African Journal of Business Management Vol.6 (22), pp. 6631-6639, 6 June, 2012
Available online at http://www.academicjournals.org/AJBM
DOI: 10.5897/AJBM12.359
ISSN 1993-8233 ©2012 Academic Journals

Full Length Research Paper

A method for detection and correction of outliers in time series data

Hausitoe Nare[1], Daniel Maposa[2]* and 'Maseka Lesaoana[3]

[1]Department of Applied Mathematics, National University of Science and Technology, P.O. Box AC939, Ascot, Bulawayo, Zimbabwe.
[2]Monash South Africa (A Campus of Monash University, Australia), Private Bag X60, Roodepoort, 1725, South Africa.
[3]University of Limpopo, School of Mathematical and Computer Sciences, Department of Statistics and Operations Research, South Africa.

Accepted 3 April, 2012

Outlier detection has become an important part of time series analysis. This paper studies the problem of detecting and correcting outliers in time series data, and proposes a method based on the Gumbel distribution as a limiting distribution for outliers. Outlier detection influences modelling, testing and inference, because outliers can lead to model misspecification, biased parameter estimation, poor forecasts and inappropriate decomposition of series. We develop an algorithm for determining when an observation can be classified as an outlier. The method is then applied to some residuals of autoregressive integrated moving average (ARIMA) models fitted to Zimbabwe Stock Exchange Indices and if any outliers are detected a correction procedure is then applied to rid the data of the outliers. A new model is then fitted to the corrected data series and some analyses are performed. The results show that the method proposed is effective in detecting outliers, and the correction procedure ensures that the correct model for the data is specified and the parameter estimates are unbiased.

Key words: Outlier detection, Gumbel distribution, algorithm, autoregressive integrated moving average (ARIMA).

INTRODUCTION

The study of outliers is not a new phenomenon. It has in fact a long history dating back to the earliest statistical analysis (Barnett and Lewis, 1994). In essence outlier methods have developed hand in hand with other statistical methods. Unfortunately, in time series analysis this expansion of outlier methods has not been as rapid and widespread. One reason for this must be that methods of time series outliers were first considered explicitly in the 1970s (Fox, 1972). However, since then the amount of papers dealing with the issue has grown steadily (Glendinning, 1998; Ane et al., 2008; Gutierrez and Gregori, 2008).

*Corresponding author. E-mail: danmaposa@gmail.com.

In addition the proposed methods have also been used in applied work in several fields, including economics (Gutierrez and Gregori, 2008). There is still controversy, however, and not everyone accepts the use of outlier methods. But at least in econometrics it is commonplace to use dummy (indicator) variables where necessary, and this can be seen as an informal use of outlier methods (Proietti, 2008). Therefore, it seems that there is no reason not to have formal methods for detecting and modelling outliers, since outliers may have complex dynamic properties, which must be taken into account.

Outliers in a data set can arise for different reasons. A distinction must be made between two types of anomalies, namely gross errors and (true) outliers. Gross errors are faulty observations and their frequency in 'routine data' varies from 1% to as high as 10%, whereas

in 'high quality data' there are virtually no errors of this kind (Hampel et al., 1986). If the observation treated as a potential outlier cannot be shown to be a gross error, it has to be considered as a true outlier. These are the real values of the data, correct but somehow suspicious or surprising. It is these observations, from now on called outliers, and their detection and modelling that we are interested in here.

There are two main reasons for outlier analysis. Firstly, outliers bias our estimates and we would like to prevent this. Secondly, we want to find potential causes of extreme scores, for example, to some subgroup of our data set (Osborne and Amy, 2004; Dehon et al., 2009a, b; Kaya, 2010; Gumedze et al., 2010). In economic theory, this separation of the two goals of outlier analysis has an interesting interpretation. First of all, outliers can help in fitting imperfect theories to complex real world phenomena. This is the traditional usage, implicit in the use of dummy variables. But on the other hand, outliers can also be used to reveal where a theory does not work, or to check what aspects of it need refining in order to better describe the real world. The examination of outliers can therefore be justified not only from the traditional data analysis perspective, but also by appealing to the interaction of theoretical and empirical economics (Zellner, 1981).

