals than the rest of the series. It is quite likely that a variance change may have occurred towards the end of the series, so instead of following the usual procedure we adopt a flexible approach and look for the possibility of a variance shift.

Using the bottom-up approach described in Section 4, we perform leave-k-out diagnostics for $k = 16, 32, 64$. The diagnostics for $k = 64$ are displayed in Fig. 8(c) and dramatically support the conjecture of non-homogeneity of variance in the series:

TABLE 2
Parameter fit to UNFTV data

Time period	*Model*	*Step*	$\hat{\theta}_1$	$\hat{\Theta}_1$	$\hat{\Theta}_2$	$\hat{\sigma}^2$	*Points deleted*
1958–80	$(0, 1, 1) \times (0, 1, 1)_{12}$	—	0.41	0.75	—	3123	—
1958–75	$(0, 1, 1) \times (0, 1, 1)_{12}$	—	0.18	0.92	—	1303	—
1976–80	$(0, 1, 1) \times (0, 0, 2)_6$	0	0.36	−0.25	−0.51	8960	—
	$(0, 1, 1) \times (0, 0, 2)_6$	1	0.49	−0.46	−1.00	5340	2/78–5/78
	$(0, 1, 1) \times (0, 0, 2)_6$	2	0.36	−0.46	−1.00	4173	9/78

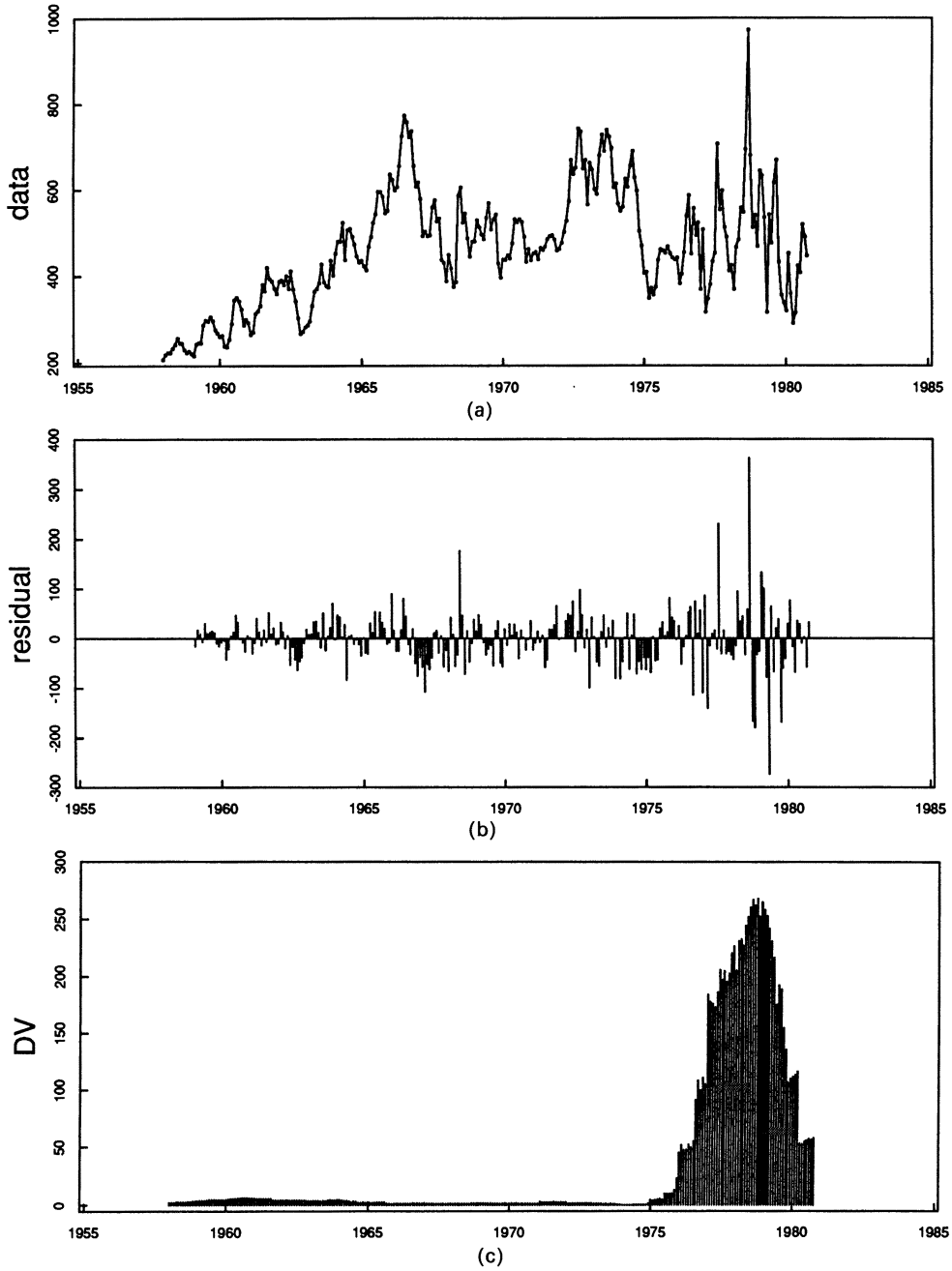

Fig. 8. Example 6, unfilled radio and television orders: (a) plot of data; (b) plot of residuals; (c) scaled leave-64-out diagnostics—innovations variance; (d) scaled leave-eight-out diagnostics—innovations variance; (e) scaled leave-one-out diagnostics—innovations variance; (f) scaled leave-four-out diagnostics—innovations variance

the maximum value for DV is over 250. The diagnostics for $k = 16$ and $k = 32$ (not shown) display a similar pattern, although achieving a smaller maximum value. It is clear that the behaviour of this series is fundamentally different towards the right-hand end of the data. The data are split into two series, and each part is analysed

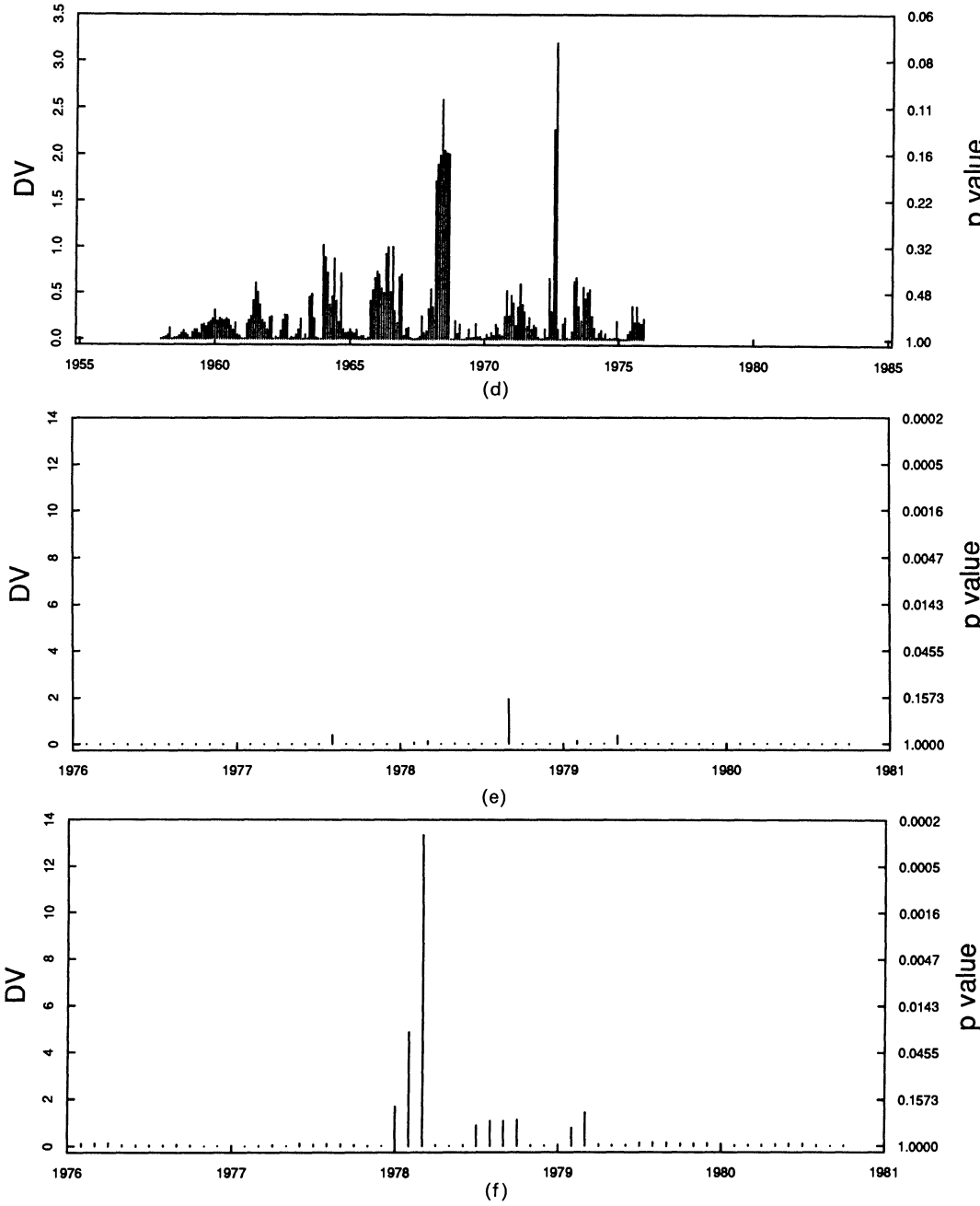

Fig. 8 (*continued*)

separately. We choose 1/76 as the change point, based on the residuals plot and on computation of DV for a few judiciously selected subsets. Specifically, patches of increasing size were truncated from the right-hand end of the data, and the data were split (approximately) according to the patch of maximum influence.

Checking the model order for the first part of the series again yielded an ARIMA(0, 1, 1) $\times$ $(0, 1, 1)_{12}$ model. The MLEs for this model are given in Table 2. Note the reduction in the estimated innovations variance from 3123 to 1303.

However, running leave-k-out diagnostics on this portion of the data reveals more problems. Leave-eight-out diagnostics displayed in Fig. 8(d) indicate several patches of influence. Two patches are especially prominent: one during 1968 and another in 1972. The patch in 1972, which shows up only in the leave-eight-out diagnostics, is associated with what appears to the eye to be a local level shift spanning from 5/72 to 10/74 (see Fig. 8(a)). The patch in 1968 corresponds to a large residual at 6/68 and has a less well-defined structure. It might be that there is a local level shift during 11/67–5/68 or during 6/68–10/69, or both. In any case, the diagnostics help us to identify problem areas in the data and show that the large residual is associated with a patch of influential points rather than an isolated outlier.

An ARIMA(0, 1, 1) $\times$ $(0, 0, 2)_6$ model was fitted to the second part of the series, though the short time span makes model selection difficult. A summary of the iterative deletion procedure is given in Table 2. The leave-one-out diagnostics, plotted in Fig. 8(e), pick up the isolated outlier at 9/78 ($p = 0.17$), which was prominent in the residual plot (see Fig. 8(b)). However, leave-k-out diagnostics for $k = 2, 3, 4$ reveal a highly influential patch at 2/78–5/78 ($p < 0.003$): see the plot for leave-four-out diagnostics in Fig. 8(f). This patch is not associated with any large residual and was not detected by leave-one-out diagnostics! Removal of the patch 2/78–5/78 has a dramatic effect on the fit, which may indicate that the model is poor. Another round of diagnostics (not shown) identified 9/78 as an influential point, though its removal primarily affects just the estimated innovations variance.

In summary, the UNFTV series shows the effectiveness of the diagnostics at revealing different types of influential points, and not just outliers. Also, this example shows the need for adopting a flexible data analytic approach, and it demonstrates, once again, the importance of looking for influential patches.

Incidentally, with regard to dealing with the variance shift in the last part of the series, the ARCH modelling approach of Engle and Kraft (1983) appears to be a viable alternative.

6. MORE ABOUT DV

6.1. *Decomposition*

An outlier can inflate the estimated innovations variance for two reasons: first, it can inflate the variance by distorting the parameter estimates of the ARIMA coefficients, and hence the fit, and secondly the outlier can inflate the variance because of an associated large residual. To reflect the first component define

$$\text{DFIT}(A) = (n - k)\left\{\frac{\hat{\sigma}_A^2(\hat{\alpha})}{\hat{\sigma}_A^2(\hat{\alpha}_A)} - 1\right\} \tag{6.1}$$

where $\hat{\sigma}_A^2(\alpha)$ is the marginal MLE of σ^2 with subset A deleted for an arbitrary *fixed* α. Note that $\hat{\sigma}_A^2 = \hat{\sigma}_A^2(\hat{\alpha}_A)$, where $\hat{\sigma}_A^2$ is a component of the joint MLE $(\hat{\sigma}_A^2, \hat{\alpha}_A)$, with subset A deleted.

Intuitively, DFIT reflects the change in fit, as measured by the innovations variance estimate, solely due to the change in the ARIMA coefficients. DFIT is not directly

influenced by a large residual caused by any outliers in a patch A, since it is based on an estimate of σ^2 with patch A removed. However, DFIT may be large if the estimated coefficients change substantially. In fact, it can be shown that asymptotically DFIT depends on $\mathbf{x}^n$ only through the empirical influence function $\mathbf{EIC}(A)$ (Bruce, 1988).

We now obtain a measure for the second component in a way which will yield a convenient decomposition of DV. Let $\mathbf{x}_A$ be a vector with elements x_t, $t \in A$ where A has k elements, and let $\mathbf{x}^n_{A^c}$ be a vector with elements x_t, $t = 1, 2, \ldots, n$, $t \notin A$. Given $\hat{\boldsymbol{\alpha}}$, the conditional expectation 'interpolate' $\hat{\mathbf{x}}^n_A = E(\mathbf{x}_A \mid \hat{\boldsymbol{\alpha}}, \mathbf{x}^n_{A^c})$ is the minimum mean-squared error estimate of $\mathbf{x}_A$ based on $\mathbf{x}^n_{A^c}$. The notation is meant to indicate that the expectations are for a process x_t generated by $\hat{\boldsymbol{\alpha}}$. The corresponding vector of interpolation residuals and their covariance matrix are

$$\begin{aligned} \hat{\mathbf{e}}^n_A &= \mathbf{x}_A - \hat{\mathbf{x}}^n_A \\ \sigma^2 \boldsymbol{\Sigma}^n_A &= \operatorname{cov}(\hat{\mathbf{e}}^n_A \mid \hat{\boldsymbol{\alpha}}, \sigma^2, \mathbf{x}^n_{A^c}). \end{aligned} \tag{6.2}$$

Define

$$\operatorname{DRES}(A) = (\hat{\mathbf{e}}^n_A)^{\mathrm{T}} \{\hat{\sigma}^2_A(\hat{\boldsymbol{\alpha}}_A)\boldsymbol{\Sigma}^n_A\}^{-1}(\hat{\mathbf{e}}^n_A). \tag{6.3}$$

DRES is a norm of the interpolated residuals $\hat{\mathbf{e}}^n_A$, which involves not only $\hat{\boldsymbol{\alpha}}_A$ but also $\hat{\boldsymbol{\alpha}}$ through equation (6.2). DRES is large whenever the interpolated residuals are large relative to their variance–covariance matrix based on $\hat{\boldsymbol{\alpha}}$. This happens, for example, when the series contains a single additive outlier at time t and DRES is evaluated at the outlier position $A = t$, since the interpolated value $\hat{\mathbf{x}}^n_t$ is based on outlier-free data. Similarly, if a patch A of additive outliers is present, then $\hat{\mathbf{e}}^n_A$ will be large.

A slight change in the definition of DV yields a linear decomposition in terms of DFIT and DRES. Define

$$\mathrm{DV}_0 = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_A} - 1\right)\sqrt{\frac{n}{2}} \tag{6.4}$$

so that $\mathrm{DV} = \mathrm{DV}_0^2$. Then the following theorem holds.

