Working Paper 365

May 1992

AN INFLUENCE METHOD FOR OUTLIER DETECTION APPLIED TO TIME SERIES TRAFFIC DATA

S M Watson, S Clark, M Tight and E Redfern

ABSTRACT

The applicability of an outlier detection statistic developed for standard time series is assessed in estimating missing values and detecting outliers in traffic count data.

The work of Chernick, Downing and Pike (1982) is extended to form a quantitive outlier detection statistic for use with time series data. The statistic is formed from the squared elements of the Influence Function Matrix, where each element of the matrix gives the influence on ρ_k of a pair of observations at time lag k. Approximate first four moments for the statistic are derived and by fitting Johnson curves to those theoretical moments, critical points are also produced. The statistic is also used to form the basis of an adjustment procedure to treat outliers or estimate missing values in the time series. Chernick et al's (1982) nuclear power data and the Department of Transport's traffic count data are used for practical illustration.

An Influence method for outlier detection applied to Time Series traffic data.

1. INTRODUCTION

Many types of transport data are collected over time and can therefore be considered as suitable candidates for time series analysis. Examples include the Department of Transport core census of traffic counts, ticket sales data and journey time information. A serious problem in analysing traffic count data (for, say, forecasting purposes) is what to do when missing or extreme values (outliers) occur in the series. This can arise for a variety of reasons, such as broken counters or other machinery. In practice, missing data are currently patched in a variety of ways, for example using grossing up factors calculated using data from other areas and some computerized algorithms. Little work has been undertaken to assess the merits of alternative methods or develop a more analytical approach.

Recently more attention has been given to the detection and treatment of outliers in a more general time series context. Statistical methods for modelling and forecasting time series data are often based on the popular Box-Jenkins (1976) ARIMA time series modelling philosophy. However this approach relies heavily upon the sample autocorrelation function, r_k, which can be severely depressed by the presence of outliers (Walker, 1960).

Despite this, until recently little attention has been given to the detection and treatment of outliers. The work of Fox (1972) and more recently Hau and Tong (1984), Tsay (1986), Tiao (1985), Tsay (1988), Abraham and Chuang (1989) and Bruce and Martin (1989) have created interest in detecting and modelling outliers. Chernick, Downing and Pike (1982) developed a method of outlier detection using a plot of the Influence Function of data points on the theoretical autocorrelation function. The method relies largely on subjective visual interpretation of a graphical plot. In this paper, the work of Chernick et al (1982) is extended to form a quantitative outlier detection statistic which in an amended form can replace the outlier detected, or may be used to estimate a missing observation in the series. Nuclear Power data is used to illustrate outlier detection and the treatment of both outlying and missing observations is illustrated using time series of traffic counts.

2. THE INFLUENCE FUNCTION AND SUBJECTIVE DETECTION OF OUTLIERS

Using similar notation to that of Chernick et al (1982) and based on Hampel's (1974) Influence Function, the influence, I, on the theoretical autocorrelation function, ρ_k, of any pair of observations k time points apart is defined by Chernick et al (1982) as:

$$I(\rho_k, (y_t y_{t+k})) = y_t y_{t+k} - \rho_k (y_t^2 + y_{t+k}^2) / 2 \qquad \ldots. (1)$$

Here y_t is the t'th observed value in a time series. An n x L Influence Function Matrix can be formed with elements given by (1) where n = sample size, L = a fixed number of Lags, and L << n.

If the j'th standardized observation of the time series is denoted by z_j, then the (j,k)'th entry of that matrix is given by $I(\rho_k, z_j, z_{j+k})$.

The matrix formed then consists of L columns and n rows, where z_j appears in each element of the j'th row and a diagonal beginning at the first column of the preceding row.

Chernick et al (1982) show that substitution and manipulation of (1) gives

$$I(\rho_k (z_i, z_{i+k})) = (1 - \rho_k^2)\, u_{ik}\ u_{ik}$$

where u_{ik}, $u_{ik} \approx$ iid N (0,1)

More concisely this may be written

$$I = G\, u_1\, u_2$$

where $G = (1 - \rho_k^2)$

Outlier detection using the Influence Function Matrix method relies on discerning a pattern of unusually "high" or "low" values of the Influence Function. Chernick et al (1982) use a critical value of +/- 1 to isolate those values that are significant. This is clearly inappropriate considering the need to estimate ρ_k in expression (1).

Those values of I found to be significant are then marked in a graphical plot (size n x L) by a "+" or "-" sign. A "+" sign indicates a value of I exceeding the critical value of + 1 and a "—" sign indicates that I is less than — 1. An outlier at the t'th point is indicated by a "clear" pattern of those signs in the corresponding horizontal and preceding diagonal of the plot.

3. OBJECTIVE DETECTION OF OUTLIERS

By taking expectations under the assumption that u_1, u_2 are two independent standard normal variates, and substituting for G gives the following moments for the Influence Function.

$$E(I) = 0.0$$

$$V(I) = (1 - \rho_k^2)^2$$

$$\sqrt{\beta_1}(I) = 0.0$$

$$\beta_2(I) = 9.0$$

where $\sqrt{\beta}$, and β_2 are the skewness and kurtosis of I.

When approximate critical points were evaluated for I, by fitting Johnson curves to the above moments, the value of + 1 was found to be appropriate in only a few cases and a critical value as high as + 2.71 would be needed for $\rho_k = 0.0$ at the 1% level of significance (see Watson, 1987). The method may also be criticised as it relies heavily on the subjective identification of a clear pattern of significant points in the horizontal and preceding diagonal of the Influence Function Matrix. The potential problems in outlier detection are illustrated by Fig 1 which refers to part of the Influence Function Matrix for eastbound traffic counts from 17:00 hours (T/B East 17:00). The series forms part of the DTp's data set and consists of 153 observations. A time series plot is shown below. Visual inspection of the Influence Function Matrix does not clearly indicate which points are suspect outliers, though some pattern exists at observation 17, 45, 52 and 72 (y_{72} being a missing value).

