Carla da Fonseca Miranda

Modelação Linear de Séries Temporais na presença de Outliers

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Março/2001

Carla da Fonseca Miranda

Modelação Linear de Séries Temporais na presença de Outliers

Tese submetida à Faculdade de Ciências da Universidade do Porto
para obtenção do grau de Mestre em Estatística.

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Março/2001

RESUMO

Na análise de séries temporais, encontram-se frequentemente outliers e mudanças estruturais, que podem estar associadas a acontecimentos inesperados ou incontroláveis como por exemplo, greves, guerras, mudanças políticas, ou podem dever-se simplesmente a erros de medição ou de registo de observações.

Estas observações podem comprometer os procedimentos usuais de modelação linear de uma série temporal, nomeadamente podem induzir a uma identificação incorrecta de um modelo ARIMA e a uma estimação enviezada dos parâmetros do modelo.

O objectivo principal deste trabalho é apresentar alguns procedimentos de modelação linear de uma série temporal na presença de outliers e de mudanças estruturais. A abordagem usualmente adoptada neste tipo de procedimentos consiste na identificação da localização e dos tipos de outliers ou mudanças estruturais e na utilização de modelos de intervenção de Box e Tiao (1975) para acomodar os seus efeitos. Esta aproximação requere iterações entre etapas de detecção, utilizando estatísticas de razão de verosimilhanças para localizar e identificar os outliers e as mudanças estruturais de acordo com o seu tipo, e de estimação de um modelo gerador destas perturbações, para acomodar os seus efeitos.

Os outliers usualmente considerados são os outliers do tipo aditivo (AO) e os outliers do tipo inovador (IO) e as mudanças estruturais são as alterações de nível permanentes e transitórias (LC) e (TC).

Uma abordagem alternativa ao uso de estatísticas de razão de verosimilhanças para detectar outliers e alterações de nível, consiste na utilização de estatísticas que se baseiam na exclusão de uma ou de um grupo de observações para medir as consequentes alterações nas estimativas dos parâmetros do modelo. Esta aproximação permite detectar observações influentes que podem ser outliers. Neste sentido, também serão apresentados neste trabalho diagnósticos indicadores de observações e de outliers influentes.

ABSTRACT

Anomalies such as outliers and structure changes are quite common in time series data. These extraordinary observations may result from a gross error, for example a recording or typing error or may be associated with identifiable events such as wars, strikes, and changes in policies.

Identifying and managing these observations is necessary because their presence poses problems for the identification and estimation of autoregressive integrated moving average (ARIMA) models.

A common approach to deal with outliers and structure changes in a time series is to identify their location and type and then use intervention models discussed in Box and Tiao (1975) to accommodate their effects. This approach requires iterations between stages of outlier detection and estimation of an intervention model.

The main goal of this work is to present some of these iterative procedures that have been developed by various authors in the context of ARIMA time series model specification in the presence of outliers and structure changes when their number and dates are unknown. To identify their presence the proposed procedures usually use likelihood ratio statistics for the various types of disturbances.

The outliers considered are the additive outlier type (AO) and the innovational outlier type (IO). The structure changes allowed for are permanent level change (LC) and transient level change (TC).

Another approach to the detection of outliers and permanent level changes is based on deletion of an observation or a group of observations to measure the change that this deletion produces in the parameter estimates values. So diagnostics to measure the influence of an observation are also described in this study.

Índice

1. Introdução

Na análise de séries temporais, encontram-se frequentemente outliers e mudanças estruturais, que podem estar associadas a acontecimentos inesperados ou incontroláveis como por exemplo greves, guerras, mudanças políticas, ou podem dever-se simplesmente a erros de medição ou de registo de observações. Estas observações, são muitas vezes classificadas como anómalas, aberrantes ou discordantes e podem comprometer os procedimentos usuais de análise de uma série temporal.

Neste trabalho serão apresentados alguns procedimentos desenvolvidos específicamente para a modelação linear de séries temporais na presença de outliers. Estes procedimentos são de natureza iterativa e são essencialmente executados em dois passos:

- Identificação da localização e dos tipos dos outliers;
- Ajustamento dos efeitos dos outliers identificados com o objectivo de identificação do modelo ou de estimação dos parâmetros.

Uma contribuição fundamental para a diferenciação dos outliers que ocorrem em séries temporais deve-se a Fox (1972), que distinguiu os outliers correspondentes a erros grosseiros de execução ou de registo, isolados e independentes das outras observações, dos outliers do tipo "inerente", que são as observações anómalas que têm influência sobre as observações sucessivas. Introduziu assim os conceitos de outlier do tipo I e do tipo II, também denominados por outliers aditivos (AO) e outliers inovadores (IO). No capítulo 2 deste trabalho serão apresentados os modelos geralmente utilizados para descrever uma série temporal com cada um destes tipos de outliers e também com uma alteração permanente de nível (LC), ou com uma alteração de nível temporária (TC).

Uma abordagem ao problema de detecção de outliers e de alterações de nível, usualmente utilizada nos diversos procedimentos iterativos de tratamento de outliers, consiste na utilização de estatísticas de razão de verosimilhanças para testar a existência de cada tipo de outlier considerado. Esta abordagem permite identificar a localização e o tipo do outlier que é detectado em cada iteração, por ordem decrescente de magnitude. Estas estatísticas serão apresentadas na secção 3.1 (detecção de outliers) no capítulo 3 deste trabalho.

Nas secções 3.2 e 3.3 serão apresentados dois métodos de detecção alternativos ao testes de razão de verosimilhanças. Na secção 3.2 , expoêm-se os fundamentos e a formulação de dois diagnósticos diferentes, baseados na exclusão de observações e que foram propostos por Bruce e Martin (1989), para medir a influência de uma observação, ou de um grupo de observações, sobre as estimativas dos parâmetros do modelo. Nomeadamente, o diagnóstico para os coeficientes, DC, mede as alterações sobre as estimativas dos coeficientes estimados e o diagnóstico para a variancia, DV, mede as alterações sobre a variância estimada do ruído. Os autores concluíram, contudo, através de um estudo de simulação, que o diagnóstico para os coeficientes não permite detectar com exactidão uma observação outlier isolada pois apresenta uma magnitude muito similar em pontos de tempo adjacentes ao da ocorrência do outlier. Por isso, o diagnóstico para a variância dos ruídos, é o utilizado no procedimento de detecção. Penã (1990) considerou o problema de detecção de outliers influentes e obteve duas medidas de influência diferentes para detectar os outliers aditivos e os outliers inovadores que alteram significativamente as estimativas dos parâmetros do modelo. Apenas será descrito em 3.3.1, o processo de construção da estatística $D_2(T)$, diagnóstico para outliers aditivos, pois o autor concluiu que a estatística para detectar outliers inovadores influentes, não os localiza claramente, o que é compreensível, pois os modelos são muito mais robustos na presença de outliers inovadores do que na de aditivos. Para detectar um LC, ou uma sequência de outliers aditivos consecutivos de impacto similar, Penã e Sánchez, (1997), desenvolveram uma medida de infuência para um LC, $DL(T)$, que será apresentada em 3.3.2.

Os vários procedimentos de tratamento de outliers desenvolvidos, requerem iterações entre etapas de detecção de outliers, utilizando as estatísticas de razão de verosimilhanças para os localizar e identificar, e de estimação de um modelo gerador dos outliers, para acomodar os seus efeitos. Esta aproximação foi adoptada por diversos autores, por exemplo Tsay (1986, 88), Chen e Liu (1993) e Penã (1997) e os procedimentos por eles desenvolvidos de acordo com objectivos específicos, serão descritos no capítulo 4. Em 4.1, descrever-se-á o procedimento de

especificação de um modelo na presença de outliers aditivos e inovadores, proposto por Tsay (1986) e na secção 4.2 uma generalização deste procedimento, também desenvolvida por Tsay em 1998, para tratar de um modo unificado os outliers e as alterações de nível.

Na secção 4.3 será exposto um procedimento iterativo de estimação conjunta dos efeitos dos outliers e dos parâmetros do modelo, desenvolvido por Chen e Liu (1993), para reduzir a influência dos outliers sobre as estimativas dos parâmetros e em 4.4 descrever-se-á um procedimento iterativo de detecção de múltiplos outliers, de Penã (1997).

2. Modelos para Séries Temporais com Outliers

Uma contribuição fundamental para o estudo de outliers em séries temporais deve-se a Fox (1972), que distinguiu a possibilidade de ocorrência de dois tipos diferentes de outliers, nomeadamente outliers do tipo I e do tipo II, numa série temporal discreta $\{Y_t\}$ gerada por um modelo autoregressivo de ordem p conhecida,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t$$

onde ϕ_i são os coeficientes autoregressivos e $\{a_t\}$é uma sequência de variáveis aleatórias independentes $N(0, \sigma_a{}^2)$.

Para descrever as perturbações da série, consequentes da presença de cada um dos tipos de outliers, num instante t=T, Fox propôs dois modelos paramétricos distintos:

$$Z_t = Y_t + wI_t^T \qquad (2.1)$$

é o modelo que caracteriza a presença de um outlier do tipo I e,

$$Z_t = \sum_{i=1}^{p} \phi_i Z_{t-i} + wI_t^T + a_t \qquad (2.2)$$

é o modelo para um outlier do tipo II.

Para ambos os modelos, I_t^T é uma variável indicadora do instante em que ocorre o outlier, e por isso, $I_t^T = 0$ para $t \neq T$ e $I_t^T = 1$ para $t = T$, w representa a magnitude do efeito do outlier e $\{Z_t\}$é a série observada.

Um outlier do tipo I caracteriza-se por ser a única observação afectada, numa série onde ocorre um erro grosseiro, normalmente de execução ou de registo. Um outlier do tipo II caracteriza-se por ter um efeito que se manifesta não só na observação Z_T mas também nas observações posteriores, $Z_{T+1},...,Z_n$.

A representação paramétrica introduzida por Fox foi adoptada por diversos autores para modelizar a presença de observações perturbadoras em séries temporais, cujo modelo subjacente é um ARMA estacionário ou não estacionário.

Genéricamente, as observaçoes perturbadoras consideradas são os outliers de cada um dos tipos, usualmente denominados por outliers aditivos (AO) e inovadores (IO).

Chang, Tiao e Chen (1998) mostraram que ambos os modelos podem ser considerados como casos particulares do modelo de análise de intervenção de Box e Tiao (1975). Adoptando esta aproximação, alguns autores por exemplo Tsay, (1988), Chen e Liu, (1993) e Balke, (1993), também consideram a modelização de alterações de nível permanentes e transitórias (LC) e (TC) respectivamente.

Seja $\{Y_t\}$ uma série temporal sem outliers gerada por um modelo ARIMA dado por:

$$\phi(B)\nabla^d Y_t = \theta(B)a_t \qquad (2.3)$$

onde, n é o número de observações, B é o operador atraso, definido por $B^i Y_t = Y_{t-i}$, $\phi(B) = 1 - \phi_1 B - .... - \phi_p B^p$ e $\theta(B) = 1 - \theta_1 B - .... - \theta_q B^q$ são polinómios em B de ordens p e q respectivamente, sem factores comuns e com as raízes fora do circulo unitário, d um inteiro não negativo, $\nabla = (1 - B)$, $\nabla^d Y_t$ é um processo estacionário, e $\{a_t\}$é uma sequência de variáveis aleatórias normais e independentes com média zero e variância σ_a^2.

Assumindo que a série $\{Y_t\}$ sofre uma perturbação com origem num período ou instante, t = T, Chang e Tiao (1983), Tsay (1986), Chang, Tiao e Chen (1988), entre outros, utilizaram os seguintes modelos para caracterizar a série observada $\{Z_t\}$:

Modelo para um outlier aditivo (AO)

$$Z_t = Y_t + w_{AO} I_t^T \qquad (2.4)$$

onde I_t^T é uma variável indicadora do período T onde ocorre o outlier, como foi referido em (2.2), e portanto $I_t^T = 1$ se $t = T$ e $I_t^T = 0$ se $t \neq T$, e w_{AO} é uma constante que representa a magnitude ou impacto inicial do efeito do outlier aditivo.
Este outlier tem um efeito imediato e localizado, isto é Y_T é a única observação afectada.

Modelo para um outlier inovador (IO)

$$Z_t = Y_t + w_{IO} \frac{1}{\pi(B)} I_t^T \qquad (2.5)$$

onde, w_{IO} representa o impacto inicial do efeito do outlier inovador e $\pi(B) = (1 - \pi_1 B - \pi_2 B^2 - ....) = \theta^{-1}(B)\nabla^d \phi(B)$.

Este outlier ocorre na série das inovações, $\{a_t\}$, em $t = T$ e por isso o seu efeito propaga-se através de $\psi(B) = \frac{\theta(B)}{\nabla^d \phi(B)} = \left(\psi_0 + \psi_1 B + \psi_2 B^2 + ...\right)$, com $\psi_{0=1}$, afectando $\{Y_t\}$ para todo o $t \geq T$.