Theorem 1.

$$\mathrm{DV}_0(A) = \frac{1}{\sqrt{(2n)}}\{\operatorname{DRES}(A) - k + \operatorname{DFIT}(A)\}. \tag{6.5}$$

Proof. Assume $\mathbf{x}^n$ behaves according to a stationary ARMA process (the proof for ARIMA processes is virtually identical). By rearranging terms, we can write

$$n\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2_A} - 1\right) = \frac{1}{\hat{\sigma}^2_A}\{n\hat{\sigma}^2 - (n-k)\hat{\sigma}^2_A(\hat{\boldsymbol{\alpha}})\} - k + \operatorname{DFIT}(A) \tag{6.6}$$

where k is the number of elements in A. Now

$$\hat{\sigma}^2_A(\hat{\boldsymbol{\alpha}}) = \frac{1}{n-k}\sum_{\substack{t=1 \\ t \notin A}}^{n} \{\hat{\mathbf{e}}^{t-1}_t(A)\}^2 / f^{t-1}_t(A)$$

where $\hat{\mathbf{e}}_t^{t-1}(A)$ and $f_t^{t-1}(A)$ are the one-step prediction residuals and their variances for an x_t process generated by $\hat{\boldsymbol{\alpha}}$ and innovations variance σ^2 with subset A missing:

$$\hat{\mathbf{e}}_t^{t-1}(A) = x_t - E(x_t \mid \hat{\boldsymbol{\alpha}}, \mathbf{x}_{A^c}^{t-1})$$

and

$$\sigma^2 f_t^{t-1}(A) = \operatorname{var}(x_t \mid \hat{\boldsymbol{\alpha}}, \sigma^2, \mathbf{x}_{A^c}^{t-1}).$$

In the case that A is null, the notation $\hat{\mathbf{e}}_t^{t-1}$ and f_t^{t-1} is used. The first term in equation (6.6) may be written as

$$\frac{1}{\hat{\sigma}_A^2}\left[\sum_{t=1}^{n} (\hat{\mathbf{e}}_t^{t-1})^2/f_t^{t-1} - \sum_{\substack{t=1 \\ t\notin A}}^{n} \{\hat{\mathbf{e}}_t^{t-1}(A)\}^2/f_t^{t-1}(A)\right]. \tag{6.7}$$

But

$$\log p(\mathbf{x}^n \mid \hat{\boldsymbol{\alpha}}, \hat{\sigma}_A^2) - \log p(\mathbf{x}_{A^c}^n \mid \hat{\boldsymbol{\alpha}}, \hat{\sigma}_A^2) = \log p(\mathbf{x}_A \mid \hat{\boldsymbol{\alpha}}, \hat{\sigma}_A^2, \mathbf{x}_{A^c}^n) \tag{6.8}$$

where

$$p(\mathbf{x}_A \mid \hat{\boldsymbol{\alpha}}, \hat{\sigma}_A^2, \mathbf{x}_{A^c}^n) = N_k(\hat{\mathbf{e}}_A^n, \hat{\sigma}_A^2 \boldsymbol{\Sigma}_A^n) \tag{6.9}$$

with $N_k(\cdot)$ a standard k-variate normal density. Using equations (6.8) and (6.9) in equation (6.7) shows that the first term in equation (6.6) is equal to DRES(A), which yields equation (6.5). □

6.1.1. *Remarks*

(a) From equations (6.8) and (6.9), the likelihood of $\mathbf{x}^n$ depends on $\mathbf{x}_A$ only through DRES(A). This gives us another way to interpret the significance of DRES.

(b) For an uncontaminated Gaussian ARIMA process, DRES(A) is asymptotically distributed as a chi-squared random variable on k degrees of freedom. Since DRES asymptotically dominates DFIT (see Section 6.2), the factor k in equation (6.5) ensures that

$$E\left[\lim_{n\to\infty} \mathrm{DV}_0(A)\sqrt{(2n)}\right] = 0.$$

(c) A more powerful diagnostic DRES(A) could be defined by evaluating the interpolated residuals and covariance of equations (6.2) using $\hat{\boldsymbol{\alpha}}_A$ instead of $\hat{\boldsymbol{\alpha}}$. However, to obtain a decomposition of DV, the definition of DFIT in equation (6.1) would have to be changed, resulting in a much less powerful diagnostic for the change in fit as measured by the innovations variance estimate.

6.2. *Dominance of DRES over DFIT*

The following result shows that DRES asymptotically dominates DFIT.

Proposition.

$$\mathrm{DV}_0(A)\sqrt{(2n)} = \mathrm{DRES}(A) - k + o_p(1).$$

Proof. The proof is based on a Taylor series expansion of $\hat{\sigma}_A^2(\hat{\boldsymbol{\alpha}})$ about $\hat{\boldsymbol{\alpha}}_A$:

$$(n - k)\{\hat{\sigma}_A^2(\hat{\boldsymbol{\alpha}}) - \hat{\sigma}_A^2\} = (n - k)(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_A)^{\mathrm{T}} \frac{\partial}{\partial \boldsymbol{\alpha}} \hat{\sigma}_A^2(\boldsymbol{\alpha})\Big|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_A} + o_{\mathrm{p}}(1). \tag{6.10}$$

Since the asymptotic value of $\sigma_A^2(\boldsymbol{\alpha})$ is minimized by the true parameter value $\boldsymbol{\alpha}_0$, and

$$\hat{\boldsymbol{\alpha}}_A \xrightarrow{\mathrm{p}} \boldsymbol{\alpha}_0,$$

we have

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \hat{\sigma}_A^2(\boldsymbol{\alpha})\Big|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_A} = o_{\mathrm{p}}(1).$$

Since $(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_A) = O_{\mathrm{p}}(n^{-1})$, it follows that

$$\begin{aligned} \mathrm{DFIT}(A) &= \frac{1}{\hat{\sigma}_A^2}(n - k)\{\hat{\sigma}_A^2(\hat{\boldsymbol{\alpha}}) - \hat{\sigma}_A^2\} \\ &= O_{\mathrm{p}}(1)\{O_{\mathrm{p}}(1)o_{\mathrm{p}}(1) + o_{\mathrm{p}}(1)\} \\ &= o_{\mathrm{p}}(1). \end{aligned}$$

However, clearly $\mathrm{DRES}(A) = O_{\mathrm{p}}(1)$, and the result follows. □

6.2.1. *Remarks*

The fact that $\mathrm{DV}_0\sqrt{(2n)}$ *is asymptotically equal to* $\mathrm{DRES} - k$ helps to explain why DV is so effective in detecting outliers. Despite the asymptotic dominance of DRES over DFIT, we have found that in finite samples DFIT is often non-negligible.

Fox (1972) proposed a likelihood ratio test (LRT) procedure for testing the presence of a single AO-type outlier at a *known* time t in AR models. The Fox test has been used as the basis for the ARIMA model estimation scheme of Chang and Tiao (1983); see also Hillmer *et al.* (1983) and Tsay (1986). It is well known that the LRT is asymptotically equivalent to a test based on the Studentized interpolation residual $\mathrm{DRES}(t)$ (see Fox (1972)). This result can be extended to the LRT for a subset A of additive outliers (see Bruce (1988)). By this result and the above proposition, a test based on $\mathrm{DV}_0(A)$ is asymptotically equivalent to the LRT for the presence of a subset AO at times $t \in A$.

6.3. *Diagnostics Based on Decomposition of DV*

The decomposition (6.5) shows how the effects of an outlier on DV can be separated into two sources: DRES, which is a measure of the size of interpolation residual, and DFIT, which reflects the change in fit. This suggests a graphical display, i.e. plotting DV_0 as DV is plotted in this paper and superimposing a line plot of DFIT. This shows both the total change in the estimated innovations variance and the change due to the change in fit.

$\mathrm{DRES}(A) - k$ and DFIT, and hence $\mathrm{DV}_0(A)$, can be negative. However, in practice all 'significant' values of DV_0 are positive. This is because deletion of a small patch of points can lead at most to a moderate increase in the estimate of σ^2 (but an almost arbitrarily large decrease).

An attractive alternative to the use of DFIT, suggested by Yohai (1987), is to base a diagnostic for fit on a highly robust (i.e. breakdown point $\frac{1}{2}$) estimate of the scale of the residuals. Computing the change in a highly robust estimate of scale for the residuals from the leave-k-out fit will clearly provide a measure primarily of how the fit is influenced by a patch of observations.

7. LIKELIHOOD AND MAXIMUM LIKELIHOOD ESTIMATES FOR MISSING DATA AND ADDITIVE OUTLIERS MODELS

The 'fixed magnitude' AO model (3.1), where the outlier times z_t are presumed *fixed* and *known*, has played a prominent role in ARIMA model estimation procedures. The parameter vector $\boldsymbol{\zeta}_A = (\zeta_{t_1}, \ldots, \zeta_{t_k})$ is used to model outliers at k time points $A = \{t_1, \ldots, t_k\}$. For example, in intervention analysis (Box and Tiao, 1975), the time points A are known and typically come in patches. A special case of this model, where $k = 1$, is used in an iterative manner by Chang and Tiao (1983), Hillmer *et al.* (1983) and Tsay (1986) to handle estimation in the presence of outliers.

We show in theorem 2 that *the likelihood function concentrated* with respect to $\boldsymbol{\zeta}_A$ is related in a very simple way to the likelihood which treats subset A as missing; the two likelihoods differ only by a factor which does not depend on the data. Thus theorem 2 clarifies the distinction between estimation with missing data and estimation using the special AO model (3.1).

7.1. *Relating Likelihood Functions*

Assume that x_t is an uncontaminated ARIMA process and we observe y_t, where

$$y_t = \begin{cases} x_t & \text{if } t \notin A \\ x_t + \zeta_t & \text{if } t \in A. \end{cases} \tag{7.1}$$

Then the likelihood for the AO intervention approach is formed by treating $\boldsymbol{\zeta}_A$ as a parameter vector to be estimated along with $\boldsymbol{\alpha}$ and σ^2 and computing the usual ARIMA model likelihood for x_t.

Let $\mathbf{y}^n$ be the vector $(y_1, y_2, \ldots, y_n)$, $\mathbf{y}_A$ be the vector with elements y_t, $t \in A$, and $\mathbf{y}^n_{A^c}$ be the vector with elements y_t, $t = 1, 2, \ldots, n$, $t \notin A$. Denote the AO model likelihood function for $\mathbf{y}^n$ by $p(\mathbf{y}^n \mid \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\zeta}_A)$ and denote the missing data likelihood for $\mathbf{y}^n_{A^c}$ by $p(\mathbf{y}^n_{A^c} \mid \boldsymbol{\alpha}, \sigma^2)$, where the parameter $\boldsymbol{\zeta}_A$ is dropped since $\mathbf{y}^n_{A^c}$ does not depend on $\boldsymbol{\zeta}_A$. Let $\hat{\mathbf{x}}^n_A(\boldsymbol{\alpha})$ be the conditional mean interpolate of $\mathbf{x}_A = \mathbf{y}_A - \boldsymbol{\zeta}_A$ given $\boldsymbol{\alpha}$, and let $\sigma^2 \boldsymbol{\Sigma}^n_A(\boldsymbol{\alpha})$ be the associated interpolation error covariance matrix.

Theorem 2. Suppose that y_t behaves according to equations (7.1). Then the MLE of $\boldsymbol{\zeta}_A$ given $\boldsymbol{\alpha}$ and σ^2 is the interpolation residual vector

$$\tilde{\boldsymbol{\zeta}}_A = \mathbf{y}_A - \hat{\mathbf{x}}^n_A(\boldsymbol{\alpha}) \tag{7.2}$$

and

$$p(\mathbf{y}^n \mid \boldsymbol{\alpha}, \sigma^2, \tilde{\boldsymbol{\zeta}}_A) = \left(\frac{1}{2\pi}\right)^{k/2} \left(\frac{1}{|\sigma^2 \boldsymbol{\Sigma}^n_A|}\right)^{1/2} p(\mathbf{y}^n_{A^c} \mid \boldsymbol{\alpha}, \sigma^2). \tag{7.3}$$

Proof. The likelihood for $\mathbf{y}^n$ can be factored as follows:

$$\begin{aligned} p(\mathbf{y}^n \mid \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\zeta}_A) &= p(\mathbf{y}_A \mid \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\zeta}_A, \mathbf{y}^n_{A^c})\, p(\mathbf{y}^n_{A^c} \mid \boldsymbol{\alpha}, \sigma^2) \\ &= p(\mathbf{y}_A - \boldsymbol{\zeta}_A \mid \boldsymbol{\alpha}, \sigma^2, \mathbf{y}^n_{A^c})\, p(\mathbf{y}^n_{A^c} \mid \boldsymbol{\alpha}, \sigma^2). \end{aligned} \tag{7.4}$$

Since the distribution of $\mathbf{y}_a - \boldsymbol{\zeta}_A$, given $\mathbf{y}^n_{A^c}$, for fixed $\boldsymbol{\alpha}$, σ^2 is $N_k(\hat{\mathbf{x}}^n_A(\boldsymbol{\alpha}), \sigma^2 \boldsymbol{\Sigma}^n_A(\boldsymbol{\alpha}))$, it is clear that equation (7.4) is maximized with respect to $\boldsymbol{\zeta}_A$ by equation (7.2), and concentrating $\boldsymbol{\zeta}_A$ out of equation (7.4) yields equation (7.3). □

We can obtain the following corollary, which shows that the difference between the MLEs of α and σ^2 obtained under the two approaches is $O_p(n^{-1})$ and converges to a constant.

Corollary 1. Let $(\tilde{\sigma}^2_A, \tilde{\boldsymbol{\alpha}}_A, \tilde{\boldsymbol{\zeta}}_A)$ be the joint MLE of $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\zeta}_A)$ for the AO model given $\mathbf{y}^n$, and let $(\hat{\sigma}^2_a, \hat{\boldsymbol{\alpha}}_A)$ be the joint MLE of $(\sigma^2, \boldsymbol{\alpha})$ for the missing data model. Under suitable regularity conditions, the difference in the MLEs of the true parameters σ^2 and $\boldsymbol{\alpha}$ is asymptotically

$$(n - k)(\tilde{\sigma}^2_A - \hat{\sigma}^2_A) = -k\sigma^2 + o_p(1) \tag{7.5}$$

and

$$(n - k)(\tilde{\boldsymbol{\alpha}}_A - \hat{\boldsymbol{\alpha}}_A) = -\tfrac{1}{2}\mathbf{I}(\boldsymbol{\alpha})\,\mathbf{c}(\boldsymbol{\alpha}) + o_p(1) \tag{7.6}$$

where $\mathbf{I}(\boldsymbol{\alpha})$ is the asymptotic information matrix for $\boldsymbol{\alpha}$ and

$$\mathbf{c}(\boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \log |\boldsymbol{\Sigma}^n_A(\boldsymbol{\alpha})|.$$

Proof. See Bruce (1988).

8. CONCLUDING COMMENTS

8.1. *Alternatives for Scaling*

The scaling used for the diagnostics has a possible shortcoming: both DV and DC converge to zero as the sample size increases to infinity. This is also true (under certain conditions) for the usual regression diagnostics, such as Cook's distance. Thus, these statistics fail to have a χ^2 distribution asymptotically as well as for finite samples. The fact that DV and DC tend to zero as $n \to \infty$ shows empirically, since we find that DV and DC tend to uncover proportionally more influential points in smaller data sets. For some applications, perhaps this is as it should be: relative to the (approximate) confidence intervals for $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$, an individual point tends to exert more influence in a small data set than in a large data set.

An alternative approach would be to scale DC and DV to obtain an asymptotically non-degenerate random variable, and then to determine approximate significance levels for testing whether the diagnostic is significant. This is similar in spirit to the approach adopted by Fox (1972) and Chang and Tiao (1983). One problem with this approach is choosing the correct significance level, since the leave-k-out procedure involves many dependent 'tests'. See Atwood (1987) for an approximation to determine the overall significance level of many dependent tests in the time series setting.

A more fundamental problem is that the hypothesis testing framework is not appropriate for diagnostics (see Hampel *et al.* (1986)). The reference distribution is constructed under the null hypothesis that no outliers are present, so a Neyman–Pearson testing approach would protect us in precisely the situation for which no protection is needed! The diagnostics are useful primarily in a *data analytic* context, and attempting to establish overall significance levels is likely to be misleading and fruitless.

The problem of finding an appropriate reference distribution remains a topic of controversy for regression diagnostics (Chatterjee and Hadi, 1986). There is no

absolute reference distribution which works well for all problems. Perhaps the most sensible approach is to combine an external reference distribution (such as that used in this paper) with an internal reference distribution based on exploratory data analysis of the diagnostic values (Welsh, 1982). These issues are not fully resolved, and deserve further investigation.

8.2. *Diagnostics versus Robust Filter and Smoother Cleaners*

Diagnostics can be obtained from robust model fitting based on robust filter cleaners or robust smoother cleaners (see Martin (1979, 1981), Martin and Thompson (1982) and Martin *et al.* (1983)). The robust model fitting diagnostics consist of looking for large values of the observation prediction residuals $\hat{e}_t^{t-1} = x_t - \hat{x}_{t|t-1}$, where $\hat{x}_{t|t-1}$ denotes the one-step-ahead robust prediction based on filter- or smoother-cleaned data. There is a close connection between the prediction residuals produced by leave-k-out diagnostics and those produced by a rejection-type filter cleaner: the diagnostics obtained from the leave-k-out procedure will often closely match those resulting from a robust procedure based on the rejection filter (see Bruce and Martin (1987)).

8.3. *Other Related Work*

Another important direction for dealing with model changes of various types has been pursued by Harrison and Stevens (1976) and Smith and West (1983), who use a Bayesian approach. In their approach a mixture of normals is used to adapt the model automatically to outliers and other local structures. West (1986) and West *et al.* (1985) propose a somewhat different method based on Bayes factors, in which a nominal model is compared with alternative models.

Other approaches to the problems of outliers and structural disturbances in time series have been explored in the literature. An approach proposed by Chang and Tiao (1983), Hillmer *et al.* (1983) and Tsay (1986) is based on iterative fitting of ARIMA models, utilizing Fox (1972) tests to decide whether an individual observation is an IO, AO or not an outlier. The approach is easily extended to cover shifts in level and shifts in variance.

ACKNOWLEDGEMENTS

The authors wish to thank William Bell, David Findley and Don Percival for helpful discussions during the preparation of this paper. This research was supported by Office of Naval Research contracts N00014-84-C-0169 and N00014-88-K-0265. The work benefited from Dr Bruce's support as a visiting intern at the US Bureau of the Census for three months.

