Working Paper 362

June 1991

DETECTION OF OUTLIERS IN TIME SERIES

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

This work was sponsored by the Science and Engineering Research Council

CONTENTS

INTRODUCTION

As part of a SERC funded project this study aims to summarise the most relevant and recent literature with respect to outlier detection for time series and missing value estimation in traffic count data. Many types of transport data are collected over time and are potentially suited to the application of time series analysis techniques, including accident data, ticket sales and traffic counts. Missing data or outliers in traffic counts can cause problems when analysing the data, for example in order to produce forecasts. At present it seems that little work has been undertaken to assess the merits of alternative methods to treat such data or develop a more analytic approach. Here we intend to review current practices in the transport field and summarise more general time series techniques for handling outlying or missing data.

The literature study forms the first stage of a research project aiming to establish the applicability of time series and other techniques in estimating missing values and outlier detection/replacement in a variety of transport data. Missing data and outliers can occur for a variety of reasons, for example the breakdown of automatic counters. Initial enquiries suggest that methods for patching such data can be crude. Local authorities are to be approached individually using a short questionnaire enquiry form in order to attempt to ascertain their current practices. Having reviewed current practices the project aims to transfer recently developed methods for dealing with outliers in general time series into a transport context. It is anticipated that comparisons between possible methods could highlight an alternative and more analytical approach to current practices.

A description of the main methods for detecting outliers in time series is given within the first section. In the second section practical applications of Box-Jenkins methods within a transport context are given. Current practices for dealing with outlying and missing data within transport are discussed in section three. Recommendations for methods to be used in our current research are followed by the appendices containing most of the mathematical detail.

Section 1 - The Detection of Outliers in Time Series

The definition of an outlier may be an observation which is unrepresentative, spurious or discordant. It may be regarded as an observation which does not come from the target population. Whilst many methods have been developed for detecting outliers in simple random samples, outlier detection in a time series context has only evolved more recently. The majority of methods assume that the time series process can be represented by a univariate Box-Jenkins (1976) ARIMA structure ("B-J model"). This class of models are frequently applied to many types of time series data. A definition of the general structure and summary of the method are contained within Appendix A.

Fox (1972) appears to be the first to consider outliers within time series, assuming an AR structure with Gaussian noise. Two broad categories of outlier are defined: "Additive Outliers (AO)" where a single point is affected and "Innovative Outliers (IO)" where an innovation to the process affects both an observation and the subsequent series. Fox's approach is basically a likelihood ratio criteria, comparing the estimated error for an observation with the estimated standard error of that discrepancy.

In practice it may be felt that there is little point in detecting outliers unless they are influential to the outcome of model fitting or forecasting. Hau and Tong (1984)

interpret the leverage of points in time series regression as the relative change in the fitted or estimated values brought about when there is a small change in the observed value.

Outlier detection methods based on the influence of an observation often refer to Hampel's (1974) influence curve. This provides a basis for measuring the influence of an observation on a particular parameter (such as the mean). It is often used for robust estimation where influential or leverage points exist.

In sampled data, this was employed by Devlin, Gnanadesikan and Kettenring (1975), to investigate how outliers affect correlation estimates in Bivariate data. They propose a first order approximation to the Sample Influence Function on the estimate r and also an Empiric Influence Function using the Empirical Distribution Function in place of the underlying Probability Distribution.

This result is extended into a time series context by Chernick, Downing and Pike (1982) with the development of an Influence Function Matrix for the detection of outliers in time series. Their methodology is largely subjective and relies upon the visual interpretation of a matrix of "+" and "-" signs, indicating the influence of an observation upon the autocorrelation function at different lags. Their method was refined and quantified by Watson (1987) to give the outlier detection statistic.

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

where $I = Y_t Y_{t+k} - \rho_k (Y_t^2 - Y_{t+k}^2) /2$

Σ_{Lt} indicates summation over the L elements in the t'th row and Σ_{Dt-1} indicates summation over all available elements in the (t-1)'th diagonal of the Influence Function Matrix, $P = L + D_{t-1}$. A further development was the proposal of an outlier replacement/adjustment based on the statistic. Tests using real and simulated time series of an econometric nature produced encouraging results.

As linear regression type models are more commonplace in a transport context than stochastic models, it may be possible to expand regression based diagnostics into time series data. For example Cook's distance statistic (Cook, 1977) is based on the influence of the i'th observation on the parameter β in the simple model

$$\underline{Y} = \underline{x}\,\underline{\beta} + \underline{\varepsilon}$$

Influence is then measured by comparing $\hat{\beta}$ with $\hat{\beta}_{(i)}$ where $\hat{\beta}_{(i)}$ is the estimate of β computed without the i'th case. In time series data, $\hat{\beta}_{(i)}$ is the parameter estimate when the i'th case is perturbed and Cook's distance statistic is then in the amended form given in the appendix.

This methodology is expanded by both Bruce and Martin (1989) and Kohn and Ansley (1986). Bruce and Martin observe that although the technique is well established in ordinary regression analysis, the dependency relations which exist in time series gives rise to a "smearing" effect when the test statistics for a model coefficient is calculated. Building on work by Kohn and Ansley (1986) they propose an alternative diagnostic based on the effect on the noise variance of the deletion of an observation. The ARIMA

model is represented in state space form. The variance based diagnostic is shown to be clearer and more sensitive than that based on model coefficients and patches of outliers are accommodated by a "leave k out" principle. Masking problems are approached using an iterative or subset deletion method.

