OUTLIER DETECTION AND MISSING VALUE ESTIMATION IN TIME SERIES TRAFFIC COUNT DATA

Susan WATSON
Research Fellow
ITS
University of Leeds-England

Stephen CLARK
Research Assistant
ITS
University of Leeds-England

Edwin REDFERN
Lecturer
Statistics
University of Leeds-England

Miles TIGHT
Lecturer
ITS
University of Leeds-England

INTRODUCTION

A serious problem in analyzing traffic count data is what to do when missing or extreme values occur. Missing data occurs for a variety of reasons, for example the breakdown of automatic counters, and current methods for dealing with this problem can be crude. Existing techniques for dealing with missing values and outliers in transport range from fitting best estimates by eye to a few computerised algorithms which are known to exist. As this paper shows, however, little work has been undertaken to assess the relative merits of alternative methods. The necessity for such statistical techniques to be developed and applied in the transport field is highlighted in the paper, in particular the need to be able to determine what are the true outliers in a time series.

1. METHODS FOR DETECTING OUTLIERS AND MISSING VALUES

1.1. By eye

A simple method still used surprisingly often when cleaning traffic count data is to identify any outliers by eye and estimate a replacement value using an estimate from the values the same time of the same day of the weeks either side of the problem value. Clearly if any of these values is a problem then they should not be included in the estimation. It is prone to subjectivity on the part of the person doing the work.

1.2. Averaging techniques

An automatic procedure used by the British Department of Transport (DTp) is to validate each observed data value against corresponding data from the same site 1, 2, or more weeks previously. A mean and standard deviation for each site*day of week*period of day*combined vehicle category is updated using an exponentially weighted moving average (EWMA), where the lagged values, x(t-k), are the previous estimates from the corresponding time of day and week for the same site and vehicle category:

$$\mu_t = (1-\theta)\mu_{t-1} + \theta\, x_t$$
$$\sigma_t^2 = (1-\theta)\sigma_{t-1}^2 + \theta(x_t-\mu_t)^2$$

Typical values used for the smoothing parameter are about $\theta = 0.3$. Any observation more than 4 standard deviations away from the current mean is rejected and replaced. Missing or rejected data, x(t), is estimated by the exponentially weighted average:

$$x(t) = \theta\, x(t-1) + \theta(1-\theta)\, x(t-2) + \theta(1-\theta)^2\, x(t-3) + \ldots$$

1.3. The Box-Jenkins approach

Box-Jenkins is the most widely applied approach and has been used with some success in modelling transport data. The principal method has been that based on the auto-regressive integrated moving average (ARIMA) model using the Box-Jenkins (1976) model building methodology. Routines for fitting and validating the models are readily available in the major statistical packages such as SAS, BMDP and SPSS.

McLeod et al (1980) report on the analysis of rail and air passenger flows between London and Glasgow using the Box-Jenkins approach. The level of aggregation in the data was found to have a marked effect on the modelling results and further model development involved the inclusion of causal variables such as journey time. Known outliers in the data were modelled using intervention analysis.

A combination of linear regression and Box-Jenkins techniques were applied by Gaudry (1975) in order to model demand for public transport. The initial linear regression model included explanatory variables such as price and service levels and the error term was then specified in terms of an ARIMA model. Although an excellent fit to the data was achieved, no mention is made of the treatment of missing or outlying observations.

Successful applications of the Box-Jenkins method are also described in work by Ahmed and Cook (1979) and Nihan and Homesland (1980), although both pieces of research use United States traffic volume data. Models were found to perform well in short term forecasting and in the case of the work by Nihan and Homesland (1980), were able to deal with several problems such as missing information and interventions.

ARIMA models have also been used to analyze accident statistics (see for example a recent review by Hakim et al, (1991)). These models were compared to the structural model approach, based on slopes and levels, by Harvey and Durbin (1986) in a study of the safety effects of changes in the seat belt law.

Outlier detection methods based on the ARIMA model utilise the whole of the data to judge whether a particular point is influential. Early work by Fox (1972), Box and Tiao (1983) and Pena (1984) to develop test statistics for outliers in general time series all use the Box-Jenkins ARIMA modelling structure. Each construct likelihood ratio statistics, and in the case of Fox (1972), the estimated error for an observation is compared with the estimated standard error of that discrepancy. Box and Tiao (1983)

and Pena (1984) concentrate on influential observations, developing test statistics which reflect the influence of an observation on the model parameters.