APPENDIX A: COMPUTATION OF ASYMPTOTIC INFORMATION MATRIX

This section derives an analytical expression for the asymptotic information matrix of a stationary and invertible ARMA(p, q) process $\Phi(B)x_t = \Theta(B)\varepsilon_t$. The formulae are easily extended to non-stationary and seasonal models. Let $g_1, \ldots, g_p$ and $h_1, \ldots, h_q$ be the roots of the polynomials $\Phi(B)$ and $\Theta(B)$, so that

$$\begin{aligned} \Phi(B) &= (1 - g_1 B)(1 - g_2 B) \cdots (1 - g_p B), \\ \Theta(B) &= (1 - h_1 B)(1 - h_2 B) \cdots (1 - h_1 B). \end{aligned} \tag{A.1}$$

Assume that the roots are distinct: $g_i \neq g_j$ and $h_i \neq h_j$ for $i \neq j$.

Let c_i and d_i be the coefficients in the expansion of $\Phi(B)^{-1}$ and $\Theta(B)^{-1}$ respectively, i.e.

$$\begin{aligned} \Phi^{-1}(B) &= \sum_{i=0}^{\infty} c_i B^i, \\ \Theta^{-1}(B) &= \sum_{i=0}^{\infty} d_i B^i. \end{aligned} \tag{A.2}$$

The asymptotic information matrix $\mathbf{I}(\alpha)$ is given by

$$\left.\begin{aligned} I_{i,j} &= \sum_{k=0}^{\infty} c_k c_{k+j-i} && \text{if } 1 \leqslant i \leqslant j \leqslant p \\ I_{i,p+j} &= -\sum_{k=0}^{\infty} d_k c_{k+j-i} && \text{if } 1 \leqslant i \leqslant p,\ 1 \leqslant j \leqslant q,\ i \leqslant j \\ I_{i,p+j} &= -\sum_{k=0}^{\infty} c_k d_{k+j-i} && \text{if } 1 \leqslant i \leqslant p,\ 1 \leqslant j \leqslant q,\ j \leqslant i \\ I_{p+i,p+j} &= \sum_{k=0}^{\infty} d_k d_{k+j-i} && \text{if } 1 \leqslant i \leqslant j \leqslant q. \end{aligned}\right\} \tag{A.3}$$

The coefficients can be computed recursively from the relation $1 = \Phi(B)\Phi(B)^{-1} = \Theta(B)\Theta(B)^{-1}$, or

$$\Phi(B)c_t = 0 \qquad \Theta(B)d_t = 0 \qquad t = 1, 2, \ldots . \tag{A.4}$$

Initial conditions for the recursions are $c_0 = 1$, $c_{-p+1} = \cdots = c_{-1} = 0$ and $d_0 = 1$, $d_{-p+1} = \cdots = d_{-1} = 0$. Hence, equations (A.3) provide an explicit expression for $\mathbf{I}(\alpha)$.

By expressing equations (A.3) in terms of the roots $\Phi(B)$ and $\Theta(B)$, the formulae can be reduced to a summation of a finite number of terms. For $t = 1, 2, \ldots$, a solution to equations (A.4) is given by (see Box and Jenkins (1976))

$$\begin{aligned} c_t &= k_1 g_1^t + \cdots + k_p g_p^t \\ d_t &= l_1 h_1^t + \cdots + l_q h_q^t \end{aligned} \tag{A.5}$$

where $k_1, \ldots, k_p$ and $l_1, \ldots, l_q$ are (possibly complex-valued) constants. Since c_t and d_t can be evaluated recursively, equations (A.5) define a system of linear equations which can be solved for $k_1, \ldots, k_p$ and $l_1, \ldots, l_q$. Substituting equations (A.5) into equations (A.3), Fubini's theorem yields

$$\begin{aligned} I_{i,j} &= \sum_{t=0}^{\infty} \left(\sum_{m=1}^{p} k_m g_m^{t+j-i} \sum_{n=1}^{p} k_n g_n^t \right) \\ &= \sum_{m=1}^{p} \sum_{n=1}^{p} \left\{ k_m k_n g_m^{j-i} \sum_{t=0}^{\infty} (g_m g_n)^t \right\} \\ &= \sum_{m=1}^{p} \sum_{n=1}^{p} k_m k_n g_m^{j-i} (1 - g_m g_n)^{-1} \qquad \text{if } i \leqslant j. \end{aligned} \tag{A.6a}$$

Similarly,

$$I_{i,p+j} = \sum_{m=1}^{p} \sum_{n=1}^{q} k_m l_n g_m^{j-i} (1 - g_m h_n)^{-1} \qquad \text{if } i \leqslant j, \tag{A.6b}$$

$$I_{i,p+j} = \sum_{m=1}^{p} \sum_{n=1}^{q} k_m l_n h_n^{j-i} (1 - g_m h_n)^{-1} \qquad \text{if } j \leqslant i \tag{A.6c}$$

and

$$I_{p+i,p+j} = \sum_{m=1}^{q}\sum_{n=1}^{q} k_m l_n h_m^{j-i}(1 - h_m h_n)^{-1} \qquad \text{if } i \leqslant j. \tag{A.6d}$$

REFERENCES

Atkinson, A. C. (1985) *Plots, Transformations, and Regressions: an Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Clarendon.

Atwood, C. L. (1987) The size of a test for a family of time series interventions. *Technical Report*. Idaho National Engineering Laboratory, EGGG Idaho Inc., Idaho Falls.

Bell, W., and Hillmer, S. (1987) Initializing the Kalman filter in the nonstationary case *Report CENSUS/SRD/RR-87/33*. Statistical Research Division, US Bureau of the Census, Washington DC.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics*. New York: Wiley.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control*, 2nd edn. Oakland: Holden-Day.

Box, G. E. P. and Tiao, G. C. (1975) Intervention analysis with applications to economic and environmental problems. *J. Amer. Statist. Ass.*, **70**, 70–79.

Brillinger, D. R. (1966) Discussion on Linear functional relationships (by P. Sprent). *J. R. Statist. Soc.* B, **28**, 294.

Bruce, A. G. (1988) *PhD Thesis*. University of Washington, Seattle.

Bruce, A. G. and Martin, R. D. (1987) Leave-*k*-out diagnostics for time series. *Technical Report 107*. Department of Statistics, University of Washington, Seattle.

Burman, J. P. (1985) *Report on ASA Fellowship 1984–5*. Statistical Research Division, US Bureau of the Census, Washington DC.

Chang, I. and Tiao, G. C. (1983) Estimation of time series parameters in the presence of outliers. *Technical Report 8*. Statistical Research Center, University of Chicago.

Chatterjee, S. and Hadi, A. S. (1986) Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.*, **1**, 379–416.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.

Denby, L. and Martin, R. D. (1979) Robust estimation of the first-order autoregressive parameter. *J. Amer. Statist. Ass.*, **74**, 140–146.

Engle, R. F. and Kraft, D. F. (1983) In *Applied Time Series Analysis of Economic Data* (ed. A. Zellner), pp. 176–177. Washington DC: US Bureau of the Census.

Fox, A. J. (1972) Outliers in time series. *J. R. Statist. Soc.* B, **34**, 350–363.

Fuller, W. A. (1976) *Introduction to Statistical Time Series*. New York: Wiley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) *Robust Statistics: the Approach Based on Influence Functions*, p. 59. New York: Wiley.

Harrison, P. J. and Stevens, C. F. (1976) Bayesian forecasting. *J. R. Statist. Soc.* B, **38**, 205–247.

Harvey, A. C. and Pierse, R. G. (1984) Estimating missing observations in economic time series. *J. Amer. Statist. Ass.*, **79**, 125–131.

Hillmer, S. C., Bell, W. R. and Tiao, G. C. (1983) Modeling considerations in the seasonal adjustment of economic time series. In *Applied Time Series Analysis of Economic Data* (ed. A. Zellner), pp. 74–100. Washington DC: US Bureau of the Census.

Jones, R. H. (1980) Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, **22**, 389–395.

Kohn, R. and Ansley, C. F. (1986) Estimation, prediction, and interpolation for ARIMA models with missing data. *J. Amer. Statist. Ass.*, **81**, 751–761.

Martin, R. D. (1979) Approximate conditional-mean type smoothers and interpolators. In *Smoothing Techniques for Curve Estimation* (eds T. Basser and M. Rosenblatt), pp. 147–176. New York: Academic Press.

——(1981) Robust methods for time series. In *Directions in Time Series* (ed. D. F. Findley), pp. 683–759. New York: Academic Press.

Martin, R. D., Samarov, A. and Vandaele, W. (1983) Robust methods for ARIMA models. In *Applied Time Series Analysis of Economic Data* (ed. A. Zellner), pp. 153–169. Washington DC: US Bureau of the Census.

Martin, R. D. and Thompson, D. J. (1982) Robust-resistant spectrum estimation. *Proc. IEEE*, **70**, 1097–1115.

Martin, R. D. and Yohai, V. J. (1986) Influence curves for time series. *Ann. Statist.*, **11**, 781–818.

Martin, R. D. and Zeh, J. E. (1977) Determining the character of time series outliers. In *Proc. Bus. Econ. Statist. Sect. Amer. Statist. Ass.*, pp. 818–823. Washington DC: American Statistical Association.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.

Smith, A. F. M. and West, M. (1983) Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, **39**, 867–878.

Storer, B. E. and Crowley, J. (1985) A diagnostic for Cox regression and general conditional likelihoods. *J. Amer. Statist. Ass.*, **80**, 139–147.

Tsay, R. S. (1986) Time series model specification in the presence of outliers. *J. Amer. Statist. Ass.*, **81**, 132–141.

Weisberg, S. (1980) *Applied Linear Regression*. New York: Wiley.

Welsch, R. E. (1982) Influence functions and regression diagnostics. In *Modern Data Analysis* (eds R. L. Launer and A. F. Siegel). New York: Academic Press.

West, M. (1986) Bayesian model monitoring. *J. R. Statist. Soc.* B, **48**, 70–78.

West, M., Harrison, P. J. and Migon H. S. (1985) Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist., Ass.*, **80**, 73–98.

Yohai, V. J. (1987) *Personal communication.*

DISCUSSION OF THE PAPER BY BRUCE AND MARTIN

Dr A. Robinson (University of Bath): It is a great pleasure to welcome this paper which makes a significant contribution in adapting diagnostic techniques for application to the field of time series. In time series modelling we are only too well aware that rarely is the series entirely consistent with any standard model. It is subject to patches of outliers, level changes, variance distortions and even complete model changes which add to the difficulties of model selection at whatever stage. I would like to begin by commenting on the interplay between model criticism and the objectives of analysis. Diagnostics have been most widely developed in the context of multiple linear regression and this technique is used for many purposes all of which are paralleled in time series. These range through data exploration, data description and prediction.

We may have in mind several or even all these objectives during the course of an investigation but at any given stage there will usually be one which is the primary concern. Starting with an exploratory aim, the requirement for sophisticated criticism of the putative model is likely to be quite modest but becomes increasingly stringent as we progress towards prediction and, for time series, forecasting. With regard to criticism we are also mindful of whether or not the model is theoretically meaningful and/or consistent with our expectations.

Recognizing this, I must express my unease concerning the application of sophisticated diagnostics to the autoregressive integrated moving average (ARIMA) class of models, which many (including myself) regard as a 'black-box' class within which a search is made to explain the data as well as possible without the analyst necessarily ever being required to attempt to understand the data.

This consideration surfaces at several points in the paper. Firstly in Section 3 the authors argue that the innovations variance rather than the set of coefficients is the natural focus for the construction of diagnostics, i.e. 'How well does it fit?' is more important than 'Does it mean anything?' I would agree, in that rarely do I feel inclined to interpret the coefficients of an ARIMA model other than to take them as some mysterious reflection of the autocorrelation in the series. In the context of multiple linear regression, there is often a case for consideration of both coefficient-based and overall-fit-based diagnostics but the former seem to have less relevance in ARIMA model fitting. One exception may be that a change in some coefficients could be important if the objective is to forecast a small number of steps ahead. Perhaps the authors would disagree?

Box and Jenkins gave us a strategy for model selection and now Bruce and Martin offer us further strategies for cleaning up the series so that we may identify which bits of the data are most consistent with the chosen model. The residue must be accounted for in some other way. What do we learn by use

of such complicated black-box tools? At most we obtain a description of the series but is it widely useful? I certainly have reservations about this approach when the objective is forecasting because we may be led, for example, to underestimate forecast error drastically. Other approaches here include the use of diagnostics which are directly related to changes in forecasts or the use of models and model selection techniques which provide resistant forecasts.

The authors may attract criticism by the use of these methods *ab initio* to discover first-magnitude defects which I consider would be better sought by other simpler, less expensive methods, e.g. robust filtering or smoothing (see Section 8.2) and especially looking at a plot of the series! The authors demonstrate (see Section 8.1) that the proposed diagnostics are less effective in long series, so that plotting should present no problem for the suggested area of application. We can compare this with the usefulness of regression diagnostics which is greatest as we move beyond one explanatory variable, where a good plot is still the best diagnostic. How do the authors envisage moving beyond univariate series? There lies the real reward.

In spite of the foregoing reservations, I feel that the suggestions in the paper are useful in uncovering the not-so-obvious problems; the strategies for patch length determination and iterative unmasking are very interesting. I look forward to when they may be incorporated rather less rigidly—the plots are good but there can be very many of them if we follow the suggested strategy, in everyday time series analyses. I hope that other workers will be persuaded to develop a diagnostic toolkit for a wider spectrum of models. I have great pleasure in proposing the vote of thanks.

Professor H. Tong (University of Kent at Canterbury): As a non-native English speaker, I sometimes find the English language inscrutable. For example, the word 'public' really means 'private' in many situations (e.g. public schools), perhaps as a result of 'privatization'. Likewise, to an outsider, the Royal Statistical Society must sometimes seem an equally inscrutable institution because I understand that the expected manner in seconding a vote of thanks to the fortunate (or unfortunate) speakers is to be critical. As one of the authors, Professor Martin, has come back for the second time, I can only assume that he enjoyed himself the last time and has somehow convinced Professor Bruce of the pleasure. I hope that they are not going to be disappointed.

The paper is mainly concerned with influential data in time series, and has given some impressive techniques. As I am not an expert in this field of outlier detection, I find it difficult to be too critical. I am in complete agreement with the authors when they say that influential observations frequently occur in patches and there are smearing effects in the time series setting, although these points are not new. For example, Künsch (1984) was probably aware of them and Hau and Tong (1984) addressed them explicitly. I would also agree with the authors that methods based on deleting one observation will be of doubtful applicability to time series. This rather reminds me of the Chinese story of a person in the period of the Three Kingdoms who deleted his toes to fit his shoes. Clearly deleting one toe will not help; deleting five toes might. Similarly, leave-one-out diagnostics will probably not help but leave-k-out diagnostics might. However, although I can appreciate the logic of deletion for independent data, my basic concern is whether it is really necessary to have deletion *for dependent data*, because there are simpler ways of finding out whether a shoe gives a good fit or not without deleting any of one's toes, e.g. by measuring the strengths with which the shoe pinches each of the toes.

Now DC(A) and DV(A) are both functions of the ith diagonal element of the hat matrix if $A = \{i\}$. It is well known that the diagonal elements of the hat matrix may be used to measure the influence of the independent observations. Künsch (1984) and Hau and Tong (1984) have independently explored the efficacy of the approach based on examining the diagonal elements of the hat matrices in autoregressive modelling. Modifications for the autoregressive integrated moving average case are possible in principle but I am not convinced that it is necessary to do so in the context of influential data.

Briefly, let $\mathbf{X} = (x_1, \ldots, x_n)'$ denote the data vector from the pure AR(p) model (with $\gamma = 0$). Then we have $\mathbf{X} = \Gamma\boldsymbol{\phi} + \boldsymbol{\varepsilon}$, where Γ is the design matrix, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Let $\mathbf{z}_i'$ and h_t denote the ith row of Γ and the tth diagonal element of the hat matrix $\Gamma(\Gamma'\Gamma)^{-1}\Gamma'$ respectively. Clearly $h_t = \mathbf{z}_t'(\Gamma'\Gamma)^{-1}\mathbf{z}_t$. Now, nh_t is the *Mahalanobis distance* between $\mathbf{z}_t$ and $E\mathbf{z}_t$ (i.e. $\mathbf{0}$). (In conventional linear regression, this property is not obtained.) Under standard conditions, we may replace $n^{-1}\Gamma'\Gamma$ by $\Sigma = \mathrm{cov}(x_i, x_j)$, and $\mathbf{z}_t'\Sigma^{-1}\mathbf{z}_t$ has a χ_p^2 distribution. Accordingly, we may use nh_t to quantify the influence of the *state vector* $\mathbf{z}_t = (x_t, \ldots, x_{t-p+1})$, with the reference value being χ_p^2. nh_t *incurs absolutely trivial computation once an AR(p) model is fitted.* Also $\{h_t\}$ is a stationary time series if $\{x_t\}$ is, which underlines the smearing effect of influential data in time series.