The identification of influential observations is also the key concern of Pena (1984), but within an ARIMA modelling context. He argues the importance of concentrating on influential observations and not specifically outliers which may not substantially affect the parameter estimates of the model. Two outlier types, namely the AO and IO are considered particularly and their effects on the estimated parameters of the model are measured.

Pena refers to the work of Chang and Tiao (1983) who, building on the work of Fox (1972), suggest a useful iterative procedure for outlier detection and parameter estimation. They recommend comparing two likelihood ratio statistics $\lambda_{1,T}$ and $\lambda_{2,T}$

$$\lambda_{1,T} = \frac{\hat{W}_I}{\hat{\sigma}_a} \qquad \text{(to test for an IO)}$$

$$\lambda_{2,T} = \frac{\hat{W}}{\hat{\sigma}_a (1+\pi_1^2 \ldots \pi_n^2)^{-1/2}} \qquad \text{(to test for an AO)}$$

where $\hat{W}_I$ and $\hat{W}_A$ are the estimated values of the outlier W, and $\pi_I (I = 1,...n)$ are the parameters of the AR process, and $\hat{\sigma}_a^2$ estimates noise variance.

Two statistics are also presented by Pena, one for each outlier type and these are given in Appendix E. From the appendix it can be seen that they are very similar to those of Chang and Tiao, and in the case of the IO, to that of Cook (1977). Simulation and 'real life' data results show the AO test statistic to be quite accurate and reliable in picking out influential points, although the IO statistic did not perform quite so well. The IO test tended to highlight several adjacent points, none of which need be the relevant observation.

A quite different basis for modelling outliers in time series is that described by Box and Tiao (1975) as "intervention analysis". Here a stochastic model is built which specifically allows for known interventions in the series, the example given here being the effects of petrol legislation on the levels of atmospheric oxidant. The general form of the model is given by:

$$Y_t = f(k, \varepsilon, t) + N_t$$

where Y_t is the time series or log series etc.

f (k, ε, t) allows for interventions,

k are unknown parameters
t represents time
ε represents the effects of interventions

N_t is a stochastic model for noise and may be a mixed ARIMA model $\phi(B)\, N_t = \theta(B)\, a_t$. This could also be amended to allow for nonstationarity, seasonality and so on.

The part of the model which allows for an intervention can be expanded into a dynamic structure representing the effects of ε as follows

$$f(\delta_1 W_1 \varepsilon_1 t) = \sum_{j=1}^{k} y_{ij} = \sum_{j=1}^{k} [W_j(B)/\delta_j(B)]\ \varepsilon_{t,j}$$

where y_{ij} represents the dynamic transfer from ε_{ij} and the parameters k (previously combined) are now denoted by δ and W.

$\varepsilon_{t,j}$ could be an exogenous time series, but as a $0 \rightarrow 1$ indicator variable could denote the presence or absence of an intervention.

For example, for a single variable k = 1, then

$$Y_t = y_t + N_t$$

$$= (W(B)/\delta(B))\varepsilon_t + (\Theta(B)/\Phi(B))a_t$$

then the transfer y_t to the output from ε_t is generated by the linear difference equation

$$\delta(B)y_t = W(B)\varepsilon_t$$

eg for a step $\varepsilon_t = S_t^{(T)}$, where $S_t^{(T)} = \begin{cases} 0 & t < T \\ 1 & t \geq T \end{cases}$

or for a pulse $P_t^{(T)} = \begin{cases} 0 & t \neq T \\ 1 & t = T \end{cases}$

So for a step change in input, an output step change would be produced according to

$$y_t = WBS_t^{(T)}$$

Box and Tiao (1975) continue to describe various responses to step and pulse changes, where the effect of the intervention may be permanent, temporary, constant or decaying in several ways. It is also important to note that y_t represents the additional effect of the intervention over the noise. When N_t is non-stationary, large changes in output can occur even with no intervention present. Interventions which occur over a series of time intervals can also be expressed using a series of pulses, for example a 3-month advertising campaign as 3 pulses.

An example is given using Los Angeles oxidant data. Here there are 2 definite interventions representing legislation concerning petrol lead levels and speed limits, together with a "winter month" effect. This is successfully modelled as 3 interventions with a noise model which reflects non-stationary and seasonal behaviour. Maximum likelihood is used as the method for parameter estimation.

This technique of intervention analysis is not dissimilar to the method of including dummy variables in ordinary regression, and many time series packages have a facility to model interventions.

The method is extended and generalised by Tsay (1988) in an investigation of outliers, level shifts and variance changes in time series. Tsay summarizes outlier detection methods to date as batch procedures, sequential methods and robust procedures. The latter include work by Denby and Martin (1979) and Martin and Yohai (1985) and are not described in much detail as they tend not to be very reliable in handling IO's.

Sequential methods, such as those by Harrison and Stevens (1976), Smith and West (1983), West, Harrison and Migon (1985) and West (1986) are described later in this report. They assume probabilistic models for outlier disturbances and require prior information of the underlying model to begin with.

Tsay describes his method as a "Batch procedure" because the full data set is used in the detection of outliers. Work by Hillmer, Bell and Tiao (1983), Chang and Tiao (1983) and Chang (1982) would also fall into this classification.

The basic univariate ARMA structure is used for model fitting with an intervention function to model outliers, the function varying according to the outlier type.