Outliers can take several forms in time series. There are additive and innovational outliers (Fox, 1972). An additive outlier affects a single observation, which is smaller or larger in value than expected. In contrast an innovational outlier affects several observations. Three other types of outliers can be defined, namely level shifts, transient changes and variance changes (Tsay, 1988). A level shift simply changes the level or mean of the series by a certain magnitude from a certain observation onwards. A transient change is a generalisation of the additive outlier and level shift in the sense that it causes an initial impact like an additive outlier but the effect is passed on to the observations that come after it. A variance change simply changes the variance of the observed data by a certain magnitude.

Outliers affect the autocorrelation structure of a time series, and therefore they also bias the estimated autocorrelation (ACF), partial autocorrelation (PACF) and the extended autocorrelation functions (EACF). The exact results of the effects are complicated and require lengthy computations (Tsay, 1986a).

Some simulation results have suggested that additive outliers, transient changes and level shifts cause substantial biases in estimated ARMA parameters, whereas innovational outliers have only minor effects (Chen and Liu, 1993b). ARMA model estimation is traditionally based on the estimated ACFs and PACFs, and will in the presence of outliers be therefore misleading, unless outliers are somehow taken into account. Some robust methods of model selection and estimation of the ACF and the PACF have been presented (Masarotto, 1987; Glendinning, 1998).

Least squares and maximum likelihood methods are both sensitive to the presence of outliers, especially additive outliers, whereas various robust estimators can handle some of the problems caused by outliers (Alkutubi and Ali, 2011). The expectation-maximization (EM) algorithm produces outlier robust approximate maximum likelihood estimates for ARMA models.

Outliers have some effects on the forecasts from ARMA models, and especially outliers near the beginning of the forecast period can have serious consequences. Point forecasts may suffer only a little from additive outliers, but the prediction intervals can become severely misleading, as outliers can inflate the estimated variance of the series (Ledolter, 1989; Hotta, 1989).

Level shifts and transient changes can have more serious effects also on point forecasts even when outliers are not close to the forecast region (Trivez, 1993). Attempts have been made to construct forecasting intervals in the presence of outliers (Chen and Liu, 1993a; Phillips, 1996; Dehon et al., 2009a, b; Harvey et al., 2010).

METHODOLOGY

According to extreme value theory, a class of extreme value distributions characterise the possible distributions of sample maxima (LeBaron and Samanta, 2004). As an illustration, let X_i be a sequence of independent and identically distributed random variables with common distribution F, and denote their partial maxima by:

$$M_n = \max(X_1, ..., X_n), n \geq 1 \quad (1)$$

The Fisher-Tippett theorem (Fisher and Tippett, 1928) then states that there exist only three location-scale families of extreme value distributions, G, namely Frechet, Weibull and Gumbel distributions for which one can find constants, $c_n > 0$ and $d \in \Re$ such that:

$$\lim_{n\to\infty} \Pr\left(\left(\frac{M_n - d_n}{c_n} \right) \leq x \right) = G(x), \; x \in \Re \quad (2)$$

and the distribution F is then said to be in the maximum domain of attraction of G $\left(F \in MDA(G)\right)$. In other words, provided a non-degenerate limit can be observed, it follows that, whatever the distribution of the original variables; the limiting distribution of the maximum belongs only to this small set of distributions. These distributions can all be written as the generalised extreme value (GEV) distribution which is given by:

$$G_\xi(x; \mu, \beta, \xi) = \exp\left(-\left(1 + \xi(x - \mu)/\beta\right)^{-\frac{1}{\xi}} \right), \; if \; \xi \neq 0 \quad (3)$$

Or

$$G_{\xi}(x;\mu,\beta,\xi)=\exp\left(-e^{(-(x-\mu)/\beta)}\right), \text{ if } \xi=0 \quad (4)$$

The generalised extreme value distribution with $\xi<0$ is called the Weibull distribution Ψ_{α}. An example of an $MDA(\Psi_{\alpha})$ is the uniform distribution on $(0,1)$. The case $\xi=0$ is the Gumbel distribution. The normal distribution belongs to its maximum domain of attraction. Third family, where $\xi>0$, is defined by the Frechet extreme value distribution Φ_{α} with a positive tail index α, where $\alpha=\frac{1}{\xi}$. Its distribution is given by $\Phi_{\alpha}(x)=\exp\left(-x^{-\alpha}\right)$ for $x>0$.