Box and Tiao (1975) describe a technique called intervention analysis in which the ARIMA model is extended to allow for known interventions such as the effect of petrol legislation on levels of atmospheric oxidant. They achieve this by including exogenous time series made of zeros and ones which allow for the absence or presence of an event. Such a technique is similar to including dummy variables in ordinary regression. This was the approach used by McLeod et al (1980) to model known outliers though no detection method was considered.

Missing values can be estimated using an intervention series consisting of a pulse at the time of the missing values. An extension of the ARIMA (p,d,q)(P,D,Q) model is used, given by:

$$\Phi(B)\phi(B)Y_t = \Theta(B)\theta(B)\varepsilon_t + \sum_{k\in M} \omega_{k0} I_t^{(k)} \quad \text{where} \quad Y_t = (1-B)^d(1-B^s)^D X_t = \Delta^d \Delta_s^D X_t$$

is the appropriately differenced time series.

Here $\phi(B)$ and $\theta(B)$ are polynomials in B of orders p and q, $\Phi(B)$ and $\Theta(B)$ are polynomials in B^s of orders P and Q where s is the period of the season and B is the backward shift operator. ε_t is a white noise sequence and $I_t^{(k)}$ is a pulse series with a 1 at time k and 0 elsewhere. M is the set of time points corresponding to the missing values and outliers.

The initial identification of the model can be done using either the autocorrelation function, partial autocorrelation function or inverse autocorrelation function estimated from the longest part of the series with no missing values, or the estimated correlation functions using the whole data calculated from Marshall's (1980) method. Large residuals are indicative of additive outliers in the data set and such values beyond 3 standard deviations are used to determine possible interventions for inclusion in the model. Once identified, replacement values can be estimated by including a pulse intervention at the appropriate point and re-fitting the time series model. Proceeding iteratively we can identify groups of points that may be influential.

Tsay (1988) described an enhancement of this process for identifying 'faults' in the model that occur as outliers, jumps and variance changes which are based on scaled values of the residuals resulting from the current fitted model. The model can be written in the form:

$$\phi(B)\Phi(B)\Delta^d\Delta_7^D X_t = \theta(B)\Theta(B)\varepsilon_t + \sum_{k\in M} f_k(t) \quad \text{where} \quad f_k(t) = \omega_{k0} \frac{\omega(B)}{\delta(B)} I_t^{(k)}$$

Two types of outlier considered were the innovative and additive. Innovative outliers

$\left(\frac{\omega(B)}{\delta(B)} = \frac{\theta(B)\Theta(B)}{\phi(B)\Phi(B)}\right)$ are those that affect the series in the future, while additive outliers

$\left(\frac{\omega(B)}{\delta(B)} = 1\right)$ are those that only have an impact at the specific time point.

The simpler model based on the large residuals is equivalent to restricting attention to modelling additive outliers. Structural changes such as level shifts

$\left(\frac{\omega(B)}{\delta(B)} = \frac{1}{(1-B)}\right)$ and seasonal shifts

$\left(\frac{\omega(B)}{\delta(B)} = \frac{(1-B)}{(1-B^s)}\right)$ can be handled.

A variance shift can be included by replacing the pulse $I_t^{(k)}$ by $\zeta_t^{(k)}$ where $\zeta_t^{(k)} = \varepsilon_t$ for $t \geq k$ and zero for $t<k$. Using these models Tsay (1988) described a series of tests that can be performed to evaluate the significance and type of any 'fault' in the data.

Algorithms based on the Kalman filter to obtain maximum likelihood estimates of the ARIMA model parameters in the presence of missing values have been developed by Jones (1980), Harvey and Pierse (1984) and Kohn and Ansley (1986). These allow the estimation procedure to skip missing or deliberately omitted values when obtaining the estimated values, and have permitted an alternative approach to outlier detection, namely the leave-k-out diagnostics introduced by Bruce and Martin (1989). They considered criteria based on the change in the estimated parameter values and the estimated innovation variance which measure the effect of the omitted points. Such methods allow the evaluation of the influence of a point or group of points on a model and its fit. The calculation process is lengthy even for short time series, hence such techniques are not yet widely available.

Overall, where Box-Jenkins models have been applied with transport time series, a good deal of success has been achieved both in modelling existing series and in short term forecasting. Some researchers (such as Nihan and Homesland (1980)) extended the basic univariate model to include behavioural variables, which not only improved model performance, but also assisted the intuitive interpretation.