Tsay describes the following outlier types:

Additive Outliers (AO)
Innovative Outliers (IO)
Level Shift (LS)
Permanent Level Change (LC)
Transient Level Change (TC)
Variance Change (VC)

A summary of Tsay's methodology to deal with each type of outlier is given within the Appendix F. Parameter estimation is described for AO, IO, LC and TC using techniques of simple linear regression, which may be extended to derive Normally distributed test statistics. Parameter estimation is slightly different for the VC model and the test statistic, which follows an F-distribution, is most powerful if the time point d is known.

The modelling process is iterative and can detect both outliers and variance changes within the series. Estimation was via the Scientific Computing Associates (SCA) package (for IO, AO, VC) and additional Fortran routines (to handle TS and VC). No further details are given within the paper about the nature of the package or its availability. Illustrative examples are given using three data sets, including the monthly air-passenger miles (216 observations). Here a log transformation and differencing were needed, 2 LC, 4 TC and 4 AO were identified by the iterative process. The forecasts from the resultant model were compared with those from a previous model fit which did not allow for outliers. A considerable reduction in the width of the 1-step ahead forecast interval was seen, although the effect on point estimates was found to be negligible.

An alternative approach to outlier detection and the treatment of missing values employs Kalman filter techniques. These are basically recursive techniques where parameter estimates are updated as new information becomes available. They generally require initial estimates to be used as "start up" values. Although these methods can be quite cumbersome to program, they have an advantage in that they

can respond fairly quickly to changes within the data series. The Kalman filter provides a well established system of equations to allow the sequential updating of parameters.

Jones (1980) uses a Kalman filter recursive method to look at ARIMA fitting to time series with missing observations in particular. A non-linear optimization procedure is used to give Maximum Likelihood estimates of the ARIMA parameters, and allows model identification based on AIC values. Jones observes that in certain circumstances the MA order of an ARIMA model may be increased by outliers within the series. Modelling the error may therefore result in a more parsimonious model than would otherwise be achieved. It should also be noted that with such a recursive process, if a large block of data is missing, information about the past contained in the 'State' vector slowly dies away. Running the recursion across a large block of missing data is therefore equivalent to restarting the recursion at the other end. A Practical illustration is given using US drought data with an outlier, where various ARIMA models are fitted and compared using the -2 Ln likelihood and AIC criteria.

Harvey and Pierse (1984) suggest that the problems of parameter estimation and estimating missing values can both be solved by setting up the model in state space form and applying the Kalman filter. A state space formulation for an ARIMA model is described and where missing observations occur within the data the corresponding updating equation can be just omitted from the recursive cycle.

Predicting future observations is achieved using Kalman filter prediction equations. Harvey and Pierse also describe a form of fixed point smoothing via the Kalman filter to compute estimates for missing values.

An illustration is given from the B-J airline data using the whole series and then deleting sections to simulate missing data. However the method does require a run of values present at either the beginning or end of the series. The authors wrote their own program to model fit and provide estimates of missing values, and the results shown indicate success in both areas.

Harvey and Pierse also describe a method by Chow and Lin (1971, 1976) for estimating missing values based on a regression approach. The method assumes that the series is related to another fully observed series or set of series. However the algorithm contains some difficulties in that it requires the construction and inversion of the covariance matrix associated with the variables. Also some measure of the covariances between missing and observed values is needed. Chow and Lin concentrate on the situation where observations are missing at regular intervals and the noise series is either uncorrelated or follows an AR(1) process. In a transport context this need not necessarily be the case, and some difficulty could be experienced in finding a wholly observed related series.

In a highly theoretical paper, Kohn and Ansley (1986) generalize the results given by Harvey and Pierse (1984). Again the Kalman Filter is used to calculate the marginal likelihood for an ARIMA model with missing observations. It is also used to predict and interpolate missing observations, obtaining the MSE of the estimate. One of the main changes is in the requirements for the initial start up vector and any pattern of missing observations is allowed (not necessarily with an unbroken run).

The technique is applied to the same Air Passenger data examined by Harvey and Pierse and the results were found to coincide. The work of Harvey and Peirse and

Kohn and Ansley formed a basis for the Bruce and Martin method described earlier.

Clearly a variety of methods have been applied to deal with missing data and treat outlying observations. Three main approaches have emerged from the relevant literature, although a degree of overlap is evident in their evolution. Chernick, Downing and Pike (1982) and Watson (1987) concentrate on the influence a particular observation has on the correlation estimate, where no model fitting is initially needed. Outliers or even missing observations may not be detected therefore unless they exert some undue influence.

Harvey and Pierce (1984), Kohn and Ansley (1986) and Bruce and Martin (1989) look at the influence of observations on the parameters of an ARIMA model fit. Here the model is represented in state space form and estimation follows on iterative process.

Finally Tsay (1988) summarizes an intervention modelling approach, where outliers are actually modelled within the broad ARIMA fit. The form of the intervention function varies according to the form of the suspect outlier and an iterative procedure can assist in unmasking hidden points. In section 2 the application of time series methods specifically in a transport context is examined in more detail.

Section 2 - Application of Box-Jenkins Time Series Models to Transport Data

To date it seems that B-J models have not been applied widely to transport data - some notable exceptions are described below. This may be due to the statistical knowledge needed to apply the models, the specialist software packages required, or the conservative approach generally apparent within transport modelling. It may further be argued that the models produced are sometimes difficult to interpret practically and so may be less popular. However research discussed in this section shows that this class of time series model can be applied to different types of transport data successfully, accurately and at much lower cost than some other models. It is important to demonstrate the successful application of B-J models here as many methodologies to treat missing or outlying observations rely upon this. Several different classes of transport data are referred to below including traffic counts, air passenger flows and journey times.

