응용통계연구 제 5권 2호
1992년 pp 139-156

Outlier Detection and Time Series Modelling in the Stationary Time Series[1)]

Jong-Hyup Lee[2)] and Ki-Heon Choi[2)]

< Abstract >

Recently several authors have introduced iterative methods for detecting time series outliers. Most of these methods are developed under the assumption that an underlying outlier-free model is known or can be identified. Since outliers can distort model identification or even make it impossible, we propose in this article a model independent procedure for detecting outliers. The proposed procedure begins with a descriptive data analysis of a time series using distance measures between two observations. Properties of the proposed test statistic are presented. To distinguish the type of an outlier are used transfer function models. An empirical example is given to illustrate the time series modeling procedure.

1. INTRODUCTION

Time series observations are often affected by unexpected events like Korean War, War in the Gulf, unusual changes in weather, and so on. Aberrant observations, which are the consequences of these interruptive events, are inconsistent with the rest of the series and are referred to as outliers. Since outliers are known to wreak havoc in time series analysis and make the resultant inference unreliable or even invalid, procedures are needed which can detect and hence remove such outlier effects. One of the first contributors to this problem is Fox(1972), who introduced two types of outliers. The first one consists of an outlier that affects only a single observation, and he referred it as an additive outlier (AO), or a Type I outlier. The second one corresponds to an outlier which affects not only a particular observation but also other subsequent observations,

1) This research was supported by NON DIRECTED RESEARCH FUND, Korea Research Fund, 1

2) Department of Statistics, Duksung Women's University, Seoul, 132-714, Korea

and he referred it as an innovational outlier (IO), or a Type II outlier.

Let X_t be a discrete time series which follows an autoregressive moving average(ARMA) model of order (p,q),

$$\phi(B)X_t = \theta(B)a_t \tag{1.1}$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots \theta_q B^q$ are polynomials in B, B is the backshift operator such that $BX_t = X_{t-1}$, and a_t is a Gaussian process of i.i.d. continuous random variables with mean zero and variance σ^2. We shall require that all the zeros of ϕ(B) and $\theta(B)$ lie outside the unit circle, and also that $\phi(B)$ and $\theta(B)$ have no common factors. In what follows, X_t will be used as the outlier-free stationary time series.

Let Z_t be the observed time series. If we assume a single outlier at time T, then the model of Z_t can be written as

$$Z_t = X_t + \omega I_t(T) \tag{1.2}$$

for an AO. For an IO, the model of Z_t can be written as

$$Z_t = X_t + \{ \omega\theta(B)/\phi(B) \} I_t(T) \tag{1.3}$$

Here, X_t is the outlier-free time series in the model (1.1), ω represents the magnitude of the outlier, and

$$I_t(T) = \begin{cases} 1 & \text{if } t = T \\ 0 & \text{otherwise} \end{cases}$$

is an indicator signifying the time occurrence of the outlier.

Following Fox(1972), several authors proposed different approaches to resolve outlier problems. For example, Abraham and Box (1979) proposed a Bayesian method, Denby and Martin (1979) introduced a robust estimation procedure, and Chang, Tiao, and Chen (1988) suggested an iterative maximum likelihood method. Other studies include Tsay (1986) and Muirhead (1986). One common assumption in these studies is that the underlying outlier-free model of X_t is known, or can be identified from the observed time series Z_t. However, in practice, the model of X_t is rarely known, and its identification through Z_t is often distorted due to the effect of outliers.

A test statistic for detecting outliers is developed in Section 2. Section 3 discusses the properties of the test statistic. Section 4 present an iterative modeling procedure based on a model independent detection method. The proposed procedure is then applied to an empirical example in Section 5. Finally some

concluding remarks are given in Section 6.

2. DEVELOPMENT OF A TEST STATISTIC FOR OUTLIER DETECTION

Both models (1.2) and (1.3) can be regarded as special cases of the following mean shift model

$$Z_t = X_t + \xi(t) \tag{2.1}$$

where X_t is a transformed covariance stationary process and $\xi(t)$ is a possible time dependent mean function. The additive and innovational outliers are a one-time shock which can be applied either directly to the observation or to the innovations of the process. In the AO case, the observed time series of n observations contains the set

$$\{ X_1, \cdots, X_{T-1}, X_T + \omega, X_{T+1}, \cdots, X_n \}$$

Similarly, in IO case, the observed time series contains the following set

$$\{ X_1, \cdots, X_{T-1}, X_T + \omega, X_{T+1} + \omega\varphi_1, \cdots, X_n + \omega\varphi_{n-T} \} .$$