1.4. The Influence Function Method

A different approach to outlier detection is based on the influence of an observation on the autocorrelation estimate. The sample autocorrelation function r_k,

k=1, ... L is used extensively in time series modelling, particularly in the widely used Box-Jenkins methodology. It can be argued that observations which have an undue influence on r_k should be identified as these may affect the success of the modelling process. Using the influence function matrix of Chernick et al (1982), Watson (1987) proposes a quantitative outlier detection statistic, IS_t, based on the influence of the t'th observation on r_k.

Firstly a matrix of influence values ($I_{t,k}$) is computed:

$$I_{t,k}=Y_tY_{t+k} - r_k \frac{(Y_t^2-Y_{t+k}^2)}{2}$$

the influence statistic for the t'th observation is then defined by

$$IS_t = \frac{1}{p} \left\{\sum_{L_t} I^2 + \sum_{D_{t-1}} I^2\right\}$$

where $\Sigma_{L_t}I^2$ indicates summation of I_{ij} over the L elements in the t'th row and $\Sigma_{D_{t-1}}I^2$ indicates summation of I_{ij} over all available elements in the (t-1)'th diagonal of the influence function matrix ($p = L_t + D_{t-1}$). In the derivation of theoretical moments and critical values for IS_t, ρ_k is assumed to be constant. In practice a global "summary sample estimate" is needed. The algorithm allows several different estimates to be used, although the following measure, r*, was found through empirical results to be appropriate:

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

Where $|\max r_k|$ and $|\min r_k|$ are the absolute values of the maximum and minimum sample autocorrelation values respectively. Further manipulation of IS_t suggests a possible replacement or estimation procedure for use with outlying or missing observations. The replacement value is given by:

$$(z_t^1\ s)+y \quad \textit{where} \quad z_t^1 = \frac{z_{t+k}}{r_k}\ (1 - \sqrt{1-r_k^2})$$

y and s being the mean and standard deviation of the original series, z_t representing the transformed data.

2. APPLICATION OF TECHNIQUES TO TRANSPORT TIME SERIES

To obtain an indication of the usefulness of these methods, four of the above are applied to two traffic flow time series from an automatic counter maintained by the DTp. The counter is located on a main trunk road which provides access to the

Pembrokeshire Coast. The two series considered here are hourly traffic flows at the same time each day for a period of five months in mid 1990. Figure 1 shows the traffic flow at 20:00 hours travelling in a westerly direction and Figure 2 shows the same hours flow in the opposite direction. Both series have obvious missing values at observations 27, 72 and 122. There is evidence of a seven day seasonal pattern in both series. For the westerly flow it is strong, the peak coming on a Friday, while for the easterly flow it is much weaker, the peak being on a Sunday.