McLeod et al (1980) report on the analysis of rail and air passenger flows between London and Glasgow using B-J methods. The aims of the study were to produce short and medium term forecasts for both types of flow and to compare the results with those obtained from conventional regression analysis. The data consisted of passenger flows from British Rail and British Airways between 1964 and 1976. Although the rail data was available in several subclasses such as first and second class tickets, the aggregation of the data into total sales was found to best reflect air/rail competition. Both models contained "irregularities" due to strikes and non-consistent sampling.

Initially, purely stochastic models were applied using intervention analysis to reflect strikes and other anomalies. Comparisons were made on the basis of 13 period and quarterly aggregation, with the results shown below:

Average % SE of 1-step ahead forecast		
TOTAL AIR	4.7	13 PERIOD
TOTAL RAIL	12.9	
TOTAL AIR	5.0	QUARTERLY
TOTAL RAIL	11.4	

Clearly the level of aggregation used has some effect on forecasting errors.

Further model development involved the inclusion of causal variables, such as journey time and several indicators of price. Although the quality of those explanatory variables was felt to be somewhat suspect, the model was improved by their inclusion and the average % SE of 1-step ahead forecasts was reduced by the order of 1%-2%.

A summary of the final model is included within the Appendix B. Although it seems that the data did not include missing values, known outliers in the data were modelled using intervention analysis.

A regression type approach is taken for rail passenger levels by Stark (1981). The main aims of this study were to assess the effects of opening a new station on rail patronage in Glasgow. Effects of fare variations and information on seasonal/random variations were also considered. As is often the case with regression analysis, dummy variables were used to model aberrant observations. The dummy variables had uncertain interpretation and were used extensively to "carry" the model through bad patches of data.

A successful application of the Box-Jenkins methodology to journey time data was made by Davies (1989). Journey times of inbound traffic on a link near Leeds had been recorded for the period 7.15am to 9.40am. Time was measured in seconds for cars travelling between two points and was averaged for cars passing the first point in successive five minute intervals.

The process of model identification consisted of testing for stationarity, applying a log transform, differencing to induce stationarity and then model fitting. The Root Mean Square Error (RMSE) was calculated for the three future data points which were forecast from the model.

A further improvement was made by the inclusion of explanatory variables to form a transfer function model. This had the effect of reducing the RMSE from 64.91 (stochastic model) to 21.09 seconds (transfer model). Overall a very successful fit to the data was achieved using this class of time series model. However it should be noted that these results refer to data for a single link collected over a limited period whereby the data would not be subject to seasonal and other variation.

Similar models were fitted to journey time data by Watson (1990). Here the time series were collected on inbound traffic on a London route as part of a separate study. The aim was to develop dynamic journey time models for inclusion in a route guidance system. Overall the fitting process was not quite as successful as that achieved by R Davies. This was almost certainly due to the "noisy" nature of the data and the difficulty in obtaining data which had been collected at equally spaced intervals. Data were recorded at the downstream end of the link and therefore referred to car journey

times of successive vehicles arriving at a point within a given time interval. Information was collected according to traffic light cycle times which in theory were constant but in practice were subject to quite large variations in length. No formal attempts were made to model outliers and no sections of missing data were included in the series. However the exclusion of a couple of visually obvious outliers from the series did result in an improved fit.

A combination of linear regression and B-J techniques were applied by Gaudry (1975) in order to model demand for public transport. The data consisted of monthly time series of aggregate demand in Montreal. Initially linear regression techniques were used to include a variety of explanatory variables such as price, service levels, comfort levels and other socio-economic variables. Gaudry then specifies the error term of the model in terms of a B-J AR series. He argues that this should be included because of the absence of some independent variables (such as parking, price indicators and visiting) known to have been significant in other studies. Their absence was expected to leave some systematic variation within the residuals following the initial regression fit.

A summary of the model is given in the Appendix C. It can be seen that the model is non linear in β_i and P_i and so was solved using non-linear estimation. Results are compared for three models which include slight variations in the predictor variables. All models seem to provide an excellent fit to the data with "r^2" $\geq$ 0.96 in each case. No mention is made of the treatment of missing or outlying observations in the study.

Difficulties in B-J modelling with extreme values in the series are reported by Davies et al (1990). Linear time series techniques were used to forecast freeway traffic congestion and were found to perform well in forecasting mean values. However the models did not seem to be able to easily accommodate the data extremes which corresponded to the onset of congestion. Loop detectors were used to give measurements of traffic volume and lane occupancy, the aim being to forecast the conditions that give rise to congestion. Data were collected at one minute intervals from 6am to 9am on five weekdays in late November/early December in order to develop and test potential forecasting routines.

Univariate and transfer function models were fitted to forecast lane occupancy and storage rates for a particular road section. It was found that the forecasts of storage tended to hover around the mean and ignore extreme values. This behaviour caused some difficulty as it was the extreme values which were of particular interest in this study. Forecast values of lane occupancy tended to track the true values but were a minute too late. The overall conclusion of the study was that to predict extreme values the time series models were not adequate and a different approach should be adopted.

A quite different conclusion was reached by Ahmed and Cook (1979) in their analysis of freeway traffic using B-J techniques. Here 166 data sets from 3 surveillance systems in LA, Minneapolis and Detroit were used to develop short term forecasts of traffic data (1-step ahead) and in particular traffic volumes and occupancy. The performance of the models was compared with that of three other methods, ie the moving average model, double exponential smoothing and the "TRIGG and LEACH adaptive model". Comparisons were made on the basis of the forecast errors produced by each model.