Como foi indicado por Chang, Tiao e Chen (1988), as perturbações causadas por ambos os tipos de outliers podem ser modeladas como casos particulares do modelo de análise de intervenção de Box e Tiao (1975):

$$Z_t = Y_t + w_i v_i(B) I_t^T \qquad (2.6)$$

Onde $i = AO, IO$ e w_i e $v_i(B)$ representam o impacto inicial e a estrutura dinâmica do efeito do outlier, respectivamente.

Especificando a estrutura $v_{AO}(B) = 1$ obtém-se o modelo para um AO, dado por (2.4) e tomando $v_{IO}(B) = \frac{1}{\pi(B)}$ obtém-se o modelo para um IO, dado por (2.5).

O modelo paramétrico (2.6) foi adoptado e generalizado por Tsay (1988), Chen e Liu (1993) que consideraram também a modelização de alterações de nível permanentes, LC, e transitórias, TC, numa série temporal. Então se $i = AO, IO, LC, TC$ e impondo estruturas específicas, também a $v_{LC}(B)$ e $v_{TC}(B)$, nomeadamente $v_{LC}(B) = \frac{1}{(1-B)}$ e $v_{TC}(B) = \frac{1}{(1-\delta B)}$ com $0 < \delta < 1$, obtêm-se os seguintes modelos:

Modelo para uma alteração de nível permante (LC)

$$Z_t = Y_t + \frac{w_{LC} I_t^T}{1-B} \qquad (2.7)$$

Modelo para uma alteração de nível temporária (TC)

$$Z_t = Y_t + \frac{w_{TC} I_t^T}{1-\delta B} \qquad (2.8)$$

Então, uma alteração de nível permanente tem um efeito abrupto e constante sobre a série $\{Y_t\}$, pois $Z_t = Y_t$ para t < T, mas $Z_t = Y_t + w_{LC}$ para t ≥ T. Uma alteração de nível temporária afecta a série, inicialmente em $t = T$, e o seu efeito desaparece gradualmente para $t \geq T$.

Em situações práticas de análise de uma série temporal, é usual que esta apresente múltiplos outliers de diversos tipos. Para descrever a série observada, $\{Z_t\}$, pode-se utilizar o seguinte modelo:

Modelo para múltiplos outliers

$$Z_t = \sum_{j=1}^{m} w_j v_j(B) I_t^{T_j} + Y_t \qquad (2.9)$$

onde m é o número de outliers, T_j é o período de ocorrência de um outlier,

$I_t^{T_j} = \begin{cases} 1 \text{ se } t = T_j \\ 0 \text{ se } t \neq T_j \end{cases}$ e $v_j(B) = 1$ para um AO, $v_j(B) = \frac{1}{\pi(B)}$ para um IO,

$v_j(B) = \frac{1}{(1-B)}$ para um LC e $v_j(B) = \frac{1}{(1-\delta B)}$ para um TC.

3. Detecção de outliers

Uma abordagem ao problema de detecção de outliers, em dados de séries temporais, consiste na utilização de estatísticas de razão de verosimilhanças para testar a existência de um outlier de acordo com o seu tipo. Este é um método utilizado em diversos procedimentos iterativos de tratamento de outliers, pois permite detectá-los um a um, por ordem decrescente de magnitude. No entanto, em séries temporais os efeitos de um outlier não podem ser avaliados isoladamente, e um aspecto importante de qualquer procedimento de detecção é localizar sequências de outliers que ocorrem em pontos adjacentes e que isoladamente poderão não ser detectáveis. Por esta razão, Bruce e Martin (1989), definiram dois diagnósticos, DC e DV, que se baseiam na exclusão de grupos de observações consecutivas e que indicam, respectivamente, as observações que afectam significativamente as estimativas dos coeficientes do modelo e as estimativas da variância dos ruídos. Partilhando esta abordagem alternativa de detecção de observações influentes, Penã (1990), também utilizou a exclusão de observações para apresentar uma estatística que permite detectar outliers aditivos influentes, e em 1997 apresentou outra estatística para detectar sequências influentes de outliers aditivos.Em 3.1, 3.2 e 3.3, serão apresentados os processos de obtenção das estatisticas de razão de verosimilhanças, dos diagnósticos de Bruce e Martin e das medidas de influência de Penã (1990,1997), respectivamente.

3.1 Estatísticas de razão de verosimilhanças para detectar e identificar outliers

Fox (1972) propôs o uso de critérios de razão de verosimilhanças para detectar um outlier do tipo I ou um outlier do tipo II numa série temporal gerada por um modelo autoregressivo puro de ordem p conhecida.Chang (1982) estendeu os resultados de Fox a modelos ARIMA e propôs um procedimento iterativo para detectar e identificar outliers, aditivos e inovadores, quando o seu número e localização são desconhecidos. Este método de detecção foi adoptado por Chang e Tiao (1983), Tsay (1986,88), entre outros e constitui uma etapa fundamental em procedimentos iterativos de tratamento de dados de séries temporais com outliers. Tsay (1988) generalizou a utilização das estatísticas de teste, para detectar também alterações de nível temporárias e transitórias e Chen e Liu (1993) incorporaram o procedimento de Tsay no procedimento desenvolvido, para estimação conjunta dos efeitos dos outliers e dos parâmetros do modelo.

Em 3.1.1 descreve-se o processo de construção das estatísticas de teste da razão de verosimilhanças e em 3.1.2 a etapa de detecção e de identificação de outliers.

3.1.1 Construção das estatísticas de teste de existência de outliers

Suponha-se que ocorre um outlier num período de tempo t=T, conhecido. Então a série $\{Z_t\}$que se observa, pode ser descrita por um dos modelos do tipo de (2.6), de acordo com o tipo do outlier.

A utilização destes modelos obriga à estimação do impacto do outlier.

Considere-se então a expressão geral dos resíduos,

$$e_t = \pi(B)Z_t \qquad (3.1)$$

que se obtém aplicando um filtro, $\pi(B)=\left(1-\pi_1 B-\pi_2 B^2-....\right)=\theta^{-1}(B)\nabla^d\phi(B)$, à série observada e considerando que os parâmetros do modelo são conhecidos.

Recorrendo novamente ao modelo (2.6), especificam-se as seguintes expressões para os resíduos:

$$e_t = w_{AO}\pi(B)I_t^T + a_t, \text{ para um AO,} \qquad (3.2)$$

$$e_t = w_{IO}I_t^T + a_t, \text{ para um IO,} \qquad (3.3)$$

$$e_t = w_{LC}\{\pi(B)/(1-B)\}I_t^T + a_t, \text{ para um LC,} \qquad (3.4)$$

e

$$e_t = w_{TC}\{\pi(B)/(1-\delta B)\}I_t^T + a_t, \text{ para um TC.} \qquad (3.5)$$

As expressões anteriores permitem escrever os estimadores dos mínimos quadrados do impacto w_{AO}, w_{IO}, w_{LC} e w_{TC} de cada um dos tipos de outliers:

$$\tilde{w}_{AO} = \rho_A^2\pi(F)e_T \qquad (3.6)$$

onde $\rho_A^2 = \left(1+\pi_1^2+\pi_2^2+\ldots+\pi_{n-T}^2\right)^{-1}$, $\pi(F) = \left(1-\pi_1 F-\ldots-\pi_{n-T}F^{n-T}\right)$ e $Fe_T = e_{T+1}$

$$\tilde{w}_{IO} = e_T \qquad (3.7)$$

$$\tilde{w}_{LC} = \rho_L^2\left(e_T - \sum_{i=1}^{n-T}\eta_i e_{T+i}\right) \qquad (3.8)$$

onde η_i é o coeficiente de B^i no polinómio $\eta(B) = (\eta_0 - \eta_1 B - \ldots) = \pi(B)/(1-B)$ e $\rho_L^2 = \left(1+\eta_1^2+\eta_2^2+\ldots+\eta_{n-T}^2\right)^{-1}$

$$\tilde{w}_{TC} = \rho_{TC}^2\left(e_T - \sum_{i=1}^{n-T}\beta_i e_{T+i}\right) \qquad (3.9)$$

onde β_i é o coeficiente de B^i no polinómio $\beta(B) = (\beta_0 - \beta_1 B - \ldots.) = \pi(B)/(1-\delta B)$ e $\rho_{TC}^2 = \left(1+\beta_1^2+\ldots.+\beta_{n-T}^2\right)^{-1}$.

A variância de cada um destes estimadores é:

$$Var(\tilde{w}_{AO}) = \rho_A^2 \sigma_a^2 \tag{3.10}$$

$$Var(\tilde{w}_{IO}) = \sigma_a^2 \tag{3.11}$$

$$Var(\tilde{w}_{LC}) = \rho_L^2 \sigma_a^2 \tag{3.12}$$

$$Var(\tilde{w}_{TC}) = \rho_{TC}^2 \sigma_a^2 \tag{3.13}$$

Estes resultados também são frequentemente obtidos considerando a expressão geral dos resíduos (3.1), definindo $x_t = \pi(B)v_i(B)I_t^T$ e reescrevendo o modelo paramétrico geral dado por (2.6), do seguinte modo:

$$e_t = w_i x_t + a_t \tag{3.14}$$

Então, através desta equação de regressão linear simples, e assumindo como anteriormente, que os parâmetros do modelo são conhecidos, as expressões para o estimador dos mínimos quadrados de w_i e para a variância deste estimador, são respectivamente:

$$\tilde{w}_i = \frac{\sum_{t=1}^{n} e_t x_t}{\sum_{t=1}^{n} x_t^2} \tag{3.15}$$

e

$$\text{var}(\tilde{w}_i) = \frac{\sigma_a^2}{\sum_{t=1}^{n} x_t^2} \tag{3.16}$$

Substituindo $v_i(B)$ em x_t de acordo com os diferentes tipos de outliers, estas expressões conduzem aos resultados (3.6), (3.7), (3.8) e (3.9).

Chen e Liu (1993) referem que é importante notar que quando o outlier coincide com a última observação da série, isto é quando T = n, $\widetilde{w}_{AO} = \widetilde{w}_{IO} = \widetilde{w}_{LS} = \widetilde{w}_{TC} = e_n$. Como resultado torna-se impossível distinguir empiricamente o tipo de outlier, que ocorre no fim de uma série temporal.

Para testar a existência de cada um dos tipos de outliers, em t = T, as hipóteses usualmente consideradas são as seguintes:

i) $H_0 : w_{AO} = w_{IO} = w_{LS} = w_{TC} = 0$ (não existem outliers)

ii) $H_A : w_{AO} \neq 0$

iii) $H_I : w_{IO} \neq 0$

iv) $H_{LC} : w_{LC} \neq 0$

v) $H_{TC} : w_{TC \neq 0}$

As estatísticas da máxima verosimilhança para testar H_0 versus H_A, H_I, H_{LC} e H_{TC}, são respectivamente:

$$\lambda_{A,T} = \frac{\widetilde{w}_{AO}}{\rho_A \sigma_a},$$

$$\lambda_{I,T} = \frac{\widetilde{w}_{IO}}{\sigma_a},$$

$$\lambda_{LC,T} = \frac{\widetilde{w}_{LS}}{\rho_L \sigma_a}$$

e

$$\lambda_{TC,T} = \frac{\widetilde{w}_{TC}}{\rho_{TC} \sigma_a}$$

Em situações práticas os parâmetros do modelo são inicialmente estimados pressupondo que não existem outliers na série temporal observada. Nestas situações os parâmetros das estatísticas de teste, são substituídos por estimativas consistentes. Sob a hipótese nula de

inexistência de outliers, as estimativas da máxima verosimilhança são consistentes e mais especificamente, as estatísticas $\lambda_{i,T}$, $(i=AO,IO,LC,TC)$, têm uma distribuição assimptótica N(0,1).

Para iniciar procedimento de detecção, utilizam-se então as estatíticas $\hat{\lambda}_{A,T}$, $\hat{\lambda}_{I,T}$, $\hat{\lambda}_{LC,T}$ e $\hat{\lambda}_{TC,T}$. Estas estatísticas são assimptóticamente equivalentes a $\lambda_{A,T}$, $\lambda_{I,T}$, $\lambda_{LC,T}$, $\lambda_{TC,T}$ respectivamente.

3.1.2 Etapa de detecção e identificação de outliers

A etapa de detecção incorporada em diversos procedimentos iterativos de tratamento de dados com outliers, pode ser descrita através dos seguintes passos:

1. Utiliza-se o modelo estimado para a série observada $\{Z_t\}$, ou para a série ajustada $\{Z_t^*\}$, e calculam-se os resíduos e a variância residual.A variância residual é utilizada como uma estimativa de σ_a^2.