The Gumbel distribution

The general formula for the probability density function of the Gumbel (maximum) distribution is:

$$f(x)=\frac{1}{\beta}e^{(-(x-\mu)/\beta)}\exp[-e^{(-(x-\mu)/\beta)}] \quad (5)$$

Where μ is the location parameter and β is the scale parameter. The case where $\mu=0$ and $\beta=1$ is called the standard Gumbel distribution (Gumbel, 1958).

Extreme value analysis

The purpose of this investigation is to determine the distribution of maximum values derived from normally distributed data with approximately mean zero and variance one. The other reason is to decide the potential cut-off where an observation can be accepted as an outlier or not (Reiss and Thomas, 2007; Gutierrez and Gregori, 2008; Dehon et al., 2009a, b; Harvey et al., 2010).

Simulation of extreme values

This was done using Microsoft Excel in the following steps:

1. A random sample of 250 observations $X_1,.....,X_{250}$ is generated from a normal distribution with mean zero and variance one.
2. The maximum value $X_{\max}=\max(X_1,.....,X_{250})$ of the random sample generated in the foregoing step is identified.
3. This is repeated a thousand (1000) times yielding a thousand extreme values.

Histogram plot of extreme values

The 1000 extreme values from each of the simulations described earlier are then plotted on a histogram to have a visual picture of the kind of distribution that the extreme values approximate.

Ranking of extreme values

This is done so as to determine, with some certainty, the magnitude of the potential cut-off point for an observation to be viewed as usual or unusual (outlier).

Probability plots

The histogram plot gives an indication of the distribution that the data might follow but a probability plot confirms the actual distribution satisfactorily. When the probability plots are done the extreme values must first of all be arranged in ascending order. For the Gumbel distribution the ordered samples are plotted against the linearised cumulative distribution function of the Gumbel distribution given by:

$$F^{-1}\left(\frac{i}{m+1}\right)=-\log\left(-\log\left(\frac{i}{m+1}\right)\right) \quad (6)$$

Where i is the position of the ith observation in the ordered sample and m is the total number of observations. If the probability is approximately linear then the extreme values follow the Gumbel distribution.

Box-Jenkins methodology

Autoregressive integrated moving average (ARIMA) models were developed by Box and Jenkins (1976). They added weight to the work done by Yule (1926), Walker (1931) and Wold (1938). The ARIMA model is a comprehensive approach to univariate time series analysis. The identification of the time series characteristics such as stationarity and seasonality require a schematic approach.

Model identification

By computing the autocorrelation coefficient one can see the autocorrelation of the time series with itself lagged 0, 1, 2 or more time periods. The pattern of the autocorrelation is used to identify the presence of seasonality in the series, thus helping to identify an appropriate model for a specific situation. By looking at the partial autocorrelations, one can identify the extent of the relationship between current values of a variable with earlier values of that same variable while holding the effects of all other time lags constant.

Diagnostic checks

The statistical adequacy of the model is checked at the diagnostic-checking stage (Sarkar, 2011). This is achieved by using various techniques that include residual analysis and checking the adequacy of the individual coefficients in the model using the t-statistic.

Properties of a good model

A good model is parsimonious. That is to say the model fits and represents the available data adequately without using unnecessary coefficients. A good model should also be stationary and invertible, that is to say the autoregressive (AR) and moving average (MA) coefficients should satisfy some inequalities. The

