Published paper

Tight, M.R., Redfern, E.J., Watson, S.M., Clark, S.D. (1993) *Outlier Detection and Missing Value Estimation in Time Series Traffic Count Data: Final Report of SERC Project GR/G23180.* Institute of Transport Studies, University of Leeds. Working Paper 401

UNIVERSITY OF LEEDS
Institute for Transport Studies

ITS Working Paper 401 *ISSN 0142-8942*

June 1993

OUTLIER DETECTION AND MISSING VALUE ESTIMATION IN TIME SERIES TRAFFIC COUNT DATA: FINAL REPORT OF SERC PROJECT GR/G23180

MR Tight, EJ Redfern, SM Watson & SD Clark

ITS Working Papers are intended to provide information and encourage discussion on a topic in advance of formal publication. They represent only the views of the authors, and do not necessarily reflect the views or approval of the sponsors.

CONTENTS

Page

ABSTRACT

1.OBJECTIVES 1

2.RESEARCH DEVELOPMENT 1
2.1Methods based on the auto-correlation 2
2.2Methods based on a parametric model 3
2.3Multi-variate models 4
2.4Non-linear models 4
2.5An illustration 4

3.PROGRAMME MANAGEMENT 5

4.SIGNIFICANCE OF RESULTS 5

5.CONCLUSIONS 5

6.REFERENCES 6

ABSTRACT

TIGHT, MR & REDFERN, EJ (1993). Outlier detection and missing value estimation in time series traffic count data

A serious problem in analysing traffic count data is what to do when missing or extreme values occur, perhaps as a result of a breakdown in automatic counting equipment. The objectives of this current work were to attempt to look at ways of solving this problem by:

1)establishing the applicability of time series and influence function techniques for estimating missing values and detecting outliers in time series traffic data;

2)making a comparative assessment of new techniques with those used by traffic engineers in practice for local, regional or national traffic count systems

Two alternative approaches were identified as being potentially useful and these were evaluated and compared with methods currently employed for `cleaning' traffic count series. These were based on evaluating the effect of individual or groups of observations on the estimate of the auto-correlation structure and events influencing a parametric model (ARIMA).

These were compared with the existing methods which included visual inspection and smoothing techniques such as the exponentially weighted moving average in which means and variances are updated using observations from the same time and day of week.

The results showed advantages and disadvantages for each of the methods.

The exponentially weighted moving average method tended to detect unreasonable outliers and also suggested replacements which were consistently larger than could reasonably be expected.

Methods based on the autocorrelation structure were reasonably successful in detecting events but the replacement values were suspect particularly when there were groups of values needing replacement. The methods also had problems in the presence of non-stationarity, often detecting outliers which were really a result of the changing level of the data rather than extreme values. In the presence of other events, such as a change in level or seasonality, both the influence function and change in autocorrelation present problems of interpretation since there is no way of distinguishing these events from outliers.

It is clear that the outlier problem cannot be separated from that of identifying structural changes as many of the statistics used to identify outliers also respond to structural changes. The ARIMA $(1,0,0)(0,1,1)_7$ was found to describe the vast majority of traffic count series which means that the problem of identifying a starting model can largely be avoided with a high degree of assurance.

Unfortunately it is clear that a black-box approach to data validation is prone to error but methods such as those described above lend themselves to an interactive graphics data-validation technique in which outliers and other events are highlighted requiring acceptance or otherwise manually. An adaptive approach to fitting the model may result in something which can be more automatic and this would allow for changes in the underlying model to be accommodated.

In conclusion it was found that methods based on the autocorrelation structure are the most computationally efficient but lead to problems of interpretation both between different types of event and in the presence of non-stationarity. Using the residuals from a fitted ARIMA model is the most successful method at finding outliers and distinguishing them from other events, being less expensive than case deletion. The replacement values derived from the ARIMA model were found to be the most accurate.

KEY WORDS: Time series, Outliers, Residuals, Missing values, Traffic, Transport

Contact: Mr E.J. Redfern, Department of Statistics (tel: 0532 335172).

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

1.OBJECTIVES

A serious problem in analysing traffic count data is what to do when missing or extreme values occur. Such 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. Hence, more analytical approaches are needed for identifying extreme values and replacing missing data. The main objectives of this current work were therefore:

1)To establish the applicability of time series and influence function techniques in:
a)estimating missing values
b)detecting outliers.
2)To make a comparative assessment of new techniques with those used by traffic engineers in practice for local, regional or national traffic count systems

2.RESEARCH DEVELOPMENT

An initial literature review (Watson et al, 1991) suggested that there were a range of alternative methods that had been developed in the statistical arena for the detection and replacement of outliers and missing values. There had however only been a few instances of time series methodology applied to transport time series and none of these covered outlier techniques. The models were usually based on the ARIMA family and the methodology developed by Box and Jenkins (1976).