Figure 1: Series plot for T/B EAST 17:00.

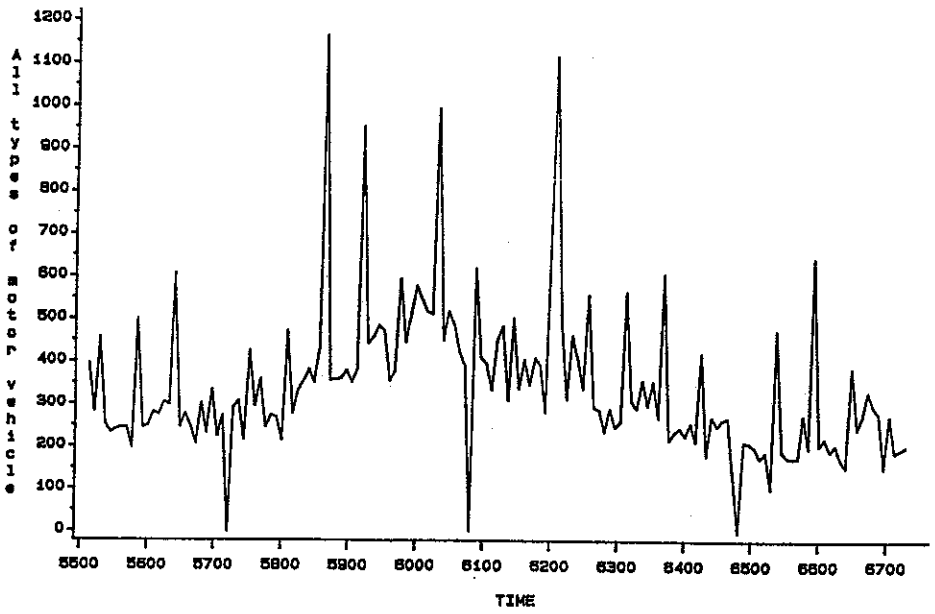


Figure 1

Influence Function Matrix and series plot for a section of T/B EAST 17:00

	LAG							
	1	2	3	4	5	6	7	8
OBS								
46								
47					-			
48				-				
49			-					
50		-						
51	-							
52					-		+	+
53								
54								
55								
56								
57								
58								
59			+	+	+	+		
60								
61					+			
62	+			+		+		
63			+					
64		+				-		-
65	+						-	+
66		+					+	
67					-			
68				-	+			
69			-					
70		-						
71	-							
72	-	-	-	-	-	-		
73								
74								

(a "+" or "—" indicates a significant I value, using $\pm$ 1 as critical values)

In an attempt to avoid some of the problems of a subjective method for outlier detection, we define a quantitative outlier detection statistic:

$$IS_t = \frac{1}{P} \{ \Sigma_{L_t} I^2 + \Sigma_{D_{t-1}} I^2 \} \qquad \text{....(2)}$$

where I is defined by (1), Σ_{L_t} indicates summation over the L_t

elements in the t'th row and $\sum_{D_{t-1}}$ indicates summation over all available elements in the (t-1)'th diagonal of the Influence Function Matrix and $P = L + D_{t-1}$

The statistic (2) treats positive and negative values in the matrix with equal weight and squaring the elements avoids the problem of large positive or negative values cancelling each other. For a given L, P increases for $t < L$, and for $t > L$, P is constant at 2L.

Equation (2) may be rewritten as

$$IS_t = \frac{M}{P} \sum_{P} (\chi_1^2)_1 \, (\chi_1^2)_2$$

where $(\chi_1^2)_j$, $j = 1,2$ are independent χ_1^2 variables and $M = G^2$. M is not strictly constant due to variation in ρ_k, but is assumed to be so here.

The first four moments of IS_t are needed to find approximate critical values for the Influence Statistic, although this is not trivial as the definition of IS_t involves the sum of squares of products of normal variates. Using the general result that the r'th moment about the origin for a χ^2 variate with γ df is given by

$$\mu_r = 2^r \prod_{i=0}^{r-1} [\, i + (\gamma / 2)]$$

straightforward but tedious algebra is needed to take expectations and deduce the moments about the origin of IS_t. These are

$$u_1 = M$$

$$u_2 = \frac{M^2}{P^2} (9P + 3P (P-1))$$

$$u_3 = \frac{15M^3}{P^2} (8 + 6P + P^2)$$

$$u_4 = \frac{105m^4}{P^3} (48 + 44p + 12P^2 + P^3)$$

Using a binomial expansion to give relationships with moments about the mean (Kendall and Stuart, 1977)

The mean of $IS_t = M$

The variance of $IS_t = \frac{6M^2}{P} + 2M^2$

The skewness of $IS_t = \frac{120 + 54P + 8P^2}{2^{3/2} P(3+P)^{3/2}}$ .... (3)

The kurtosis of $IS_t = \frac{5040 + 4140P + 963P^2 + 60P^3}{4P(3 + P)^2}$

Since M must be estimated and is not strictly constant, the mean and variance derived above will be approximate. In order to assess how good these approximations (3) are, and how much sampling variability can be expected in the higher order moments of IS_t, simulation was used to evaluate empirical moments.