where the φ_j are the generating weights of the process, i.e., $\psi(B) = \theta(B)/\phi(B) = 1 + \varphi_1 B + \varphi_2 B^2 + \cdots$. When one begins with a set of time series data, one has no idea about the underlying outlier-free model. Therefore a reasonable method to check a possible outlier in a series is to examine the distances between adjacent observations, i.e., $D_1(t) = Z_t - Z_{t-1}$, for t = 2, ···, n. If indeed the series contains an AO, then from (1.2) we get

$$D_1(t) = Z_t - Z_{t-1} = \begin{cases} X_t - X_{t-1} & \text{if } t \neq T, T+1 \\ X_T - X_{T-1} + \omega & \text{if } t = T \\ X_{T+1} - X_T - \omega & \text{if } t = T+1 \end{cases} \tag{2.2}$$

Likewise, if the series contains an IO, then from (1.3) we have

$$D_1(t) = Z_t - Z_{t-1} = \begin{cases} X_t - X_{t-1} & \text{if } 2 \le t \le T-1 \\ X_T - X_{T-1} + \omega & \text{if } t = T \\ X_t - X_{t-1} + \omega(\varphi_{t-T} - \varphi_{t-T-1}) & \text{if } T+1 < + t \le n \end{cases} \tag{2.3}$$

where $\varphi_0 = 1$. On the other hand, if there is no outlier in the series, we have $D_1(t) = Z_t - Z_{t-1} = X_t - X_{t-1}$ for $t = 2, 3, \cdots, n$. Equation (2.2) implies that if Z_t is contaminated by an AO at time T, then the distance measures at time T and

T+1, i.e., $D_1(T)$ and $D_1(T+1)$ will be increased or decreased by the magnitude of the outlier ω. When Z_t is contaminated by an IO at time T, the behavior of $D_1(t)$ is different. $D_1(T)$ still reflects the effect of an outlier ω, as shown in (2.3). But $D_1(t)$ for $t \geq T+1$ contains the outlier effect ω through its interaction with the φ_j weights of the generating mechanism of the underlying outlier-free process. This difference is useful in distinguishing the type of outlier.

However, there is a problem in using only $D_1(t)$ in (2.2) and (2.3); unless ($X_T - X_{T-1}$) and ω are of the same sign, the cancellation effect may result in a small value of $D_1(t)$ at time T, which will obscure the detection of outliers. This leads us to consider more generally the distance measure of observations l periods apart,

$$D_l(t) = Z_t - Z_{t-l} \tag{2.4}$$

For a series with an AO at time T, we have

$$D_l(t) = Z_t - Z_{t-l} = \begin{cases} X_t - X_{t-l} & \text{if } t \neq T, T+l \\ X_T - X_{T-l} + \omega & \text{if } t = T \\ X_{T+l} - X_T - \omega & \text{if } t = T+l \end{cases} \tag{2.5}$$

and for a series with an IO at time T, we have

$$D_l(t) = Z_t - Z_{t-l} = \begin{cases} X_t - X_{t-l} & \text{if } l+1 < +t \leq T-1 \\ X_T - X_{T-l} + \omega & \text{if } t = T \\ X_t - X_{t-l} + \omega(\varphi_{t-T} - \varphi_{t-T-l}) & \text{if } T+1 \leq t \leq n \ , \end{cases} \tag{2.6}$$

where $\varphi_{j=0}$ for $j<0$ and $\varphi_0 = 1$. The behavior of $D_l(t)$ in (2.5) and (2.6) is similar to that of $D_l(t)$ in (2.2) and (2.3). However, the simultaneous use of $D_l(t)$ for l=1,2,3 will minimize the chance of the above mentioned cancellation effect. In addition, the different behavior of $D_l(t)$ for l= 1,2,3 in the AO and IO cases is useful in distinguishing the type of outlier. Since the value of $D_l(t)$ can be either positive or negative, the absolute value of $D_l(t)$ is considered.

The other problem with the $D_l(t)$ is that they are unit dependent. Thus a standardization is needed. The expectation of $D_l(t)$ is known to be zero. So let

$$D_l'(t) = \begin{cases} D_l(t) & \text{if } |D_l(t)| \neq \max\limits_{l+1 \leq t \leq n} |D_l(t)| \\ 0 & \text{if } |D_l(t)| = \max\limits_{l+1 \leq t \leq n} |D_l(t)| \ . \end{cases} \tag{2.7}$$

Then, we consider the following standardized measure of distances:

$$D_l^*(t) = D_l(t)/S_l \ , \tag{2.8}$$

where $S_l = (\sum_{t=l+1}^{n} D_l'(t)^2/m)^{1/2}$ and $m = n-l$.