A good feature about real data is that almost surely there is at least one 'outlier' to every model fitted. Fig. 9 is a plot of the nh_t versus t for the export data in example 5. An AR(3) model is fitted to the

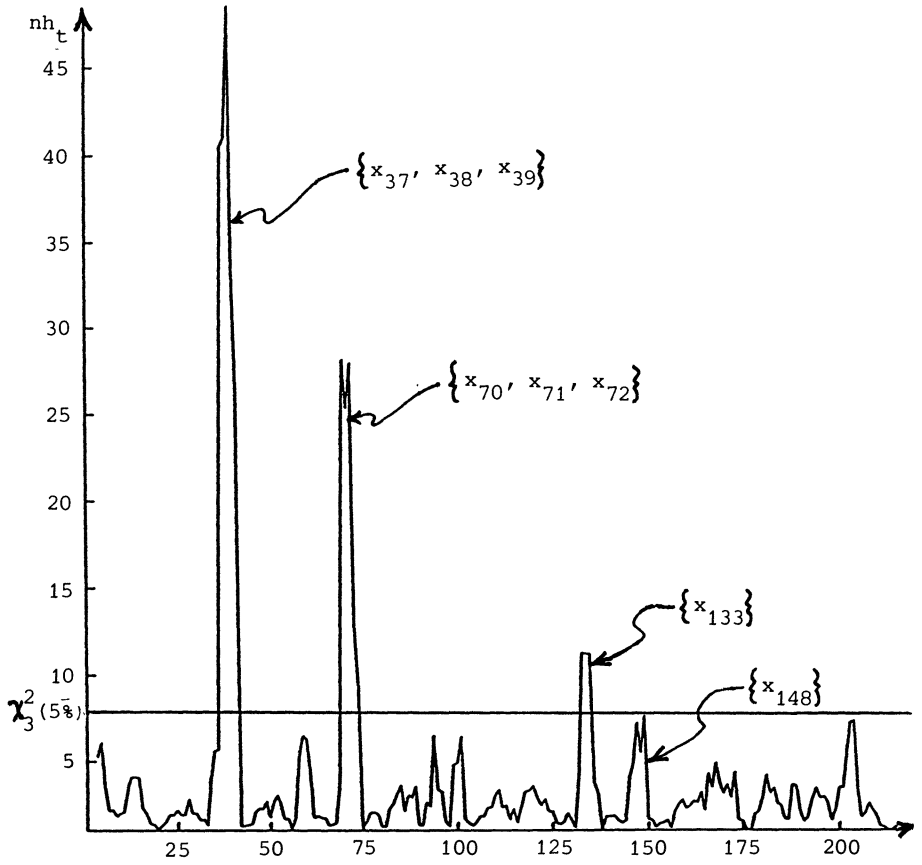

Fig. 9. nh_t versus t for the export data

first differences of the logged data, say x_t, the order being determined by one of the popular methods. The plot shows three patches in excess of $\chi^2_3(5\%)$. In order of decreasing influence, they lead to $\{x_{37}, x_{38}, x_{39}\}$, $\{x_{70}, x_{71}, x_{72}\}$ and $\{x_{133}\}$ as influential data, broadly in agreement with the paper's conclusion. Fig. 10 is a plot of h_t versus t for the famous (logarithmically transformed) Canadian lynx data using the AR(11) model of Tong (1977). Here, about a third of the data are influential, suggesting the necessity of non-linear time series models.

Chan *et al.* (1988) have illustrated that the h_t plots are useful even in non-linear time series modelling. We may also weight the data with h_t to obtain robust estimates of autoregressive parameters. (See Hau (1984) and Künsch (1984).)

This paper has led me to the following question: is there a simpler but equally effective way of detecting outliers in time series without deletion? It therefore gives me great pleasure in seconding the vote of thanks.

Professor A. J. Lawrance (University of Birmingham): Professor Martin mentioned some earlier related work; for instance, Peña in his discussion to Cook's (1986) paper on local influence referred to his own use of the 'missingness' idea. However, there was not the emphasis on leave-k (> 1)-out or patch diagnostics and the associated problems of smearing, masking and scaling, which I found most interesting in the present paper. Perhaps at a more basic level, can I raise the question about whether missingness is an appropriate way to handle diagnostics for time series? It seems to destroy the essential connectivity of the time series and to be a rather drastic assumption. Perhaps, going back to local influence, it would be more sympathetic to the time series structure to perturb some assumptions or data values just slightly, leading to 'leave-in' diagnostics.

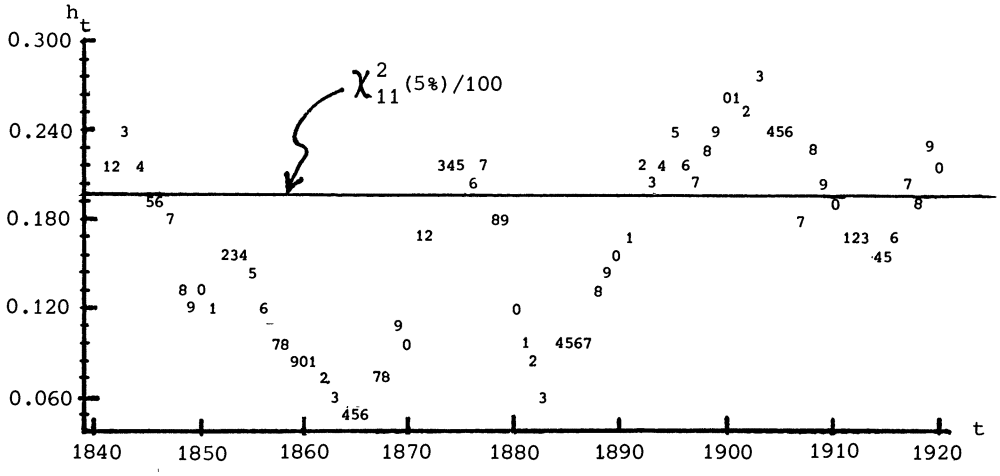

Fig. 10. h_t versus t for the Canadian lynx data

I have a brief comment on the interpretation of the statistics relative to p values—I agree with the authors that the significance testing framework is not appropriate, so I prefer the complementary description of displacing parameter estimates out of $100(1 - p)\%$ confidence regions. Also, I have found in regression that the guidance value of $p = 0.5$ is too stringent and hence too small and could be up to 0.9, corresponding to movement out of the much smaller 10% confidence regions. This would not demean the leave-one-out diagnostics quite so badly.

The topic of masking is important, but difficult to handle analytically both in regression and in time series. To gain a little insight into regression, we may consider Cook's distance for the ith data case, both before and after deletion of the jth data case. Before deletion, it is well known, for instance (Atkinson (1985), equation 3.2.4) that

$$C_i = r_i'^2 h_i/(1 - h_i)p,$$

where r_i' is the ith standardized residual, h_i is the ith leverage coefficient and p is the number of regression parameters. After jth-case deletion, this becomes

$$C_{i(j)} = (r_i' + \tilde{h}_{ij} r_j')^2 \left(\frac{h_i}{1 - h_i} + \tilde{h}_{ij}\right) \Big/ p(1 - \tilde{h}_{ij}^2)^2 \left(\frac{n - p}{n - p - 1} - \frac{r_j'^2}{n - p - 1}\right)$$

where

$$\tilde{h}_{ij} = h_{ij}/\sqrt{\{(1 - h_i)(1 - h_j)\}},$$

and h_{ij} are the off-diagonal terms of the leverage matrix. The effect of masking can be assessed as $C_{i(j)}/C_i$; it can occur for several reasons. Important contributing factors are seen to be the off-diagonal elements $\tilde{h}_{ij}$ related to the leverage matrix, the size and sign of the ratio of the two standardized residuals, and leverage. Are there any similar analytical insights for time series? Computationally, they might arise in the estimation procedures.

A. C. Harvey (London School of Economics and Political Science): It is a pity that the authors have based their approach on autoregressive integrated moving average (ARIMA) models. Specifying a suitable ARIMA model is difficult at the best of times, and when outliers are present it becomes even more difficult. As the authors themselves observe, the message conveyed by statistics such as the correlogram can be seriously distorted. As we shall see, this point is well illustrated by their treatment of the Latin-American export (LAX) series.

The advantages of structural time series models have been argued elsewhere; see, for example, Harvey and Durbin (1986). In the possible presence of outliers these advantages become even more apparent. The leave-k-out diagnostics of the present paper could be applied just as easily with structural models as with ARIMA models. However, rather than following this route, I have chosen the much simpler

procedure of analysing the LAX series by fitting a structural model and then using the smoothed estimates of the irregular component to identify outliers. These estimates provide a better means of identifying outliers than do the one-step-ahead prediction errors, even though the latter have better statistical properties in a correctly specified model; compare Burman (1983). Since the series has a clear upward trend and consists of monthly observations, the obvious starting point is to fit a basic structural model. This was carried out on a personal computer using the STAMP package. The outliers were clearly detected by the Bowman–Shenton normality test, the value of which was 362, and a study of the estimates of the irregular component indicated outliers at 69m1, 69m2 and 71m9–m11. Re-estimating the model by treating these observations as though they were missing suggested possible outliers at 68m10, 76m12 and 77m1. However, the model appeared to be reasonably satisfactory with these observations included. The seasonal component is not particularly strong, but it is clearly present and when it was omitted there was a marked autocorrelation in the residuals at lag 12.

As regards model selection within the ARIMA framework, the correlogram of first differences could be regarded as being roughly consistent with a stationary process. The autocorrelations at lags 1 and 2 are $r(1) = -0.26$ and $r(2) = -0.24$ respectively: hence the ARIMA(0, 1, 2) model adopted by Bruce and Martin. However, the correlogram for the period after the outliers, i.e. 71m12 onwards, has $r(1) = -0.30$ and $r(2) = -0.05$, showing the identification of a second MA term to be completely spurious. Furthermore, the seasonal autocorrelation, $r(12)$, is 0.54, as opposed to the earlier 0.26, thus indicating the need for including seasonal effects in the model.

Dr I. T. Jolliffe (University of Kent at Canterbury): This is an interesting paper, not least for the questions it raises regarding what is meant by an 'influential' observation and the differences between 'outliers' on the one hand and 'influential observations' on the other. The authors' main interest appears to be in detecting outliers, but they tackle this objective by assessing the influence of observations. In general, there is no guarantee that an outlier will be influential, or vice versa. The paper says little about this problem except in the context of different measures of influence having different powers to detect outliers. I would welcome the authors' thoughts on the more general question of whether examining influence is the most appropriate way of looking for outliers.

The results in the paper suggest a rethink of what is meant by influential. In Fig. 2(b), for example, observation 27 *is* clearly influential for coefficients in the sense that if it is omitted then the coefficients change greatly, much more so than for any other observation. However, further investigation shows that the influence is apparent rather than real, being caused by the dependence between observation 27 and the outlier 28. The time series nature of the data makes clear the reasons for this misleading apparent influence, and I assume that such behaviour cannot occur for independent observations. Are the authors aware of any other, less structured, situations where calculation of influence can similarly mislead?

The term influence needs to be qualified—influential for what? The paper examines influence on coefficients, and on the innovation variance. Have the authors considered influence for the autocorrelation function or partial autocorrelation function, or for the choice of model orders?

Julian Besag and Allan Seheult (University of Durham): Anomalies also occur in data from agricultural field trials and are catered for in the classical framework by randomization theory. However, this can be very inefficient unless a sophisticated design is used and, unfortunately, most variety trials worldwide employ nothing more efficacious than randomized complete blocks, in which case a model-based analysis may be preferred.

Here we consider the usual situation with several replicates, each arranged as a 'column' of contiguous long narrow plots, so that a one-dimensional model is adequate. Then the simplest plausible assumption is that first differences of adjacent plot values are uncorrelated and have equal variance. It can be shown that the corresponding estimates of variety contrasts, based on an ordinary least squares analysis of first differences of yields, are approximately invariant to locally linear trends in fertility (Besag and Kempton, 1986). However, anomalies in the data, usually obscured by the variety effects, can seriously distort the estimates.

Two particular problems are outliers, due to abnormal plots or false records, and abrupt jumps in fertility, caused by underlying geological changes (see also Papadakis (1984)). We tackle these problems, in this context, by first applying resistant regression to detect and remove anomalies before reverting to the ordinary least squares analysis: note here that, having removed an outlier, we subsequently allow information about fertility to flow between its neighbours but treat jumps as complete breaks in the series.

We briefly consider two examples. The first, from Papadakis (1984), is a uniformity trial on which 13 dummy varieties in five blocks have been superimposed. Fertility is therefore indicated by the yields

themselves and, assuming treatment additivity, rigorous comparisons between different additive analyses can be made. The variety mean square (VMS), ideally zero, is 130 in a classical analysis. Without consideration of outliers, the simple first-differences formulation is rejected in favour of the extended model in Besag and Kempton (1986), with a corresponding VMS of 84. However, a preliminary resistant phase suggests 11 anomalies, whose removal vindicates the simpler formulation and further reduces the VMS to 47.

The second example, from Wilkinson (1984), involves 50 varieties of wheat in three replicates of a randomized complete block design. Strong fertility effects are evident. Resistant analysis identifies three outliers and 24 jumps in fertility, whose removal again supports the simple first-differences formulation rather than the extended one and also produces a substantially different ranking of the varieties.

Do the authors see a role for their methods in contexts such as this?

Professor R. M. Loynes (University of Sheffield): Time series have two properties, one of which makes the task of providing appropriate methods more difficult and the other easier than in ordinary regression. The estimating equations are usually non-linear in the parameters and the estimates themselves are all non-linear in the observations, and, apart from ordinary computational complexities caused by this, simple (and indeed elegant) formulae relating, for example, leave-one-out residuals to ordinary residuals no longer exist. However, because of the ordered nature of time it is much easier to contemplate leave-k-out quantities for $k > 1$: those that are likely to be of interest are fewer in number, which both reduces the computing needed and makes interpretation and presentation easier. As far as this second issue is concerned I merely point out that there are other models with somewhat similar characteristics: in particular design models, in which it is natural to omit a treatment or perhaps a block completely, and I have some results in this area. (These models, incidentally, have another interesting and complicating feature: the rank is reduced if a treatment is omitted completely.)

To return to the first point, it is true that computing is much less of a problem than it used to be, but it can still impose a heavy burden, and, if it can be simplified, so much the better. The use of one-step or related approximations (see Cook and Weisberg (1982) and Pregibon (1981)) offers a useful alternative, or at least a preliminary, approach: in these we find, by Taylor expansions, that $\hat{\alpha}_A - \hat{\alpha}$ is approximately some function of $\hat{\alpha}$ which can be calculated without excessive difficulty, for example. I have some theoretical results here, but have not yet attempted to see how well they work on data. Have the authors tried methods of this kind directly for diagnostics? They have apparently used such approximations for a rather different purpose in Section 3.2.

Professor J. P. Burman (University of Kent at Canterbury): The problem that the paper addresses can be illustrated from the method that I am using for outlier detection, which is a modified version of that proposed in Hillmer *et al.* (1983): if $\pi(B)\mathbf{z}_t = \mathbf{a}_t$ is the autoregressive form of the model, $\pi(B)\,\pi(F)\mathbf{z}_t$ is a symmetric, doubly infinite, filter of $\mathbf{z}_t$. This series is easily computed, using signal extraction, and has the character of an irregular component, so it can be searched for outliers—values above a given threshold. The search is applied iteratively, as suggested by Hillmer *et al.* (1983), i.e. each time new outliers are detected, their approximate magnitudes are obtained by linear regression, and the irregular series modified accordingly. The threshold is also reduced.

This method is measuring only the DRES effect and not the disturbance of the parameter estimates (DFIT), but it seems to work well in detecting isolated outliers. My current program was applied to the exports to Latin-America (ELA) series and detected three additional outliers in 11/68, 10/70 and 12/76, apart from the five associated with dock strikes and forestalling; these are not quite the same as those identified in this paper.

There is a problem with the Hillmer *et al.* method about when to stop iterating: in some series, the detection process becomes unstable, with increasingly more outlier patches appearing. This may cast doubt on the model; for example, a transformation is needed. At the same time, there may be evidence that not all outliers have been identified. The threshold is calculated at each step from a robust mean absolute deviation estimate of the innovation standard error, and not directly from the standard error which is usually larger. As more outliers are removed, the two estimates should converge. For the ELA series, the root-mean-square estimate is still 30% above the robust estimate, when the detection iteration is stopped.

I congratulate the authors on devising a workable and theoretically sound method of deciding the size of the patches. I find the treatment of the ELA series entirely convincing, but I am not so sure about the unfilled orders series. The authors demonstrate a break around 1976, but I am puzzled by the dramatic

change in the seasonal operator from $(0, 1, 1)_{12}$ to $(0, 0, 2)_6$. The final values give a model of seasonality with moving average roots on the unit circle, i.e. it changes completely after 1 year. A more balanced model with a 12th difference added would be more plausible.

Sir David Cox (Nuffield College, Oxford): I am uneasy at two aspects, one the very central role put on autoregressive integrated moving average (ARIMA) processes and the other, somewhat related, the relative failure to distinguish incisively between different objectives of time series analysis.

It would not be feasible to list all the objectives one might have but these include the following:

(a) the wish to isolate patches of anomalous structure against a broadly regular background and for this the authors' approach seems very suitable;
(b) to estimate ARIMA parameters, probably, however, fairly rarely the ultimate objective because often ARIMA processes lack physical significance;
(c) to study qualitative aspects of structure such as cointegration;
(d) to study regression parameters in contexts where the error has a time series structure, where we should surely concentrate on the regression parameters and their standard errors;
(e) prediction.