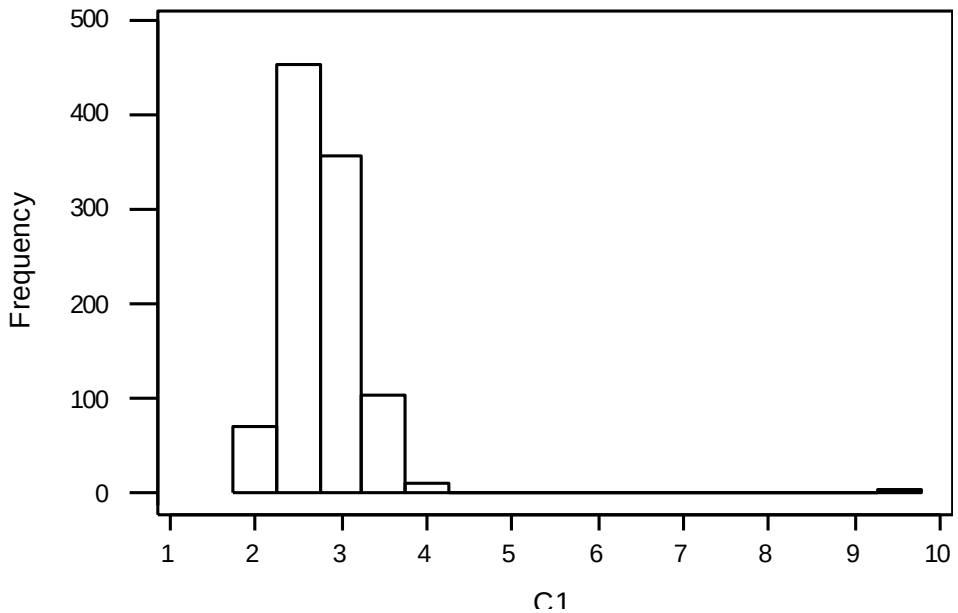

Figure 1. Histogram of extreme values.

model has to have uncorrelated residuals, as this is sufficient evidence that we cannot improve the model anymore by additional AR or MA terms.

Simulation of ARIMA models

Microsoft Excel was used to simulate ARIMA models. The simplest ARIMA models were considered, that is the moving average processes of order one and two (MA(1) and MA(2) and autoregressive processes of order one and two (AR(1) and AR(2)). The general structure of these processes is well known (Box and Jenkins, 1976). The series length is 1000 observations for each of the processes.

Outlier detection

In order to understand the stages involved in the outlier analysis process let us refer to the general outlier model, which is given as follows:

$$Y_t = Z_t + I_t(d)\omega_0\omega(L)/\delta(L) \quad (7)$$

Where Z_t is a regular ARIMA model, $\omega(L)$ and $\delta(l)$ are lag polynomials and $I_t(d)$ is an indicator variable and d is the timing of an outlier (Ané et al., 2008). For our purposes we will look at the case where $\frac{\omega(l)}{\delta(L)} = 1$ giving rise to the following model:

$$Y_t = Z_t + I_t(d)\omega_0 \quad (8)$$

This is known as the additive outlier model. The next stage will entail a situation where the simulated series discussed earlier are contaminated with an additive outlier at the midpoint, say t=500, the additive outliers are of magnitude $\omega = 3\sigma, \omega = 4\sigma \; and \; \omega = 5\sigma$. We then fit the 'best' ARIMA model to each series (both contaminated and uncontaminated).

Outlier detection and correction steps

1. From the fitted models, using the Box-Jenkins methodology, we standardise the residuals so that they follow a normal distribution with mean zero and variance one and then inspect the magnitudes of the residuals.
2. If the standardised value of any of the residuals is above the cut-off value determined from the ranking of the residuals, then consider the observation at that position to be an outlier.
3. We then eliminate the effect of the outlier by applying the correction procedure depending on the type of outlier at the position where the outlier has been detected.
4. Steps 1 to 3 are repeated until all further outliers are identified.

Parameter estimation

When the effects of the outlier have been eliminated, the Box-Jenkins methodology is then applied to estimate the parameters of the time series models without the outliers. The results are then compared with those of the original series to see if there are any differences.

RESULTS OF EXTREME VALUE ANALYSIS

The histogram plot of the extreme values (Figure 1) is skewed to the left indicating that the generated extreme values do not follow a normal distribution but a totally