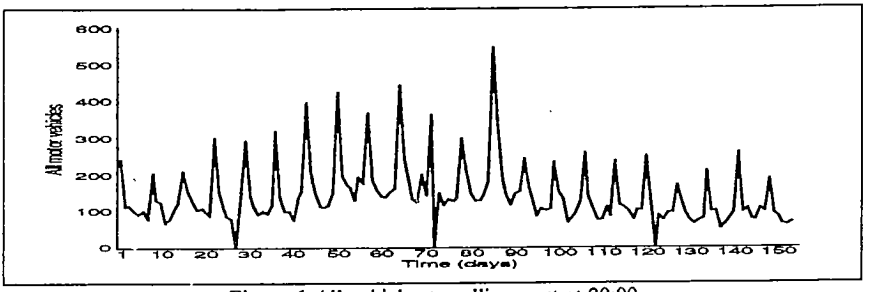

Figure 1 All vehicles travelling west at 20:00

Source: DTp

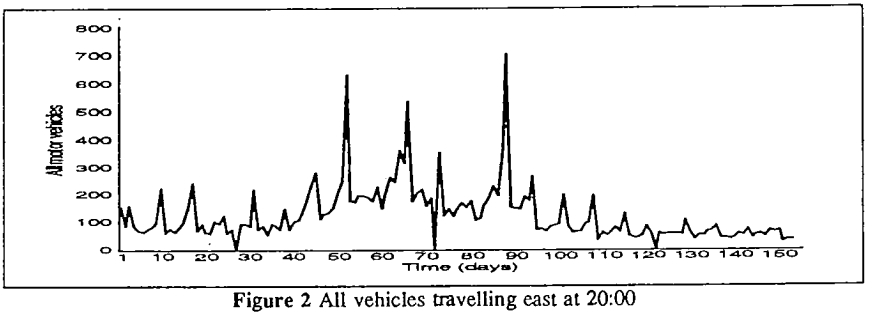

Figure 2 All vehicles travelling east at 20:00

Source: DTp

Tables 1 and 2 show the results of analyses of the time series using the four different methods described above. The Box-Jenkins approach suggests an initial ARIMA $(1,0,0)\ (0,1,1)_7$ model for the westerly series with three pulse intervention variables for the missing values at time 27, 72 and 122. The estimated parameters are $\phi = 0.57$ and $\Theta = 0.62$. This model has a large standardised residual (+ 4.8) at observation 85. The ARIMA model is refitted with four intervention variables, the three associated with the missing values and a fourth at time 85. This process of fitting a model and examining the residuals is carried out a further two times to give an ARIMA model, outliers being successively identified at times 22 and 86. The parameter

estimates of the intervention variables associated with the missing values provide replacement values. The three outlying observations are 22, 85 and 86. The final parameter values were $\phi = 0.43$ and $\Theta = 0.36$ which gives an indication of the influence of these observations. The day corresponding to observation 88 was an August bank holiday Monday. Large outliers are detected at observations 85 and 86 (Friday and Saturday) in the westerly series, caused by traffic heading towards the coast at the start of the holiday weekend. In the easterly series there is a large outlier at observation 88, caused by the returning traffic.

Table 1 Westerly traffic flow summary

Missing value replacements	Eye	Average value	Influence	ARIMA
27	100	69	119	108
72	170	192	174	208
122	110	124	121	107
Outliers detected	**Eye**	**Average value**	**Influence**	**ARIMA**
Observation	85	44	64	22
		53	71	85
		55	77	86
		86	85	

Table 2 Easterly traffic flow summary

Missing value replacement	Eye	Average value	Influence	ARIMA
27	110	41	115	66
72	150	78	184	217
122	140	205	106	128
Outliers detected	**Eye**	**Average value**	**Influence**	**ARIMA**
Observations	52	41 62	10 66	52
	66	44 65	17 73	66
	88	48 66	31 87	88
		51 69	44 88	(64)
		52 76	45 94	(66)
		53 83	52 95	(73)
		55 88	64 101	(80)
		58 90	65 108	(87)

For the easterly series the parameters changed from $\phi = 0.56$ and $\Theta = 0.80$ to $\phi = 0.77$ and $\Theta = 0.61$ if the outliers at 52, 66 and 88 (detected by the large residuals) are included. There was some evidence to support a non-seasonal differencing, but the

resulting model was a much poorer fit. For both series the average and influence techniques have produced a large number of outliers, which may indicate a lack of homogeneity in the structure of the series.

If the two series are modelled using the methods developed by Tsay (1988), then in addition to the outliers described above there is a gradual change to a new level in both series at about time 32 and a drop to a new level at time 95, the changes being larger in the case of the easterly flow. In the case of the westerly flow no additional outliers are found. The rise over the summer months is an average of 10, with a drop of about 40 at the end of the summer. For the easterly flow there are additional significant outliers at 64, 66, 73, 80 and 87. The observations at 52, 66, 73, and 87 are the summer Sundays, 64 is a Friday. The Sunday of 59 is not significantly higher than expected while the 80 value is depressed, not increased (did it rain?). There is a jump in level over the summer months rising gradually through July to an increase in the base flow of 23. At the end of the summer there is a drop of 100, the base flow now being below that before the holiday period. There is also a change in the seasonal pattern at the end of September (about time 127) with the Sunday peak, the main feature of the seasonal pattern almost disappearing.

In general this approach gives an interesting series of insights into the likely outliers and missing values in the datasets. It is not possible, however, to determine from this information alone which of the techniques give the best indication of the actual outliers and missing values in the series. The next step, therefore, was to undertake a series of simulation studies.

3. SIMULATION STUDIES

To evaluate the effectiveness of the various methods in detecting additive outliers a time series of vehicle flows were taken in which none of the methods detected any departures from the model. Certain observations were then replaced with known outliers of varying significance in a variety of patterns, including those typical of the ones observed in studying other traffic count series. Hence, the simulations include one or more outliers at randomly selected points in the series, typical bank-holiday pattems such as high flow on the Friday, and finally a depressed flow throughout a week which may be typical of a local event such as road works.