4. SIMULATION RESULTS

Series were simulated 2000 times for various values of n and the empirical moments calculated for IS_t. The empirical moments were then averaged for $t > (L + 1)$ to give a representative value to compare with theoretical moments found from (3).

We begin with the special case $\rho_k = 0$ for all k. The comparison is shown in Table 1. It is clear that good agreement exists between theoretical and empirical moments, particularly for the mean and variance. More discrepancy is observed between theoretical and simulated skewness and kurtosis, but this would be expected even for standard normal variables (see for example Bowman & Shenton, 1975). As L increases this discrepancy is seen to decrease. Varying the sample size n does not affect simulated moments appreciably.

Simulations have shown the agreement between theoretical and simulated moments for IS_t to be good when $\rho_k = 0$. We now extend to AR and MA series where $\rho_k \neq 0$, by simulating ARMA series and calculating empirical moments for IS_t.

TABLE I

Theoretical (in brackets) and Averaged Empirical Moments of IS_t (2000 Simulations), $\rho_k = 0$

MAX LAG L	n	MEAN	VAR	SKEW	KURT
1	200	0.99 (1.00)	5.03 (5.00)	6.16 (5.81)	67.50 (87.72)
2	200	0.99 (1.00)	3.39 (3.50)	4.79 (4.43)	42.31 (51.55)
3	500	0.99 (1.00)	3.45 (3.50)	4.79 (4.43)	41.66 (51.55)
5	200	1.00 (1.00)	2.61 (2.60)	3.90 (3.48)	29.87 (29.59)
5	500	0.99 (1.00)	2.59 (2.60)	3.77 (3.48)	26.34 (29.59)

In Tables II, III and IV, the model parameters, theoretical ρ_k and a comparison of theoretical and empirical moments are shown for simulated AR(1), MA(1) and MA(2) processes. Note that from expression (3) the theoretical skewness and kurtosis depend on the number of elements summed, not ρ_k and so appear as constants in Tables II, III and IV.

From Table II the empirical skewness and kurtosis is particularly high in the AR process for $\phi_1 > 0.6$ although Tables III and IV show generally good agreement between theoretical and empirical values for these moments.

It can be seen from Table IV that all simulated and empirical moments are close for the MA(2) process for the range of Θ_1, Θ_2 chosen. For both the MA(1) and MA(2) the theoretical mean and variance have been somewhat over estimated.

Repeated simulations for AR, MA and ARMA series up to order 2 were made with varying values for the maximum lag (see Watson, 1987). Overall the agreement between theoretical and empirical moments was good. We found (by plotting the relative positions of the theoretical Θ_1, Θ_2) that those series with theoretical parameters closest to non-stationarity give the largest discrepancy between simulated and expected moments of IS_t.

TABLE II

Theoretical (in brackets) and Empirical Moments of IS_t for AR (1), 2000 simulations, n=200, maximum lag=10

ϕ_1	MEAN	VAR	SKEW	KURT
ϕ.1	0.99	2.21	3.29	19.93
	(0.98)	(2.21)	(3.15)	(22.26)
0.2	0.93	1.96	3.29	20.29
	(0.92)	(1.95)	(3.15)	(22.26)
0.3	0.83	1.59	3.46	23.45
	(0.83)	(1.58)	(3.15)	(22.26)
0.4	0.69	1.12	3.42	21.83
	(0.71)	(1.15)	(3.15)	(22.26)
0.5	0.55	0.73	3.70	26.17
	(0.56)	(0.73)	(3.15)	(22.26)
0.6	0.39	0.39	3.82	26.97
	(0.41)	(0.39)	(3.15)	(22.26)
0.7	0.24	0.15	4.15	32.14
	(0.26)	(0.16)	(3.15)	(22.26)
0.8	0.11	0.04	5.41	59.29
	(0.13)	(0.04)	(3.15)	(22.26)
0.9	0.028	0.004	6.55	78.04
	(0.036)	(0.003)	(3.15)	(22.26)

TABLE III

Theoretical (in brackets) and Empirical Moments of IS_t for MA(1), 2000 simulations, n=200 maximum lag=10

θ_1	ρ	MEAN	VAR	SKEW	KURT
0.1	0.099	0.99	2.21	3.31	20.47
		(0.98)	(2.21)	(3.15)	(22.26)
0.2	0.19	0.93	1.96	3.31	20.35
		(0.93)	(1.98)	(3.15)	(22.26)
0.3	0.28	0.84	1.16	3.35	21.23
		(0.85)	(1.68)	(3.15)	(22.26)
0.4	0.34	0.75	1.27	3.42	22.52
		(0.78)	(1.39)	(3.15)	(22.26)
0.5	0.4	0.63	0.92	3.39	21.79
		(0.71)	(1.15)	(3.15)	(22.26)
0.6	0.44	0.53	0.67	3.56	23.57
		(0.65)	(0.97)	(3.15)	(22.26)
0.7	0.47	0.45	0.47	3.52	23.07
		(0.61)	(0.85)	(3.15)	(22.26)
0.8	0.49	0.37	0.31	3.51	23.73
		(0.58)	(0.78)	(3.15)	(22.26)
0.9	0.49	0.29	0.21	3.57	23.83
		(0.57)	(0.74)	(3.15)	(22.26)
1.0	0.5	0.25	0.14	3.55	23.69
		(0.56)	(0.73)	(3.15)	(22.26)

Since the first four theoretical moments given by (3) are close to the empirical distribution of IS_t, these may be used to evaluate critical points for the statistic for use in outlier detection. The algorithm of Dodgson (1987) was used to calculate approximate critical points for IS_t by passing the first four theoretical moments to Algorithm AS99 (1976) which finds the type and parameters of an appropriate Johnson curve. Critical values are returned for the specified significance levels (α = 0.09 to α = 0.99 here) and according to P.