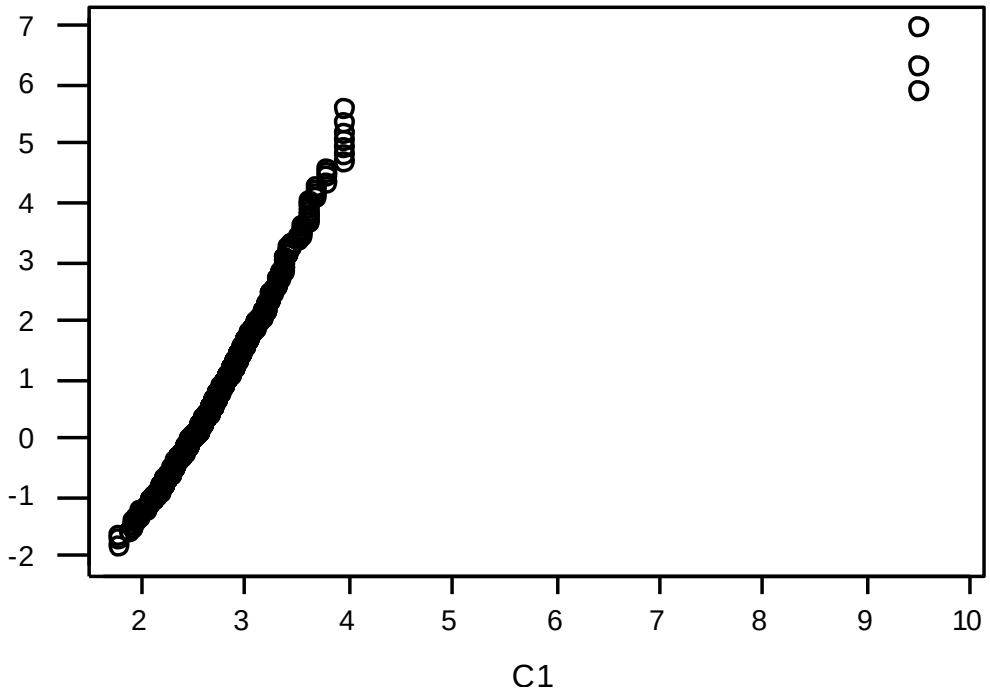

Figure 2. Gumbel probability plot for extreme values.

different family of distributions that we wish to investigate. The probability plot (Figure 2) is generally linear indicating that the extreme values follow a Gumbel distribution.

Ranking of extreme values

The ranking of the extreme value yielded 3.66 as a cut-off point for an observation to be classified as an outlier if its standardised residual from the fitted model is above this threshold.

Applications to real data

The daily end of day industrial index data from the Zimbabwe Stock Exchange for the period 3 January 2005 to 31 July 2006 is used to illustrate the effectiveness of the algorithm for detecting outliers, correcting for their effects and then estimating the model for the data that has been corrected for outliers.

Model identification

Stationarity of industrial index data

The time series plot of the log transformed data (Figure3) reveal an upward trend of industrial index data series. This is an indication of the non-stationarity of the data. Taking regular first differences might help in making the data stationary. A time series plot of the differenced data series is given in Figure 4.

The time series plot (Figure 4) shows that the data has become somewhat stationary and hence some autocorrelation function (ACF) and partial autocorrelation function (PACF) might actually be made to identify possible models for the data.

Autocorrelation structure of data series

ACF and PACF plots of the data strongly suggest an autoregressive model of order one and a seasonal moving average term at lag 12, that is, ARIMA (1 1 0)*(0 0 1)$_{12}$.

Parameter estimation

The model postulated in the foregioing is then fitted to the data and the two model parameters are highly significant with autoregression having a t-ratio of 9.99 and the seasonal moving average having a t-ratio of -2.81.

Diagnostic checks

The ARIMA (1 1 0)*(0 0 1)$_{12}$ model is then checked for its suitability as a fit for the data. The Box-Pierce statistic

Figure 3. Time series plot of logarithm transformed data.