Summarizing the previous discussions, we conclude the followings:

(1) If the series contains an AO at time T, then the value of $|D_l^*(t)|$ are relatively large at t=T and t=T+l for some l = 1, 2, or 3.

(2) If the series contains an IO at time T, then the values of $|D_l^*(t)|$ for $t \geq T$ are all affected, and the maximum value of $|D_l^*(t)|$ occurs at t=T, for some l= 1, 2, or 3.

Therefore, for detecting outliers, we will use the following maximum absolute values of $D_l^*(t)$ as a test statistic:

$$M_l = \max_{l+1 \leq t \leq n} |D_l^*(t)| \ , \text{ for } \ l = 1,2,3. \tag{2.9}$$

An outlier will be declared if the value of M_l is large.

3. PROPERTIES OF THE TEST STATISTIC

To perform the test we need to know the sampling behavior of the test statistic M_l under the null hypothesis that there is no outlier. The derivation of the exact sampling distribution of M_l is difficult, but for large samples we can consider the asymptotic distribution of M_l .

We first consider the following lemma.

Lemma 1. Let W_t be a Gaussian random variables with mean zero, variance one, and $E(W_iW_j) = \rho(i,j)$ for all i,j. Let $M = \max_{1 \leq t \leq n} |W_t|$ and $\nu_n = \sup_{|i-j| \geq n} |\rho(i,j)|$ .If $\nu_1 < 1$, and one of the following two conditions holds:

(I) $\sum_n \nu_n^2 < \infty$ and

(II) $\nu_n (\log_e n)^{2+\delta} \to 0$ for some $\delta > 0$, then

$$\Pr\{ e_n^{-1}(M - b_n) \le x \} \to \exp(-e^{-x}),$$

where $-\infty < x < \infty$, $e_n = (2 \log_e n)^{-1/2}$ and

$b_n = 2(\log_e n)^{1/2} - (8 \log_e n)^{-1/2}(\log_e \log_e n + \log_e \pi)$.

Proof. see Deo(1972).

Suppose that X_t follows a model in (1.1). Under the null hypothesis of no outlier, model (2.1) becomes (1.1) and can be expressed as

$$\begin{aligned} X_t &= a_t + \varphi_1 a_{t-1} + \varphi_2 a_{t-2} + \cdots \\ &= \sum_{j=0}^{\infty} \varphi_j a_{t-j} \ , \end{aligned}$$

where $\psi(B) = \theta(B)/\phi(B) = 1 + \varphi_1 B + \varphi_2 B^2 + \cdots$ such that $\sum_{j=0}^{\infty} \varphi_j^2 < \infty$. Thus, $D_l(t)$ can be written as

$$\begin{aligned} D_l(t) &= X_t - X_{t-l} \\ &= \sum_{j=0}^{\infty} \varphi_j (a_{t-j} - a_{t-j-l}) \ . \end{aligned}$$

which has a normal distribution with

$$E[D_l(t)] = 0 \ .$$

$$Var[D_l(t)] = 2\sigma^2 \sum_{j=0}^{\infty} (\varphi_j^2 - \varphi_j \varphi_{j+l}) \ .$$

$$Cov[D_l(t), D_l(t-k)] = -\sigma^2 \sum_{j=0}^{\infty} (\varphi_j \varphi_{j+k-l} - 2\varphi_j \varphi_{j-k} + \varphi_j \varphi_{j+k+l}) \ .$$

It follows that $D_l(t)$ is a stationary zero mean Gaussian process. Now since $D_l^*(t)$ is the standardized variable of $D_l(t)$, $D_l^*(t)$ is also a Gaussian process with $E[D_l^*(t)] = 0$, $Var[D_l^*(t)] = 1$, and

$$\begin{aligned}\rho(k) &= E[D_l^*(t)D_l^*(t-k)] \\ &= -\sum_{j=0}^{\infty}(\varphi_j\varphi_{j+k-l} - 2\varphi_j\varphi_{j+k} + \varphi_j\varphi_{j+k+l})/2\sum_{j=0}^{\infty}(\varphi_j^2 - \varphi_j\varphi_{j+l}).\end{aligned} \tag{3.1}$$

We summarize our discussion in the following lemma.

Lemma 2. Let X_t be a outlier-free time series which follows an ARMA model in (1.1). Let $D_l^*(t)$ be the standardized variable of $D_l(t)$ given in (2.9). Then, $D_l^*(t)$ is a Gaussian process with mean 0, variance 1, and correlation $\rho(k)$ of (3.1).