An example of a plot of ρ_k against critical values of IS_t for α = 0.9, α = 0.95 and α = 0.99 for the case when P = 10 (L=5) is shown in Fig 2. From Fig 2, assuming ρ_k = 0.2 for all k, an outlier would be indicated at the t'th point if IS_t > 6.6 using the 1% significance level. The estimation of ρ_k in practice is discussed in section 5.

As the simulated moments of IS_t were found to differ marginally from the theoretical moments Johnson curves were therefore also fitted to the empirical moments. This would enable a check to be made on whether the empirical critical values differed appreciably from the critical values found from the theoretical moments. Critical values were produced for ρ_k = 0.0 and for P = 2, 4, 6, 10. The discrepancy with theoretical critical points was small (no more

than 1 unit in the first decimal place of entries). As P increased the discrepancy also increased marginally but did not appear to be substantial.

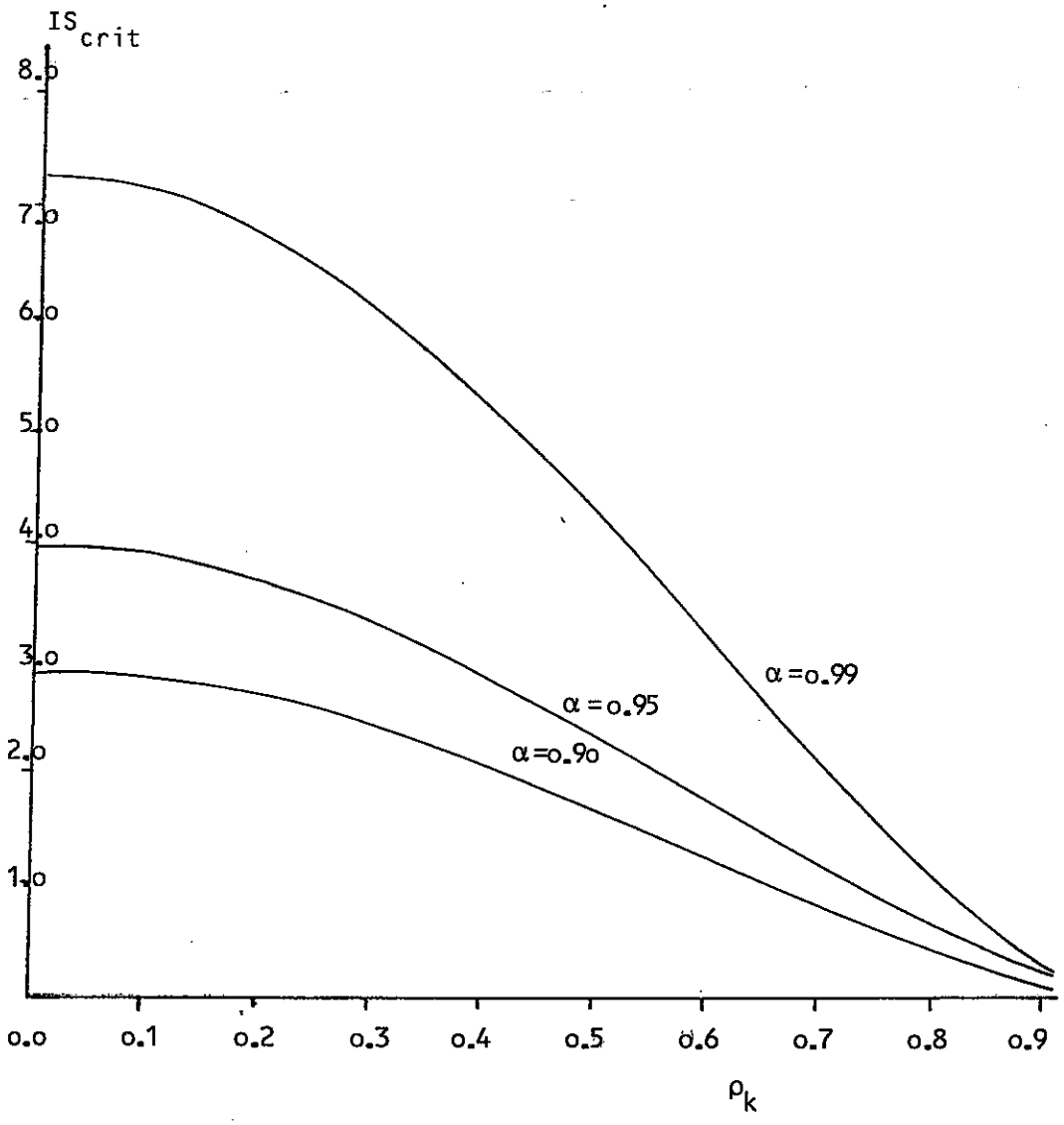
TABLE IV

Theoretical (in brackets) and Empirical moments of IS_t for MA(2) Process, 2000 simulations, n=200, maximum lag = 10