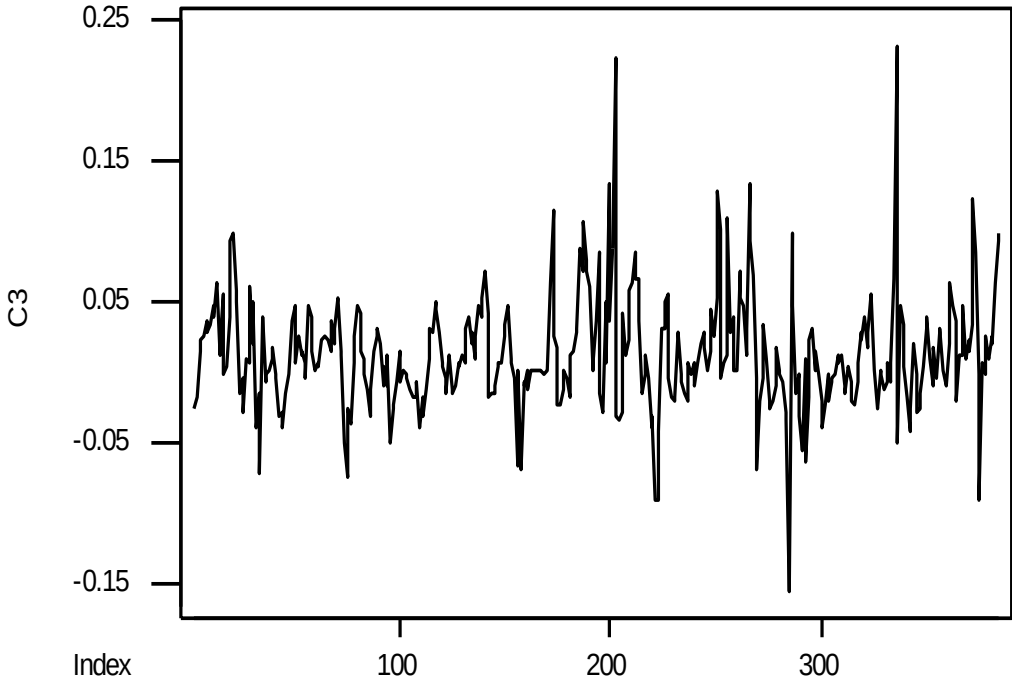

Figure 4. Time series plot of differenced and logarithm transformed data.

indicates that the model is an adequate fit for the data. Histogram and probability plots of the residuals are used to check for the normality of the residuals (Figures 5 and 6).

The histogram in Figure 5 shows a violation of the normality assumption as the histogram is not symmetric about the mean zero. This is reinforced by the non-linearity of the probability plot (Figure 6).

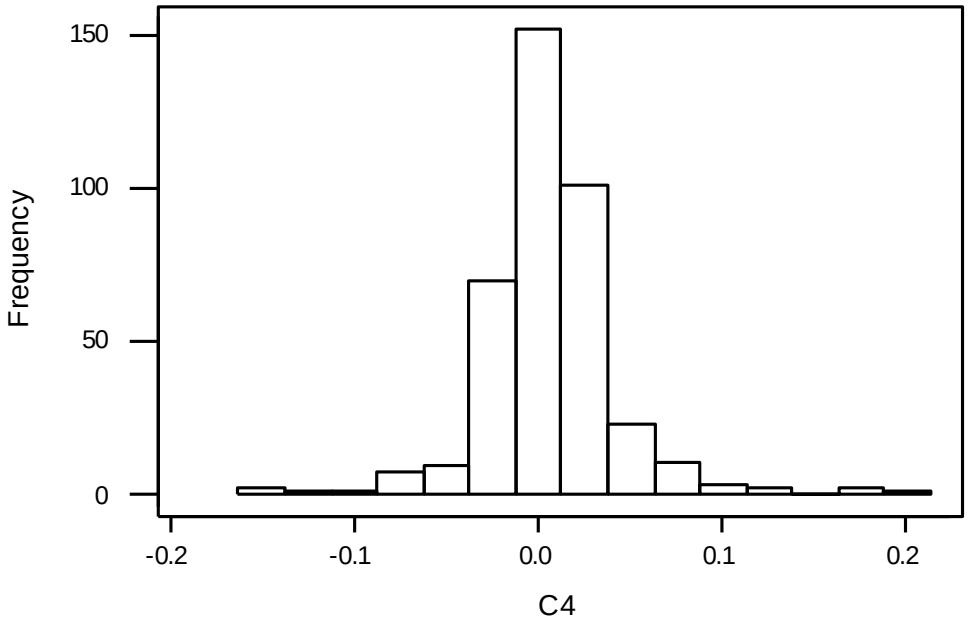

Figure 5. Histogram plot of residuals for model fitted to industrial index data.