In (b)–(e) should not the diagnostics be focused on the ultimate objective?

For more exploratory aspects can there not be methods based directly on the estimated spectrum or autocorrelation or autocovariance, rather than necessarily viewing these via the ARIMA structure?

There are many possibilities for generalization, e.g. to the statistical analysis of point processes.

Mr E. F. Harding (University of Cambridge) **and Dr K. K. K. Y. Sew-Hee** (Novaction UK Ltd): Our contribution is inspired by two reasons.

First, many discussants have raised the possibility of an approach different from simply leaving out observations. A recent doctoral thesis by one of us (Sew-Hee, 1988) deals with the real-time detection of change points in time series, i.e. moments at which the series changes from obeying one model to obeying another model, without being too specific *a priori* about what are the models in question. In other words, we were seeking a rather broad spectrum test for detection of change points.

Thus, for example, the variation in a series might change from a régime of slow oscillation to one of rapid oscillation, from small to large variance, from downward trend to upward trend.

The approach was to fix a *reference window* at the start of the series and to slide a *test window* along the series, and to make a comparison of suitable statistics computed for each as a function of the displacement of the test window. Statistics might be a serial correlation, or covariance matrix, or a spectrum, and the comparison in terms of a criterion such as likelihood ratio, Kolmogorov–Smirnov or Bayes factor. As the test window moves through the series away from the reference window, the criterion fluctuates; then, as the test window crosses a change point, the criterion moves to a new value about which it again fluctuates. It is fairly easy to detect the change point when the movement between levels is reasonably linear.

Some of the series we looked at had outlying observations in them. The effect on the procedure described was to produce a 'temporary blip' in the criterion. The thesis describes a back-tracking technique for rapid recognition of such false alarms, which were not the main object of analysis. However, the work had, as a by-product, an approach to the detection of outliers in time series, either individual observations or short-lived aberrations.

The second reason is that both of the examples of real data given by the authors (Figs 7 and 8) in their respective ways quite closely resemble examples of real data considered in the thesis.

Consider Fig. 8, which has large fluctuations towards the end. We had an example that behaved in a very similar way: the test criterion simply moved rapidly to the end of the scale at this point and so the problem was clearly detected.

In series analogous to Fig. 7, individual outliers were picked out as transients using these methods. Our technique for handling them was simply to interpolate linearly across them, and then to continue with the rest of the analysis.

Dr Frank Critchley (University of Warwick): In welcoming the paper and, in particular, its plea for a flexible approach to modelling, I would like to mention two other general approaches to assessing the joint influence of k observations.

The authors consider deleting influential observations. Cook's (1986) paper on local influence considers a less drastic change. He monitors the likelihood displacement under local perturbations.

The second general approach is based on what, in work in progress, I term the joint influence function. This methodology requires independent, identically distributed observations when used in the context of a probability model. However, it can be applied outside such a context, e.g. in exploratory data analysis. It has the advantage of handling masking naturally and thereby avoids the potential dangers of a sequential approach. It is not limited to local perturbations, although it includes them, and it generalizes easily to simultaneous influence for several aspects of a model. It is based on one simple idea: replace one perturbation by k.

Returning to the paper, I note the following.

(a) 'Influence for what?' is an important question. Observations influential for one aspect of a model need not be so for another. The authors discuss influence for α and for σ^2. Other possibilities meriting study include influence for prediction. In this connection see Johnson and Geisser (1983).

(b) In a similar vein, whatever their value as overall summaries, norms confound detailed information. For example, DC confounds influence for the r parameters in α. In the present context this increases the smearing effect in cases where $r > 1$.

(c) In considering asymptotics, it is vital to distinguish the cases where k is or is not fixed as $n \to \infty$. In the former case it is clear intuitively why diagnostic measures such as DC and DV converge to zero. Parameter estimates with and without a fixed number of observations both converge to the true value more rapidly than the estimated asymptotic covariance matrix converges to its true value. An entirely different asymptotic theory applies when k/n is kept fixed.

(d) Perhaps more importantly, we need to consider the joint null distribution of our diagnostic measures in finite samples. How else are we to know that the values observed are not just the result of sampling fluctuations? In the present context this seems particularly necessary when the values are near the boundary of the stationary invertible region. This distribution will rarely be known exactly so subtle approximations and/or simulations will be required.

Dr C. R. Muirhead (National Radiological Protection Board, Chilton): In standard regression diagnostics, it is helpful to distinguish between an outlying measurement on the response variable—detected for example by a residual plot—and a set of covariates that has a large influence on the fit—detected say by examining leverages. In time series, however, as has been pointed out in the paper, there are difficulties in separating these effects since the 'covariates' are now functions of observations at other time points. This is particularly evident in the plots of DC. Another way of looking at the problem may be as follows.

Under an AR(p) model, standard diagnostics based on 'response variable' x_t and 'covariates' $x_{t-1}, \ldots, x_{t-p}$ can be calculated. This in a sense corresponds to deleting *innovations* rather than observations. In particular, as also suggested by Professor Tong, examination of leverages based on $x_{t-1}, \ldots, x_{t-p}$ will provide an idea about those time points at which the one-step-ahead prediction of x_t has a large influence on the fit. Such high influence may not simply be due to outliers in the earlier observations but may arise for particular combinations of the observations used to predict x_t. This situation may not be as easy to identify using observation deletion.

Plots of residuals based on this formulation can be helpful in detecting innovation outliers. Also, constructing an analogue of DRES corresponding to innovation deletion and comparing this with DRES based on observation deletion (in each case based on an estimate of α that is robust with respect to outlier type) provides a basis for distinguishing additive and innovation outliers. Such a test has been studied by Muirhead (1986) for a single outlier.

The following contributions were received in writing after the meeting.

William R. Bell (US Bureau of the Census, Washington DC): Bruce and Martin's interesting leave-k-out approach can be contrasted with the approach of Chang and Tiao (1983) which uses indicator variables. A conceptual difference for additive outliers (AOs) is that use of an indicator variable lets the mean at the given time point be anything while still assuming that the observation is normal with the same variance and covariances as other observations, whereas omitting the observation makes no assumptions at all about it. The latter approach is preferable theoretically, but the paper notes that the two approaches give approximately the same results for AOs. Other differences between the paper's approach and the Chang and Tiao approach (as implemented in Hillmer *et al.* (1983) or Bell (1983)), such as looking for outlier patches, could be mimicked with appropriate use of indicator variables, though with a computational effort closer to what is required under Bruce and Martin's approach.

To amplify, I shall mention two problems that I have been concerned with recently in our implementation of the Chang and Tiao approach. First, the outlier test statistics approximate a generalized least squares (GLS) regression by doing an ordinary least squares regression of filtered data (model residuals) on filtered indicator variables. The filtering is done conditionally, corresponding to conditional maximum likelihood estimation, though it can be done exactly, yielding the exact GLS regression (see Findley *et al.* (1988)). Exact filtering is preferable theoretically but is computationally intensive since, for moving average models, a different filtering must be done for the outlier indicator at each of the time points. The second problem is that, if there are multiple outliers *or* other regression variables in the model, then we must do a multiple regression (note Burman (1983)), because the filtered outlier and regression variables are not orthogonal. For multiple outliers this can lead to the masking problem referred to in the paper. Bruce and Martin's approach handles both these problems almost automatically; their iterative deletion strategy is a modification to address the problems with multiple outliers that are not sequential.

I have two final comments.

(a) I have found an extension to the Chang and Tiao procedure to handle level shift (LS) outliers (Bell, 1983) to be extremely useful in practice. Have the authors tried to apply the diagnostics of the paper to the differenced data to handle LS outliers?

(b) For models with regression terms it might be useful to consider another diagnostic—say DR, analogous to DC but based on changes in the regression coefficients.

Ronan Bradley and John Haslett (Trinity College, Dublin): We should not underrate simpler views than model-based views of the data, such as histogram and time series plots of the data and residuals from a model. Thus in all the authors' examples most of the 'strange' data points could, we argue, have been spotted rather directly with simpler views, particularly if made available in the context of 'linked' windows in a mouse-and-window environment. What types of 'strangeness' could *only* be spotted with the authors' tool?

In this paper, the authors have pointed out the significant problems of smearing, and clustering of outliers which occur in time-referenced data. These problems also appear in spatially referenced data, in which our interests lie, and indeed they are added to by the lack of natural directionality in the data and the possibility of irregularly spatially distributed data. For these reasons, the whole question of outlier classification and identification for spatial data becomes much more complex and, to the best of our knowledge, is as yet untouched.

In view of the added difficulty of analysis we also propose an exploratory strategy and a new spatial view to accompany more basic views. We illustrate this in the context of the variogram cloud (Chauvet, 1982) but it has clear analogies in time series. This cloud (Fig. 11) for a set of data $Z(\mathbf{x})$ at $\mathbf{x} = x_1, \ldots, x_n$ (shown in Fig. 12, as the most important view, namely the data locations) is a plot of $\gamma_{ij} = \frac{1}{2}\{Z(x_i) - Z(x_j)\}^2$ against $d_{ij} = \text{dist}(x_i, x_j)$. Its average for a given distance h is an estimator of the variogram $\gamma(\mathbf{h}) = \frac{1}{2}\text{var}\{Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})\}$. In Fig. 11 we select γ_{ij} values which are large for a given

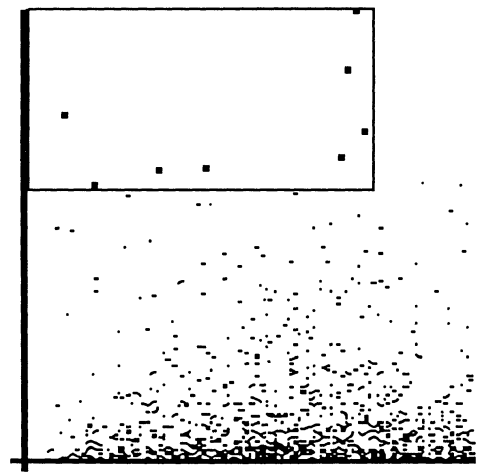

Fig. 11. Variogram cloud showing the selected region

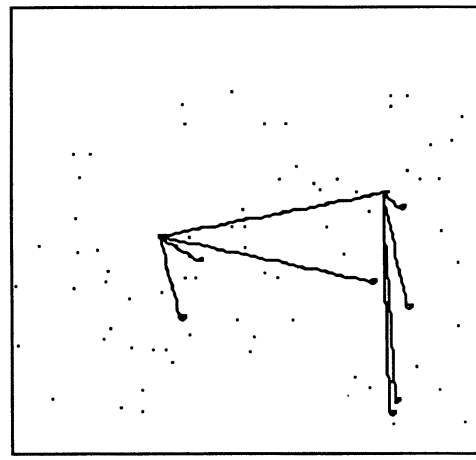

Fig. 12. Data map showing corresponding point pairs

range of d_{ij} and highlight the corresponding pairs in the linked window shown in Fig. 12. Attention is immediately drawn to two points, each of which is different from many of its neighbours, and suggests that the surrounding regions are worthy of further investigation, perhaps with simpler views. A 'histogram view', not shown, establishes that the two points are definitely global inliers. They may, however, be local outliers.

Dr C. Chatfield (University of Bath): It seems natural to extend the use of regression diagnostics to time series analysis, and this clearly presented paper achieves this successfully. Even so I am rather dubious about whether I would want to *use* these methods. The proponents of diagnostic procedures often underemphasize the importance of using background information to predict or explain unusual observations and in my experience this often obviates the need for formal diagnostics. Alternatively, in time series analysis, a visual examination of the time plot will often prove adequate and this is arguably the case in most, if not all, of the authors' examples. Likewise, in Fig. 8, I suspect that most experienced analysts will immediately notice the increasing variance. If this sounds overly negative, let me emphasize that I do appreciate the authors' examples and, if appropriate software becomes readily available, then it is certainly possible that these rather complicated methods could become part of the superanalyst's toolkit.

R. Dennis Cook (University of Minnesota, St Paul):

Calibration

Deciding exactly what is influential based only on the value of an influence measure can be a problem. I earlier proposed (Cook, 1977a, b) that in linear regression D_i be compared with the quantiles of an F distribution and referred to the result as the 'descriptive level of significance'. This entails nothing more than a monotonic transformation to a more familiar scale and, as Bruce and Martin emphasize, is not a test of significance and does not have the familiar p value interpretation. Although the idea is helpful, the associated use of terms such as 'descriptive significance level' and 'p value' may be didactically unsound since the procedure is misinterpreted as a formal test too frequently. It is wise to avoid the use of phrases associated with testing in the context of influence studies.

The exact distribution of influence measures, including DC and DV, is largely irrelevant. An influential observation is just as interesting when we expect it to be influential than when we are surprised that it is influential, although the same cannot be said for outliers which I regard as related but distinctly different. Thus I cannot understand why Bruce and Martin think it is useful to '. . . scale DC and DV to obtain an asymptotically non-degenerate random variable, and then determine approximate significance levels . . .'.

Alternatively, an influence measure may be used only for ranking observations. Influence is then assessed by deleting the observations with the larger ranks and recomputing the analysis. Although difficult to automate, this procedure seems to avoid the need for careful calibration of the measure itself.

K_{max}

Upper bounds have often been used to induce identifiability in outlier and influence studies. Such methods are not very satisfactory: the required bound is rarely based on firm information and it can be very difficult to determine when the 'best' answer has been obtained.

Local influence methods (Cook, 1986), which do not require the specification of a bound, are effective at identifying multiple influential cases in studies with independent observations. Such methods may also avoid masking in time series problems. For example, one implementation could be based on using normal curvatures to study the influence of perturbations of the innovation variances. Because of the emphasis that Bruce and Martin place on the innovations variance, this implementation may be particularly useful. Work along these lines is available in Tsai (1986).

EIC and EIV

Norming an empirical influence function is one way of developing an influence measure. An alternative is to use likelihood displacement (Cook and Weisberg, 1982; Cook, 1986). In some cases they will be essentially the same, but how would influence measures based on likelihood displacement compare with EIC and EIV?

F. D. J. Dunstan (University of Wales College of Cardiff): One disadvantage of the method is the amount of computation involved and a method which avoided having to leave out values would be preferable if it could be shown to work equally well. The disadvantage of such a method, however, is that the potential outliers will be used in model fitting and their values can easily grossly distort the model selection and fitting. I once used as an example in a lecture a series of working days lost due to industrial action. A semi-automatic analysis, using the autocorrelation function (ACF) and partial ACF to choose the type of model without close examination of the data, strongly suggested that a moving average model of order 1 would be appropriate. A closer look, however, showed that two consecutive observations were considerably larger than the rest and this induced an artificially high value for the ACF at lag 1 and smaller values at higher lags, hence the selection of an MA(1) model. If these observations were deleted, then the model selected was an AR(1) model. For this reason, even though this example is somewhat extreme, a method which deletes observations in the way suggested has much to commend it.

There is one case in which the amount of computation required can be reduced, namely when the models considered are purely autoregressive. In this case the EM algorithm handles the problem of missing data efficiently both computationally and also in terms of the estimates produced; although the theoretical properties of its convergence are not entirely satisfactory, in practice it works well and an earlier study (Barham and Dunstan, 1982) suggested that its performance was superior to that of a method based on the Kalman filter. In this case, since the parameter estimates for the whole series give a good starting point for the algorithm, it works very quickly to give revised estimates when data are omitted, each iteration consisting of solving a set of least-squares-type equations. I have tried the method on series of 200 observations and have found that it has worked well.

Marc Hallin and Guy Mélard (Université Libre de Bruxelles): In spite of the obvious importance in all statistical domains of the detection of influential observations, interest has been centred, almost exclusively, on linear model situations. The reason why time series analysis has so far been avoided is twofold:

(a) the concept of an influential observation in a time series context is more complex than in linear model situations—see for example the *smearing effect* in Section 3.2—and
(b) the numerical implementation in a time series context of regression diagnostic procedures requires powerful computing techniques.

The present contribution therefore constitutes a significant and timely breakthrough.

Section 7 explores the close relationship between additive outliers and missing data likelihood methods. This suggests replacing the proposed missing data procedures with simpler and more widely available intervention analysis methods (also allowing for outlier patches at the start of the series). Estimation then can be achieved using fast filtering algorithms, which are an order of magnitude faster than ordinary Kalman filtering algorithms (Pearlman, 1980; Mélard, 1984). Experience has shown that the penalty caused by interventions is small (at least when avoiding binary variable generation and transfer function specification: Mélard (1981)).