A traditional route of identification, estimation and diagnostic checking was followed, with the ACF, PACF and BOX and Pierce Q-statistic being used to test for residual autocorrelations. The initial ACF/PACF plots showed an ARIMA (0,1,3) to be an appropriate form, and diagnostic checks on the residuals supported the adequacy of this model. The transferability of the model was checked by applying the same broad ARIMA (0,1,3) structure to different time series from the three collection systems. Almost inevitably variations in the values of the parameters were found, but the overall structure did seem appropriate for several series.

The effect of the sampling interval was also checked by aggregating 20 second data to 60 seconds and refitting. As the model parameters varied a little it was felt that the sampling interval had had some effect but not a particularly significant one. The study concludes overall that the model performed very well in short term forecasting.

Perhaps the most enthusiastic proposers of B-J methods in traffic forecasting are Nihan and Homesland (1980). Their research uses monthly volumes of traffic on a freeway segment between 1966 and 1976 in the USA - the aim being to forecast for 1977. As the actual volumes were available for that year it was then possible to compare actual and forecast volumes. They comment that forecasting using traditional socio-economic models can be expensive and inaccurate with large errors even in short term prediction. Instead they suggest the use of a univariate time series model which could then be further amended to include behavioural variables. A previous study by Homesland (1979) describes the use of some popular existing techniques - decomposition, moving average and exponential smoothing and concludes that the response of these techniques to significant traffic changes is inadequate.

Nihan and Homesland also comment on the appropriateness of B-J models for time series with problems, such as missing information and interventions. In fact their own data, which consists of traffic flows across a bridge has an overall upward trend and is subject to usual random disturbances such as incidents, accidents, weather and maintenance work. However they are also aware of four additional major shocks, these being redundancies, the gas crunch, priority treatment for high occupancy vehicles and major surface repairs. There were also problems with missing and partial data.

A standard B-J package finally produced an ARIMA (12, 1, 7) model and despite all the problems with the data, forecasting up to 12 months ahead gave errors in the order of only 5%. This is a fairly high order model, and normally a simpler model would be preferred. Slightly reduced errors were produced by more complex multiplicative models, but the slightly better results were felt to be not warranted by the increased complexity of the model. Overall the study was extremely encouraging about the use of this type of model in terms of the cost, accuracy and simplicity. The authors see the only real difficulty in the expertise needed by the modeller. Expertise is not needed particularly in the use of relevant computer packages but in the overall methodology of model building and assessment and in the interpretation of results obtained from Packages.

In conclusion, the Box-Jenkins class of models have been applied to a selection of transport data with generally good results. Where difficulties have been reported these may in some part be attributable to problems with the quality of data available. However a succesful model fit does seem to rely upon a measure of experience or expertise by the practitioner.

Section 3 - Practical Solutions to Missing Value/Outlier Detection with Transport Data

Little documentation seems to exist on how missing data and outlier problems are solved by the transport practitioner. Informal discussions with consultants and local authorities suggest two categories of solution. One is the "Black Box" situation where computer packages used to process the data allow for such observations, but it is not clear how they are treated, if at all. It has been difficult to obtain any further details about such packages. An alternative solution is the "Rule of Thumb" where practitioners use their subjective experience to "guestimate" an appropriate value or may apply a simple procedure such as averaging adjacent values.

One computer package known to be used in practice is the WODIN suite (Hackett, 1981). This allows the user to flag 'Special events' (which may be the cause of outliers) and add a small amount of text for description. Holiday affected days can also be indicated and a particular symbol is used for estimated flows. Missing data may also be flagged. Output statistics are then produced concerning the overall quality of the data, where the number of missing or patched items is recorded. However the manual does not explain how such observations are treated within the program.

The procedure used by the DTp in London is known to be currently under review and existing FORTRAN routines are to be replaced by new SAS macros.

The current procedure involves validating new traffic count data against old data from the same site collected 1,2,.. weeks previously. A mean and standard deviation are defined by the breakdown.

Site/day of week/period of day/vehicle cat.

The new mean and variance are then given by:-

new mean = (0.7 x old mean) + (0.3 x new obs.)
new var = (0.7 x old var) + (0.3 x (new obs - new mean)^2)

Error messages are produced for an observation outside 4 standard deviations from the old mean, although the limit can be changed for example with bank holidays. Missing or rejected data are then estimated by an experientially weighted mean, so the estimated count at time t, $\hat{x}$ (t) is

$$\hat{x}(t) = 0.3x(t-1) + 0.21x(t-2) + 0.147x(t-3) + ...$$

No details are available at present about how the detection/replacement procedure may be updated.

A more detailed and theoretically sound approach is given by Southworth et al (1989). They describe a computerized traffic analysis system operating in the U.S.A. which was developed primarily to record and predict large traffic flows at the time of a major evacuation such as in war time, although it also has practical applications for smaller localized traffic events. The software is designed to process traffic counts in massive quantities, for example 5 minute interval counts. The software contains the following

units:

GURN - an expert systems for integrating the software
AUTOBOX - a time series modelling package
CROSSTALK - to load data from the counters into the computer
DYNAMIC NETWORK TRAFFIC ASSIGNMENT MODEL
TRAFFIC PATTERN RECOGNITION PROGRAM.