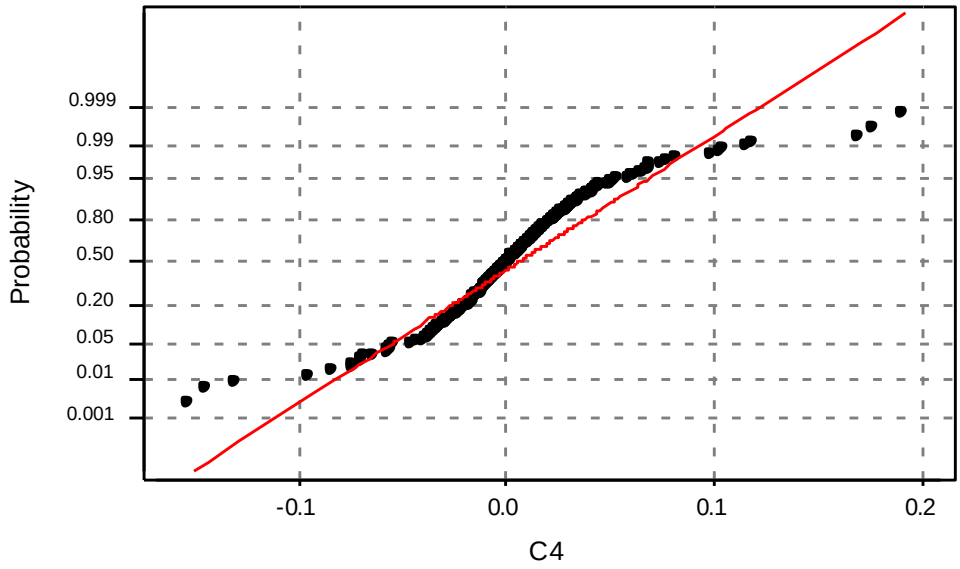

Figure 6. Normal probability plot of residuals.

Outlier detection and correction

We now implement the outlier detection and correction algorithm on the fitted model by comparing the absolute values of the standardised residuals with the cut-off value of 3.66 (determined by ranking extreme values) to detect outliers in the series. Outliers were detected at t=202, which appears to be a level shift, t=286, which appears to be an additive outlier, and t=337 which appears to be a level shift as well.

Table 1. Final estimates of parameters.

Type	Estimate	St. Dev.	t-ratio
MA 1	-0.5372	0.0472	-11.39
MA 2	-0.3998	0.0475	-8.41
SMA 12	-0.1343	0.0530	-2.54

Outlier correction at t=202

The standardised residual has a magnitude of 5.00458 and all the residuals had a standard deviation of 0.03543. The product of these two values was then added to all the observations before t=202 to correct for the effect of level shift. A new model is then fitted to the data and the procedure implemented again. An outlier is detected at t=286, which is additive in nature.

Outlier correction at t=286

The standardised residual has a magnitude of 4.23 and all the residuals had a standard deviation of 0.03375. The observation seems to be smaller relative to the neighbouring observations and hence the value is added to the observation at t=286 to correct for the effect of the additive outlier. A new model is fitted to the data and the procedure implemented again. An outlier is detected at t=337.

Outlier correction at t=337

The standardised residual has a magnitude of 5.90736 and all the residuals have a standard deviation of 0.03204. The product of these two is then added to added to all the observations before t=337. A new model is then fitted and procedure applied.

Final model for outlier corrected series

A combination of ACF and PACF plots helped to identify the ARIMA (0 1 2)*(0 0 1)$_{12}$ as a possible model for the outlier corrected series. Parameter estimation yielded the following results in Table 1.

It is apparent that all the parameters are significant in the model (Table 1) and hence the model is a suitable fit for the data. Implementation of the outlier detection procedure yielded no outliers hence we can conclude that the data series is now outlier free.

Conclusion

The study supports the claim that outliers do result in model misspecification as they affect the autocorrelation structure of any time series (Kaya, 2010; Li, 2011; Sarkar, 2011). In our case it is illustrated by the fact that initially we had the ARIMA (1 1 0)*(0 0 1) $_{12}$ as the best model that could be fitted to our data. Testing the residuals for normality and constant variance showed that both assumptions were violated although the parameters in the model were significant. Using this model for forecasts would have given misleading figures for a decision maker. This is possibly attributed to the presence of outliers.

The best model was found to be ARIMA (1 1 2)*(0 0 1) $_{12}$ after correcting the series for outliers and all the parameters were significant in the model. Diagnostic checks also showed that the assumptions of normality and constant variance were not violated.

This therefore demonstrates that the procedure is useful in detecting and correcting for outliers. It can be applied to all invertible ARIMA models. Moreover, it is flexible and easy to interpret. The procedure must be used with other diagnostic tools for time series to produce even better results. Further study is needed to investigate the variances and other sampling properties of the resulting parameter estimates.

The message from this study is that when examining economic time series data any potential outliers should be taken seriously, no matter what the ultimate aim or the model used may be. Outliers have already been shown to be potentially harmful, and there is also increasing evidence that the dangers are not only theoretical.

Other possible models that might be useful for modelling time series must be explored such as GARCH and ARCH models. These are non-linear forms of time series that might be used to model data that has got a lot of fluctuations in it. Non-linearity tests are normally done on the data before the previous models can be applied.

REFERENCES

Alkutubi HS, Ali HM (2011). Maximum likelihood estimators with complete and censored data. Eur. J. Sci. Res., 54(3): 407-410.