θ_1	θ_2	ρ_1	ρ_2	MEAN	VAR	SKEW	KURT
0.5	0.3	0.26	0.22	0.55	0.73	3.69	26.51
				(0.24)	(0.13)	(3.15)	(22.26)
0.3	0.2	0.21	0.18	0.78	1.41	3.47	23.04
				(0.91)	(1.91)	(3.15)	(22.26)
-0.7	-0.1	0.51	0.07	0.43	0.44	3.58	24.20
				(0.54)	(0.68)	(3.15)	(22.26)
1.0	-0.95	-0.67	0.33	-0.12	0.03	3.44	22.27
				(0.30)	(0.21)	(3.15)	(22.26)
-1.0	-0.95	0.67	0.33	0.12	0.03	3.38	21.66
				(0.30)	(0.21)	(3.15)	(22.26)
0.2	-0.7	-0.22	0.46	0.41	0.39	3.62	26.36
				(0.21)	(0.10)	(3.15)	(22.26)
0.5	-0.1	0.44	0.08	0.63	0.90	3.36	21.01
				(0.66)	(0.99)	(3.15)	(22.26)
0.7	-0.1	-0.40	0.29	0.32	0.24	3.49	23.52
				(0.70)	(1.14)	(3.15)	(22.26)
0.1	0.7	-0.02	-0.47	0.44	0.45	3.49	22.79
				(0.61)	(0.86)	(3.15)	(22.26)

FIGURE 2
Critical values for $IS_t, \rho=10$

FIGURE 2
Critical values for $IS_t, \rho=10$



5. OUTLIER DETECTION IN PRACTICE

Chernick et al (1982) use seven time series which refer to inventory differences of quantities of Nuclear materials to illustrate the use of the Influence Function Matrix. From Fig 1 it is clear that detecting spurious points using this method can be difficult in practice due to confused patterns. Here the Nuclear Power and DTp data sets are re-examined using the Influence Statistic defined by (2) together with critical points to indicate outlying / influential data.

Data from The Department of Transport, consists of traffic counts for a single trunk road with information for Eastbound and Westbound traffic. Four time series are available in each direction, beginning at 08:00, 12:00, 17:00 and 20:00 in each case, making 8 time series of traffic counts in total. Information is also available for other routes which will be analysed later in the project.

So far we have assumed ρ_k to be "constant". Therefore some global "sample summary" estimate of ρ_k is needed for practical application. Watson (1987) looked at three possible representative measures, these being:

i) $r = \max |r_k|$

ii) $$r^* = \frac{|\max r_k| + |\min r_k|}{2}$$

iii) $$r^+ = \frac{|\max r_k| + |\max_{-1} r_k| + |\max_{-2} r_k|}{3}$$

where $\max_{-1} r_k$ and $\max_{-2} r_k$ are the second and third largest sample autocorrelations respectively. Some measure is needed which is representative across the range of values expected to occur in the autocorrelation function, because of the assumption of constant r_k. As the theoretical basis for IS_t lies in the influence outliers have on the sample autocorrelation function, the possibility of using a robust estimator of ρ_k is rejected.

Simulations of AR(1) and MA(1) processes with varying n, ϕ and Θ values were used to assess the effect of the use of the three possible estimators on the detection rate of IS_t. Comparing the empirical significance levels with the nominal levels of 10%, 5% and 1%, the estimator r* gave the most satisfactory results from the three estimators considered. In Fig 3, the proportion of outliers found at the three nominal significance levels is shown for simulated AR(1) processes, using r* in calculations. The number of outliers detected using r* only rose consistently above that expected at the nominal levels for $\phi_1 > 0.6$ approximately. Clearly a "better" summary measure than r* may well exist, but for our purposes r* has proved adequate.

In practice we have found a maximum lag value of L = 5 gave satisfactory empirical outlier detection rates for most of the series examined. In Table V the points detected by IS_t (using r* in calculations) and those detected by Chernick's Influence Function Matrix are compared. Although fewer are now detected at the 1% level, some consistency in the points picked out is apparent.

In table VI outlier detection results are shown for East and West bound traffic counts for part of the DTp's data. A comparison is made between points detected by the influence statistic and those given by the straightforward measure of residuals outside 3 standard deviation limits. The residuals were calculated from a Box-Jenkins Multiplicative Seasonal ARIMA with a season of period 7.

For each series observations 27, 72 and 122 were missing and coded as zero.

In column 2 of table VI the observations highlighted by both methods are shown, whilst columns 3 and 4 indicate points detected by the high residual and influence statistic exclusively. As each method utilizes different criteria in the selection of "significant" points it is impossible to categorize results as correct or incorrect. However it is clear from table VI that for each series a number of points were picked out by both methods, including missing observations. Several other data points were also highlighted by each criteria. Note that for the DTp data a maximum lag of L=8 was used together with the r* estimate of ρ_t. The increased value of L was felt to be appropriate because of the pattern of spikes shown in the series ACF, which would not have been accommodated using L=5.

6. A SIMPLE ADJUSTMENT PROCEDURE FOR OUTLIERS

Further examination of the Influence Function considered suggests possible replacement values for any outlying or influential data found. For data without highly influential or outlying points, $I(p_k, Z_tZ_{t+K})$ will be a random variable with zero expectation. If the sample estimate of (1) takes its expected value of zero, then approximately

$$Z_t Z_{t+k} - {}^1/_2 \ (Z_t^2 + Z_{t+k}^2)\, r_k = 0 \qquad \ldots. (4)$$

Consider Y_t as the quantity to be estimated (i.e. the missing data point or outlier), then from knowledge of Z_{t+k} and r_k solving expression (4) suggests a replacement value. Note that the calculation of r_k will initially have been made with Y_t included. Manipulation shows the solution of (4) to be

$$Z_t = \frac{Z_{t+k}}{r_k} \ (1 - \sqrt{1 - r_k^2})$$

If the approximation given by (4) holds, then the replacement value $(Z_t \times S) + \bar{Y}$ may be practically applied to replace identified outliers. Simulation studies with series of normal variates showed that the replacement procedure works very well (see Watson, 1987).