2. Calculam-se as estatísticas de teste para todos os períodos de tempo *t*, pois o tempo T onde ocorre o outlier é desconhecido, e localiza-se o máximo em valor absoluto de cada uma. Por exemplo,

$$\hat{\lambda}_{A,\max} = Max\left\{\left|\hat{\lambda}_{A,t}\right| \quad 1 \le t \le n\right\}$$

e seja T_A o período de tempo correspondente a $\hat{\lambda}_{A,\max}$. Do mesmo modo determinam-se T_I e $\hat{\lambda}_{I,\max}$, T_{LC} e $\hat{\lambda}_{LC,\max}$, T_{TC} e $\hat{\lambda}_{TC,\max}$.

3. Determina-se $\hat{\lambda} = Max\left\{\hat{\lambda}_{A,\max}, \hat{\lambda}_{I,\max}, \hat{\lambda}_{LC,\max}, \hat{\lambda}_{TC,\max}\right\}$ e compara-se este valor com um valor crítico C. Se $\hat{\lambda} < C$, não se detecta nenhum outlier. Se $\hat{\lambda} \geq C$ detecta-se um outlier. Neste caso, ficam determinados o tipo, o período de ocorrência e o impacto $\hat{w}_i$ do outlier identificado.

O valor crítico C, assume usualmente os valores 3.0, 3.5 e 4.0. Estes valores foram seleccionados por diversos autores, tendo por base alguns resultados de simulações efectuadas por Chang (1982).

3.2 Diagnósticos DC e DV

Bruce e Martin (1989) consideraram o ajustamento de um modelo ARIMA(p,d,q) a uma série temporal e definiram dois diagnósticos diferentes para medir a influência de uma ou várias observações consecutivas, sobre a estimação dos parâmetros do modelo. A formulação de ambos os diagnósticos tem por base a exclusão de observações, e a medição das consequentes alterações nas estimativas dos parâmetros. Deste modo, o diagnóstico para os coeficientes, DC, serviria para detectar os outliers que alteram significativamente as estimativas dos coeficientes do modelo e o diagnóstico para a variância dos ruídos, DV, os outliers que alteram a variância estimada dos ruídos.

A construção de cada um dos diagnósticos será descrita em 3.2.1 e em 3.2.2, respectivamente. Os autores concluiram contudo, através de um estudo de simulação, que o diagnóstico para os coeficientes não permite detectar com exactidão uma observação outlier isolada pois apresenta uma magnitude muito similar em pontos de tempo adjacentes ao da ocorrência do outlier. Por isso, o diagnóstico para a variância dos ruídos, é o utilizado no procedimento de detecção de outliers isolados e de grupos de outliers que ocorrem em períodos de tempo adjacentes. Em 3.2.3 serão expostos os exemplos utilizados para concluir a superioridade de DV sobre DC na detecção de outliers isolados e em 3.2.4 uma estratégia para determinar o número de observações adjacentes que constitui um grupo influente presente numa série temporal. Esta estratégia resulta da análise do comportamento dos valores do diagnóstico DV, baseado na exclusão de k observações adjacentes.

Os problemas de "masking" surgem frequentemente na detecção de outliers que ocorrem num grupo influente, mas também na detecção de outliers com impacto moderado e que ocorrem em períodos de tempo relativamente próximos. Para evitar este tipo de problemas, os autores desenvolveram um procedimento de eliminação iterativa de grupos de observações. Em 3.2.5 será exposto o procedimento e um exemplo ilustrativo.

3.2.1 Diagnóstico para os coeficientes:DC

Seja $\hat{\alpha}$ a estimativa da máxima verosimilhança de $\alpha = (\phi_1,...,\phi_p,\theta_1,...\theta_q)^T$ e considere-se que $A = \{t_1,t_2,...,t_k\}$ é um subconjunto de k períodos de tempo $t_i \in \{1,2,...,n\}$ adjacentes e centrados em t e $t > [(k-1)/2]$ e $t \leq n-[k/2]$.

Utilizando a aproximação de Harvey e Pierse (1984), quando existem observações desconhecidas numa série temporal é possível obter estimativas da máxima verosimilhança dessas observações desconhecidas e após a sua substituição, obter as estimativas da máxima verosimilhança para os coeficientes do modelo.

Deste modo, seja $\hat{\alpha}_A$ a estimativa da máxima verosimilhança de α quando se supõe que as observações $Z_{t_1}, Z_{t_2},...,Z_{t_k}$ são desconhecidas. Se alguma das observações que ocorre em A tiver uma influência excessiva na estimativa de α, então isso revelar-se-á por uma diferença significativa entre $\hat{\alpha}$ e $\hat{\alpha}_A$. Os autores definem a influência empírica sobre os coeficientes por:

$$EIC(A) = -n(\hat{\alpha}_A - \hat{\alpha}) \qquad (3.17)$$

Utilizando uma forma quadrática da influência empírica obtém-se o seguinte diagnóstico:

$$DC(A) = \frac{1}{n} EIC^T(A)\hat{C}^{-1}EIC(A)$$

onde $\hat{C}$ é uma estimativa da matriz de covariância C de $\hat{\alpha}$ que é obtida através das seguintes considerações:

Sob condições de regularidade de que $\hat{\alpha}$ é assimptoticamente normal, então $(\hat{\alpha}-\alpha)\sqrt{n} \rightarrow N_r(0,C(\alpha))$, Fuller (1976), e a matriz de covariância assimptótica $C(\alpha)$ está relacionada com a matriz de informação assimptótica $I(\alpha)$, por $C(\alpha)^{-1} = I(\alpha)$.

Se $\hat{I}(\alpha)$ é um estimador consistente de $I(\alpha)$, então o teorema de Mann-Wald implica que:

$$n(\hat{\alpha}-\alpha)^T \hat{I}(\alpha)(\hat{\alpha}-\alpha) \rightarrow \chi^2_r$$

onde $r = p+q$.

Escolhe-se $\hat{I}(\alpha)$ para $\hat{C}^{-1}$ e a informação esperada $I(\hat{\alpha})$ calculada no estimador da máxima verosimilhança, é um estimador para $I(\alpha)$. Apesar de não ser muito comum na literatura, existe uma expressão para $I(\alpha)$ em termos de α, e utilizando essa expressão os autores definem o diagnóstico para os coeficientes, baseado na exclusão de k observações de A, do seguinte modo:

$$\begin{aligned} DC(A) &= \frac{1}{n} EIC^T(A) I(\hat{\alpha}) EIC(A) \\ &= n(\hat{\alpha} - \hat{\alpha}_A)^T I(\hat{\alpha})(\hat{\alpha} - \hat{\alpha}_A) \end{aligned} \qquad (3.18)$$

A distribuição de DC(A) não é conhecida, mas a utilização da distribuição χ^2_r, permite vizualizar DC(A) numa escala familiar. Um guia grosseiro para se julgar um subconjunto A de pontos, como sendo influente é verificar se o valor probabilístico (valor p) de DC(A), baseado na distribuição de referência χ^2_r é menor que 0.5.

3.2.2 Diagnóstico para a variância dos ruídos: DV

A influência de um subconjunto A de pontos tempo adjacentes, também pode ser medida avaliando as alterações nas estimativas da variância dos ruídos, $\hat{\sigma}^2$, consequentes da remoção das observações correspondentes. Para isso os autores definiram a função de influência empírica sobre a variância dos ruídos, da seguinte forma:

$$EIV(A) = -n\left(\hat{\sigma}^2_A - \hat{\sigma}^2\right) \qquad (3.19)$$

onde $\hat{\sigma}^2_A$ é o estimador da máxima verosimilhança de σ^2 tratando as observações nos tempos $t_i \in A$ como sendo desconhecidas e utilizando a aproximação de Harvey e Pierce (1984).

O diagnóstico de exclusão para a variância dos ruídos é obtido através de uma teoria assimptótica similar à utilizada na construção do DC(A) e é definido do seguinte modo:

$$DV(A) = \frac{n}{2}\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_A^2} - 1\right)^2 \qquad (3.20)$$

Sendo a distribuição de referência uma qui-quadrado com um grau de liberdade (χ_1^2). Do mesmo modo, suspeita-se que o subconjunto A, é influente, quando o valor p para DV(A) é menor do que 0.5 utilizando uma distribuição χ_1^2.

3.2.3 Superioridade de DV sobre DC

Quando a influência de um subconjunto de pontos, se deve à ocorrência de um outlier num desses pontos, o diagnóstico DV permite detectar exactamente a localização do outlier, enquanto que o diagnóstico DC apresenta valores similares em todos os pontos influentes, devido à propagação do efeito do outlier a observações adjacentes. Para ilustrar esta conclusão, os autores recorreram a cada um dos modelos (2.4) e (2.5) para simularem um AR(1) com ruído branco Gaussiano, $\phi = 0.4$, $\sigma^2 = 1$ e com um outlier aditivo de impacto +4 em T=28, no 1° caso e com um outlier inovador, também em T=28 e com o mesmo impacto no 2° caso. Ambas as séries têm 100 observações e cada um dos diagnósticos, de exclusão de uma observação, foi calculado excluindo a observação correspondente a cada ponto t, para t = 1,2, ... n e reestimando os parâmetros do modelo. Deste modo, DC(t) e DV(t) representam a influência da observação, que ocorre no ponto t. A representação dos valores dos diagnósticos e as conclusões são apresentadas para cada um dos casos nos exemplos 3.1 e 3.2.

Exemplo 3.1: O ajustamento da máxima verosimilhança de um modelo AR(1), com $\phi = 0.4$, $\sigma^2 = 1$ considerando todas as observações, produz $\hat{\phi} = 0.17$ e $\hat{\sigma}^2 = 1.09$. O cronograma está representado na Fig. 3.1 e o outlier aditivo está assinalado.

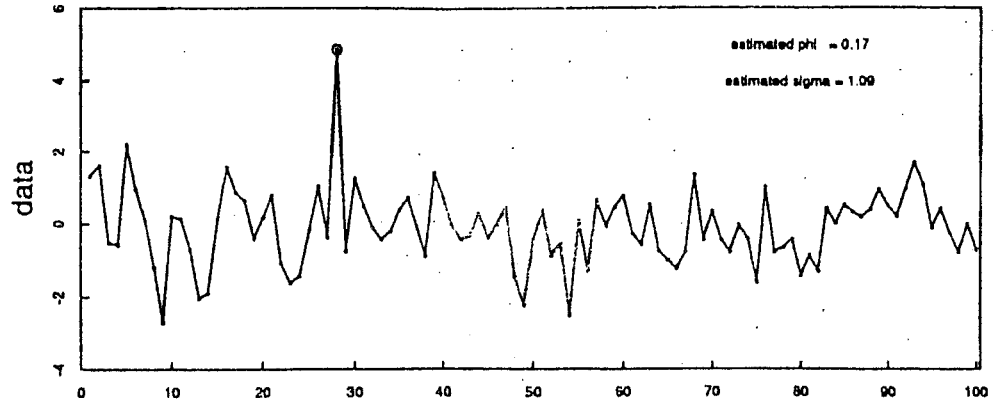

Fig 3.1- modelo AR(1) simulado com um AO isolado

Os diagnósticos DC(•) para $\hat{\phi}$ e DV(•) para $\hat{\sigma}^2$ estão representados nas Figs 3.2 e 3.3 respectivamente, e os valores p associados que correspondem a uma distribuição χ_1^2 estão marcados no eixo direito.

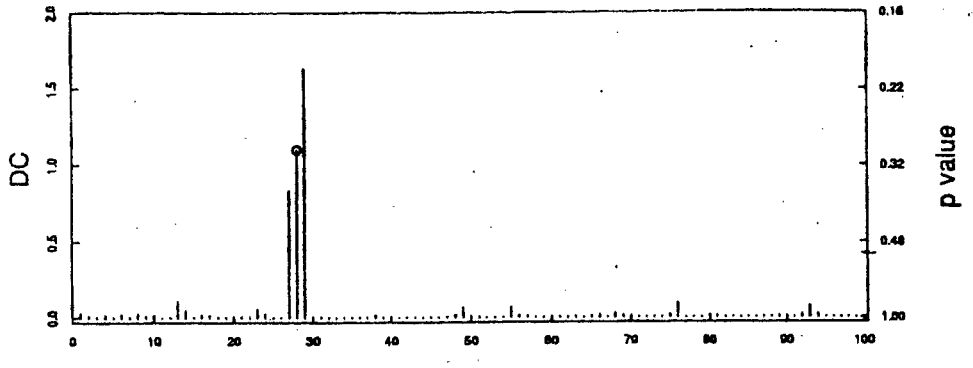

Fig 3.2- diagnóstico para os coeficientes

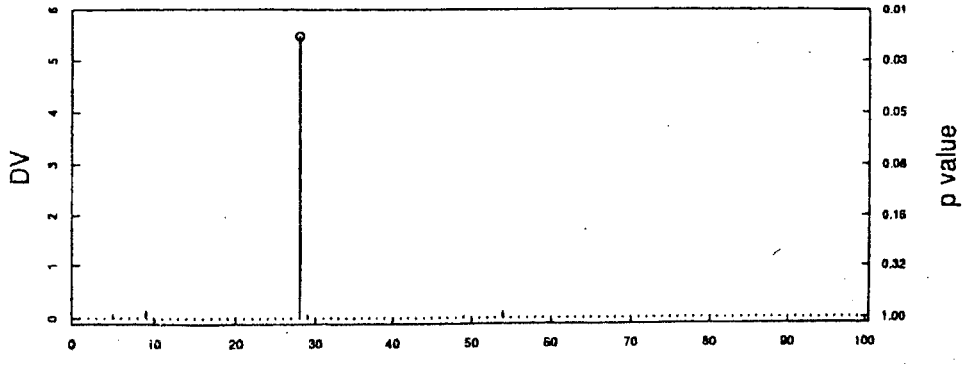

Fig 3.3- diagnóstico para a variância dos ruídos

A Fig.3.2 permite observar a propagação do efeito do outlier pois vários valores do diagnóstico DC(•) são significativos. Os valores p para DC(t) em t=27,28 e 29 são todos menores do que 0.5, indicando portanto que as observações Z_{27}, Z_{28} e Z_{29} são outliers. Pelo contrário, de entre todos os valores de DV(t) na Fig 3.3, apenas DV(28) é significativo indicando exactamente qual é a observação influente.