Using lemmas 1 and 2, we have the following theorem.

Theorem 1. For a given Gaussian process $D_l^*(t)$ in Lemma 2, let $M_l = \max_{l+1\le t\le n}|D_l^*(t)|$. Then

$$\Pr\{e_m^{-1}(M_l - b_m) \le x\} \to \exp(-e^{-x}), \tag{3.2}$$

where $e_m = (2\log_e m)^{-1/2}$, $b_m = (2\log_e m)^{-1/2} - (8\log_e m)^{-1/2} + (\log_e\log_e m + \log_e \pi)$ and $m = n - l$.

Proof. Clearly $\nu_1 < 1$. We will prove this theorem by showing that the condition (I) in Lemma 1 holds for the Gaussian process $D_l^*(t)$

Let $\nu_m = \sup_{|k|\ge m}|\rho(k)|$. Note that

$$|\sum_{j=0}^{\infty}(\varphi_j\varphi_{j+k-l} - 2\varphi_j\varphi_{j+k} + \varphi_j\varphi_{j+k+l})| \le \sum_{j=0}^{\infty}|\varphi_j\varphi_{j+k-l}| + 2\sum_{j=0}^{\infty}|\varphi_j\varphi_{j+k}| + \sum_{j=0}^{\infty}|\varphi_j\varphi_{j+k+l}| \tag{3.3}$$

By Schwarz' inequality,

$$\begin{aligned}\sum_{j=0}^{\infty}|\varphi_j\varphi_{j+k-l}| &\le \{\sum_{j=0}^{\infty}|\varphi_j|^2\}^{1/2}\{\sum_{j=0}^{\infty}|\varphi_{j+k-l}|^2\}^{1/2} \\ &\le \sum_{j=0}^{\infty}|\varphi_j|^2 < \infty\,. \text{ since } \varphi_{j=0} \text{ for } j<0\end{aligned}$$

Thus, each term in (3.3) is absolutely convergent for all k, and so is the numerator of ρ(k). Now

$$|\sum_{j=0}^{\infty}(\varphi_j^2-\varphi_j\varphi_{j+l})| \le \sum_{j=0}^{\infty}|\varphi_j|^2+\sum_{j=0}^{\infty}|\varphi_j\varphi_{j+l}|<\infty$$

So, $\Sigma(\varphi_j^2-\varphi_j\varphi_{j+l})$ is absolutely convergent. Also $\Sigma\varphi_j^2>\Sigma\varphi_j\varphi_{j+l}$. As a result, $\rho(k)$ is absolutely convergent for all k and the condition (I) holds.□

Therefore, under the null hypothesis that there is no outlier, Z_t = X_t and the asymptotic distribution of M_l is given as in Theorem 1. More specifically. let V be a random variable associated with the limiting distribution of $e_m^{-1}(M_l-b_m)$. We have from (3.2)

$$E(V) = \int_{-\infty}^{\infty} v\ \exp(-v-e_{-v})dv = -\int_0^{\infty}(\log_e v)e^{-v}dv = \gamma\ (\text{Euler's Constant})$$

and

$$Var(V)=\int_{-\infty}^{\infty}v^2\exp(-v-e^{-v})dv-\gamma^2=\pi^2/6\ .$$

Hence, the corresponding asymptotic mean and variance of M_l can be expressed as

$$E(M_l)=\frac{\sqrt{2\log_e m}-(\log_e\log_e m+\log_e\pi-2\gamma)}{2\sqrt{2\log_e m}}$$

and

$$Var(M_l)=\pi^2/12\log_e m\ .$$

From the asymptotic distribution of M_l given in (3.2), we can easily calculate the upper tail significant points $C_{\alpha,l}$ such that

$$\Pr(M_l>C_{\alpha,l})=\alpha \qquad (3.4)$$

for l= 1, 2, 3 and various samples of size n.