Perhaps of most interest is the AUTOBOX package which follows a traditional path of B-J model identification and fitting, also allowing input series to be included in order to form generalized transfer function models. The system also allows for the identification and treatment of outliers of three types - a pulse, step and seasonal pulse. The methodology seems to be an intervention analysis approach and points are identified individually. A series of multiple regressions are performed using the residuals from the original B-J model as the dependent variable. Each intervention is then adjusted and a final model produced for the adjusted series.

This is an interesting example as it illustrates a successful and very practical application of B-J models with "real-life" count data and describes the on-line treatment of outliers.

Wright (1974) describes a method of estimating traffic flows from incomplete data. Here data is not treated as a time series as such, but is grouped across sites according to similar patterns of traffic flow variation. The "missing" flows are then estimated as proportional to those measured during corresponding hours at the remaining sites. The technique described can also be used for the related problem of estimating average daily total flow at a single site from one hour flow samples taken at intervals over a period of several days, weeks or months. Sampling fluctuations, including the effects of outliers and seasonal variation can then be allowed for.

Overall it has been difficult to find detailed information on current practices in patching or treating transport data. Although a small number of computer packages are available (such as the WODIN suite), it is not clear how they treat missing and aberrant data. The "rule of thumb" or expert judgement may be the most popular approach, suggesting room for a more analytic methodology.

Section 4 - Conclusions and Recommendations

Although the detection and treatment of outliers in a time series context is relatively new, it is clear that several main approaches have now emerged. These could be classed as follows:

(1) Methods based on the influence of observations
(2) Intervention type models
(3) Recursive processes using the Kalman filter

Each method has its own advantages and problems and it is not clear yet which would perform most successfully with transport data. It is therefore proposed that the methods of Watson (1987), Tsay (1988) and Kohn and Ansley (1986) are used initially as typifying each approach. Clearly the ease of application and interpretation of the method would ultimately influence its use by the transport practitioner.

However it can be seen from Sections 2 and 3 that whilst little analytic work has been carried out in a transport environment to solve the problem of missing and abberrant observations, classic time series models can be applied accurately and economically with great success.

The next stage of our research will therefore involve the development of computer algorithms to apply each of the above methodologies. If any of the detection and treatment process were to be adopted, in practice it would be essential for the algorithms to be robust and "user friendly". It will be interesting to form a comparison between the performance of each approach, though difficult to categorize results as "correct" or "incorrect" due to the different fundamental basis in each case.

It is anticipated that several types of transport data may be used to assess each approach, although initially we will concentrate on traffic counts obtained from HETS and the Department of Transport.

REFERENCES

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

BOX, G.E.P. and JENKINS, G.M. (1976) "Time Series Analysis, Forecasting and Control". Holden-Day

BOX, G.E.P. and TIAO, G.C. (1975) "Intervention Analysis with Application to Economic and Environmental Problems. JASA, 70, 70-79.

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

COOK, R.D. (1977) "Detection of Influential Observations in Linear Regression". Technometrics, Vol. 19, 15-18.

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, University of Chicago, Statistics Research Centre.

CHERNICK, M.R. DOWNING, D.J. and PIKE, D.H. (1982). "Detecting Outliers in Time Series Data". J.A.S.A. Vol 77, no.380, 743-747.

CHOW, G.C. and LIN, A. (1971) "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series By Related Series". Review of Economics and Statistics, 53, 372-375.

CHOW, G.E. and LIN, A (1976) "Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series". J.A.S.A, 71, 719-721.

COOK, R.D. (1977) "Detection of Influential Observations in Linear Regression". Technometrics, Vol.19, 15-18.

DAVIES et al (1990) "Adaptive Forecasting of Freeway Traffic Congestion". TRB Annual Meeting 1990.

DAVIES, R. (1989) BOX-JENKINS. "Analysis of Journey Times for Leeds Radial Route" Unpublished Technical Note, Institute for Transport Studies, Leeds.

DENBY, L and MARTIN, R.D. (1979) "Robust Estimation of the First Order Autoregressive Parameter". J.A.S.A. Vol.74, no. 365, 140-146.

DEVLIN, S.J. GNANADESKAN, R and KETTENRING, J.R. (1975) "Robust Estimation and Outlier Detection with Correlation Coefficients". Biometrika Vol. 62, no.3, 531-545.

FOX, A.J. (1972) "Outliers in Time Series". J.R. Statist. Soc. Ser.B. 34, 350-63.

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

HACKETT, T (1981), "WODIN; Traffic Information Storage Tabulation and Analysis". User Guide-Version 1.0.

HAMPEL, F.R. (1974). "The Influence Curve and ITS Role in Robust Estimation". J.A.S.A. Vol.69, 383-93.

HARRISON, P.J. and STEVENS, C.F. (1976). "Bayesian Forecasting". J.R.S.S. SER. B, 205-247.

HARVEY, A.C. and PIERSE, R.G. (1984). "Estimating Missing Observations in Economic Time Series". J.A.S.A., 79, 125-131.

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.

HILLMER, S.C., BELL, W.R. and TIAO, G.C. (1983) "Modelling considerations in the Seasonal Adjustment of Economic Time Series". In Applied Time Series Analysis of Economic Data (Ed. A. Zellner), pp. 74-100. Washington D.C.: US Bureau of The Census.

HOLMESLAND, K.O. (1979) "Use of the Box and Jenkins Time Series for Forecasting Commuter Traffic Volumes". Research Report No.79-9, Seattle: Urban Transportation Program, Depts. of Civil Engineering and Urban Planning, University of Washington.