The DV(A) diagnostic method apparently yields good performance. However, there is something disturbing about suspecting a patch A of k outliers if $S = \hat{\sigma}^2/\hat{\sigma}_A^2 - 1 > S_{BM} = 0.954n^{-1/2}$, irrespective

of k. One way to take k into account consists of substituting information criteria for the likelihood criterion. A patch A is then suspected if AIC $= \log\hat{\sigma}^2 + 2r/n >$ AIC$_A$—i.e., approximately, if $S > S_{\text{AIC}} = 2k/n$—or if SBIC $= \log\hat{\sigma}^2 + r\log n/n >$ SBIC$_A$—i.e. if $S > S_{\text{SBIC}} = k\log n/n$. For $k = 1$, $n = 100$, $S_{\text{BM}} = 0.095 > S_{\text{SBIC}} = 0.046 > S_{\text{AIC}} = 0.020$. To illustrate this, consider a white noise series with a single outlier e_τ. Here $\hat{\sigma}_A^2 \approx \hat{\sigma}^2 - e_\tau^2/n$, and the DV($A$), AIC and SBIC methods identify e_τ as an outlier if $|e_\tau| > \hat{\sigma}n^{1/4}$, $|e_\tau| > \hat{\sigma}\sqrt{2}$ or $|e_\tau| > \hat{\sigma}\sqrt{\log n}$ respectively. In example 5.1, these cut-off values are $3.8\hat{\sigma}$, $1.4\hat{\sigma}$ and $2.3\hat{\sigma}$ respectively. Information criteria thus are likely to be more efficient in detecting outliers than the proposed DV(A) method—despite the apparently liberal 50% significance level.

Douglas M. Hawkins (University of Minnesota, St Paul): One drawback of the leave-k-out methods in regression has been the need to specify the value of k. Too small an estimate leads to problems of masking; too large leads to problems of swamping. Presumably the same will hold (though less severely) for the time series application also.

Have the authors considered adapting another method from the regression setting—one that we might contrast with their approach by calling it the 'put-P-in' approach? In a linear regression on P predictors, the approach is to take repeated subsets of size P from the original data set. Each provides an exact fit to some set of values of the parameters and may be used to obtain a predicted residual for every observation not in the subsample. This framework encompasses the least median of squares algorithm of Rousseeuw and Leroy (1987), the L1 norm and regression quantiles (see for example Portnoy (1987) and Koenker and D'Orey (1987)) and the 'median elemental predicted residual' approach of Hawkins *et al.* (1984). All these methods can identify arbitrary numbers of outliers (provided that these are fewer than the breakdown point) and resist masking and swamping.

Adapting this method to the time series analysis of say an autoregressive integrated moving average model with $P = p + q$ parameters, we would take repeated subsets of size P from the original data. Each such subset provides an exact fit to the P unknown parameters, and estimated values and hence residuals for the remaining points. Two possible approaches would then be

(a) concentrating on the parameters ϕ and θ, to turn the estimates from the different calibrating subsamples into a single consensus estimate, or
(b) concentrating on the points, to turn the different estimated residuals for each point into a single consensus residual for that point and to use these to provide case diagnostics.

We have only recently started experimenting with this approach, but the results so far hold promise. For example, a simulation like that of examples 1 and 2 of the paper but including both an innovative and an additive outlier provided quite a clear identification of both. At least, we hope that this method would be a useful adjunct to leave-k-out diagnostics—perhaps in a framework where put-P-in is used to identify good candidate subsets of the data which may then be confirmed with leave-k-out.

Have the authors any experience of this approach, or views on its potential in time series?

Robert Kohn (University of New South Wales, Kensington) **and Craig F. Ansley** (University of Auckland): Intuitively statistic DV should be superior to DC because

(a) it is very close to a likelihood ratio test for additive outliers or innovations outliers when dummy variables are used and
(b) autoregressive integrated moving average models with different coefficient values often give nearly the same forecasts and forecast standard errors, so that in many cases DC gives much the same information as DFIT in the decomposition of Section 6.

For example, if data are generated by the ARMA(1, 1) model $x_t + 0.2x_{t-1} = e_t + 0.1e_{t-1}$, then for most realizations an AR(1), MA(1) or ARMA(1, 1) model will give very similar forecasts and estimates of σ^2.

There are three issues that need further clarification. First, we need to know more about the operating characteristics of the deletion rule based on p values less than 0.5. Although it is pointed out in Section 8 that a hypothesis testing framework is unsuitable for diagnostics, the proposed heuristic has the same form, and because it is just a heuristic it is important to know the probability of rejecting a good observation and the probability of not detecting an outlier for a given size and type. Second, if the authors' procedure is to be adopted widely then it must be computationally efficient, especially for seasonal models. The Hillmer *et al.* (1983) and the Tsay (1986) approaches require only $O(n)$ operations compared with $O(n^2)$ for the authors' approach, although they are not as robust. Recently (Kohn and Ansley, 1988) we have managed to extend the Hillmer *et al.* (1983) regression

approach to general state space models with missing or unequally spaced data using only $O(n)$ operations. Third, what is the distribution of the parameter estimates after deflection of points using the authors' method?

Professor Hans R. Künsch (Eidgenössiche Technische Hochschule Zürich): The method requires a substantial subjective appraisal by the data analyst. In particular it seems impossible to estimate the variance of the final estimator after all outliers have been deleted. Since the robust filters and smoothers mentioned in Section 8.2 produce comparable results, I would prefer to use these. Another approach which might lead to some simplification is to interpolate x_t, $t \in A$, from the remaining observations and then to approximate the difference between the estimator based on all data and the estimator based on the interpolated values by the influence function, see formula (1.27) of Künsch (1984). The interpolation is not trivial either, but I conjecture that a suboptimal interpolator will be sufficient.

Next I would like to make two comments on the diagnostic quantity DC. First, masking occurs if there are two or more additive outliers of the same magnitude, but well separated. The estimated model will then be close to independence regardless of whether we use all data or delete one outlier or a block around one outlier. With DV no masking occurs in this case. I was unable to construct a situation where a similar masking effect would occur for DV, but can it be proved that DV always works well? Furthermore with DC there is the problem of standardization. It would be advantageous to use $C(\hat{\alpha})^{-1}$ with a robust estimator $\hat{\alpha}$. Otherwise $C(\hat{\alpha})^{-1}$ can be completely inaccurate if outliers are present.

In the independently identically distributed setting deletion of single observations also gives, through the jackknife, an estimate of the variance of the estimator. As is shown in Künsch (1989), deletion of blocks of consecutive observations is the right thing to do in the time series setting. To obtain consistency, we must let the block size increase with the sample size, however. Moreover the definition of an estimator other than the maximum likelihood estimator after deleting a block is less clear. If the estimator is defined through an estimating equation I leave out the corresponding terms in the estimating equation. This works well for variance estimation. It would be interesting to study its diagnostic capabilities.

Johannes Ledolter (University of Iowa, Iowa City): I agree with the authors that a flexible approach to the treatment of outliers is beneficial and that there is no single approach that can handle all outlier situations. The objective of the analysis should determine the type of influence diagnostics that we are considering. If the interest lies in the time series coefficients, then a measure such as DC should be used. If the innovation variance is of interest, then the measure DV is appropriate. If the interest is in prediction, then we should study the influence of observations on point predictions and prediction intervals. It can be shown (Ledolter, 1989) that point forecasts are largely unaffected by additive outliers, provided that the outliers are not too close to the forecast origin. The conclusion concerning the width of prediction intervals is much different, however. Outliers inflate the innovation variance estimate and, as a consequence, affect the width of prediction intervals.

The authors have shown how to check whether a certain observation, or a group of observations, has an influential effect on the innovation variance estimate. The more difficult question, however, is what to do with this information if our interest is in forecasting. For example, it is not obvious why we should calculate the prediction intervals for future exports to Latin-American republics from an innovation variance estimate that is obtained after omitting suspected outliers. It can be argued that the resulting prediction intervals are too narrow, as they do not reflect the variability that is normally associated with the series. Also, my research on the effect of outliers on point predictions leads me to speculate that the point forecasts for 1984 will not change much if the suspected outliers are replaced by missing observations.

The diagnostics DC and DV are related to likelihood displacements if observations are treated as *completely missing*. A general method for assessing the *local influence* of minor perturbations of a statistical model is discussed in Cook (1986). Its application to the time series context is given in Ledolter (1988). The results of this analysis show which particular functions of the data lead to the largest local change in the likelihood displacement. For example, in the additive outlier model it is the differences between the observations and the interpolated values that determine the sensitivity of the innovation variance estimate. For innovation outliers it is the one-step-ahead prediction errors that are of interest.

Jack C. Lee (Bellcore, Morristown): It is not surprising that in a time series setting a diagnostic based on the innovations variance is far superior to a coefficient-based diagnostic. This is especially clear from the perspective of stationarity, as non-stationarity of a time series can be caused by either a shift in the

level or a change in the innovations variance. The innovations variance estimate $\tilde{\alpha}^2$, which is a function of the coefficient estimate $\hat{\alpha}$, will reflect both types of non-stationarity, if they have occurred. However, the coefficient estimate will not reflect a shift in the innovations variance. This consideration is crucial in the context of forecasting future observations.

As indicated by the authors, the leave-k-out diagnostic for the innovations variance, DV(A), is also useful in detecting level shift or variance change. This is an important application of the technique because it will result in better forecasting. However, although σ_A^2 is estimated by deleting set A, the diagnostic and its statistic do not fully reflect the size of the deleted set. It is not clear how DV(A) will perform if the size of A is relatively large. An alternative procedure is to use an F statistic reflecting both the size of A and its complement.

This can be accomplished by first transforming the data so that they are uncorrelated. The ratio of the variance estimates based on A and its complement can then be used as the statistic which is distributed as an F statistic with appropriate degrees of freedom. This idea is being exploited in a current study of forecasting telephone circuit counts with Nabil Ahmed and S. C. Horng of Bellcore.

R. J. Martin (University of Sheffield): I have recently (Martin, 1989a) been considering the generalization of the regression concepts of residuals, leverage and influence to the regression model with a known dependence structure. That is $E(y) = X\beta$ and $\text{var}(y) = V\sigma^2$ with known $V \neq I$, and β is estimated by generalized least squares. Although the generalizations are relatively straightforward, care is necessary. In particular with a deletion set A it is necessary to differentiate between the deviations $y - X\hat{\beta}$ and $y - X\hat{\beta}_A$, and the prediction residuals—the differences between y and the estimated (substituting $\hat{\beta}$ or $\hat{\beta}_A$ for β) predicted values of y in A. The difference $\hat{\beta} - \hat{\beta}_A$ is most simply expressed using the prediction residuals. Influence can be measured using a generalization of Cook's distance such as

$$(\hat{\beta} - \hat{\beta}_A)'X'V^{-1}X(\hat{\beta} - \hat{\beta}_A)/p\hat{\sigma}^2,$$

whereas $\hat{\sigma}^2/\hat{\sigma}_A^2 - 1$ can be used to check for outliers.

If it is assumed that V is unknown, then this model contains the usual regression model and the time series models as special cases. With a given time series model for V these statistics can be calculated substituting $\hat{V}$ or $\hat{V}_A$ for V, or the authors' statistic DV(A) based on estimates of the innovations variance could be used. Do the authors have any suggestions? Presumably appropriate diagnostics would depend on what ways are felt to be most likely for the model to be inappropriate. Recursive residuals may be most appropriate for detecting a change in the mean over time. For spatial data, the lack of a natural ordering means that it may not be sensible to think in terms of innovations. Perhaps the generalized Cook statistic would be useful in segmenting images. Residuals and some other diagnostics for this model have been considered in some applications, e.g. econometrics. An example in another area is given by Martin (1989b).

Professor Daniel Peña (Universidad Politécnica de Madrid): Building a measure of influence for time series models requires

(a) a method of extending the deletion approach used in static models to dependent observations and
(b) a measure of the parameter change for autoregressive integrated moving average (ARIMA) models.

In 1984 I showed (Peña, 1984) how to generalize the deletion approach used in regression to study influence in ARIMA models assuming that each observation was missing, and I related this technique to outlier estimation and to intervention analysis. Since then, I have worked on procedures to compute missing observations, to compare influence measures and to extend these procedures to transfer function analysis Peña (1985, 1986, 1987a,b). During this time I have often found that many competent statisticians did not believe that the missing value approach was a fruitful way to deal with influence data in time series. I am happy to see that, nowadays, this idea seems to be obvious. However, I found the authors' statement about Brillinger (1966) surprising. The phrase: 'this may be achieved on occasion by applying a missing values technique' is a clever suggestion to generalize the jackknife to a specific problem, but not an articulate method to deal with influence observations in time series.

On the second problem, the measure of distance between ARIMA models used by the authors has three problems:

(a) it does not allow us to compare ARIMA models with different degrees of differencing;

(b) it does not take into account the possibility of almost cancellation between the autoregressive (AR) and moving average (MA) operators;
(c) it does not allow for the duality between the AR and MA forms.

For instance, with their definition of empirical influence, model $Z_t = (1 - 0.4B)a_t$ is 'closer' to the model $Z_t = (1 - 0.8B)a_t$ than to $(1 + 0.4B + 0.16B^2 + 0.06B^3 + 0.02B^4)Z_t = a_t$, whereas clearly the third model is virtually identical with the first. Similarly, the model $(1 - 0.6B)Z_t = (1 - 0.59B)a_t$ is closer to $(1 - 0.6B)Z_t = a_t$ than to $Z_t = a_t$. A way to avoid this problem, Peña (1984, 1987), is to define the empirical influence as

$$\mathrm{EIC}(A) = -n(\hat{\pi}_A - \hat{\pi})$$

where the π coefficients are obtained from $\theta(B)\pi(B) = \phi(B)\nabla^d$.

In regression, influence analysis provides information that cannot be obtained by the usual outlier tests, because, as is well known, some very influential points (which produce large parameter changes) may not be outliers (which do not produce large variance changes). The authors have decided to choose as diagnostic for influence the variance change which is, as they recognize, equivalent to looking for outliers, and so the advantage of their procedure over the well-known methods for identifying outliers, level shift or variance changes (Tsay, 1988) is not clear.

Dr R. Schall (Institute for Biostatistics of the South African Medical Research Council, Tygerberg) **and Dr T. T. Dunne** (University of Cape Town): The authors are to be congratulated for their development of the swamping-free diagnostic DV(A), and exhibiting its relationship to DRES(A) and DFIT(A). The authors have pointed out the specific problems of 'smearing' or swamping not encountered in the usual independent observation setting. Their proposed solution, use of DV(A) instead of DC(A), is striking. They have also situated their diagnostic within an extensive and flexible array of strategies for exploring multiple outliers, and other phenomena, iteratively. We ask whether DV(A) can be shown to be swamping free in general autoregressive integrated moving average models, as the theoretical results in the paper and an earlier technical report discuss only autoregressive models.

Could the authors comment on the following? Influence of one or more observations or innovations can be assessed in terms of likelihood perturbations, and the corresponding likelihood displacement, in the manner of Cook (1986). In that context Schall and Dunne (1988) distinguish between diagnostics for the local (potential) influence of perturbations and their actual (observed) influence. Diagnostics for the local or potential influence of perturbations are developed by applying the general methodology of Cook (1986) to the time series context. In contrast, we proposed to measure the actual or observed influence of perturbations by the observed likelihood displacement due to the set of perturbations in question. A second-order approximation of the observed likelihood displacement *due to additive observation outliers* leads to the leave-k-out diagnostics proposed by the authors. Thus the diagnostics proposed by the authors can be motivated using an approach based on the likelihood displacement.

However, leave-k-(observations)-out diagnostics are not appropriate diagnostics under an *innovative outlier* perturbation scheme, where a 'leave-k-(innovations)-out' approach (treat k *innovations* as missing) should rather be applied. Further, variance shift perturbation schemes lead to diagnostics which involve a *downweighting* of observations or innovations (depending on the perturbation scheme), rather than their removal.

In summary, specific diagnostics are available for specific perturbation schemes, while the authors apply their leave-k-(observations)-out diagnostic to both the observation and innovation outlier situation, and suggest that we use the diagnostic even for the detection of variance shifts.

Dr J. Q. Smith (University of Warwick, Coventry): Other types of disruptive events particularly in economic and business series such as shifts in mean and changes in growth as well as parameter changes occur with similar frequency to outliers. The procedures described here are essentially for the detection of outliers and despite claims to the contrary by the authors do not stand on their own. They are possibly helpful as one of many diagnostic tools.

What disturbs me most about this paper is the authors' apparent advocation that we should fit poor models to data, look for 'outliers' which we intend to ignore and then refit. In example 5 to choose a model whose first differences have zero expectation seems eccentric both from the context of the series and the actual model. The assumption of constant predictive variance (not to mention additive seasonality and first differencing) in example 6 which is so obviously false just from the graph of the data is even more dubious. Surely fortune will dictate whether outliers defined relative to bad models can be usefully omitted.

Only when the modelling process is taken seriously might k-out diagnostics have some value. They can be incorporated into Bayesian forecasting models in the following very simple way without fancy asymptotics. Suppose we have built a Bayesian forecasting system (Dawid, 1984) where the next observation Y_t given the past $y_1 \ldots, y_{t-1}$ (after the required parameters have been integrated out) has a continuous distribution function F_t. Then $\{U_t\}_{t\geqslant 1}$ where $U_t = \Phi^{-1}[F_t(Y_t)]$ is a sequence of *independent* normal variates with zero mean and unit variance, where Φ is the standard normal distribution function (see Smith (1985)). This will remain true after we have modified F_t to allow for outliers. So effects of various sets of observations, on generalized residuals $\{U_t\}$, can be straightforwardly calculated and their distributions known, both under null and alternative hypotheses. These results do not depend on the original model having a linear structure or that the process is Gaussian, so are quite general. They also have the desirable property of centring diagnostics on forecast performance rather than parameter estimation.