A detailed study of a large data base of traffic counts (Redfern et al, 1993b) and other time series suggested that there is usually a strong weekly cycle in the data, the shape of which may vary according to season, time and type of road. The underlying volume of traffic also changes and so in addition to outliers the performance of the various methods in the presence of changes in level, seasonality and variability would need to be evaluated.

Two alternative approaches were identified as being potentially useful and these were evaluated and compared with methods currently employed for `cleaning' traffic count series. These were:

1)Diagnostics based on the estimated autocorrelation function:
a)using an influence function approach in which the effect of each observation on the estimated autocorrelation is determined;
b)the average change in the autocorrelation resulting from removing observations from its estimation.
2)Events influencing a parametric model (ARIMA):
a)diagnostics based on residuals resulting from the fitted model;
b)diagnostics based on the changes in components of the parametric model that result from allowing for events in the data set.

These were compared with the existing methods which included visual inspection and smoothing techniques such as the exponentially weighted moving average in which means and variances are updated using observations from the same time and day of week. Observations outside 4 standard deviations of current mean are rejected and replaced using a weighted average of previous values.

Both strategies also allow missing values to be estimated. Using the auto-correlation function these are based on its estimated values, while for the ARIMA parametric model including an intervention at the time of a missing value allows its estimation.

The methods were first applied to a range of traffic count data (Clark et al, 1993). Then to allow more detailed comparison of the methods a simulation study was carried out in which the events in simulated data from both seasonal and non-seasonal ARIMA models were controlled (Redfern et al, 1993a). Events superimposed on traffic count series in which no real events could be detected by any methods were also examined (Watson et al, 1992a).

The types of event included in the study were the two types of outliers that can occur in a time series sense (Innovative Outliers caused by a change in the underlying noise and Additive Outliers which are a change in the value), and structural changes such as change in level, gradual change to new level, transient (non-permanent) change, seasonal change and variance change. Such a study allowed the evaluation of the performance of the methods in distinguishing between outliers and other events. Problems such as the correct location of events and the effect of non-stationarity were also assessed. The accuracy of methods for estimating missing values were also studied using simulated data.

2.1METHODS BASED ON THE AUTO-CORRELATION

Based on the estimated autocorrelation function two statistics were considered:

1)The influence of a pair of values (y_t,y_{t+k}) k time points apart on the kth lag of the estimated autocorrelation function r_k is given by:

$$I_{t,k} = y_t\, y_{t+k} - r_k \frac{(y_t^2 - y_{t+k}^2)}{2}$$

from which the influence of the tth observation on all lags of the estimated autocorrelation function is defined by (Watson, 1987) as:

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

where the L_t indicates the set of values in the tth row and D_{t-1} the trailing diagonal from time t-1.

2)The global change in the estimated autocorrelation function resulting from deleting a point or group of points (A) from the data set can be assessed using statistics such as:

$$\Delta_A = \sum_{k=1}^{l} (r_k - r_k^A)^2$$

where r_k^A is the autocorrelation at lag k omitting the points in A.

The advantage of these statistics is that no parametric model needs to be identified, while the chief disadvantage is that it is difficult to identify the type of event once a point or set of points is identified as influential. Both criteria assume stationarity in the series and if this is achieved by differencing then interpretation is clouded because a single outlier is replaced by one or more outliers.

2.2METHODS BASED ON A PARAMETRIC MODEL

The parametric model used is the ARIMA family, extended where necessary by adding interventions to allow for outliers and other events. The model is:

$$X_t = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)\Delta^d \Delta_s^D}\varepsilon_t + \sum^{k\varepsilon A} f_k(t)$$

Where $BX_t = X_{t-1}$, $\Delta = 1\text{-}B$, s is the length of the season, and $\varphi(B)$, $\Phi(B^S)$, $\theta(B)$ and $\Theta(B^s)$ are polynomials in B and B^s and:

$$f_k(t) = \omega_k 0 \frac{\omega(B)}{\delta(B)} I_t^k \quad with\ \omega(B), \delta(B)\ being\ polynomials\ of\ B$$

$I_t^{(k)}$ (= 1 if t = k, 0 elsewhere) indicates the time of the event and choice of $f_k(t)$ allows the complete range of events identified above to be modelled.