Exemplo 3.2: Os estimadores da máxima verosimilhança dos parâmetros são $\hat{\phi} = 0.27$ e $\hat{\sigma}^2 = 1.06$ e os dados estão representados na Fig. 3.4.

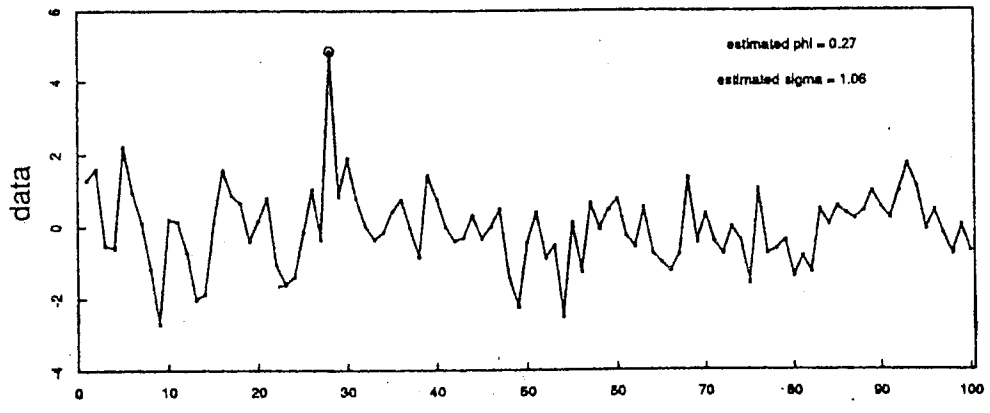

Fig 3.4- modelo AR(1) simulado com um IO isolado

Os diagnósticos de exclusão de uma observação, DC(•) e DV(•), representados nas Figs 3.5 e 3.6, permitem observar que a propagação do efeito do outlier sobre os valores de DC(t) é ainda mais problemática do que a do exemplo anterior, pois o valor do diagnóstico só é significativo em t=27 e o valor p em t=28, tempo onde ocorre o outlier, é insignificante. O diagnóstico DV(•) permite identificar exactamente o outlier em t=28 pois $DV(28)$ tem uma magnitude muito maior do que as de $DV(t)$ para qualquer outro t e o valor p correspondente é muito significativo.

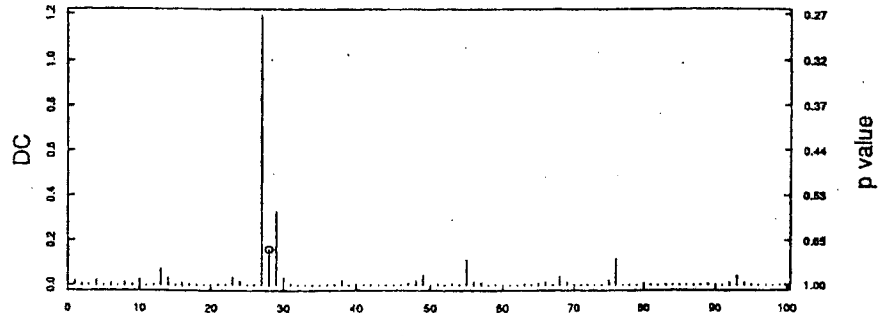

Fig 3.5- diagnóstico para os coeficientes

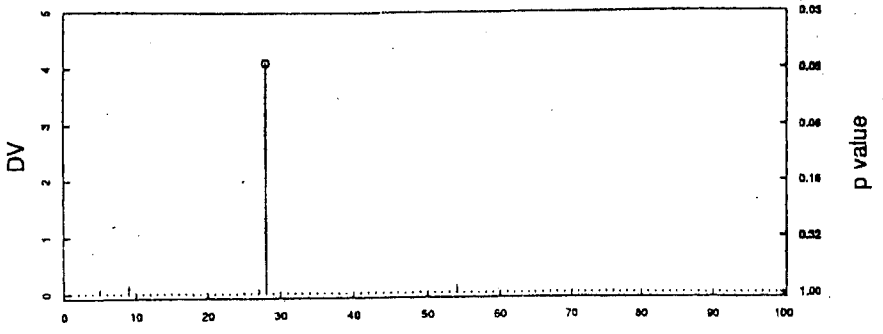

Fig 3.6- diagnóstico para a variância dos ruídos

3.2.4 Determinação do número de observações de um grupo influente

Os diagnósticos DC e DV, foram concebidos para evitar o problema de masking que surge na detecção de um subconjunto de k pontos adjacentes, que é influente devido à ocorrência de um grupo de k outliers nesses pontos.Nestas situações, os diagnósticos DV de exclusão de uma única observação em cada ponto t, podem não assinalar com exactidão os pontos T onde ocorrem os outliers, pois o efeito de um outlier pertencente a um grupo, pode não ser isoladamente significativo. Os diagnósticos DV de exclusão de k observações consecutivas, que ocorrem em pontos centrados em t, representam-se usualmente por $DV(k,t)$. Para k par, t é o ponto imediatamente à esquerda do centro. Por exemplo, quando $k=2$, $DV(k,t)$ corresponde ao valor do diagnóstico calculado sem as observações Z_t e Z_{t+1} e representa a influência de um grupo de 2 pontos adjacentes centrados em t.

Em situações práticas o número de outliers presente numa série temporal, é desconhecido e os autores recorreram novamente a uma simulação para estudar o comportamento dos valores $DV(k,t)$, e estabelecerem uma regra de determinação do número de pontos de um grupo influente, presente numa série temporal. Quando o diagnóstico é aplicado a uma série temporal com um outlier isolado e um grupo k_0 outliers adjacentes. Os comportamentos observados são os seguintes:

$k-1$ valores de $DV(k,\cdot)$, ao redor do ponto T onde ocorre o outlier isolado, são significativos, e aproximadamente iguais a $DV(k,T)$;

Para $k \geq k_0$, existem $k - k_0 + 1$ subconjuntos de k tempos centrados em t, que contêm o grupo de k_0 tempos adjacentes centrados em t onde ocorrem os outliers. Os valores p para $DV(k,t)$, respeitantes à exclusão das observações de cada um desses subconjuntos, são significativos (p<0.5) e aproximadamente iguais a $DV(k_0,T)$.

Tendo por base estes resultados, os autores proposeram a seguinte estratégia para determinar o número de observações de um grupo influente presente numa série temporal:

> *Obtêm-se os diagnósticos de exclusão de k observações, para k crescente, $k = 1,2,...$, até que a magnitude de $DV(k,t)$ não aumente significativamente para algum t. O número de observações de um grupo influente é estimado subtraindo uma unidade ao primeiro k, para o qual se evidencie um comportamento aproximadamente uniforme do valor do diagnóstico.*

3.2.5 Procedimento iterativo de eliminação de outliers dos dados

Por vezes, a presença de um outlier grosseiro, tem influência suficiente para que a exclusão de outras observações aberrantes da série temporal não tenha algum efeito significativo nas estimativas. Outros tipos, mais subtis de masking surgem quando ocorrem outliers de impacto moderado, em tempos relativamente próximos. Para lidar com estas situações, os autores propuseram um procedimento de eliminação iterativa, que consiste na remoção de possíveis outliers dos dados e na reformulação dos diagnósticos.O procedimento é constituído pelas seguintes etapas:

a) Utilizando a estratégia delineada em 3.2.4, obtêm-se os valores de $DV(k,t)$, para $k = 1,2,...$, até ser determinado o número de pontos do grupo mais influente, presente nos dados. Alternativamente, pode-se utilizar $k = K_{\max}$, onde $K_{\max}$ é o número de outliers, que na óptica do utilizador, constitui o grupo influente mais extenso, presente. Para séries com menos de 250 observações, os autores aconselham a utilização de $K_{\max} = 5$.

b) Se não se descobrem pontos influentes, então conclui-se a análise. No caso contrário, excluem-se os pontos mais influentes, identificados em a), e removem-se dos dados, os outliers correspondentes para recalcular as estimativas da máxima verosimilhança dos parâmetros do

modelo. Isto é, retoma-se a etapa a) do procedimento, para medir a influência dos restantes pontos, com os novos parâmetros do modelo para a nova série observada.

Exemplo3.3: Exportações dos EUA para a República Latino Americana no período entre 1966 e 1983.

Os autores consideraram a série dos logaritmos dos dados, que está representada na Fig.3.7, e ajustaram-lhe um modelo ARIMA(0,1,2).

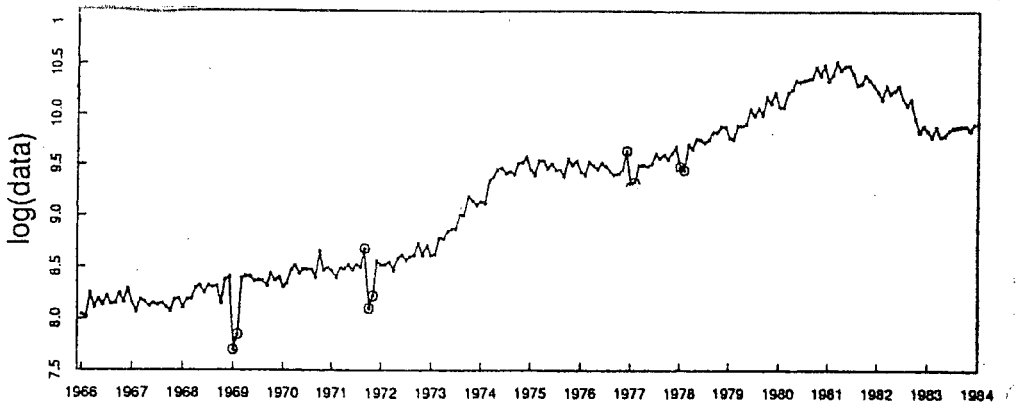

Fig 3.7- logaritmos dos dados de exportações dos EUA para a República Latino Americana no período entre 1966 e 1983.

Obtiveram os diagnósticos DV apenas para $k = 1,2$ e 3, representados nas Figs. 3.8, 3.9 e 3.10, pois o diagnóstico para $k = 3$ não produziu uma maior magnitude. Além disso o comportamento dos valores do diagnóstico reflecte a presença de um grupo influente de 2 outliers. Detectaram-se os outliers em 1/69 e 2/69 (Janeiro e Fevereiro de 1969) e pode-se observar que $DV(2,\cdot)$também é significativo noutros pontos, como por exemplo 10/71, mas sofre o efeito masking por os valores p em 1/69 e 2/69 serem tão mais significativos.