Table 1. Critical Values $C_{\alpha,l}$ for M_l obtained from the asymptotic distribution given in (3.2)

n	α	$C_{\alpha,1}$	$C_{\alpha,2}$	$C_{\alpha,3}$
50	.05	3.4058	3.4010	3.3961
	.01	3.9901	3.9868	3.9835
100	.05	3.5710	3.5686	3.5662
	.01	4.1086	4.1069	4.1050
150	.05	3.6670	3.6654	3.6638
	.01	4.1822	4.1810	4.1798
200	.05	3.7346	3.7334	3.7322
	.01	4.2355	4.2346	4.2337
250	.05	3.7866	3.7857	3.7848
	.01	4.2773	4.2766	4.2758

4. TIME SERIES MODELING IN THE PRESENCE OF OUTLIERS

Given an observed time series $Z_1, Z_2, \cdots, Z_n$, we first consider the case of a single outlier in the series. The detection procedure consists of the following steps:

Step 1. Detect the existence of the outlier. For this, we first compute $D_1(t)$, S_1, $D_1^*(t)$, and M_1. For a given n and α, an outlier is declared if $M_1 \geq C_{\alpha,1}$. If M_1 is not significant, we compute $D_2(t)$, S_2, $D_2^*(t)$, and M_2, and then declare the existence of an outlier if $M_2 \geq C_{\alpha,2}$. If M_2 is not significant, we compute $D_3(t)$, S_3, $D_3^*(t)$, and M_3, and then declare that there is an outlier in the series if $M_3 \geq C_{\alpha,3}$. If none of the M_l for $l = 1, 2$ and 3 are significant, then we conclude that there is no outlier in the series, and stop the detection procedure.

Step 2. Find the timing of the outlier. For a given l, suppose that the maximum M_l occurs at time T. Then the contaminated observation will be either Z_T or Z_{T-l}. Compute the sample mean of the Z_t except Z_T and Z_{T-l} :

$$\bar{Z}' = \sum_{t \neq T, T-l} Z_t/(n-2)$$

Then, the timing of the outlier is T if $|Z_T - \bar{Z}'| > |Z_{T-l} - \bar{Z}'|$; otherwise, the

timing is (T−l).

Step 3. Determine the type of the outlier. For this, we first obtain a preliminary outlier adjusted series. Let T be the timing of the outlier detected in Step 2. The preliminary outlier adjusted series Z_t^* is obtained as follows:

$$Z_t^* = \begin{cases} Z_t & \text{if } t \neq T \\ Z_t - \omega_p & \text{if } t = T \end{cases} \tag{4.1}$$

where $\omega_p = \sum_{l=1}^{3} D_l(T)/3$.

The reason for the above preliminary adjustment is that if the unknown outlier is indeed an AO, then according to equation (2.5), ω_p is an estimate of ω. On the other hand, if the unknown outlier ia an IO, then according to (2.6), ω_p will also correctly adjust Z_t at t = T. Although in the latter case, the use of ω_p will leave Z_t for t > T unadjusted, it will be detected in the next iteration. Now we treate Z_t^* as an outlier-free time series and use it to identify the underlying outlier-free model using the standard identification procedure. Based on Z_t^*, suppose that the underlying outlier-free model is identified to be

$$\phi(B)X_t = \theta(B)a_t \tag{4.2}$$

where ϕ(B), θ(B), and the a_t are defined as in (1.1). Having obtained model (4.2), we note that (1.2) and (1.3) can be written as

$$Z_t = \omega\beta(B)I_t(T) + \{\theta(B)/\phi(B)\}a_t \tag{4.3}$$

where β(B) = 1 for an AO and β(B) = θ(B) / ϕ(B) for an IO. As a result, we can determine the type of the outlier by fitting model (4.3) with $\beta(B) = \theta(B)/\phi(B)$ If none of the parameters in β(B) is significant, then the outlier at time T is declared as an AO. On the other hand, if some of the parameters in β(B), as well as ω, are significant, then the outlier is declared as an IO.

Step 4. Once the type of the outlier is determined, we can specify β(B), re-estimate model (4.3), and obtain the modified outlier adjusted series. Specifically, if the outlier is an AO, we can estimate

$$Z_t = \omega_A I_t(T) + \{\theta(B)/\phi(B)\}a_t \ . \tag{4.4}$$

Let the estimate of ω be ω_A; the modified outlier adjusted series is given by

$$\hat{Z}_t = Z_t - \omega_A I_t(T) . \tag{4.5}$$

Similarly, if the outlier is an IO, we can estimate

$$Z_t = \{ \theta(B)/\phi(B) \} (\omega_I I_t(T) + a_t) \tag{4.6}$$

and obtain the modified outlier adjusted series as

$$\hat{Z}_t = Z_t - \{ \theta(B)/\phi(B) \} \omega_I I_t(T) , \tag{4.7}$$

where ω_I is the estimate of ω , $\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$ and $\theta(B) = (1 - \theta_1 B - \cdots - \theta_q B^q)$.