From a study of a large data base of traffic count time series (Redfern et al, 1993b) one model in the ARIMA family was identified that seemed to adequately describe the majority of the series. This was the multiplicative seasonal model with a non-seasonal auto-regressive parameter, a seasonal moving average parameter and a seasonal differencing (ARIMA (1,0,0) $(0,1,1)_7$). After all outliers and other events had been allowed for, based on 50 traffic counts, the autoregressive parameter had a mean value of 0.397 (with standard error 0.033) and the seasonal moving average value had a mean value of 0.597 (0.023). There was no significant difference in the values of the auto-regressive parameter between road types, but the moving average parameter was higher for non-built up and B roads than roads in built up areas.

Using the residuals to the fitted model two methods are available for identification of outliers. The first identifies them from large residuals (greater than 3 standard deviations from zero was felt to be appropriate) which is very easy to do but it is not always easy to decide what type of event has caused the large residual. An alternative approach (Tsay, 1988) is to regress residuals on interventions at each point in turn from which a weighted average of the residuals at **and after** the point of interest allows a set of likelihood ratio statistics to be constructed. The maximum over all the significant values of this set identifies both the time point and the type of event that has occurred. It is a method that is computationally efficient and varying the form of the weighting applied to the residual allows the full range of events to be considered.

Results based on analysing real series and working with simulated data suggests that the method is poor at identifying events covering several time points, such as a block of outliers, and transient changes in which the return to the normal level is over a period of a few observations. It is more robust in identifying harder events such as outliers or permanent steps. There is some evidence of smearing resulting in inaccuracies in pinpointing the time of the event.

A second approach was considered based on changes in the estimated parametric model that results from removing observations from the time series called **leave-k-out diagnostics** (Bruce and Martin, 1989). This technique uses the measures of influence of a point or points on parameters and the residual variance. They reported that the diagnostic based on residual variance is less susceptible to smearing and the method allows evaluation of outliers or patches of outliers. It is however computationally intensive and uses algorithms that are not widely available in existing statistical software.

As an alternative therefore the idea of **add-in diagnostics** was introduced in which an intervention can be added to the series at each time point in turn. Such a method can be done using most time series packages and has the advantage that a full range of possible events can be considered. It is, however, still computationally intensive as a complete model has to be evaluated for each time point and each type of event that needs to be considered.

2.3MULTI-VARIATE MODELS

If a site consists of a short time series, but there are several nearby sites, the collective analysis of changes in data over time for all the related sites may provide a better basis for modelling than univariate analysis of information at that site alone. In order to investigate such spatial temporal relationships the NAG library of algorithms was used to apply multivariate models to three neighbouring links using data from the SCOOT system in Leicester. The modelling process was found to be computationally very intensive and of limited success. This was almost certainly because some investigation into the prior differencing or other pre-treatment of the data was required.

2.4NON-LINEAR MODELS

One way of modelling data which exhibits problems of non-normality, cyclicity, asymmetry and time irreversibility (all of which are evident in traffic count data) is to abandon the linearity assumption and use a non-linear structure. One such structure is the Self-Exciting Threshold Autoregressive (SETAR) model, and as part of this project its applicability to several traffic count series was investigated. Results were very encouraging with relatively few parameters needed and good measures of fit achieved (i.e. low AIC values and sample noise standard deviation). It was felt that this was one area worthy of further investigation although some considerable programming effort would be needed.

2.5AN ILLUSTRATION

To illustrate the problems and the success of the method an example of a typical series in which the outliers identified by the different methods are marked and summarised is shown in Figure 1. The table below indicates the actual outliers identified by each of the methods.

Table 1: Outliers detected at time points

Average	41	44	48	51	52	53	55	58
	62	65	66	69	76	83	88	90
Influence	10	17	31	44	45	52	64	65
	66	73	87	88	94	95	101	108

ARIMA (large residual)	52	66	88					
ARIMA (Structure)	52	64	66	73	80	87	88	

The exponential weighted moving average and influence function illustrate the problems caused by non-stationarity. The first detects outliers only over the period when the average traffic flow is raised while the latter identifies all the large seasonal peaks. Using only the large residuals the ARIMA model avoids these problems and identifies the three clear outliers. However, allowing for the changes in level (about time 40 and 95) and seasonality (time 95) further outliers are identified which have a rational explanation as being different from the underlying structure.

3.PROGRAMME MANAGEMENT

Management of the project took place through regular minuted meetings of all staff involved with the project, on research and administrative matters. Such meetings were held approximately once every two to three weeks throughout the period of the project.

4.SIGNIFICANCE OF RESULTS

The significance of the results of this project for current engineering practice is largely through the provision of new tools and the extension of the range of tools currently available to the engineer which enable him/her to increase their understanding of traffic patterns and how these change over time.