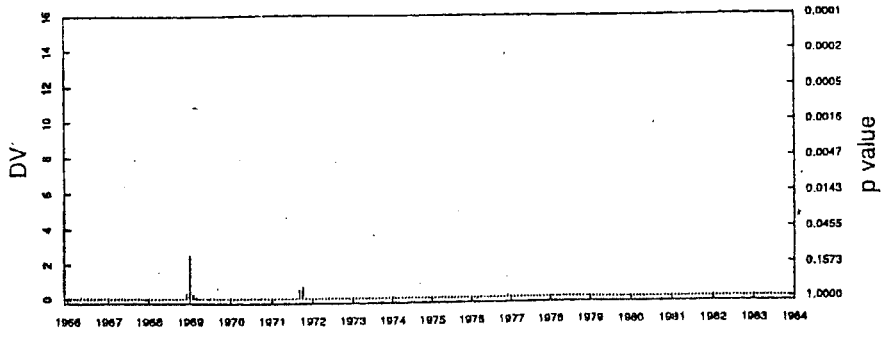

Fig 3.8-diagnóstico DV para k = 1

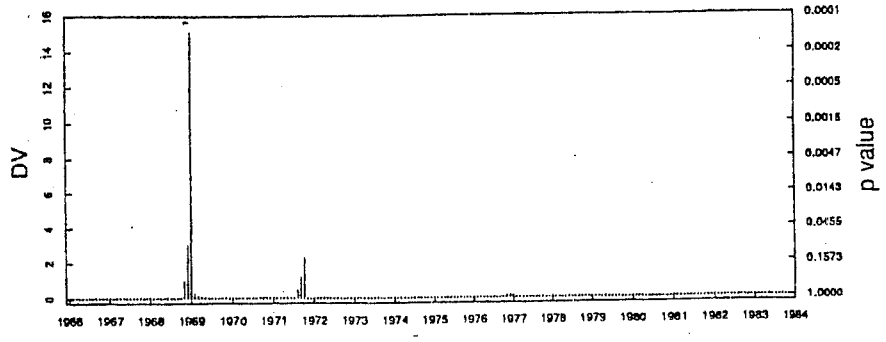

Fig 3.9-diagnóstico DV para k = 2

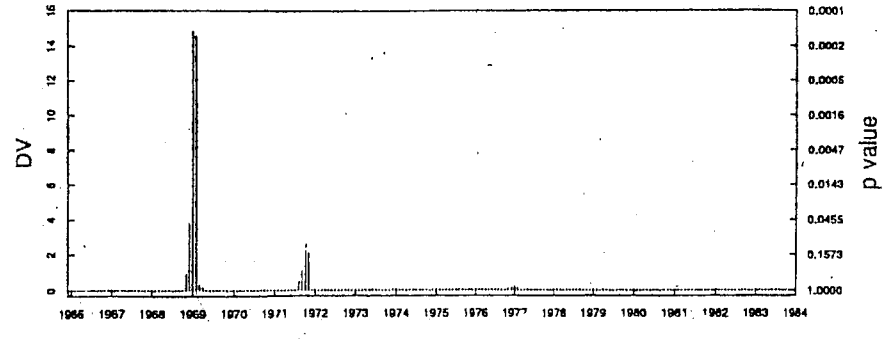

Fig 3.10-diagnóstico DV para k = 3

Seguindo o procedimento delineado em 3.2.5, as observações em 1/69 e 2/69 foram removidas para o cálculo dos novos diagnósticos, para $k = 1,2,3$ e 4. Deste modo, os resultados do 1° ciclo do procedimento de exclusão iterativa, estão representados na 2ª linha da tabela 3.1 e pode-se observar que o efeito, resultante da remoção destas observações, sobre a variância é muito significativo. Os valores dos novos diagnósticos estão representados nas Figs 3.11 a 3.14. Utilizando a estratégia exposta em 3.2.4, identificou-se o grupo influente de outliers em 9/71, 10/71 e 11/71, com $p < 0.01$.

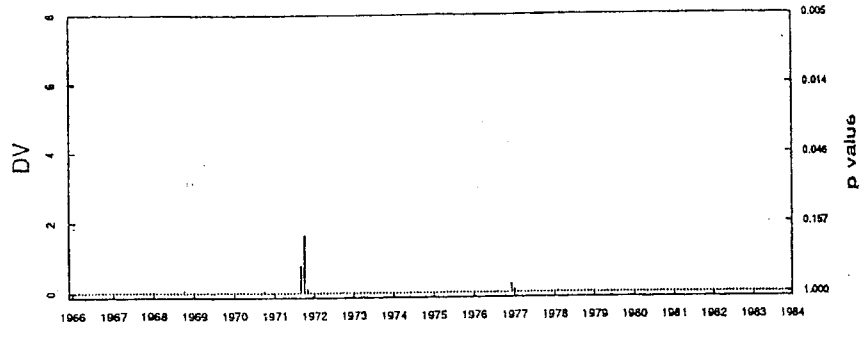

Fig 3.11-diagnóstico DV para k = 1

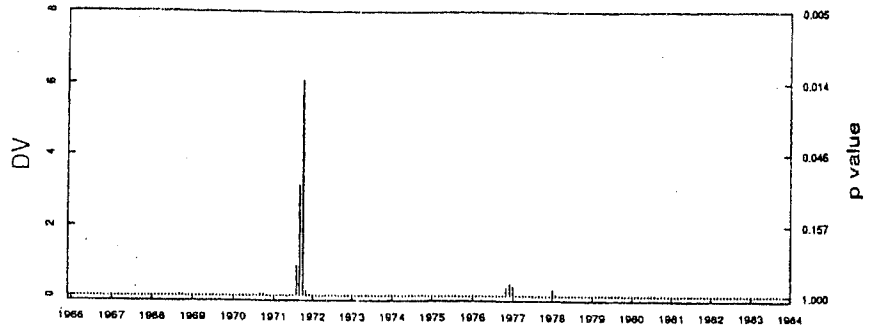

Fig 3.12-diagnóstico DV para k = 2

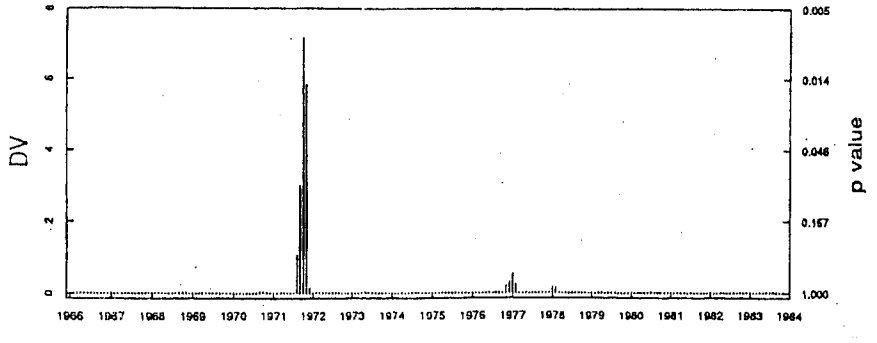

Fig 3.13-diagnóstico DV para k = 3

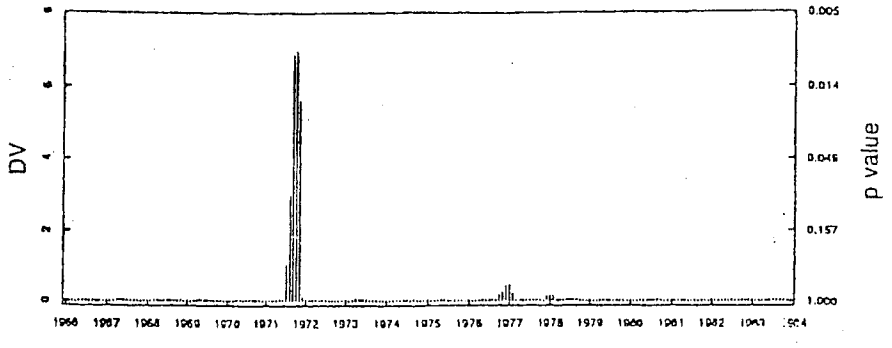

Fig 3.14-diagnóstico DV para k = 4

A inclusão de 9/71 no grupo influente é menos evidente do que a de 10/71 e 11/71 pois o aumento de $DV(3,t)$relativamente a $DV(2,t)$ para t=10/71 é pequeno. Contudo, as Figs. 3.12 a 3.14 permitem observar o padrão de propagação que reflecte mais a presença de um grupo de 3 outliers, do que de 2 outliers. Por isso estes pontos são removidos dos dados.

Na 2ª iteração do procedimento, os diagnósticos são reformulados e os $DV(k,t)$ para $k = 1,2,3$ e 4 estão representados nas Figs 3.15 a 3.18 . Identifica-se um grupo influente em 12/76, 1/77 e 2/77. Este grupo é menos influente do que os anteriormente detectados, pois $p = 0.29$, é menos significativo.

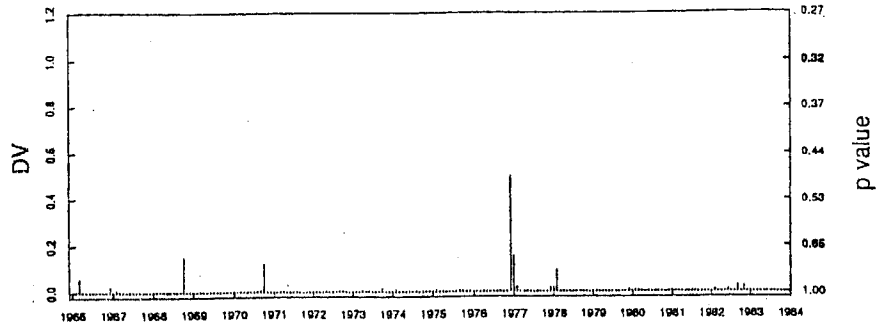

Fig 3.15-diagnóstico DV para k = 1

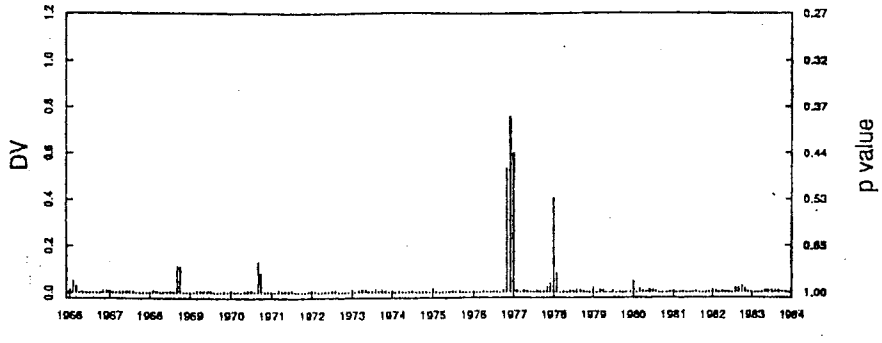

Fig 3.16-diagnóstico DV para k = 2

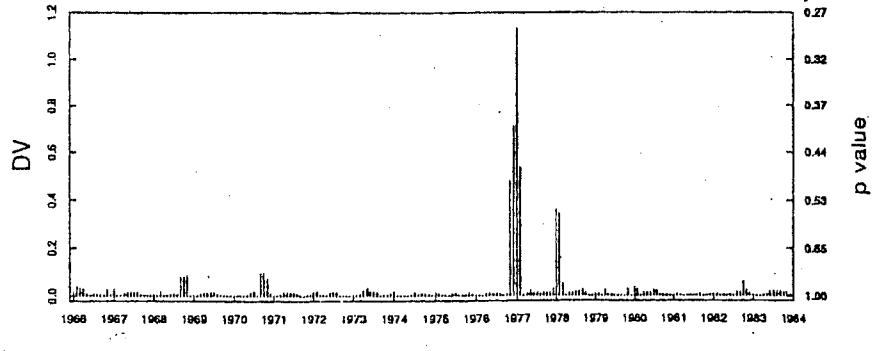

Fig 3.17-diagnóstico DV para k = 3

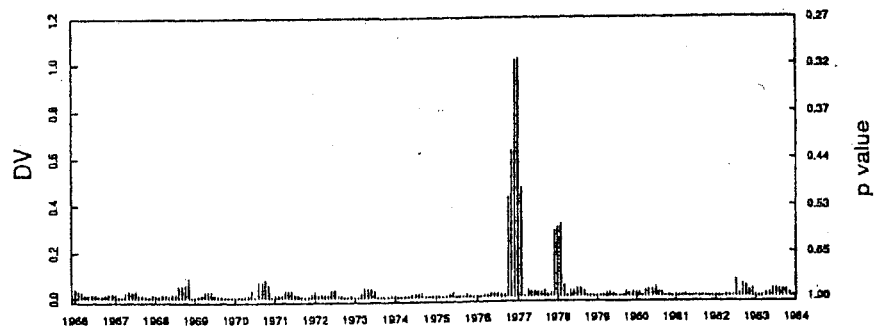

Fig 3.18 diagnóstico DV para k = 4

O procedimento foi iterado novamente e como a representação dos valores dos novos diagnósticos é similar á dos da iteração prévia, foi omitida. Detectou-se um grupo influente em 1/78 e 2/78 e, importa realçar que os diagnósticos baseados na exclusão de uma observação não o assinalam. Foram necessários os diagnósticos de exclusão de duas observações para assinalar estes pontos como influentes.

A análise final, pode-se resumir do seguinte modo:

O procedimento iterativo de eliminação de outliers, foi iterado três vezes e permitiu identificar e remover quatro grupos de outliers. Os pontos que foram eliminados em cada um dos ciclos do procedimento e os estimadores da máxima verosimilhança correspondentes, são dados na tabela 3.1.