Cleraly, in practice, the number of outliers in the series is unknown. However, by using the above obtained modified outlier adjusted series, Z_t, as new observations, we can repeat the Step 1 through Step 4 iteratively until no more outliers are detected. More generally, if a total of k outliers are detected at times $T_1, \cdots, T_k$ after k iterations, the final modified outlier adjusted series is given by

$$\hat{Z}_t = \hat{Z}_t^{(k-1)} - \omega_k \beta_k(B) I_t(T) . \tag{4.8}$$

Here $\hat{Z}_t^{(k-1)}$ is the modified outlier adjusted series obtained at Step 5 in iteration (k-1), ω_k is the estimate of ω, $\beta_k(B) = 1$ for an AO and $\beta_k(B) = \theta(B)/\phi(B)$ for an IO. The unknown underlying outlier-free model will then be re-identified based on $\hat{Z}_t$ series using standard identification procedure, and the following model will be used to re-estimate the outlier-free model in $\theta(B)$ and $\phi(B)$. That is,

$$Z_t = \sum_{i=1}^{k} \omega_i \beta_i(B) I_t(T) + \{ \theta(B)/\phi(B) \} a_t \tag{4.9}$$

where $\beta_i(B) = 1$ for an AO and $\beta_i(B) = \theta(B)/\phi(B)$ for an IO.

5. AN EMPIRICAL EXAMPLE

We now illustrate the suggested iterative detection procedure by considering the data of annual consumption of spirits in the united Kingdom from 1870 to 1938. The model fitted by Prest(1949) for the spirits data is

$$Y_t = 2.14 + 0.69 X_{1t} - 0.63 X_{2t} - 0.00095 t - 0.00011 (t-35)^2 + Z_t , \tag{5.1}$$

where Y_t is the annual per capita consumption of spirits, X_{1t} and X_{2t} are per

capita income and price of spirits respectively, Z_t is the residual series, and t is an index of time defined by t = actual year - 1869.

Fuller (1976) found that the residual series Z_t obtained from the time series regression model (5.1) is correlated and suggested an AR(1) model for the series. These could be shown by the sample autocorrelation function (SACF) and sample partial autocorrelation function (SPACF) of Z_t given in Table 2. The residual mean square of Z_t was found as .000417. Tsay (1986) re-analyzed this data by using his iterative identification procedure and suggested an ARMA(1,1) model for Z_t with six outliers.

Table 2. SACF and ESACF of Z_t in Model (5.1)

Lag	1	2	3	4	5	6	7	8	9	10	11	12
SACF	.72	.46	.25	.15	.00	-.13	-.18	-.27	-.34	-.50	-.50	-.47
S.D.	.12	.17	.19	.19	.19	.19	.20	.20	.20	.21	.23	.24
Lag	1	2	3	4	5	6	7	8	9	10	11	12
SPACF	.72	-.13	-.06	.05	-.18	-.12	.05	-.24	-.09	-.34	.02	-.15
S.D.	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12

We will treat the residual series Z_t as an observed time series assuming that the model (5.1) is adequate for the spirits data. In addition to other possible outliers, we will also assume that an additive outlier occurs at t=20. That is, Z_{20} = - .053786 is changed to Z_{20} = .53786 due to a typing error. The contaminated series Z_t for t = 1, ···, 69 is obtained from the SAS system.

The SACF and the extended sample autocorrelation function (ESACF) for this contaminated series Z_t are calculated and given in Table 3. Clearly, the information of the SACF, shown in Table 2, is completely washed out by the additive outlier Z_{20}. The phenomenon is also true in the case of ESACF. This will cause a serious difficulty in using any detecting procedure which depends on the assumption that an outlier-free model can be identified.

Table 3. SACF anf ESACF of Z_t

Lag	1	2	3	4	5	6	7	8	9	10	11	12
SACF	-.00	.00	.00	-.02	-.04	-.06	-.04	-.03	-.03	.01	.03	.05
S.D.	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12

ESACF

p \ q	0	1	2	3	4	5	6	7	8	9
0	-.00	.00	.00	-.02	-.04	-.06	-.04	-.03	-.03	.01
1	.15	.00	.00	-.01	-.00	-.02	-.00	.01	-.03	.01
2	-.16	.01	.00	-.01	.00	-.02	-.00	.00	-.02	.01
3	.03	.06	-.24	-.01	-.00	-.02	-.01	-.00	-.02	.01
4	-.43	.09	.25	-.13	.00	-.01	.00	.01	-.01	.01
5	-.47	.07	.22	-.16	-.12	-.02	.00	.01	-.00	-.00
6	-.47	.01	-.16	-.12	-.02	.00	.01	.01	-.00	.00

Indicator Symbols

0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	x	0	0	0	0	0	0	0	0	0
5	x	0	0	0	0	0	0	0	0	0
6	x	0	0	x	0	0	0	0	0	0

However, our proposed method does not require an assumed underlying outlier-free model to find an outlier. We now use the contaminated series Z_t to illustrate the procedure.