As the authors point out the existence of serial correlation often makes k-out diagnostics preferable to one-out diagnostics. However, there is another reason why they might be preferred. In the context of volumes of sales (e.g. example 5) a blockage in recording sales may give rise to those numbers being added to sales figures in the immediate future. Alternatively the blockage may continue for several periods; thus outliers might naturally occur in patches. When this is the case it is dubious whether these 'outlying' values should be ignored altogether in the modelling process. They need not be in the method outlined here.

T. Subba Rao (University of Manchester Institute of Science and Technology): Does the assumption that the innovations $\{\varepsilon_t\}$ of model (2.1) need to be Gaussian for the diagnostic criteria (2.8) and (2.11) to be valid? We could estimate the parameters of the linear model $\boldsymbol{\alpha}$ and the variance σ^2 without this assumption and by imposing suitable conditions we could prove asymptotic normality of $\hat{\boldsymbol{\alpha}}$. Now, coming to the criterion DV(A), is it not more appropriate to take logarithms of the estimates $\hat{\sigma}^2$ and $\hat{\sigma}_A^2$ and to compare these, e.g. $\log\hat{\sigma}^2 - \log\hat{\sigma}_A^2$, suitably scaled. This is for the simple reason that $\log\hat{\sigma}^2$ tends to normality faster than σ^2. An alternative way of looking at the criteria is as follows. We have

$$\log\sigma^2 \propto \int \log f(\omega)\, \mathrm{d}\omega,$$

where $f(\omega)$ is the second-order spectral density function of the process. Therefore, we could consider $\int\{\log\hat{f}(\omega) - \log\hat{f}_A(\omega)\}\,\mathrm{d}\omega$ for $\log\hat{\sigma}^2 - \log\hat{\sigma}_A^2$, where $\hat{f}(\omega)$ and $\hat{f}_A(\omega)$ are nonparametric estimates of the spectral density function (here $\hat{f}_A(\omega)$ is the estimate of the spectral density function when the set A is omitted). More appropriate criteria may be either $\int\{\log\hat{f}(\omega) - \log\hat{f}_A(\omega)\}^2\,\mathrm{d}\omega$ or

$$\int\left\{\frac{\log\hat{f}(\omega) - \log\hat{f}_A(\omega)}{\log\hat{f}_A(\omega)}\right\}^2 \mathrm{d}\omega.$$

Since $\hat{f}(\omega)$ and $\hat{f}_A(\omega)$ are nonparametric estimates, the problem of the order determination which is quite essential for the diagnostic tests does not arise. What do the authors think about using such criteria for diagnostics?

Here the authors are testing for the influential points in the form of outliers, under the assumption that the underlying model is linear (and Gaussian). After more than 10 years of research in the field of non-linear time series, I believe that we must first test whether a series is linear or non-linear before any further analysis is carried out, for the simple reason that what looks like an outlier may be a part of the underlying non-linear structure. I pointed this out in a contribution to the discussion of a paper presented by Kleiner *et al.* (1979) to this society. Unfortunately, the authors have not pointed out this possibility.

Professor Ruey S. Tsay (Carnegie–Mellon University, Pittsburgh): The present paper provides yet another useful approach for checking various types of disturbance in time series analysis. In particular, the ideas of using innovation variance and graphical display are of great interest to me. However, I have some concerns and comments about the proposed approach.

(a) The leave-k-out diagnostics appear to be relatively inefficient in handling variance changes in a time series. Consider the unfilled radio and television orders series of example 6. A simple time plot of the first-differenced series, i.e. $(1 - B)x_t$, shows clearly the non-homogeneous nature of the process; see Fig. 13. Similarly, the simple technique of Tsay (1988) for detecting variance changes also clearly shows a variance change. I used the logarithm of the data and still detected

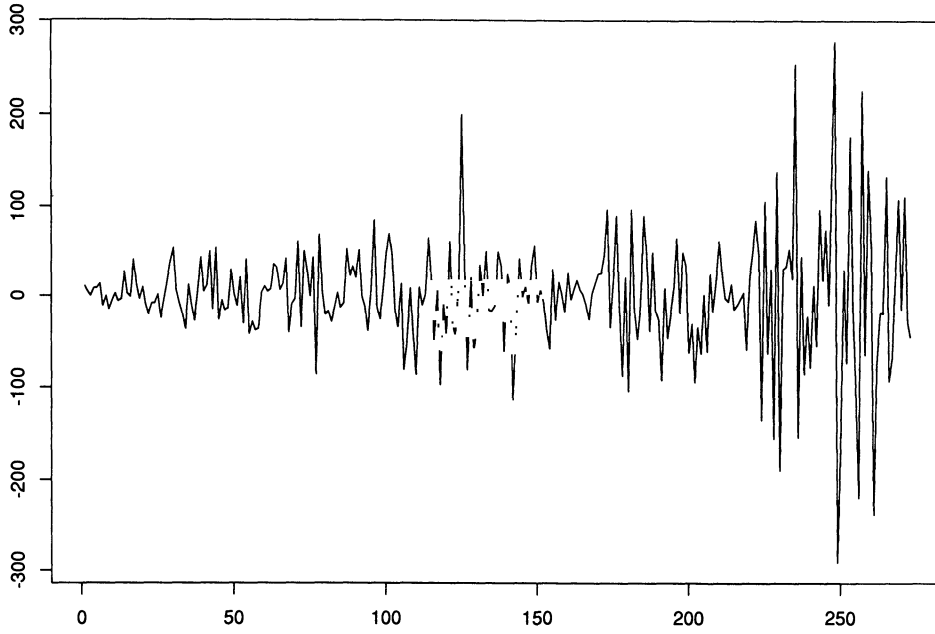

Fig. 13. First-differenced unfilled radio and television orders series

a variance change in 1976. It seems that leaving-k-out in *computing* variance estimates, not in estimating the model, is a simpler method for detecting variance change.

(b) Deleting all aberrant observations might not be the optimal way in time series analysis, especially for patches of outliers. In some situations, we are interested in assessing the impact of the disturbance. Thus, using the intervention analysis of Box and Tiao (1975) in conjunction with time series diagnostics seems more useful. This is the approach taken by Tsay (1988).

(c) The reason that DC is not satisfactory when the process is nearly non-stationary or non-invertible is the failure of the normal approximation used in Section 2. Consider the autoregressive models, for instance. The least squares estimates of the autoregressive parameters are not asymptotically normal when some of the characteristic roots are on the unit circle. See Chan and Wei (1988), and Tsay and Tiao (1989), among others. Consequently, in studying scaling problems the asymptotic results of non-stationary time series might be useful.

(d) The proposed model for the unfilled radio and television orders series from 1976 to 1980 appears to be non-invertible; see Table 2. Is this not in conflict with the assumptions of the paper?

(e) The effects of outliers in time series analysis were investigated by Chang (1982) for the general autoregressive integrated moving average models. In particular, the effects of outliers on the sample autocorrelation function and least squares estimates of autoregressive regressions were given.

The **authors** replied later, in writing, as follows.

Professor Tong hoped that presenting the paper would be pleasurable for us. For us, the pleasure of such an experience is measured by confidence that our efforts have produced at least some advance in the understanding of time series diagnostics, a bearable amount of justified criticism and the provocation of a wide range of new thoughts and proposals for further research. By this criterion, we were far from disappointed!

A sizable number of substantial issues have been raised. We address some of these issues below.

Appropriateness of 'missingness' for time series

Professor Lawrance, Professor Tong and Professor Tsay question the appropriateness of subset deletion diagnostics in which the deleted data are treated as missing. To this we reply that the missing

data approach is at a fundamental level the most natural, transparent and easy to understand concept for the broadest range of applied statisticians. In addition the leave-k-out approach based on the diagnostics DV and DFIT have merit well beyond its conceptual simplicity.

(a) DV clearly displays the influence caused by simple outliers. This is an important property for any time series diagnostic to have, and one which DC lacks. This property is analogous to the ability of Cook's distance to reveal clearly the 'all-but-one point on a line' problem in bivariate regression (i.e. all observations lie roughly on a regression line except for one outlying value).
(b) The leave-k-out strategy is general in that it can detect a variety of potentially troublesome situations. It is very powerful for detecting both isolated outliers and patches of outliers, and is reasonably sensitive towards level shifts and variance changes (we comment further on level shifts later).
(c) As Dr Bell points out, the missing data approach is maximally non-committal towards various kinds of influential data segments, and no additional modelling assumptions are made. This is especially important in the exploratory stages of an analysis.
(d) Finally, missing data methods can help the analyst to understand the data better. Data values have significance which can be related to the underlying problem.

Likelihood ratio tests

Several discussants (Burman, Bell, Hallin and Mélard, Kohn and Ansley, Peña, Schall and Dunne and Tsay) mention alternative methods based on the use of likelihood ratio tests (LRTs) of specific parametric interventions, as in the work of Chang and Tiao (1983), Hillmer *et al.* (1983), Peña (1987) and Tsay (1988). LRTs have been proven to be a valuable and useful diagnostic tool. However, there is room for both LRTs and leave-k-out methods.

As discussed in the previous section, missing data methods have some fundamental advantages not shared by diagnostics based on LRTs. In addition, we emphasize that DV is *not* equivalent to testing for the presence of additive outliers for finite sample sizes, as is clear from the decomposition of equation (6.5).

We apparently overstressed the asymptotic results in Sections 6 and 7. In finite samples, DFIT can be and often is non-negligible. Furthermore, as Critchley points out, the 'fixed k/n' case leads to quite different asymptotics than the 'fixed k' case. In particular, the missing data approach and the additive outliers approach are no longer asymptotically equivalent (see the equations in Section 7). Indeed, the fixed k/n case is likely to be more relevant for finite sample size inference.

To illustrate finite sample realities, we give an example in which DFIT dominates DRES. Quarterly Argentine wood production data from 1970 to 1985 are plotted in Fig. 14(a). The series exhibits some unusual features, including a hump in 1974 and a possible level shift in 1980. Based on Akaike's information criterion (AIC) and the usual Box–Jenkins methodology, an ARIMA$(1, 0, 0) \times (0, 1, 2)_4$ model was fitted. Leave-three-out diagnostics are given in Fig. 14(b). Instead of plotting DV, we have plotted DV_0, given by the vertical lines, and $DFIT/\sqrt{(2n)}$, given by the curve passing through the vertical lines. The straight horizontal line corresponds to the rough guideline used in the previous examples. Clearly, DFIT is the dominant term in the most influential patches. A large value of DFIT means that the model parameters are heavily influenced by the deleted data. Correspondingly, one often finds that the model has been misspecified. In the present context this gives rise to large values of DFIT (and DV_0) around 1977 rather than during 1974. The remedy is careful model selection along the lines that we suggest in the model fitting questions section (see p. 421). See Bruce (1988) for a more complete discussion of this example.

A final word on LRT methods: an important aspect of any diagnostic procedure for time series is to expose locally influential patches of observations. This is a potential shortcoming of current methodology based on using LRTs, and we would like to see further development in this area. Furthermore, most of the proposed LRT methods for handling outliers one at a time lack power relative to leave-one-out diagnostics because they use parameter estimates based on all the data.

Exclusive focus on autoregressive integrated moving average models

Concern is expressed by Professor Cox, Professor Harvey, Dr Robinson and Dr Smith over our emphasis on autoregressive integrated moving average (ARIMA) models. This concern is well taken. The main reason for such focus is simply that ARIMA modelling is widely used and very popular. However, we agree that ARIMA modelling is considerably overused and often abused.

In view of this, Harvey's emphasis on using structural models for time series seems quite appropriate. However, we question whether using a simple smoothing of the irregular component is, in general,

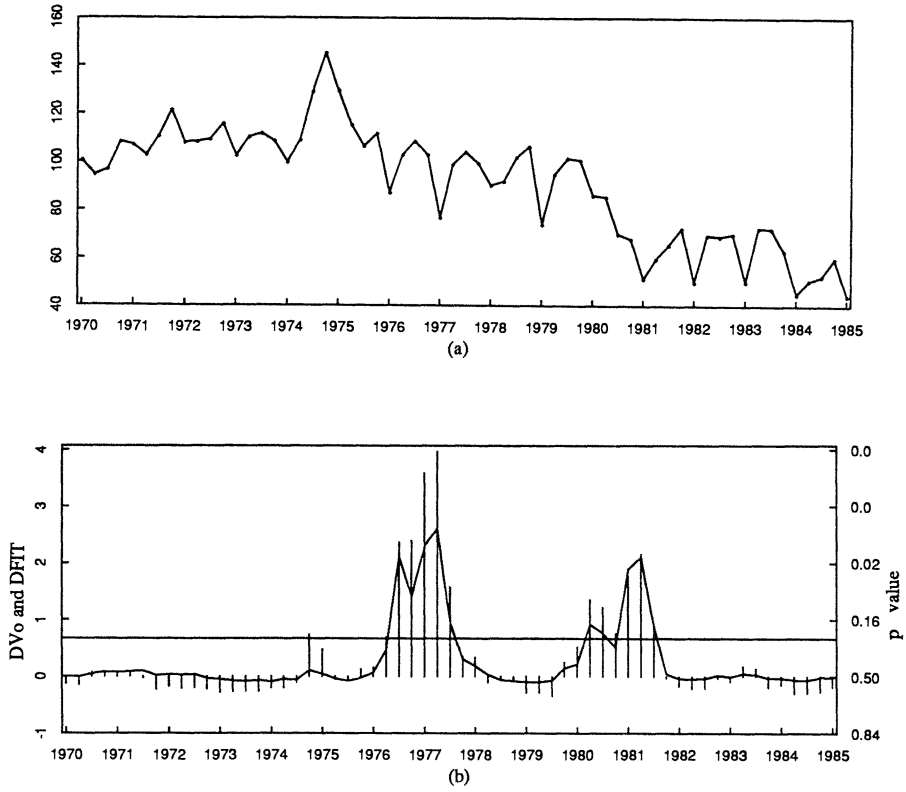

Fig. 14. Quarterly Argentine wood production data from 1970 to 1985: (a) plot of the data; (b) plot of DV_0 (vertical lines), $DFIT/\sqrt{(2n)}$ (curve) and a rough significance guide (horizontal line)

sufficient to identify outliers. At the very least, it would seem wise to use a robust smoothing procedure, such as those referenced in Section 8.2.

We note that an AR approximation will often do quite well as a substitute for ARMA modelling. Whenever AR approximations are sufficient, we can construct leave-k-out diagnostics for other quantities of interest (e.g. correlation, autocorrelation, spectrum) in a straightforward manner. The statistic of interest is computed with the missing data interpolated using the AR approximation, and then compared (using an appropriate metric) with the statistic computed using the complete data set.

In this regard, Professor Subba Rao's suggestion for basing a diagnostic on estimated spectral densities, with and without subsets deleted, seems quite promising. To this end one could construct prewhitened spectral density estimates via the AR missing data approach, much in the spirit of Kleiner *et al.* (1979).

As for Professor Cox's query about possible 'direct' methods for other quantities of interest, we do not yet see a method so nearly appealing as that based on interpolation using the AR approximation.

Focused objectives and influence

Professor Cox, Dr Critchley, Professor Ledolter and Dr Robinson quite correctly point out the need for basing the diagnostics on the final objective of the analysis. In particular, Ledolter (1989) has provided some very nice work in the area of forecasting. As Professor Ledolter indicates in his contribution, except for points near the end of the series, the leave-k-out procedure based on DV is adequate for detecting influence in the forecasting problem.

Dr Robinson and Professor Ledolter caution that deletion of outliers may result in underestimation of the innovations variance for forecasting. However, a clearly inadequate approach for incorporating the increased uncertainty caused by outliers is to use wider Gaussian prediction intervals (reflecting an inflated innovations variance). It is important to have a good measure of the scale for the core

(outlier-free) process; how we decide to model the aberrant values is often a separate and more difficult issue. For instance, in example 5, constructing prediction intervals for the core (strike-free) process is relatively straightforward. However, attempting to predict the increased variability due to possible future strikes is a highly subjective task depending heavily on external factors.

Local influence and likelihood displacement

Dr Critchley, Dr Cook, Professor Ledolter and Dr Schall and Dr Dunne mention local influence methods based on likelihood displacements. These methods may serve as a useful complement to leave-k-out diagnostics. Of particular interest is the suggestion by Critchley and Cook that local influence may avoid the masking problem in the time series setting. At the moment this does not seem likely to us.

In answer to Cook's query about influence measured by likelihood displacement in relation to EIC and EIV, we note the following. One could use the diagnostic

$$\mathrm{LD}(A) = 2\{L(\hat{\alpha}, \hat{\sigma}^2) - L(\hat{\alpha}_A, \hat{\sigma}_A^2)\}.$$

As Cook suggests, this is likely to produce roughly the same results as DC and DV, and asymptotically it can be shown that

$$\mathrm{LD}(A) = \mathrm{DC}(A) + \mathrm{DV}(A) + o(1/n).$$

In view of the effects of smearing, it might be better to consider likelihood displacement caused only by innovations variance

$$\mathrm{LDV}(A) = 2\{L(\hat{\alpha}_A, \hat{\sigma}^2) - L(\hat{\alpha}_A, \hat{\sigma}_A^2)\}$$

which is asymptotically equivalent to $\mathrm{DV}(A)$. One might then decompose LDV in a similar fashion to the decomposition given for DV in Section 6.