The methods considered in this study are applied to series containing the above patterns. Series 1-10 have a single outlier, 11-15 have a bank holiday pattern, 16-19 have a random pattern of 3 outliers while 20-24 have a depressed or raised series of values over a block of working days. Table 3 shows the preliminary results of our study but already a pattern is beginning to emerge. The standard deviation (σ) of the underlying error in the series was 50. The table indicates where outliers were detected and the difference between the estimated and the original value. It is clear from this that the ARIMA model is the best, and only fails to detect outliers in the case of a group of observations departing from the general pattern. The EWMA estimate is virtually always an underestimate of the true value.

Table 3 Comparison of the error in replacement estimates of detected outliers

Series	Single outlier	EWMA	Influence	ARIMA
1 - 2	4σ	-10 *	-212 *	21 5
3 - 6	6σ	* * -113 *	-67 -102 -48 *	-23 1 9 26
7 - 10.	>6σ	-139 -43 -142 *	-123 8 * *	1 34 -56 -23
Series	Bank Holiday	EWMA	Influence	ARIMA
11	>6σ	* *	* *	60 -22
12	>6σ	* -139	* *	52 1
13	>6σ	-94 -113	-3 *	13 -8
14	4σ	* *	* *	51 1
15	4σ	* *	* -215	39 -18
Series	Random	EWMA	Influence	ARIMA
16	8σ	* * -29	* * 129	128 -20 -24
17	6σ	* 5 -121	* * *	-72 53 -2
18	4σ	-54 * *	* -100 *	-6 -20 -54
19	4σ	* * *	-56 * *	-50 -25 27
Series	Blocks	EWMA	Influence	ARIMA
20	4σ	* * * * *	* * * * *	-158 * * * *
21	4σ	* * * * *	* * * * *	* * * * *
22	8σ	* * * * *	* * * * *	5 -4 17 -22 9
23	12σ	-208 * * -140 -164	* * * * *	-58 -100 -74 -7 -29
24	6σ	* * * *	-216 * * -140	-56 -98 -70 -10

Note : * denotes an undetected outlier

Finally it is instructive to look at the results obtained using the ARIMA models in more detail. The base series was adequately modelled using the model of order $(1,0,0)(0,1,1)_7$ described by the equation:

$$x_t - \phi\, x_{t-1} = \varepsilon_t - \Theta\, \varepsilon_{t-7} \quad \textit{where} \quad x_t = y_t - y_{t-7}$$

The estimated values of the parameters were $\phi = 0.56$ and $\Theta = 0.72$. Table 4 summarizes the results that were obtained by fitting models which made no allowance for outliers and then iteratively identifying additive outliers until no more were detected.

Table 4 Comparison of the outliers detected using the ARIMA procedure

	Initial model		Final model		Outliers found	Outliers missed
	ϕ	Θ	ϕ	Θ		
1	0.53	0.69	0.56	0.72	44	
2	0.52	0.79	0.55	0.72	45	
3	0.42	0.80	0.56	0.73	24	
4	0.49	0.79	0.56	0.73	59	
5	0.47	0.74	0.56	0.73	80	
6	0.39	0.82	0.57	0.73	38	
7	0.12	0.99	0.56	0.72	59	
8	0.22	0.99	0.56	0.72	66	
9	0.23	0.99	0.57	0.73	73	
10	0.23	0.99	0.56	0.73	24	
11	0.31	0.91	0.57	0.73	21 24	
12	0.13	0.98	0.55	0.73	56 59	
13	0.20	0.89	0.56	0.73	77 80	
14	0.25	0.89	0.55	0.73	56 59	
15	0.38	0.82	0.57	0.73	42 45	
16	0.31	0.84	0.60	0.65	27 41 71	
17	0.42	0.95	0.57	0.72	34 56 82	
18	0.38	0.86	0.56	0.74	30 45 87	
19	0.38	0.73	0.56	0.73	31 47 54	
20	0.69	0.66	0.70	0.67	45	46 47 48 49
21	0.63	0.85	-	-	-	80 81 82 83 84
22	0.71	0.96	0.58	0.76	59 60 61 62 63	
23	0.76	0.98	0.57	0.70	45 46 47 48 49	
24	0.53	0.89	0.57	0.70	45 46 47 49	

The Box-Jenkins modelling process appears to be adequate for identifying rogue outliers in the various series but does less well when there is a block. The effect of the outliers in the data is to reduce the value of the fitted auto-regressive parameter and increase the value of the seasonal moving average parameter.