We simulated a series of iid N(0,1) variables and created artificial outliers by gradually increasing the magnitude of one of the data points. A plot of the raw series, together with the first few lags of the autocorrelation function is shown in Fig 4. By visual inspection, the series has no distinctive features other than a large negative value at Y_{12}.

FIGURE 3

Empirical significance levels using r*, max lag=5

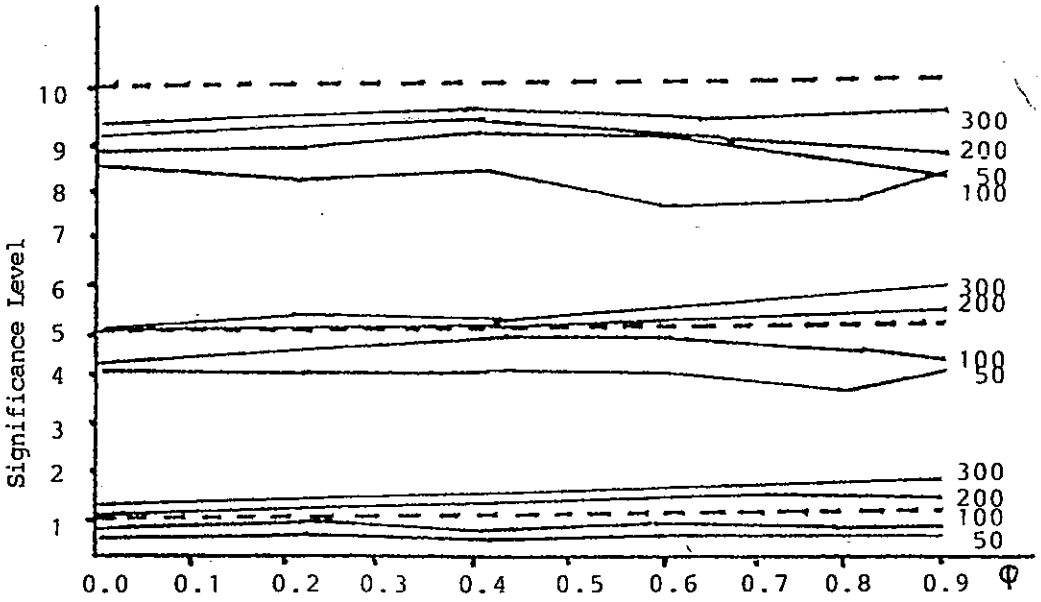


TABLE V
Comparison between points found significant by IS_t and the subjective Influence Matrix methods in the Nuclear power data

		POINTS SIGNIFICANT			
		IS_T			INFLUENCE MATRIX
SERIES	r*	10%	5%	1%	(ALL LEVELS)
LASLP	0.0572	Y_{10}, Y_{24}	-	-	Y_6, Y_{10}, Y_{24}
LASLU	0.1582	Y_{24}	Y_{28}	Y_{26}	Y_{26}, Y_{28}
RICH	0.0147	Y_{13}, Y_{21}	Y_{17}	-	Y_{12}, Y_{13}, Y_{16}, Y_{17}
ORNLP	0.0545	-	Y_6, Y_7, Y_{20}	-	Y_6, Y_7, Y_{20}, Y_{21}, Y_{22}
ORNLP2	0.0680	Y_{17}	Y_{24}	-	Y_{17}, Y_{24}
U233	0.0044	-	-	-	Y_8, Y_{26}
SAVV	0.1242	-	-	-	Y_{12}

TABLE VI

Influence statistic and extreme residual results for DTp data.

Route T/B 912 East

Data	Both Methods	High Residual Only	Influence Statistic Only
08:00	27, 65, 72	88, 122, 128	24, 73, 79, 80
12:00	27, 57, 72, 88, 89, 90, 122		1, 2, 66, 76, 79, 86, 92, 93
17:00	45, 52, 66, 72	27	17, 88, 108, 122, 128, 129, 136
20:00	52, 66, 72, 88	27, 122	10, 17, 31, 44, 45, 64, 65, 73, 87, 94, 95, 101, 108