Ané T, Ureche-Rangau L, Gambet JB, Bouverot J (2008). Robust outlier detection for Asia-Pacific stock index returns. J. Int. Financ. Mark. Inst. Money, 18: 326-343.

Barnett V, Lewis T (1994). Outliers in statistical data. 3rd Edition. John Wiley & Sons, Chichester, p. 584.

Box GEP, Jenkins G (1976). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, p. 575.

Chen C, Liu LM (1993a). Forecasting time series with outliers. J. Forecast., 12: 13-55.

Chen C, Liu LM (1993b). Joint estimation of model parameters and outlier effects in time series. J. Am. Stat. Assoc., 88: 284-297.
Dehon C, Gassner M, Verardi V (2009a). Beware of good outliers and overoptimistic conclusions. Oxf. Bull. Econ. Stat., 71(3): 437-452.
Dehon C, Gassner M, Verardi V (2009b). A Hausman-type test to detect the presence of influential outliers in regression analysis. Econ. Lett., 105: 64-67.
Fisher RA, Tippett LHC (1928). Limiting forms of frequency distribution of the largest or smallest member of a sample. Camb. Philo. Soc., 24: 180-190.
Fox JA (1972). Outliers in time series. J. Royal Stat. Soc., Series B, 34: 350-363.
Glendinning RH (1998). Determining the order of an ARMA model from outlier contaminated data. Commun. in Stat.-Theory Meth., 27: 13-40.
Gumedze FN, Welham SJ, Gogel BJ, Thompson R (2010). A variance shift model for detection of outliers in the linear mixed model. Comput. Stat. Data Anal., 54: 2128-2144.
Gumbel EJ (1958). Statistics of Extremes. Columbia University Press, New York, NY, USA, p. 375.
Gutierrez JMP, Gregori JF (2008). Clustering techniques applied to outlier detection of financial market series using a moving window filtering algorithm. Euro. Centr. Bank Working Paper Series, Number 948, October 2008.
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986). Robust statistics: The approach based on influence functions. John Wiley and Sons, New York.
Harvey DI, Leybourne SJ, Taylor AMR (2010). Robust methods for detecting multiple level breaks in autocorrelated time series. J. Econom., 157: 342-358.
Kaya A (2010). Statistical modelling for outlier factors. Oze. J. Appl. Sci., 3(1): 185-194.
Masarotto G (1987). Robust identification of autoregressive moving average models. Appl. Stat., 36: 214-220.
Osborne JW, Amy O (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Res. Eval., 9(6). Available online at: http://PAREonline.net/getvn.asp?v=9&n=6 , accessed 03 February 2012.
LeBaron B, Samanta R (2004). Extreme value theory and fat tails in equity markets, Technical report, International Business School, Brandeis University, Waltham, MA 02453 U.S.A.
Ledolter J (1989). The effects of additive outliers on the forecasts from ARIMA models. Inte. J. Forecast., 5: 231-240.
Li Y (2011). Wavelet based outlier correction for power controlled turning point detection in surveillance systems. CREATES Research paper 2011-29.
Proietti T, (2008). Missing data in time series: A note on the equivalence of the dummy variable and the skipping approaches. Stat. Prob. Lett., 78: 257-264.
Reiss RD, Thomas M (2007). Statistical analysis of extreme values with application to Insurance, Finance, Hydrology and other Fields. 3rd Edition. Birkhäuser Verlag, Basel, 6: 189-204.
Sarkar SK, Midi H, Rana S (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study.
Trivez FJ (1993). Level shifts, temporary changes and forecasting. J. Forecast., 14: 543-550.
Tsay R (1986a). Time series model specification in the presence of outliers. J. Am. Stat. Soc., 81: 132-141.
Tsay R (1988). Outliers, level shifts and variance changes in time series. J. Forecast., 7: 1-20.
Walker G (1931). On periodicity in series of related terms. Proceedings of the Royal Society Series A, 131: 518-532.
Wold H (1938). A study in the analysis of stationary time series. 2nd Edition, 1954. Almqvist and Wiksell, Uppsala.
Yule GU (1926). Why do we sometimes get nonsense-correlations between time series? A study in sampling and the nature of time series. J. Royal Stat. Soc., 89(1): 1-63.
Zellner A (1981). Philosophy and objectives of econometrics. In: Currie D, Nobay R, Peel D (1981). Macroeconomic analysis: Essays in Macroeconomics and economics. Croom Helm, London. pp. 24-34.