4. SUMMARY AND CONCLUSIONS

Experience with many series of traffic counts and other transport time series suggests that difficulties discussed in the above examples are not untypical. The EWMA method tends to be too inflexible picking up either a larger number of outliers

when there is greater probability of a structural change, and missing other extreme values that are having an influential effect on the model. In particular it seems that they are particularly poor at locating groups of values when the flow is depressed or enhanced over several days.

The influence function approach seems to have similar problems although the identification of groups of outliers not detected by the ARIMA model may indicate areas where the correlation structure is variable and the global ARIMA may break down.

Work with the ARIMA family suggests that traffic count series can be typically modelled by the $(1,0,0)(0,1,1)_7$ model. The number of outliers selected is usually more conservative than the other methods considered here, yet it is more successful than the others at identifying single additive outliers. The process used here performs indifferently when searching for a group of outliers. The ARIMA approach is more flexible, however, and can be used to explore the possibility of structural changes (Tsay, 1988) and multiple outliers (Bruce and Martin, 1989), although these lead to more complicated modelling processes which are more computationally expensive.

ACKNOWLEDGEMENTS

This research was undertaken with the financial support of the Science and Engineering Research Council, under grant GR/G/23180. The authors would like to thank Annabelle Payne for her contribution to this work.

BIBLIOGRAPHY

Ahmed, M.S. and Cook, A.B.. "Analysis of Freeway Traffic Time Series Data by Using Box - Jenkins Techniques". Transpn. Res. Rec. 1979. 722, 1-8.

Box, G.E.P. and Jenkins, G.M.. "Time Series Analysis, Forecasting and Control". Holden-Day, 1976.

Box, G.E.P. and Tiao, G.C.. "Intervention Analysis with Application to Economic and Environmental Problems". JASA 1975. 70, 70-79.

Bruce, A.G. and Martin, R.D. "Leave-k-out Diagnostics for Time Series". JRSS. B 1989. 51, 363-424.

Chemick, M.R. Downing, D.J. and Pike, D.H.. "Detecting Outliers in Time Series Data". J.A.S.A. Vol 77 1982, no.380, 743-747.

Fox, A.J.. "Outliers in Time Series". J.R.S.S. Ser B. 1972. 34, 350-63.

Gaudry, M.. "An Aggregate Time Series Analysis of Urban Transit Demand: The Montreal Case". Transpn. Res. Vol.9 1975. 249-258.

Hakim, S., Shefer, D., Hakkert, A.S. and Hocherman, I.. "A critical review of macro models for road accidents". Acc. Anal. and Prev., 1991, 23(5), 379-400.

Harvey, A.C. and Pierse, R.G.. "Estimating Missing Observations in Economic Time Series". J.A.S.A. 1984. 79, 125-131.

Harvey, A.C. and Durbin J.. "The effect of seat belt legislation on British Road Casualties: A case study in Structural time series modelling", J.R.S.S. Ser A 1986, 149, 187-227

Jones, R.H.. "Maximum Likelihood Fitting of ARIMA Models to Time Series With Missing Observations". Technometrics 22 1980. 389-95.

Kohn, R. and Ansley, C.F.. "Estimation, Prediction and Interpolation for Arima Models with Missing Data". J.A.S.A. 81 1986. 751-761.

Marshalls, R.J.. "Autocorrelation estimation of time series with randomly missing observations", 1980

McLeod, G., Everest, J.T. and Paulley, N.J.. "Analysis of Rail and Air Passenger Flows Between London and Glasgow Using Box-Jenkins Methods". TRRL Supplementary Report 524 1980.

Nihan, N, and Homesland, K.. "Use of the Box-Jenkins Time Series Technique in Traffic Forecasting". Transpn 9 1980. 125-143.

Pena, D.. "Influential Observations in Time Series". Technical Report 2718 1984. Mathematics Research Centre, Univ. of Wisconsin, Madison.

Tsay, R.S.. "Outliers, Level Shifts, and Variance Changes in Time Series". J. Forecast, 7 1988, 1-20.

Watson, S.M.. "Non-normality in Time Series Analysis". Unpub. PhD. Thesis, Trent Poly. 1987.