Ciclo do procedimento	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}^2$	Pontos eliminados	valor p
0	0.367	0.160	0.0114	_	_
1	0.460	0.003	0.0083	1/69, 2/69	0.00010
2	0.448	-0.041	0.0066	9/71,10/71,11/71	0.0075
3	0.431	-0.058	0.0060	12/76, 1/77, 2/77	0.29
4	0.43	-0.08	0.0056	1/78, 2/78	0.45

Tabela 3.1-Resultados do procedimento iterativo de eliminação de outliers dos dados de exportações dos EUA para a República Latino Americana

3.3 Medidas de influência para outliers aditivos e para alterações de nível

Penã (1990) considerou o problema de detecção de outliers influentes numa série temporal gerada por um modelo ARIMA (p,d,q) e obteve duas medidas de influência diferentes para detectar os outliers aditivos e inovadores que alteram significativamente as estimativas dos parâmetros do modelo. A construção da estatística para detectar outliers aditivos, $D_2(T)$, baseou-se na substituição de uma observação anómala por um valor interpolado utilizando todos os restantes elementos da amostra enquanto que a estatística para outliers inovadores, $D_1(T)$, resulta da eliminação da observação anómala. O processo de construção da estatística $D_2(T)$ será descrito em 3.3.1, pois o autor, concluiu que a estatística para detectar outliers inovadores influentes, não os localiza claramente, o que é compreensível, pois os modelos são muito mais robustos na presença de outliers inovadores do que na de aditivos. Quando uma série tem um LC, ou uma sequência de outliers aditivos consecutivos de impacto similar e que produz um efeito similar ao de um LC, o autor observou que a estatística $D_2(T)$, detecta poucas dessas observações. Para corrigir esta situação, Penã e Sánchez, 1997, desenvolveram uma medida de influência para um LC, $DL(T)$, que será apresentada em 3.3.2.

3.3.1 Medida de influência para outliers aditivos

O autor utiliza a representação autoregressiva do processo Y_t, dada por:

$$Y_t = \sum_{l=1}^{h} \pi_l Y_{t-l} + a_t$$

e supõe que ocorre um outlier aditivo no período T. Então considerando o modelo (2.4) para um AO, em vez de se observar Y_t observa-se Z_t, onde $Z_t = Y_t$ para $t \neq T$ e $Z_T = Y_T + w_T$.

Seja $\Pi_T = (\pi_{1,t}, \ldots\ldots, \pi_{h,T})$ o vector dos coeficientes considerando que a observação em t=T está omissa. A estimativa da máxima verosimilhança condicionada de Π_T é dada por:

$$\hat{\Pi}_T = \left(\hat{X}^T{}_y \hat{X}_y \right)^{-1} \hat{X}^T{}_y \hat{Y} \qquad (3.21)$$

sendo,

$$\hat{X}_y = \begin{bmatrix} \hat{Y}_h & \hat{Y}_{h-1} & \ldots . \hat{Y}_1 \\ \vdots & \vdots & \vdots \ \vdots \\ \hat{Y}_{n-1} & \hat{Y}_{n-2} & \ldots . \hat{Y}_{n-h} \end{bmatrix}, \hat{Y} = \begin{bmatrix} \hat{Y}_{h+1} \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

onde $\hat{Y}_j = Z_j$ para $j \neq T$ e $\hat{Y}_T = Z_T - \hat{w}_T$ é obtido através de:

$$\hat{Y}_T = \hat{Z}_{T/n} = \sum_{i=1}^{h} \delta_i \left(Z_{T+i} + Z_{T-i} \right), \qquad (3.22)$$

$$\hat{w}_T = Z_T - \hat{Z}_{T/n}$$

$$\text{e, } \delta_i = \frac{\left(\hat{\pi}_{i,T} - \sum_{l=1}^{h} \pi_{l,T} \pi_{l+i,T} \right)}{\left(\sum_{l=0}^{h} \hat{\pi}_{l,T}^2 \right)}, \hat{\pi}_{0,T} = -1$$

O sistema de equações(3.21) e (3.22) tem que ser resolvido iterativamente. Começando com um valor inicial $\hat{\pi}_T(0)$ para $\hat{\pi}_T$, calculam-se os δ_i e estima-se o efeito $\hat{w}_T(0)$. Este valor é utilizado para obter $\hat{Y}_T(0) = Z_T - \hat{w}_T(0)$, o que conduz a uma nova estimativa $\hat{\pi}_T(1)$. O processo é repetido até que tenha sido retirado o outlier que é substituído pelo seu valor interpolado, de modo que os parâmetros resultantes estejam livres do seu efeito.

A estatística apresentada por Penã (1990), para medir a influência de uma observação isolada $\hat{Z}_T$, resulta da análise das alterações que essa observação provoca nas estimativas dos coeficientes do modelo.

Seja $\hat{\Pi} = \left(X_z' X_z\right)^{-1} X_z' Z$ o estimador da máxima verosimilhança de Π, supondo que não existem outliers, e seja $\hat{\Pi}_T$ o estimador da máxima verosimilhança considerando que a observação em t=T está omissa. A medida de influência, $D_2(t)$, é a distância de Mahalanobis entre os vectores $\hat{\Pi}$ e $\hat{\Pi}_T$.

$$D_2(T) = \frac{\left(\hat{\Pi} - \hat{\Pi}_t\right)' \left(X_z' X_z\right) \left(\hat{\Pi} - \hat{\Pi}_T\right)}{h\hat{\sigma}_a^2} \qquad (3.23)$$

onde $X_z = \begin{bmatrix} Z_h & Z_{h+1} \ldots . Z_1 \\ Z_{h+1} & Z_h \quad \ldots . Z_2 \\ \vdots & \vdots \quad \ldots . Z_{t-h} \end{bmatrix}$ e $\left(X_z' X_z\right)^{-1} \hat{\sigma}_a^2$ é a matriz de variância-covariância do vector $\hat{\Pi}$ estimado e h é o número de parâmetros estimados.

A estatistica em (3.23) também pode ser escrita do seguinte modo:

Se $\hat{Z} = X_z \hat{\Pi}$ é o vector de previsão estimado e $\hat{Z}_T = X_z \hat{\Pi}_T$ é o vector de previsão assumindo que a observação em t=T está omissa, a medida de influência pode ser escrita como,

$$D_2(T) = \frac{\left(\hat{Z} - \hat{Z}_T\right)' \left(\hat{Z} - \hat{Z}_T\right)}{h\hat{\sigma}_a^2}$$

pois $\left(\hat{Z} - \hat{Z}_T\right)' \left(\hat{Z} - \hat{Z}_T\right) = \left(\hat{\Pi} - \hat{\Pi}_T\right)' \left(X_z^T X_z\right) \left(\hat{\Pi} - \hat{\Pi}_T\right)$.

Se $\hat{Z}_T^{INT}$ representar, o vector de previsão, estimado quando se assume que a observação em $T-th$ é um outlier aditivo, Penã (1991), sugere que a medição da influência de uma observação isolada, se faça através da seguinte estatística:

$$DZ(T) = \frac{\left(\hat{Z} - \hat{Z}_T^{INT}\right)' \left(\hat{Z} - \hat{Z}_T^{INT}\right)}{h\hat{\sigma}_a^2} \qquad (3.24)$$

O autor constatou, que esta estatística é mais eficiente na detecção de outliers que têm forte infuência sobre o modelo, e utilizou-a em 1997, para retirar todos os pontos influentes de uma série temporal numa etapa inicial de um procedimento de detecção de múltiplos outliers.

3.3.2 Medida de influência para alterações de nível e sequências de outliers

Assuma-se que está presente um LC em t=T, sendo $\{Z_t\}$ a série observada, que é representada pelo modelo (2.7).

Sejam $\hat{\Pi}$ e $\hat{\Pi}_L$ as estimativas de Π, assumindo-se respectivamente que não existem outliers e que existe um LC em t=T . A medida proposta por Penã e Sánchez(1997), para medir a influência de um LC, considerando as alterações no vector de previsão, é dada por:

$$DL(T) = \frac{\left(\hat{Z} - \hat{Z}_T^{ILC}\right)^T \left(\hat{Z} - \hat{Z}_T^{ILC}\right)}{h\hat{\sigma}_a^2} \qquad (3.25)$$

4. Procedimentos iterativos de modelação ARIMA na presença de outliers

Os dados de séries temporais estão frequentemente sujeitos a acontecimentos inesperados e incontroláveis dos quais resultam vários tipos de observações outliers. Dependendo da sua natureza, os outliers, podem afectar moderamente ou significativamente, a eficiência da metodologia usual de modelação linear de uma série temporal, nomeadamente as etapas de identificação do modelo e de estimação dos parâmetros. Neste sentido, foram propostos vários procedimentos que requerem iterações entre etapas de detecção de outliers, utilizando as estatísticas de razão de verosimilhanças expostas em 3.1 para os localizar e identificar, e de estimação de um modelo do tipo de (2.6), para acomodar os efeitos dos outliers com o propósito de modelação linear de uma série temporal na presença de outliers. Esta aproximação foi adoptada por diversos autores, por exemplo Tsay (1986, 88), Chen e Liu (1993) e Penã (1997) e os procedimentos por eles desenvolvidos de acordo com objectivos específicos, serão descritos neste capítulo. Em 4.1, descrever-se-á o procedimento de especificação de um modelo na presença de outliers aditivos e inovadores, proposto por Tsay (1986), em 4.2 uma generalização deste procedimento também desenvolvida por Tsay (1988), para tratamento unificado de outliers e de alterações de nível, em 4.3 um procedimento iterativo de estimação conjunta dos efeitos dos outliers e dos parâmetros do modelo, desenvolvido por Chen e Liu (1993), para limitar a influência dos outliers sobre as estimativas dos parâmetros e em 4.4 será apresentado um procedimento iterativo de detecção de múltiplos outliers, de Penã (1997).

4.1 Especificação de um modelo na presença de outliers

Constitui objectivo principal da análise de dados de uma série temporal encontrar um modelo adequado que permita a sua descrição. Existem vários modelos teóricos para modelizar séries temporais, sendo indispensável, a utilização da função de autocorrelação amostral (SACF) , da função de autocorrelaçao amostral parcial (SPACF) e da função de autocorrelação estendida (ESACF), para a identificação da ordem p de um modelo AR, da ordem q, de um MA e das ordens (p,q) de um modelo ARMA.

O trabalho de investigação desenvolvido por Chang (1982), sobre os efeitos dos outliers na SACF e na SPACF e por Tsay (1984) na ESACF, permitiu-lhes concluir que a presença de outliers provoca enviezamentos nas estimativas das autocorrelações e pode induzir a uma especificação incorrecta do modelo.

Como resultado, Tsay (1986) propôs um procedimento iterativo de especificação de um modelo para uma série com outliers. Este procedimento baseia-se essencialmente no procedimento iterativo de estimação dos parâmetros de Chang e Tiao (1983), e na ESACF de Tsay e Tiao (1984). No entanto, para obter as estimativas dos parâmetros do modelo, Tsay utilizou o método dos mínimos quadrados em vez do método da máxima verosimilhança, utilizado por Chang e Tiao. O método de estimação utilizado neste procedimento será exposto em 4.1.1 e o procedimento em 4.1.2.

4.1.1 Estimação dos parâmetros

Tsay e Tiao (1984), demonstraram que, podem ser obtidas estimativas consistentes, dos parâmetros AR através da iteração q da regressão AR(p). Para a maioria das séries contaminadas, supondo que as ordens, p e q do modelo (2.3) são conhecidas,Tsay (1984, corolário 2.6, teorema 3.3(b) e (2.1.5)), mostrou que ainda se consegue obter estimativas consistentes para os parâmetros AR , através da j-ésima regressão iterativa, de AR(p), apesar do número de iterações necessárias, ter que ser maior do que q. Isto significa que, para uma série contaminada, as estimativas AR iteradas, começam a ter um padrão constante, a partir da iteração $j > q$.

Tsay (1986), ilustra o resultado, utilizando dados de temperaturas, de uma série, de Box e Jenkins(1976).

A tabela 4.1 mostra as estimativas AR iteradas de uma regressão AR(2), para as séries dos dados originais e dos dados contaminados artificialmente. O conjunto de dados contaminado é obtido alterando y_{71}= 25.8, da série original, para 28.5. Este valor é um outlier aditivo, de efeito w_{AO} = 2.7 em t = 71 e pode resultar de um erro de registo.

Iteração j	$\hat{\phi}_1$ e $\hat{\phi}_2$ para a série contaminada		$\hat{\phi}_1$ e $\hat{\phi}_2$ para a série original	
0	1.12	-.15	1.81	-.82
1	3.28	-2.27	1.83	-.84
2	1.91	-.92	1.84	-.85
3	1.81	-.82	1.89	-.90
4	1.89	-.90	1.92	-.93
5	1.79	-.80	1.92	-.92

Tabela 4.1- estimativas dos parâmetros AR iteradas de uma regressão AR(2), para as séries dos dados originais e dos dados contaminados artificialmente

As estimativas iteradas da regressão AR(2), revelam um padrão constante a partir de j=0, enquanto que para os dados contaminados, esse padrão é obtido a partir da iteração j=3.