Step 1. We compute $D_1(t)$ for t = 2, ···, 69, S_1 = .072209, and $D_1^*(t)$ for t=2, ···, 69. For n=69 and α = .05, we calculate M_1. Since M_1 = 8.3990, which occurs at time 20, is larger than $C_{.05,1}$ = 3.4823, which is obtained from equation (3.4), we declare that there is an outlier in the series.

Step 2. To find the timing of the outlier, we note that M_1 occurred at time 20. Therefore the contaminated observation will be Z_{20} or Z_{19}. We compute $\overline{Z}'$ = -.011472. Since $|Z_{20} - \overline{Z}'|$ = .549332 $> |Z_{19} - \overline{Z}'|$ = .049224, the timing of the outlier is determined to be T = 20.

Step 3. The preliminary outlier adjusted series Z_t^* is obtained as follows:

$$Z_t^* = \begin{cases} Z_t & \text{if } t \neq 20 \\ Z_{20} - \omega_p & \text{if } t = 20 \end{cases}$$

where

$$\omega_p = \sum_{l=1}^{3} D_l(20)/3 = .587377 \ .$$

We treat Z_t^* as an outlier-free time series and then identify the underlying outlier-free model using the standard identification procedure. From the SACF and SPACF of Z_t^* given in Table 4, we entertain an AR(1) model. Employing model (4.3) with $\beta(B) = 1/(1-\phi B)$, we estimate the ω and ϕ, whose estimates are .5897(.0914) and -.0082(.0332), respectively. The values in the parenthesis denote the standard error of each estimate. Since ϕ is not significant, we decide that the outlier is an AO.

Table 4. SACF and SPACF of Z_t^* at Iteration 1

Lag	1	2	3	4	5	6	7	8	9	10	11	12
SACF	.72	.46	.25	.15	.00	-.13	-.18	-.27	-.34	-.50	-.50	-.47
S.D.	.12	.17	.19	.19	.19	.19	.20	.20	.20	.21	.23	.24
Lag	1	2	3	4	5	6	7	8	9	10	11	12
SPACF	.72	-.13	-.06	.05	-.18	-.12	.05	-.24	-.09	-.34	.02	-.15
S.D.	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12

Step 4. Having found an AO at time 20, we next estimate the following model

$$Z_t = \omega_A I_t(T) + a_t/(1-\phi B)$$

from which we obtain ω_A = .5833 and the modified outlier adjusted series

$$\hat{Z}_t = \begin{cases} Z_t & \text{if } t \neq 20 \\ Z_{20} - \omega_A & \text{if } t = 20 \end{cases}$$

By using $\hat{Z}_t$ as new observations, we repeat the Steps 1 through 4 until no more outliers are detected. The detection procedure is terminated at iteration 5 since none of the M_l for l = 1,2,3 are significant. Totally, we detect four outliers at times 20, 40, 46, and 49. The detailed results are given in Table 5.

Table 5. Results of the Proposed Detecting Procedure

Iteration	l	S_l	$C_{.05,\,l}$	M_l		Outlier		
				t	Value	time	type	ωp
1	1	.0722	3.4823	20	8.3990	20	AO	.5874
2	1	.0192	3.4823	40	4.6849	40	IO	-.0845
3	1	.0164	3.4823	50	4.8297	49	AO	-.0894
4	1	.0134	3.4823	46	4.3208	46	AO	.0696

The modified outlier adjusted series obtained at Step 5 in iteration 4 is the final outlier adjusted series $\hat{Z}_t$. We now use it to re-identify the underlying outlier-free model. The SACF and SPACF of $\hat{Z}_t$ are given in Table 6.

Table 6. SACF and SPACF of Z_t

Lag	1	2	3	4	5	6	7	8	9	10	11	12
SACF	.84	.62	.41	.22	.07	-.04	-.14	-.23	-.29	-.34	-.36	-.37
S.D.	.12	.19	.21	.23	.23	.23	.23	.23	.23	.24	.25	.25
Lag	1	2	3	4	5	6	7	8	9	10	11	12
SPACF	.84	-.25	-.13	-.05	-.06	-.01	-.13	-.11	-.02	-.11	-.03	-.11
S.D.	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12

The results seem to suggest an AR(2) model. Therefore, the model for Z_t becomes

$$Z_t = \omega_1 I_t^{(20)} + \omega_2 I_t^{(40)}/(1-\delta B) - \omega_3 I_t^{(60)} + \omega_4 I_t^{(49)} + a_t/(1-\phi_1 B - \phi_2 B^2) \tag{5.2}$$

The maximum likelihood estimation results of model (5.2) using the SAS system along with the corresponding residual sample autocorrelation functions (RSACF) are shown in Table 7.