JONES, R.H. (1980). "Maximum Likelihood Fitting of Arma Models to Time Series With Missing Observations". Technometrics 22, 389-95.

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

MARTIN, R.D. and YOHAI, V.J. (1985). "Robustness in Time Series and Estimating Arima Models". Handbook of Statistics. Vol.5. Elsevier Science Publishers. B.V. 119-155.

MCLEOD, G., EVEREST, J.T. and PAULLEY, N.J. (1980) "Analysis of Rail and Air Passenger Flows Between London and Glasgow Using Box-Jenkins Methods". TRRL Supplementary Report 524.

NIHAN, N. and HOLMESLAND, K. (1980). "Use of the Box-Jenkins Time Series Technique in Traffic Forecasting". Transpn, 9, 125-143.

PENA, D. (1984). "Influential Observations in Time Series". Technical Report 2718. Mathematics Research Centre, University of Wisconsin, Madison.

SMITH, A.F.M. and WEST, M. (1983) "Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter". Biometrics 39, 867-878.

SOUTHWORTH, F, CHIN, S.M. and CHENG, P.D. (1989) "A Telemetric Monitoring and Analysis System for use during Large Scale Population Evacuations". IEE, 2nd International Conference on Road Traffic Monitoring, Conference Publication 299.

STARK, D.C. (1981) "Time Series Analysis of Glasgow Suburban Rail Patronage". TRRL Supplementary Report 649.

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

WATSON, S.M. (1987) "Non-normality in Time Series Analysis". Unpublished PhD. Thesis, Trent Polytechnic.

WATSON, S.M. (1990). "Time Series Modelling of Journey Times for Route Guidance" Institute for Transport Studies WP331, University of Leeds.

WEST, M., HARRISON, P.J. and MIIGON, H.S. (1985) "Dynamic Generalized Linear Models and Bayesian Forecasting". J.A.S.A., 80,73-98.

WEST, M. (1986) "Bayesian Model Monitoring". J.R. Statist. Soc. Ser B, 48,70-78.

WRIGHT, C. (1974). "A New Method of Estimating Daily Traffic Flows and Its Use in Verifying Predicted Flows in Whitstable, Kent", PTRC New Year Meeting, Jan. 1974.

WRIGHT, C. "Estimating Daily Traffic Totals From Incomplete Data", Research Group in Traffic Studies, University College London, unpublished.

APPENDIX A
Summary of Box-Jenkins Methodology

The aim of univariate Box-Jenkins modelling is to progressively remove all the systematic variation in the series until only a series of random fluctuations remain. This may be thought of as applying a series of filters to the data to increasingly "sift out" more variation. The two main filters that are applied are the Autoregressive filler (AR) and Moving Average filter (MA). The data may first require a transformation such as differencing or integation (I) to achieve stationarity. A series is considered to be weakly stationary if the mean and variance are constant independent of time. The model is represented as an (ARIMA) structure of order (p,d, q), and is given by:

$$\underbrace{Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} \ldots - \phi_p Y_{t-p}}_{\text{AR order P}} = \underbrace{a_t + \Theta_1 a_{t-1} + \Theta_2 a_{t-2} \ldots \Theta_q a_{t-q}}_{\text{MA order q}}$$

Were Y_t may be the d'th difference of the original series. It is assumed that the series $\{a_t\}$ are distributed N $(0, \sigma^2)$, $\{\phi_i\}$ and $\{\phi_j\}$ are parameters to be estimated from the data. Model identification relies upon the pattern of spikes seen in the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF) and Inverse Autocorrelation Function (IACF) as indicated below.

	MA(q)	**AR(p)**	**ARMA (p,q)**	**WHITE NOISE**
ACF	D(q)	T	T	0
PACF	T	D(p)	T	0
IACF	T	D(p)	T	0

D(q) - function drops off the zero after lag q
T - function tails off exponentially
O - function equals zero at all non zero lags

A more detailed explanation is given in Box and Jenkins (1976)

APPENDIX B

Final models for B-J application to rail/air passenger flows

Air Flows

$$\nabla\nabla_4 Y_{10t} = \underset{(10)}{40}\ \nabla X_{5t-1} - \underset{(1310)}{(4720 + 1520B)}\ \nabla\nabla_4 X_{2t} + \underset{(0.18)}{(1 - 0.67B)}\ \underset{(0.19)}{(1 - 0.83B^4)}\ a_{10t}$$

Average % SE = 3.5%

Rail Flows

$$\nabla\nabla_4 Y_{9t} = \underset{(3512)}{-\ 12066}\ \nabla\nabla_4 \varepsilon_{13t} \quad \underset{(2472)}{-\ 4881}\ \nabla\nabla_4 X_{3t} - \underset{(49190)}{129990}\ \nabla\nabla \log X_{4t-1} + \underset{(0.19)\quad(0.18)}{\frac{1}{(1+0.78\,B^4 + 0.52B^8)}}\ a_{9t}$$

Average % SE = 8.9%

Where B is the backward shift operator, $[a_t]$ is the error series, bracketed figures are 1 x SE.