Depois de se seleccionarem as estimativas AR consistentes, considera-se que são os verdadeiros valores dos parâmetros autoregressivos e transforma-se Y_t em W_t,

$$\text{com} \quad W_t = X_t - \sum_{i=1}^{p} \hat{\phi}_i X_{t-i} \qquad (4.1)$$

Como os $\hat{\phi}_i$ são consistentes, W_t seguirá um modelo MA(q) puro e então aplica-se o método de Durbin(1959), para calcular as estimativas dos parâmetros MA. O método de Durbin baseia-se na consideração de que qualquer MA invertível pode ser aproximado por um AR,

$$W_t = \beta_1 W_{t-1} + ... + \beta_k W_{t-k} + f_t \qquad (4.2)$$

Os verdadeiros parâmetros MA, são funções dos coeficientes AR, β_i , em (4.2). Isto implica que na prática podemos começar por ajustar a W_t um AR(k), com k suficientemente grande e depois

resolver um sistema de equações lineares para obter as estimativas dos parâmetros MA. As equações lineares baseiam-se na relação entre parâmetros MA e os coeficientes β_i de (4.2).

4.1.2 Procedimento iterativo de especificação de um modelo na presença de outliers

O procedimento iterativo, proposto por Tsay, tem como objectivo final, a especificação do "melhor modelo" para uma série temporal, quando os dados contêm um número desconhecido de outliers. A escolha do modelo final, resulta de sucessivas modificações dos dados e das correspondentes alterações aos vários modelos, que vão sendo especificados em cada uma das iterações. Cada iteração é constituida por quatro etapas:

I) Identificação

II) Estimação dos parâmetros MA

III) Detecção

IV) Síntese e averiguação.

I) Identificação: Utiliza-se a ESACF para identificar uma possível ordem(p,q) para os dados da série, que se supõe sem outliers e obtêm-se as estimativas dos mínimos quadrados dos parâmetros AR utilizando o método de Tsay e Tiao (1984).

II) Estimação dos parâmetros MA: Utilizando as estimativas AR seleccionadas na etapa I, transformam-se os dados X_t em W_t, calculando $W_t = X_t - \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}$. A seguir calculam-se as estimativas MA através do método de Durbin.

III) Detecção: Utilizam-se apenas as estatísticas de teste $\hat{\lambda}_A$ e $\hat{\lambda}_I$ na etapa de detecção descrita em 3.1.2. Se não existem outliers, isto é, se $\hat{\lambda} < C$, o procedimento de especificação do modelo, termina e prossegue-se para a etapa IV. Se,
pelo contrário, é detectado algum outlier, então remove-se o efeito desse outlier obtendo-se a série ajustada e itera-se o procedimento. Isto é, retoma-se a etapa I de identificação, utilizando as observações que foram modificadas por uma das equações:

$$Z_t^* = Z_t - \hat{w}_{AO} I_t^{T_A}, \qquad (4.3)$$

$$Z_t^* = Z_t - \hat{w}_{IO} \frac{1}{\pi(B)} I_t^{T_I} \qquad (4.4)$$

IV) Síntese e averiguação: O modelo final, resume-se do seguinte modo:

a) A ordem(p,q) identificada na última iteração da etapa I, é a ordem possível para o modelo da série $\{Z_t\}$, sem outliers.

b) O número de outliers coincide com o número de iterações, já que em cada iteração, excepto na última, identifica-se apenas um outlier. Depois de identificado um modelo do tipo (2.9), o analista deve examinar os dados originais, procurando as possíveis razões da presença dos outliers, e se necessário, modificar o modelo de acordo com respostas obtidas as seguintes questões: Porque são estas observações diferentes das restantes? Existem razões suficientes para desconfiarmos que são outliers? O comportamento destas observações dever-se-á a alguma intervenção exógena? Qual a origem dessas intervenções?

Exemplo 4.1: Para ilustrar o método iterativo de especificação de um modelo na presença de outliers, Tsay considerou os dados do consumo anual de bebidas alcoólicas, entre 1870 e 1983, no Reino Unido.

Este conjunto de dados já havia sido examinado, por Prest (1949), Durbin e Watson (1951) e Fuller (1976). Os dados constituem uma série temporal, e o modelo da série é composto por uma variável dependente e quatro variáveis explicativas. A variável dependente, Y_t, é o consumo anual de bebidas, e duas das variáveis explicativas, X_{1t} e X_{2t}, são o ordenado per capita e o preço das bebidas. Estas variáveis estão em logaritmos. As outras duas variáveis explicativas, são os termos linear e quadrático de tendência.

Prest ajustou aos dados, o seguinte modelo:

$$Y_t = 2.14 + .69X_{1t} - .63X_{2t} - .0095t - .00011(t-35)^2 + e_t,$$
$$\sigma_e^2 = .000983$$

onde, σ_e^2 é a média dos quadrados dos resíduos e t = ano actual – 1869.

Tsay iniciou o procedimento, considerando que a série dos resíduos e_t, é uma série observável. Estas observações estão representadas na figura 4.1.

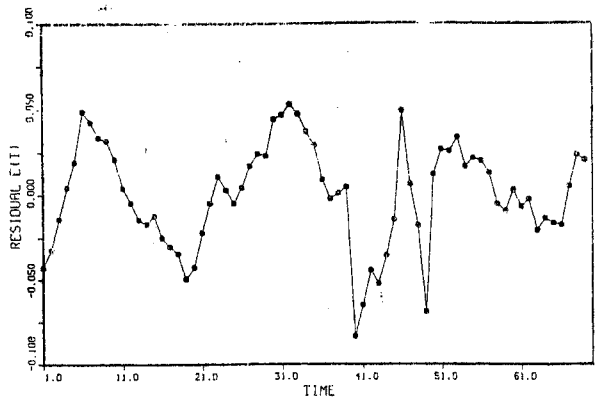

Fig 4.1- série dos resíduos, e_t, do modelo $Y_t = 2.14 + .69\,X_{1t} - .63\,X_{2t} - .0095\,t - .00011\,(t - 35)^2 + e_t$, $\sigma_e^2 = .000983$

Especificou um AR(1) para os e_t, calculando as ESACF (veja-se tabela 4.2).

	q									
p	0	1	2	3	4	5	6	7	8	9
					a. ESACF					
0	.72	.46	.25	.15	.00	-.13	-.18	-.27	-.34	-.51
1	.17	.14	.00	.16	.01	-.17	.00	-.07	.04	-.32
2	-.44	.13	.00	.17	.09	-.16	.01	-.04	-.01	-.34
3	.39	-.11	-.06	.15	.11	-.16	-.06	.05	.01	-.33
4	.17	-.23	-.42	.04	.11	-.17	-.03	.03	.04	-.23
5	-.50	.39	-.45	.18	.06	-.03	.16	-.07	.13	-.26
6	.27	-.24	-.47	-.22	-.04	.01	.14	.10	.08	-.25
					b. Indicator Symbols					
0	X	X	O	O	O	O	O	X	X	X
1	O	O	O	O	O	O	O	O	O	X
2	X	O	O	O	O	O	O	O	O	X
3	X	O	O	O	O	O	O	O	O	X
4	O	O	X	O	O	O	O	O	O	O
5	X	X	X	O	O	O	O	O	O	O
6	X	O	X	O	O	O	O	O	O	O

Tabela 4.2- ESACF para os dados de bebidas alcoólicas

Estimou $\hat{\phi}_1 = 0.72$ e seguidamente testou a existência de outliers, utilizando C = 3.5 e k = q+5, sendo k ordem AR do método de Durbin. Identificou um IO ,em t = 40 com impacto $\hat{w} = -.08663$, e $\hat{\lambda}_{I,40} = -4.22$. Os dados foram então modificados através de (4.4) e o processo foi iterado. O processo de identificação terminou na iteração 7 e podem observar-se os resultados, na tabela 4.3.

Iteração	Modelo	Estimativas dos parâmetros		Outlier		
		AR	MA	Tipo	Tempo	Magnitude
1	AR(1)	.72		IO	40	-.08663
2	AR(1)	.72		AO	49	-.06628
3	AR(1)	.78		IO	41	-.06772
4	AR(2)	1.03, -.30		AO	46	.04662
5	AR(2)	1.19, -.40		IO	42	-.04653
6	ARMA(1,1)	.72	-.30	IO	43	-.05379
7	ARMA(1,1)	.74	-.28			

Tabela 4.3- resultados do procedimento iterativo de identificação de um modelo para os dados de bebidas alcoólicas

Na tabela 4.4, estão representadas as ESACF dos dados, após terem sido retirados os 6 outliers, identificados na tabela 4.3.

	q									
p	0	1	2	3	4	5	6	7	8	9
					a. ESACF					
0	.82	.59	.38	.20	.02	−.10	−.20	−.29	−.37	−.43
1	.33	.23	.20	.23	.05	−.11	−.14	−.10	−.15	−.12
2	−.35	−.06	−.08	.26	−.13	.02	−.04	.07	−.06	.01
3	−.49	.07	−.08	.20	−.12	−.03	.00	.08	−.06	.02
4	−.30	−.25	−.47	.14	−.13	−.05	−.01	.10	.06	.04
5	−.03	−.33	−.33	−.27	.00	.07	.09	.10	−.03	.04
6	−.05	−.15	−.09	−.22	−.01	.06	−.04	.02	.03	.05
					b. Indicator Symbols					
0	X	X	X	O	O	O	O	X	X	X
1	X	O	O	O	O	O	O	O	O	O
2	X	O	O	X	O	O	O	O	O	O
3	X	O	O	O	O	O	O	O	O	O
4	X	O	X	O	O	O	O	O	O	O
5	O	X	X	X	O	O	O	O	O	O
6	O	O	O	O	O	O	O	O	O	O

Tabela 4.4- ESACF para os dados de bebidas alcoólicas

As ESACF, sugeriram um ARMA(1,1). Os parâmetros iniciais deste modelo ARMA(1,1) são dados na última linha da tabela 4.3. Então, o procedimento iterativo identificou um ARMA(1,1) com 6 outliers de vários tipos para a série residual e_t do modelo inicial.

Posteriormente, Tsay examinou os seis outliers identificados.

Será razoável suspeitar destas observações como outliers? Porquê? Se são outliers quais são as suas possíveis causas? Porque há quatro IO consecutivos em t=40, 41, 42, 43? Porque parecem os impactos dos primeiros três outliers caír exponencialmente? Como podemos especificar um modelo apropriado para e_t depois de percebermos as razões da sua presença?

Uma observação mais detalhada da figura 4.1, mostrou que:

a) e_{40} e e_{49} eram de facto os pontos inferiores de e_t.

b) a queda brusca em e_{40} teve um impacto em vários pontos, enquanto que a de e_{49} não.

c) A estrutura de e_t deve ter-se alterado após t=40.

Estes factos pareceram sustentar a utilização do procedimento iterativo proposto. Na procura de possíveis causas para os outliers identificados, Tsay considerou alguns factos históricos do Reino Unido. Os AO e_{46} e e_{49} correspondem respectivamente a 1915 e 1918 e podem reflectir o início e o fim da 1ª Guerra Mundial. e_{40} corresponde a 1909, ano em que Loyd George, 1º ministro do Reino Unido deu início a um programa de reforma social. A implementação de impostos elevados sobre as bebidas alcoólicas teve um impacto substancial sobre o consumo das mesmas. Além disso, é de esperar que este impacto diminua gradualmente. Isto explica em parte porque existem quatro IO consecutivos a partir de e_{40}, que vão diminuíndo exponencialmente. Estes factos históricos constituem razões suficientes que nos levam a confirmar que os outliers identificados são realmente diferentes das outras observações. Logo, o procedimento iterativo aplicado neste exemplo identifica as observações que requerem mais atenção.

Finalmente, baseando-se nas possíveis explicações para os seis outliers, sugeriu o seguinte modelo para os dados das bebidas alcoólicas:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 t + \beta_4 (t-35)^2 + e_t$$

$$e_t = \left[(w_1 + w_2 B^3)/(1-\delta B)\right]\xi_t^{(40)} + w_3 \xi_t^{(46)} + w_4 \xi_t^{(49)} + \left[(1-\theta B)/(1-\Phi B)\right]a_t$$

4.2 Especificação de um modelo na presença de outliers e alterações de nível

Tsay (1988) generalizou os procedimentos de Chen e Tiao (1983) de estimação dos parâmetros e de Tsay (1986) de especificação de um modelo na presença de outliers e propôs um tratamento unificado para outliers e alterações de nível permanentes (LC) e transitórias(TC).