Table 7. Estimation of Results of Model (5.2)

Parameters	ω1	ω2	δ	ω3	ω4	φ1	φ2
Estimate	.5888	-.0898	.8035	.0468	-.0607	1.1934	-.3620
S.D.	.0073	.0110	.0977	.0072	.0074	.1210	.1207

Lag	1	2	3	4	5	6	7	8	9	10	11	12
RSACF	-.08	.01	.01	-.06	.02	.01	-.03	-.08	.06	-.02	.03	-.05
S.D.	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12

The residual autocorrelations indicate that the model is adequate. The residual mean square is reduced to .000137, which is about one third of that of AR(1) model suggested by Fuller(1976).

6. CONCLUDING REMARKS

We presented a procedure which does not depend on any assumed or identified outlier-free model for detecting outliers in time series. The procedure begins with an expository data analysis of time series using a simple concept of distance measures. The proposed statistics $D_t(t)$, $D_t^*(t)$ and M_t are simple and their computations are straightforward. The method can be also be applied easily to a stationary seasonal time series with period = s by simply adding a seasonal lag in computing $D_t^*(t)$. For example, we may compute $D_t^*(t)$ for =1, 2, s. The procedure works well for numerous real and simulated series which we have tried.

The real value to this procedure is that it enables at each stage an independent model-free check on a possible outlier that is conditioned on adjustment made for outliers already discovered, but does not depend on the model for detection of the next outlier.

We believe that it deserves consideration for use in time series outlier analysis, particularly when outliers may obscure or even completely wash out the information in sample identification statistics such as sample autocorrelation, sample partial autocorrelation, and extended sample autocorrelation functions.

REFERENCES

[1] Abraham, B, and Box, G.E.P. (1979),"Bayesian Analysis of Some Outlier Problems in Time Series," *Biometrika*, 66, 229-236.

[2] Chang, I., Tiao, G.C., and Chen, C. (1988),"Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, 30, 193-236.

[3] Cramer, H. (1951), *Mathematical Methods of Statistics*, Princeton University Press.

[4] Denby, L. and Martin, R.D. (1979),"Robust Estimation of the First Order Autoregressive Parameters," *Journal of the American Statistical Association*, 74, 140-146.

[5] Deo, C.M. (1972),"Some Limit Theorems for Maxima of Absolute Value of Gaussian Sequences," *Sanhkya*, Ser. A, 34, 289-292.

[6] Fox, A.J. (1972),"Outliers in Time Series," *Journal of Royal Statistical Society*, Ser. B, 43, 350-363.

[7] Fuller, W.A. (1976), *Introduction to Statistical Time Series*, New York: John Wiley.

[8] Muirhead, C.R. (1986),"Distinguishing Outlier Types in Time Series," *Journal of Royal Statistical Society*, Ser. B, 48, 39-47.

[9] Prest, A.R. (1949),"Some Experiments in Demand Analysis," *Review of Economics and Statistics*, 31, 33-49.

[10] Tsay, R.S. (1986),"Time Series Model Specification in the Presence of Outliers," *Journal of the American Statistical Association*, 88, 132-141.

정상 시계열에서의 이상치 발견과 시계열 모형구축

이종협[1], 최기헌[1]

< 요 약 >

최근에 시계열에서의 이상치 발견을 위한 여러가지 반복적인 방법들이 소개되었으나 이들 대부분은 시계열의 기저모형이 알려져 있거나 식별될 수 있다는 가정하에서 개발되었다.그렇지만 실제로 이상치들이 모형식별을 왜곡 시키거나 심지어는 불가능하게 만드는 경우가 발생한다.

본 논문에서는 두 개의 시계열 관측치 사이의 거리에 근거한 새로운 척도를 이용한 이상치 탐색 방법을 제시하였다. 특히 이방법은 이상치를 발견하는데 시계열 모형에 의존하지 않는다. 제안된 통계량에 대한 여러가지 성질을 밝혔으며 이상치의 형태를 구별하기 위해 전이함수모형을 이용하였다. 그 밖에 이상치를 포함하고 있는 시계열의 모형을 구축하기 위한 반복적인 절차를 제안했다.

1) 132-714, 서울특별시 도봉구 쌍문동 419, 덕성여자대학교 통계학과,