Y_{10t} = London-Glasgow Air Passenger Flow
Y_{5t} = Net Acquisition of Financial Assets
X_{2t} = London-Glasgow Real Round Trip Fare
Y_{9t} = London-Glasgow Total Rail Passenger Flow
ε_{13t} = London-Glasgow Average Journey Time
X_{3t} = London-Glasgow Average Total Rail Fare
X_{4t} = Index of Industrial Production

APPENDIX C

Urban transport demand model - Gaudry

The regression-type demand model:

$$Y_t^d = \beta_0 + \beta_1 TW_t + \sum_{i=2}^{n} \beta_i X_{it} + U_t$$

where the noise model is

$$U_t = \sum_{i=1}^{r} \rho_i U_{t-i} + e_t$$

Y_t^d is the demand at time t

TW_t is a waiting time function

X_{it} are other explanatory variables

Overall this gives:

$$Y_t = \sum \rho_l Y_{t-l} = \sum_{i=0}^{n} \beta_i X_{it} - \sum_{i=0}^{n} \sum_{l=1}^{r} \beta_i \rho_l X_{i,t-l} - e_t$$

which may be written

$$Y_t^* = \sum_{i=0}^{n} \beta_i X_{it}^* + e_t$$

where

$$Y_t^* = (y_t - \sum_{l=1}^{r} \rho_t Y_{t-l}) \text{ and}$$

$$X_{it}^* = (X_{i,t} - \sum_{l=1}^{r} \rho_l x_{i,t})$$

APPENDIX D

Amended form of Cook's Distance Statistic

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^1 (X^1X) (\hat{\beta}_{(i)} - \hat{\beta})}{\rho\sigma^2}$$

referring to the simple model

$$Y = \underline{x}\,\underline{\beta} + \underline{\varepsilon}$$

where:-

$\underline{Y}$ is an n-vector of observed data
$\underline{x}$ is an nxp matrix of constants
$\underline{\beta}$ is a p-vector of unknown parameters
$\underline{\varepsilon}$ is an n-vector of iid errors
$\underline{\hat{\beta}}_{(i)}$ is the parameter estimate when the i'th case is perturbed

APPENDIX E

AO and IO outlier tests given by Pena (1987)

For the AO:

$$D_2(T) = \frac{\lambda_{2,T}^2}{h(\sum_{k=0}^{h} \hat{\pi}_{o,k}^2)} \quad \tilde{E}^1_T (X'_Z X_Z)^{-1} \tilde{E}_T$$

where: $\lambda_{2,T}^2 = \frac{\tilde{W}_0^2}{\tilde{\sigma}_a^2 (\Sigma \tilde{\pi}_{0,1}^2)^{-1}}$

$$\tilde{W}_0 = Z_t - \sum_{i=1}^{h} \hat{\delta}_{o,i} (Z_{T+i} + Z_{T-i})$$

$$\hat{\delta}_{o,i} = (\hat{\pi}_{o,i} - \sum_{k=1}^{h-i} \hat{\pi}_{o,k} \hat{\pi}_{o,k-i}) / (\Sigma \hat{\pi}^2_{o,k})$$

$\hat{\pi}_{o,i}$ = the j'th component of $\hat{\pi}_0$

$\tilde{E}_T = S_T - A_T \hat{\pi}_o$ (vector of pseudo residuals)

Z_t is the t'th observation (observed series)

Y_t = true "underlying" series

σ_a^2 is the variance of the noise process, a

h = lag

π_i are the parameters of the process

$\pi^1 = (\pi_1, \ldots \pi_h)$

$X^1_t = (Z_{t-1}, \ldots Z_{t-h})$

$\pi_0 = -1$

$S_T^1 = e\ (Z_{t+1} + Z_{t-1}, Z_{t+2} + Z_{t-2}, \ldots Z_{t+h} + Z_{t-h})$

For the IO:

$$D_1(T) = 1/h \quad \frac{e_T^2}{\sigma^2_a (1 - d_T)} \quad \left(\frac{d_T}{1-d_T}\right)$$

APPENDIX F

Modelling Outliers - Tsay (1988)

The basic model follows the form

$$Y_t = f(t) + Z_t$$

where f(t) is a deterministic or stochastic function to represent outliers or level changes and Z_t is an ARMA process. Y_t is the observed series.

For a deterministic model

$$f(t) = W_o \frac{W(B)}{\delta(B)} \varepsilon_t^{(d)}$$

where $\varepsilon_t(d) = 1$ if $t = d$ for a disturbance at time d
$= 0$ if $t \neq d$

W_o is a constant denoting the initial impact of the disturbance.

(NB this is a similar f(t) to that of the intervention model by Box and Tiao)

For a Stochastic Model

$$f(t) = W_o \frac{W(B)}{\delta(B)} e_t^{(d)}$$

$e_t^{(d)} = o \quad t<d \quad \{e_t^{(d)} \mid t \geq d\}$ iid $(o, /\sigma_e^{2)}$

f (t) may affect both the model and the variance.

Five cases of f (t) are considered, 4 deterministic and one stochastic.

a) $W_o = W_I$ and $W(B)/\delta(B) = \Theta(B)/\phi(B)$
IO model, outliers affect Z_t for every $t \geq d$

b) $W_o = W_A$ and $W(B)/\delta(B) = 1$

AO model, outlier affects Z_d only.

c) $W_o = W_L$ and $W(B)/\delta(B) = 1/(1-B)$

LC model, $Y_t = Z_t$ for $t<d$ $\quad Y_t = Z_t + W_L$ for $t \geq d$

d) $W_o = W_T$ and $W(B)/\delta(B) = 1/(1-\delta B)$ $o<\delta<1$

TC model, Z_t affected for $t \geq d$. Effect decays exponentially and finally disappears.

e) $W_o = W_v$, $W(B)/\delta(B) = \theta(B)/\Phi(B)$ $e_t^{(d)} = a_t$ for $t \geq d$
VC model, $-1 < W_v < \infty$