Calibration, p values and formal inference

Calibration of the diagnostics has been discussed by many of the contributors (Critchley, Cook, Hallin and Mélard, Kohn and Ansley, Künsch and Lee), and some (Critchley, Kohn and Ansley, and Künsch) are concerned about distributional properties of the final estimate after deletions. These problems present considerable intractability, and we wish well to those who might attack them. Should one seriously approach the distributional problem, then clearly attention must be given to the 'fixed k/n' case.

Professor Hallin and Professor Mélard and Dr Lee contend, quite correctly, that we should adjust for the size of the patch. Hallin and Mélard's suggestion to use the AIC or B information criterion (BIC) guideline is attractive. We caution that the AIC and BIC are based on the log-likelihood function, which is not quite identical with the logarithm of the estimated innovations variance, and perhaps an adjustment is needed for using information criteria in the missing data setting.

In spite of having some interest in these issues, we agree with Dr Cook in that we should not be too worried about having in hand the distribution of influence measures. We would avoid the term 'p value' in the future as quite misleading. The proposal by Cook to use an influence measure only for 'ranking' the observations seems to be the most sensible and useful approach to the problem.

Computational complexity and modern graphics workstations

As several of the discussants point out, the computing burden demanded by a full implementation of leave-k-out diagnostics is very intensive, and for practical application (especially for seasonal models) faster algorithms are needed. However, with some improvement in computational efficiency, the computational complexity will not be a limiting factor when implemented on currently available workstations.

The primary way to gain computational efficiency is to base the diagnostics on one-step estimates from the maximum likelihood solution with all the data. Preliminary studies indicate that one-step estimates are often sufficient (Bruce and Martin, 1987), though we recommend taking further iterations whenever a large initial step is taken. In most of the examples given in this paper, one-step estimates would suffice. Further, an $O(n)$ algorithm is available to produce one-step estimates needed to speed up the leave-k-out diagnostics. Details of the procedure are forthcoming.

Graphics workstations offer the ideal environment to make use of interactive graphical techniques (such as those mentioned in Section 4.3 and by Bradley and Haslett) in conjunction with such techniques as leave-k-out diagnostics. In particular, the LISP-based program for interactive time series and spectral

analysis system (Kerr and Percival, 1988) is a powerful environment for time series analysis, blending sophisticated interactive tools with advanced spectral analysis methods. As such it would support the use of leave-k-out diagnostics quite nicely.

Desire for simple methods

The desire for 'simple' methods is strong in all areas of statistics where more complex methodologies are being proposed. This is also true in the present context, as is evidenced by the pleas for simple methods by Dr Robinson, Professor Tong, Mr Bradley and Dr Haslett, Dr Chatfield and Professor Tsay. Any good data analyst would agree that simple methods play a fundamental role, and there is no substitute for careful exploratory graphical analysis.

However, even in the univariate case, the dependencies in time series data can make it difficult to determine influential values from simple plots. Only in very simple situations is the exact nature of the influence of observations in an outlier patch fully transparent; see Kleiner *et al.* (1979) for a striking example of 'hidden' influential values. Furthermore, even for obvious model changes or outliers, it is reassuring to have one's intuition supported by a more formal tool (this is especially so for less experienced analysts).

Simplicity must be balanced with the proper blend of generality and power. Complexity at some level may be acceptable, provided that it is offset by transparency or power at another level. Excessive emphasis on simplicity may lead to a much weaker diagnostic. For example, this seems to be the case with Tong's proposal to apply the leverage/Mahalanobis-distance-based diagnostic (i.e. the diagonals of the hat matrix). We would not advocate the use of leverage in general since there is a clear smearing effect: a single outlier will show up at several time points with different amplitudes which depend on the estimated covariance structure. Further, this covariance structure can itself be highly influenced by outliers to an extent that renders the diagnostic useless (as in ordinary regression). To illustrate the smearing problem, in the AR(1) model, note that

$$h_t = y_{t-1} / \sum_t y_t^2.$$

Therefore for example 1, the leverage diagnostic would identify point 29 as the overwhelmingly most influential point, which is clearly misleading.

Model fitting questions

Burman, Harvey and Smith have criticized the actual models that we fitted in our examples. For the most part, we have no defence here and greater care should have been taken. Indeed, we should have included a slope term for example 5, as pointed out by Smith, and Burman is correct in that the seasonal operator that we chose in example 6 is probably not the best. We were more preoccupied with the development of the methodology.

At the same time the fact that we may not have fit the best models does not in any way invalidate the proposed diagnostics. Examples in which more attention was paid to model fitting aspects validate the results presented here (see Bruce (1988)). It is a strength of the approach that influential points in isolation or patches are found with not quite correct models. There is quite a difference between a truly bad model and a model which has a reasonable degree of predictive power, though not the best predictive power obtainable. The former would generally be useless for diagnostics, but not the latter.

Harvey's point that a different model holds after the identification of outliers is quite pertinent. Following our suggestions for model identification in Section 4.4.1, we do end up fitting a seasonal term as suggested by Harvey. At the same time we cannot caution too strongly that we may not end up at an acceptably good model by this approach. In general, it will be necessary to entertain a range of models at the outset, and to use leave-k-out diagnostics on each of them to arrive at a good model. This is the case for the Argentine wood example of Fig. 14.

Other issues

Having addressed these frequently occurring concerns, we turn to several other interesting points raised by one or more discussants.

Missing data approaches and level shifts. We have stated previously that DV has reasonable power to detect a level shift. An intuitive rationale for this is that leaving out a sufficiently long patch of observations surrounding a level shift allows the data to interpolate through the level shift. It would be useful to have more analytical results in the spirit of Section 3.2 on this issue.

Checking for non-linearities. We agree with Professor Subba Rao that it is important to check for non-linearities in a time series. However, it is not so clear that we should show a preference for first testing for non-linearity before checking for outliers. For instance, exploratory methods for fitting non-linear additive regression models can be extremely sensitive and unreliable in the presence of outliers (Breiman and Friedman, 1985). What is evident is that we should, at some stage in the analysis, check for both outliers and non-linearity.

Put P in. The 'put-*P*-in' idea raised by Professor Hawkins is particularly interesting to us. With V. Yohai, we have spent time looking into the use of such an approach for obtaining initial parameter estimates for autoregressions which have high breakdown point with high probability (an approach which has been pursued in regression by several investigators). Preliminary investigations for autoregressions indicated that, for all but low order autoregressions, the sample size required to obtain a high breakdown point was rather large. Thus we focused on another approach for obtaining high breakdown point autoregression estimates, which we hope to publish elsewhere. At the same time, put-*P*-in methods remain very attractive, and we welcome Hawkins's novel suggestion for use in a diagnostics setting.

Multivariate case. We could treat a multivariate time series just one series at a time using the current leave-k-out approach. However, this will sometimes be inadequate. A better approach is to mimic the univariate procedure by developing a multivariate analog for DV. The multivariate setting is more complicated since we must look at C and C_A, the estimated innovations vector variance–covariance matrices for the full data set and with subset A treated as missing. A reasonable start might be to define diagnostics based on

(a) changes in 'scale', e.g. as measured by $|\det(C)/\det(C_A)|$, and
(b) changes in the 'shape', e.g. as measured by appropriate functions of the eigenvalues.

Künsch's jackknife and sliding spans. Künsch's (1989) work on the jackknife for time series and the idea to delete blocks (of increasing length) are quite important. This idea is related to the 'bottom-up' method discussed in Section 4.3; we concur with his closing suggestion that it would be interesting to investigate the diagnostic capabilities of this method further.

Peña's measure (and other comments). Professor Peña proposes an empirical influence measure based on the π weights. We caution that this approach is clearly subject to smearing effects: the resulting diagnostic is equivalent to DC for AR models. With regard to Peña's questioning of our reference to Brillinger (1966), we refer to Brillinger (1986) in which he proposes a diagnostic similar in spirit to DC, referencing his own 1966 paper as setting the stage for the missing data approach in time series. We would also like to clarify that Peña's criticisms of the 'measure of distance between ARIMA models used by the authors' apply *only* to DC, and have no bearing on the diagnostic DV which we recommend.

Invertibility assumption. Professor Tsay is correct that we violated the assumption that the process is invertible in example 6; however, this assumption is apparently not required and may be dropped (Kohn and Ansley, 1986).

Questions not answered

Many other important questions and comments were raised in which we either lacked a good response or time and space limitations prohibited a reply. These issues include an analytical formula for masking (Lawrance), existence of other less structured situations where smearing is a problem (Jolliffe), regression models with dependence structure (Besag and Seheult, Loynes and Martin), innovation deletion (Muirhead), spatial statistics (Bradley and Haslett, and Martin) and Bayesian forecasting models (Smith).

REFERENCES IN THE DISCUSSION

Atkinson, A. C. (1985) *Plots, Transformations, and Regressions: an Introduction to Graphical Methods of Diagnostic Regression Analysis.* Oxford: Clarendon.

Barham, S. Y. and Dunstan, F. D. J. (1982) Missing values in time series. In *Time Series Analysis: Theory and Practice 2* (ed. O. D. Anderson), pp. 25–51. Amsterdam: North-Holland.

Bell, W. R. (1983) A computer program for detecting outliers in time series. In *Proc. Amer. Statist. Ass., Bus. Econ. Statist. Sect.*, pp. 634–639. Washington DC: American Statistical Association.

Besag, J. E. and Kempton, R. A. (1986) Statistical analysis of field experiments using neighbouring plots. *Biometrics*, **42**, 231–251.

Box, G. E. P. and Tiao, G. C. (1975) Intervention analysis with applications to economic and environmental problems. *J. Amer. Statist. Ass.*, **70**, 70–79.

Breiman, L. and Friedman, J. H. (1985) Rejoinder to Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Ass.*, **80**, 614–619.
Brillinger, D. R. (1966) Discussion on Linear functional relationships (by P. Sprent). *J. R. Statist. Soc.* B, **28**, 294.
——(1986) Discussion on Influence curves for time series (by R. D. Martin and V. J. Yohai). *Ann. Statist.*, **11**, 781–818.
Bruce, A. G. (1988) Diagnostics for time series models. *PhD Thesis*. University of Washington, Seattle.
Bruce, A. G. and Martin, R. D. (1987) Leave-*k*-out diagnostics for time series. *Technical Report 107*. Department of Statistics, University of Washington, Seattle.
Burman, J. P. (1983) Comments on Modeling considerations in the seasonal adjustment of economic time series (by S. C. Hillmer, W. R. Bell and G. C. Tiao). In *Applied Time Series Analysis of Economic Data* (ed. A. Zellner), pp. 101–105. Washington DC: US Bureau of the Census.
Chan, K. S., Moeanaddin, R. and Tong, H. (1988) Some difficulties of non-linear time series modelling. *Technical Report*. Institute of Mathematics, University of Kent at Canterbury.
Chan, N. H. and Wei, C. Z. (1988) Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.*, **16**, 367–401.
Chang, I. (1982) Outliers in time series. *PhD Thesis*. Department of Statistics, University of Wisconsin, Madison.
Chang, I. and Tiao, G. C. (1983) Estimation of time series parameters in the presence of outliers. *Technical Report 8*. Statistical Research Center, University of Chicago.
Chauvet, P. (1982) The variogram cloud. *Internal Report N-725*. Centre de Géostatistique.
Cook, R. D. (1977a) Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.
——(1977b) Letter to the Editor. *Technometrics*, **19**, 349.
——(1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.
Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.
Dawid, A. P. (1984) Statistical theory—the prequential approach (with discussion). *J. R. Statist. Soc.* A, **147**, 278–292.
Findley, D. F., Monsell, B. C., Otto, M. C., Bell, W. R. and Pugh, M. G. (1988) Toward X-12 ARIMA. *Proc. 4th A. Res. Conf.*, pp. 101–105. Washington DC: US Bureau of the Census.
Harvey, A. C. and Durbin, J. (1986) The effects of seat belt legislation on British road casualties: a case study in structural time series modelling (with discussion). *J. R. Statist. Soc.* A, **149**, 187–227.
Hau, M. C. (1984) Some problems in time series modelling. *MPhil Thesis*. Chinese University of Hong Kong, Hong Kong.
Hau, M. C. and Tong, H. (1984) Outlier detection in autoregressive time series modelling. *Technical Report 15*. Department of Statistics, Chinese University of Hong Kong, Hong Kong.
Hawkins, D. M., Bradu, D. and Kass, G. V. (1984) Location of several outliers in multiple regression data, using elemental sets. *Technometrics*, **26**, 197–203.
Hillmer, S. C., Bell, W. R. and Tiao, G. C. (1983) Modeling considerations in the seasonal adjustment of economic time series. In *Applied Time Series Analysis of Economic Data* (ed. A. Zellner), pp. 74–100. Washington DC: US Bureau of the Census.
Johnson W. and Geisser S. (1983) A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Ass.*, **78**, 137–144.
Kerr, R. K. and Percival, D. B. (1987) Use of object-oriented programming in a time series analysis system. *SIGPLAN Notices*, **21**, 1–10.
Kleiner, B., Martin, R. D. and Thomson, D. J. (1979) Robust estimation of power spectra (with discussion). *J. R. Statist. Soc.* B, **41**, 313–351.
Koenker, R. W. and D'Orey, V. (1987) Algorithm AS 229: Computing regression quantiles. *Appl. Statist.*, **36**, 383–393.
Kohn, R. and Ansley, C. F. (1986) Estimation, prediction, and interpolation for ARIMA models with missing data. *J. Amer. Statist. Ass.*, **81**, 751–761.
——(1988) A fast method of estimating outliers and level shifts in state space models. To be published.
Künsch, H. R. (1984) Infinitesimal robustness for autoregressive processes. *Ann. Statist.*, **12**, 843–863.
——(1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, in the press.
Ledolter, J. (1988) Outlier diagnostics in time series analysis. *Technical Report 145*. Department of Statistics and Actuarial Science, University of Iowa.
——(1989) The effect of additive outliers on the forecasts from ARIMA models. *Int. J. Forecast.*, **5**, in the press.
Martin, R. J. (1989a) Leverage, influence and residuals when errors are correlated. *Research Report 326/89*. Department of Probability and Statistics, University of Sheffield.
——(1989b) The use of time series models and methods in the analysis of agricultural field trials. *Commun. Statist. Theor. Meth.*, to be published.
Mélard (1981) On an alternative model for intervention analysis. In *Time Series Analysis* (eds O. D. Anderson and M. R. Perryman), pp. 345–354. Amsterdam: North-Holland.
——(1984) Algorithm AS 197: A fast algorithm for the exact likelihood of autoregressive–moving average models. *Appl. Statist.*, **33**, 104–114.
Muirhead, C. R. (1986) Distinguishing outlier types in time series. *J. R. Statist. Soc.* B, **48**, 39–47.
Papadakis, J. (1984) Advances in the analysis of field experiments. *Proc. Acad. Ath.*, **59**, 326–342.
Pearlman, J. G. (1980) An algorithm for the exact likelihood of a high-order autoregressive–moving average process. *Biometrika*, **67**, 232–233.

Peña, D. (1984) Influential observations in time series. *Technical Report 2718*. Mathematics Research Center, University of Wisconsin, Madison.

——(1985) A measure of influence in autoregressive models. *Bull. Int. Statist. Inst.*, **2**, 481–482.

——(1986) Discussion on Assessment of local influence (by D. R. Cook). *J. R. Statist. Soc.* B, **48**, 164–165.

——(1987a) Measuring the importance of outliers in ARIMA models. In *New Perspectives in Theoretical and Applied Statistics* (eds M. Puri *et al.*), pp. 109–118. New York: Wiley.

——(1987b) Measuring influence in dynamic regression models. *Technical Report*. Statistics Research Center, University of Chicago.

Portnoy, S. (1987) Using regression fractiles to identify outliers. In *Statistical Data Analysis Based on the L1-norm and Related Methods* (ed. Y. Dodge). Amsterdam: Elsevier.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression & Outlier Detection*. New York: Wiley.

Schall, R. and Dunne, T. T. (1988) Diagnostics and robust estimation for regression ARMA time series. *Technical Report 4/88*. Institute for Biostatistics of the South African Medical Research Council, Tygerberg.

Sew-Hee, K. K. K. Y. (1988) Detection of changes in time series. *PhD Thesis*. University of Cambridge.

Smith, J. Q. (1985) Diagnostic checks of non-standard time series models. *J. Forecast.*, **4**, 283–291.

Tong, H. (1977) Some comments on the Canadian lynx data (with discussion). *J. R. Statist. Soc.* A, **140**, 432–436, 448–468.

Tsai, C. L. (1986) Score test for the first-order autoregressive model with heteroscedasticity. *Biometrika*, **73**, 455–460.

Tsay, R. S. (1988) Outliers, level shifts, and variance changes in time series. *J. Forecast.*, **7**, 1–20.

Tsay, R. S. and Tiao, G. C. (1989) Asymptotic properties of multivariate nonstationary processes with applications to autoregressions. *Ann. Statist.*, to be published.

Wilkinson, G. N. (1984) Nearest neighbour methodology for design and analysis of field experiments. *Proc. 12th Int. Biometrics Conf., Tokyo*, pp. 64–79.