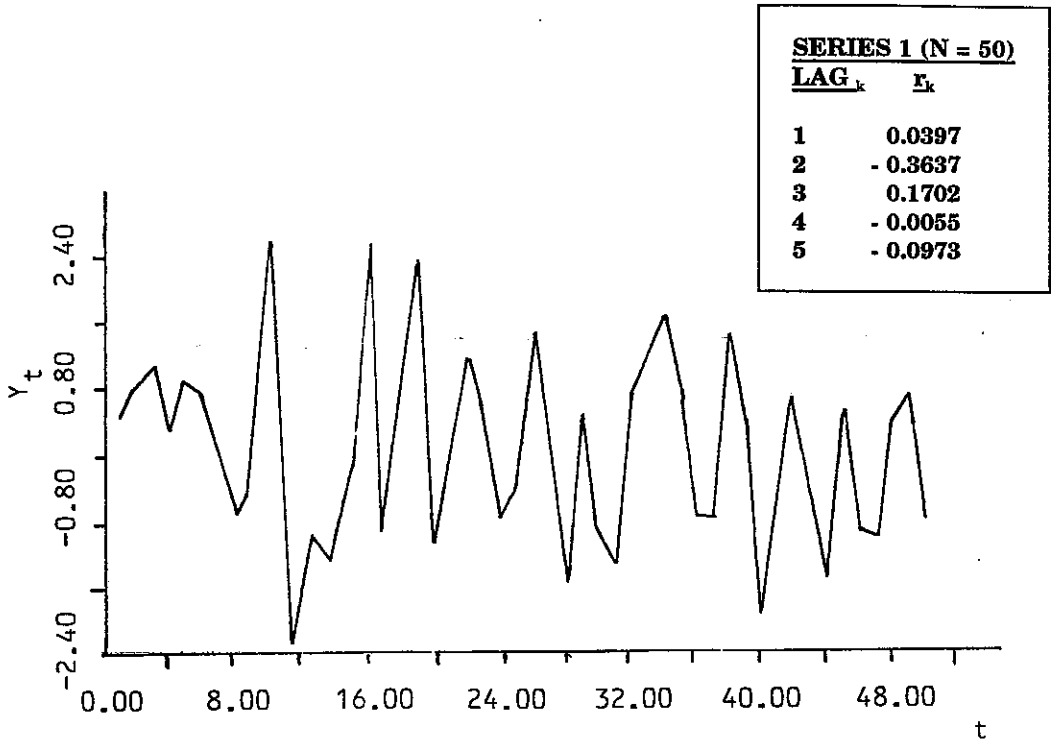
Route T/B 912 West

Data	Both Methods	High Residual Only	Influence Statistic Only
08:00	27, 51, 65, 72	93, 96	44, 58, 62, 75, 86
12:00	72	18, 24, 27, 89, 122	2, 23, 30, 37, 44, 45, 51, 58, 65, 66, 79, 86, 87, 88, 93
17:00	27, 36, 40, 72, 122	13	6, 7, 30, 35, 79, 85, 86
20:00	85	2, 22, 27, 86, 122	64, 71, 72, 77

(for each series 27, 72 and 122 are missing values)

FIGURE 4

Time series plot of series 1



Using the Influence Statistic (2), we identified the points Y_{12} and Y_{16} as potential outliers at the 5% level of significance. We concentrated on Y_{12}. Six increments of size 0.2 were consecutively subtracted from Y_{12} to investigate the corresponding effect on IS_{12} and we also used the outlier adjustment in expression (5). From Table VI, the values of IS_t before and after adjustment may be compared.

It is clear that as the size of the outlier is increased by subtracting units of 0.2, the Influence Statistic also becomes increasingly larger. After the artificial outlier was replaced with the adjustment given by (5), the outlier detection statistic was no longer significant. The adjustment procedure (5) does of course depend upon the sample acf at lag k. We decided to use only <u>one</u> value, namely the lag for which r_k (k = 1, 2, ...) was largest. Another idea would be to calculate adjusted data points using all of $r_1, r_2, \ldots r_{k\,(max)}$ in (5). The "final" adjusted value could be obtained by smoothing these k(max) adjustments. An alternative solution may be to take an exponentially weighted sum of adjustments up to a predetermined maximum lag.

TABLE VII

Value of IS_t for Y_{12} before and after adjustment

	Before	After
SERIES	IS_{12}	IS'_{12}
Y_{12}	5.6	0.21
Y_{12} - 0.2	6.36	0.13
Y_{12} - 0.4	7.13	0.19
Y_{12} - 0.6	7.89	0.19
Y_{12} - 0.8	8.64	0.19
Y_{12} - 1.0	9.38	0.18
Y_{12} - 1.2	10.09	0.18

Similar ideas have been proposed by Abraham and Chuang (1989). Their approach to model building is to "clean" the series, replacing outliers by an adjustment dependent upon parameters of a fitted autoregressive model. The influence statistic approach, of course, only depends upon the acf of the observed series and not on some fitted model.

7. A PRACTICAL APPLICATION OF THE ADJUSTMENT PROCEDURE

The DTp's traffic count data provides an ideal practical application for the adjustment procedure, containing both missing data and suspect outliers. The procedure was successfully applied to each of the 8 series and sample results for two of the series are given in Table VIII. Series plots for the data are shown by figs 1 and 5, and from these it is clear that both time series have several potential outliers and missing values (coded as zero). The original value of the outlying observation is shown in column 5 of table VIII with missing observations allocated a zero value. The observation number and corresponding IS_t value are shown in columns 2 and 3 respectively, whilst the suggested replacement figure is given in column 4. Several applications of the procedure were made, as the replacement of an initial set of outlying points uncovered a further group for treatment. The r* value is indicated in column 1, together with the critical value (CV) for IS_t and it is clear from table VIII that the r* value rises as each group of points is replaced. This is as expected and is indicative of the depressing quality of maverick data on the autocorrelation function. Some variation is found in the suggested replacement values although they do not fluctuate substantially.