5.CONCLUSIONS

The application of three fundamentally different ideas to a range of traffic count series, other transport data and an extensive set of simulated data produced some interesting results. The exponentially weighted moving average method tended to detect unreasonable outliers and also suggested replacements which were consistently larger than could reasonably be expected.

Methods based on the autocorrelation structure were reasonably successful in detecting events but the replacement values were suspect particularly when there were groups of values needing replacement. The methods also had problems in the presence of non-stationarity, often detecting outliers which were really a result of the changing level rather than extreme values. In the presence of other events, such as a change in level or seasonality, both the influence function and change in autocorrelation present problems of interpretation since there is no way of distinguishing these events from outliers. Some evidence of smearing, resulting in small errors in the exact location of the extreme values, was also detected. The advantage of such methods is that they require no identification of a parametric model and are computationally efficient.

It is clear that the outlier problem cannot be separated from that of identifying structural changes as many of the statistics used to identify outliers also respond to structural changes. Using the intervention extension to the ARIMA family it is possible to develop techniques that enable us to distinguish with a much greater reliability between extreme values and other events. The ARIMA $(1,0,0)(0,1,1)_7$ was found to describe the vast majority of traffic count series which means that the problem of identifying a starting model can largely be avoided with a high degree of assurance. Methods based on the residuals are easier and quicker to apply although again there is evidence of smearing and mis-identification in some cases. The disadvantages though are small relative to those in the other methods.

The degree of computation required to apply the methods based on observation deletion almost certainly means that they are not a viable proposition for use as a data validation tool for the large number of traffic count series that would need to be handled. The accuracy achieved in detecting events was no greater than that gained by working with the residuals to the fitted model. The estimation of replacement values for both extreme values and missing values was more efficiently done using the parametric ARIMA model.

Unfortunately it is clear that a black-box approach to data validation is prone to error but methods such as those described above lend themselves to an interactive graphics data-validation technique in which outliers and other events are highlighted requiring acceptance or otherwise manually. An adaptive approach to fitting the model may result in something which can be more automatic. Thus the ARIMA model could be cast into state-space form and the parameters updated as each observation is made, the validity of each observation being compared to projections from the currently estimated model. This would allow for changes in the underlying model to be accommodated. The type of event would need to be assessed retrospectively on the basis of succeeding points once a problem has been identified. Another such approach is to use the dynamic linear model (Harrison and Stevens, 1976) updating the estimation using the Kalman filter and some form of tracking function to identify the events. The evaluation of such a method was beyond the scope of the present study.

In conclusion it was found that methods based on the autocorrelation function are the most computationally efficient but lead to problems of interpretation both between different types of event and in the presence of non-stationarity. Using the residuals from a fitted ARIMA model is the most successful method at finding outliers and distinguishing them from other events, being less expensive than case deletion. The replacement values derived from the ARIMA model were found to be the most accurate.

6.REFERENCES

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

BRUCE, A.G. and MARTIN, R.D. (1980) Leave-k-out diagnostics for time series. *Journal of the Royal Statistical Society*. Series B, 51, 363-424.

CLARK, S.D., WATSON, S.M., REDFERN, E.J. and TIGHT, M.R. (1993a) Application of outlier detection and missing value replacement techniques to various forms of traffic count data. *ITS Working Paper **384***, University of Leeds.

HARRISON, P.J. and STEVENS, C.F. (1976) Bayesian forecasting. Jnl of the Royal Statistical Society. Series B, 205-247.

REDFERN, E.J., WATSON, S.M., CLARK, S.D. and TIGHT, M.R. (1993a) Identifying outliers and other events in seasonal time series. *ITS Working Paper*, University of Leeds (in preparation).

REDFERN, E.J., WATSON, S.M., CLARK, S.D., TIGHT, M.R. and PAYNE, G.A. (1993b) Modelling outliers and missing values in traffic count data using the ARIMA model. *ITS Working Paper **395***, University of Leeds (in preparation).

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

WATSON, S.M. (1987) Non-Normality and Non-Linearity in Time Series Analysis. PhD thesis, Trent Polytechnic, Nottingham.

WATSON, S.M., CLARK, S.D., REDFERN, E.J. and TIGHT, M.R. (1992a) Outlier detection and missing value estimation in time series traffic count data. *Proceedings of the 6th World Conference on Transport Research*, Lyon, France.

WATSON, S.M., TIGHT, M.R., CLARK, S.D. and REDFERN, E.J. (1991) Detection of outliers in time series. *ITS Working Paper **362***, University of Leeds.