Figure 5

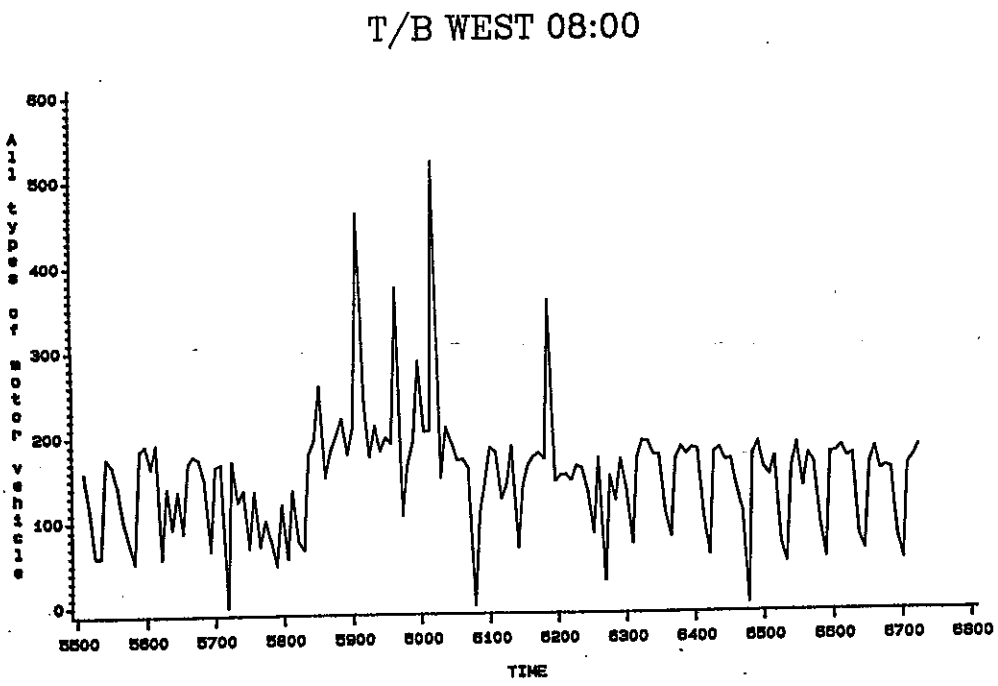

T/B WEST 08:00
All types of motor vehicle
0
100
200
300
400
500
600
5500
5600
5700
5800
5900
6000
6100
6200
6300
6400
6500
6600
6700
6800
TIME

Table VIII
Application of adjustment procedure to DTp time series traffic counts

T/B East @ 17:00

Statistic	Obs	Influence Statistic	New Value	Original Value
CV = 5.6 r* = 0.345	45 66 72 88	11.7 7.3 10.1 13.2	530 430 338 333	1164 995 0 1118
CV = 4.5 r* = 0.439	122 136	4.9 12.3	387 354	0 643
CV = 4.4 r* = 0.478	52	5.2	444	952
CV = 4.4 r* = 0.477	17 108	6.1 5.8	333 370	607 607
CV = 4.0 r* = 0.508	129	4.9	339	476
CV = 3.9 r* = 0.52	(STOP)			

T/B West @ 8:00

Statistic	Obs	Influence Statistic	New Value	Original Value
CV = 6.1 r* = 0.262	51 58 62 65 72	8.2 14.5 6.6 41.1 27.1	151 147 154 120 161	462 374 287 522 0
CV = 6.5 r* = 0.214	27 44 86	6.5 7.4 9.7	150 148 144	0 259 357
CV = 6.4 r* = 0.226	72	7.9	155	0
CV = 6.3 r* = 0.249	(STOP)			

8. CONCLUSIONS

The Influence Function Matrix proposed by Chernick et al (1982) has been developed to form an objective statistic, IS_t for outlier detection in time series. By assuming p_k is approximately constant, the approximate first four moments for the statistic have been obtained. Simulation studies have shown the first four theoretical moments to be in good agreement with empirical moments for the statistic with standard normal and ARMA series. By fitting Johnson curves to the theoretical moments, critical points have been produced to determine which data values are influential or outlying in a time series.

Practical applications of the statistic have been given using Chernick et al's (1982) nuclear power data and time series traffic counts from DTp. The statistic agreed with the points highlighted by Chernick et al (1982) as possible outliers and was successful in detecting extreme traffic counts. An adjustment procedure for outliers has been described derived from the statistic and shown to have practical application with both outliers and missing observations in traffic counts.

REFERENCES

ABRAHAM, B and CHUANG A (1989). Outlier Detection and Time Series Modelling. Technometrics, 31, 241-248.

BOWMAN, K.O. & SHENTON, L.R. (1975). Omnibus test contours for departures from normality based on $ß_1$ and $ß_2$. Biometrika. Vol 62, No 2, 211-251.

BOX, G.E.P. & JENKINS, G.M. (1976). Time series analysis, forecasting and control. Holden-Day.

BRUCE, A.G. and MARTIN, R.D. (1989). "Leave-K-Out Diagnostics for Time Series". J.R. Statist Soc. Ser. B, 51, 363-424.

CHERNICK, M.R., DOWNING, D.J. & PIKE, D.H. (1982). Detecting outliers in time series data. Journal of American Statistical Association, Vol 77, No 380, 743-747.

DODGSON, J.H. (1983). The effect of a preliminary test of normality using $\sqrt{b}$, on students t distribution. Phd thesis, Loughborough University of Technology.

FOX, A.J. (1972). Outliers in time series. Journal of the Royal Statistical Society (B), Vol 34, 340-363.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. Journal of American Statistical Association, Vol 69, 383-93.

HAU, M.C. & TONG, H (1984). Outlier detection in autoregressive time series modelling. Technical Report No 15. Dept Statistics, The Chinese University of Hong Kong.

HILL, I.D., HILL R and HOLDER, R.L. (1976) ALGORITHM AS 99. Fitting Johnson Curves by moments. Applied Statistics, Vol 25, No 2, 180-189.

KENDALL, M.G. & STUART, A (1977). The advanced theory of statistics, Vol 1, 58. Charles Griffen & Co. Ltd.

TIAO, G.C. (1985). Autoregressive moving average models, intervention problems and outlier detection in time series. Handbook of Statistics, Vol 5, 85-118. Elsevier Science Publishers B.V.

TSAY, R.S. (1986). Time series model specification in the presence of outliers. Journal of American Statistical Association, Vol 81, No 393.

TSAY, R.S. (1988). Outliers, level shifts and Variance changes in time series. Journal of Forecasting, Vol 7, 1-20.

WALKER, A. (1960). Some consequences of superimposed error in time series analysis. Biometrika Vol 47, 33-43.

WATSON, S.M. (1987). Non-normality in Time Series analysis. Phd. Thesis, Trent Polytechnic.

smw\outlier.jun