O procedimento descrito por Tsay (1988) é constituído por ciclos de especificação e estimação e de detecção e remoção, para construir um modelo para uma série temporal na presença de outliers e alterações de nível.

4.2.1 Procedimento iterativo de especificação de um modelo na presença de outliers e alterações de nível

Uma modificação introduzida neste procedimento, consiste em dividir cada iteração num ciclo exterior e num ciclo interior, para reduzir o número de vezes de estimação dos parâmetros do modelo.

Ciclo exterior:

Passo A - Assume-se que a série observada $\{Z_t\}$, não tem outliers, identifica-se um modelo ARMA e obtêm-se as estimativas da maxima verosimilhança dos parâmetros do modelo.

Ciclo interior:

Os passos, A_1, A_2 e A_3 de um ciclo interior, coincidem com os passos 1,2 e 3 da etapa de detecção exposta em 3.1.2.

Passo A_4 - Remove-se das observações z_t o efeito do outlier ou da alteração de nível identificado(a), e obtém-se a série ajustada $\{Z_t^*\}$. No caso em que é detectado um outlier, o efeito é removido utilizando uma das equações (4.3) ou (4.4). Se é detectada uma alteração de nível, a série é ajustada através de uma das seguintes equações:

$$Z_t^* = Z_t - \hat{w}_{LC} \frac{I_t^T}{1-B} \qquad (4.5)$$

$$Z_t^* = Z_t - \hat{w}_{TC} \frac{I_t^T}{1-\delta B} \text{ e } 0 < \delta < 1 \qquad (4.6)$$

Retoma-se o passo A_1 do mesmo ciclo, tratando a série $\{Z_t^*\}$como se fosse a observada.O ciclo é repetido até não ser detectado mais nenhum outlier.

Passo B - Se não foi detectada nenhuma perturbação signifitiva no ciclo interior, termina-se o procedimento. Caso contrário, o procedimento é reeniciado no passo A, tratando a última série ajustada em A_4 como se fosse a observada.

4.3 Estimação conjunta dos parâmetros do modelo e dos efeitos dos outliers ou alterações de nível

O procedimento de Tsay (1988) é bastante eficaz na detecção da localização e na estimação dos efeitos de outliers e alterações de nível isolados. Chen e Liu (1993) destacaram duas situações problemáticas neste procedimento:

a) Mesmo que o modelo seja correctamente identificado, os outliers e as alterações de nível podem distorcer as estimativas dos parâmetros afectando a eficiência da detecção.

b) Quando existem múltiplos outliers, as estimativas dos efeitos de cada outlier que ocorre num período de tempo T_i, podem não ser estimativas centradas, devido à influência de outros outliers.

O procedimento proposto por Chen e Liu (1993) foi desenvolvido com o intuito de estimar conjuntamente os parâmetros do modelo e os efeitos dos outliers. O método utilizado para este efeito será exposto em 4.3.1 e o procedimento iterativo de detecção e estimação em 4.3.2.

4.3.1 Método de estimação conjunta na presença de múltiplos outliers

Suponha-se que a série $\{Y_t\}$sofre m intervenções em $T_1, T_2, ...T_m$ que resultam em vários tipos de outliers ou alterações de nível. A série observada $\{Z_t\}$pode ser representada por um modelo do tipo de (2.9).

Os resíduos resultantes do ajustamento de um modelo ARMA a $\{Z_t\}$podem ser representados por,

$$\hat{e}_t = \sum_{j=1}^{m} w_j \hat{\pi}(B) v_j(B) I_t^{T_j} + a_t , \qquad (4.3.1)$$

assumindo que o modelo subjacente está correctamente especificado, mas não são tomados em consideração os efeitos dos outliers.

Então, no caso em que estão disponíveis o efeito e a localização de cada um dos outliers, os efeitos podem ser ajustados às observações, através da equação (2.9) e depois estimam-se os parâmetros do modelo. Por outro lado, quando são conhecidos os parâmetros do modelo, pode-se identificar os outliers ou as alterações de nível e estimar os seus efeitos através da equação (4.3.1). Como a aplicação deste método necessita de mais do que um passo, os autores desenvolveram um procedimento iterativo constituído por três etapas. Na etapa I, utilizam-se estimativas preliminares dos parâmetros do modelo para identificar todos os possíveis outliers ou alterações de nível e a sua localização. Como nesta etapa, os resíduos ainda podem estar contaminados, os autores propõem três métodos para estimar σ_a em vez da utilização usual do desvio padrão amostral. Na etapa II, estimam-se conjuntamente os efeitos dos outliers e os parâmetros do modelo e na etapa III, utilizam-se as estimativas dos parâmetros obtidas na etapa II, para identificar e localizar novamente os outliers ou alterações de nível e estimar os seus efeitos.

4.3.2 Procedimento iterativo de detecção e estimação

Etapa I : Estimação inicial dos parâmetros e detecção de outliers ou de alterações de nível

Passo I_1- Obtêm-se as estimativas da máxima verosimilhança dos parâmetros do modelo, utilizando na 1ªiteração, a série original e nas iterações posteriores a série ajustada.

Passos I_2 e I_3- Estes passos constituem o ciclo interior do procedimento iterativo de especificação proposto por Tsay (1988). No entanto, a estimativa de σ_a é obtida utilizando um dos três métodos considerados pelos autores:" the median absolute deviation"," the α% trimmed" e "the omit-one"[1].

Passo I_4- Se foram detectados outliers ou alterações de nível nos ciclos internos anteriores, e se não é detectado nenhum no ciclo interno actual, então prossegue- se para o passo II_1, da etapa II.

1 - a estimativa do desvio padrão residual obtida pelo MAD define-se por: $\hat{\sigma}_a = 1.483 \times mediana\{|\hat{e}_t - \tilde{e}|\}$. Para calcular a estimativa pelo método α% trimmed, primeiro removem-se os α% maiores valores absolutos e depois calcula-se o desvio padrão amostral. O método "omit one" calcula o desvio padrão residual omitindo o resíduo em T.

Etapa II : Estimação conjunta dos efeitos dos outliers e dos parâmetros do modelo

Passo Π_1 - Supondo que foram identificados m outliers ou alterações de nível, em $T_1, T_2, ..., T_m$, estimam-se conjuntamente os efeitos destes outliers, utilizando a equação de (4.7).

Passo Π_2 - Calculam-se as estatísticas $\hat{\lambda}_j = \frac{\hat{w}_j}{std(\hat{w}_j)}, j = 1,...,m$. $std(\hat{w}_j)$ é a estimativa do desvio padrão dos efeitos estimados dos outliers em T_j. Se $\min|\hat{\lambda}_j| = \hat{\lambda}_v \leq C$, C onde é o mesmo valor crítico utilizado na etapa I, elimina-se do conjunto dos outliers identificados, o outlier ou a alteração de nível em T_v e retoma-se o passo Π_1 com os restantes outliers. Se $\hat{\lambda}_v > C$, prossegue-se para o passo Π_3.

Passo Π_3 - Removem-se os efeitos significativos dos outliers ou das alterações de nível seleccionadas em Π_2, utilizando as últimas estimativas dos efeitos, calculadas em Π_1, e obtém-se a série ajustada.

Passo Π_4 - Calculam-se as estimativas da máxima verosimilhança dos parâmetros do modelo, utilizando a série ajustada. Se a alteração do erro padrão residual é maior do que ξ, retoma-se o passo Π_1. Se se verifica o contrário, prossegue-se para o passo $\Pi\Pi_1$. O valor de ξ é escolhido pelo utilizador para controlar a precisão das estimativas dos parâmetros.

Etapa III: Detecção de outliers utilizando as estimativas finais dos parâmetros do modelo

Passo $\Pi\Pi_1$ - Utilizando as estimativas dos parâmetros obtidas no passo Π_4, calculam-se os resíduos.

Passo $\Pi\Pi_2$ - Utilizando os resíduos obtidos em $\Pi\Pi_1$, itera-se o procedimento através das etapas I e II, com as seguintes modificações: a) as estimativas dos parâmetros a utilizar no ciclo interior da etapa I, são as obtidas em Π_4 e b) omitem-se da etapa II os passos Π_3 e Π_4, sendo as estimativas finais dos efeitos dos outliers detectados, as obtidas em Π_1 da última iteração.

Exemplo 4.2: Considere-se a série temporal dos logaritmos de totais mensais de vendas a retalho, depois de ajustados os efeitos de feriados. Os dados correspondem ao período compreendido entre Janeiro de 1967 e Setembro de 1979. Hillmer (1983), utilizou esta série para ilustrar a aplicação do procedimento iterativo de Chang (1982) de detecção de outliers aditivos e inovadores e obteve o seguinte modelo para a série:

$$\nabla\nabla_{12}Z_t = \frac{\left(1-\theta_{12}B^{12}\right)}{\left(1-\phi_1 B-\phi_2 B^2\right)}a_t$$

Para comparar os resultados obtidos por Hillmer (1983), com os da aplicação do procedimento de estimação conjunta dos efeitos dos outliers e dos parâmetros do modelo, os autores, numa fase inicial, aplicaram o novo procedimento considerando apenas a possível existência de outliers aditivos e inovadores. As tabelas 4.5 a 4.7, resumem os resultados de cada etapa do procedimento iterativo de estimação conjunta, proposto. O valor crítico, utilizado para detectar os outliers é 3.0 e a estimativa do desvio padrão residual é obtida pelo método "the 5% trimmed".

A tabela 4.5, permite visualizar os resultados de estimação e de detecção obtidos em cada uma de três iterações da etapa I. Nesta etapa, são detectados seis outliers.

Iteração 1 da etapa I

$\hat{\theta}_{12}$ = .8407 $\hat{\phi}_1$ = -.3969 $\hat{\phi}_2$ = -.2673 $\hat{\sigma}_a$ = .02501

Ciclo interior (passos I_2 e I_3) da etapa I		
t	Estimativa do efeito	Tipo do outlier
112	-.15	IO
96	-.08	AO
113	-.09	IO
45	.08	IO
103	-.07	AO

Iteração 2 da etapa I

$\hat{\theta}_{12} = .8407 \quad \hat{\phi}_1 = -.3969 \quad \hat{\phi}_2 = -.2673 \quad \hat{\sigma}_a = .02501$

Ciclo interior (passos I_2 e I_3) da etapa I		
t	Estimativa do efeito	Tipo do outlier
73	-.07	IO

Iteração 3 da etapa I

$\hat{\theta}_{12} = .6864 \quad \hat{\phi}_1 = -.5409 \quad \hat{\phi}_2 = -.3161$

Ciclo interior (passos I_2 e I_3) da etapa I
Não se detectaram outliers

Tabela 4.5

Nos passos II_1 e II_2 da etapa II, os efeitos dos outliers são estimados conjuntamente, e os outliers menos significativos são removidos.No exemplo foram removidos os outliers em $t = 103$ e em $t = 73$. Utilizando os últimos efeitos estimados em II_1, isto é os efeitos de cada um dos quatro outliers não eliminados, a série é ajustada no passo II_3, isto é removem-se os efeitos dos quatro outliers e seguidamente estimam-se os parâmetros do modelo, no passo II_4. Os resultados desta etapa estão expostos na tabela 4.6.

Estimação conjunta dos efeitos dos outliers (passos $II_1 - II_3$) da etapa II		
t	Estimativa do efeito	Tipo do outlier
45	.091	IO
96	-.079	AO
112	-.150	IO
113	-.121	IO

Passo II_4 (estimação final)

$\tilde{\theta}_{12} = .6202 \quad \tilde{\phi}_1 = -.6062 \quad \tilde{\phi}_2 = -.3828 \quad \hat{\sigma}_a = .02656$

Tabela 4.6

No passo III_2 da etapa III, retomam-se as etapas I e II do procedimento, mas os parâmetros do modelo não são reestimados. Deste modo, utilizando as estimativas obtidas no passo II_4, são detectados nove outliers, nos passos intermédios. Contudo após a estimação conjunta dos efeitos dos outliers, apenas seis são considerados significativos.

ciclo interior (passos I_2 e I_3) da etapa III

detecção de outliers através das estimativas finais

t	Estimativa do efeito	Tipo do outlier
112	-.15	IO
113	-.13	IO
96	-.08	AO
45	.09	IO
124	.08	IO
114	-.08	IO

ciclo interior (passos I_2 e I_3) da etapa III

detecção de outliers através das estimativas finais

t	Estimativa do efeito	Tipo do outlier
103	-.05	AO
73	-07	IO
136	.07	IO

Ciclo interior (passos I_2 e I_3) da etapa III

Não se detectam outliers

Etapa III		
Resultados finais da detecção de outliers		
t	Estimativa do efeito	Tipo do outlier
45	.94	IO
96	-.079	AO
112	-.148	IO
113	-.135	IO
114	.084	IO
124	.085	IO

Tabela 4.7

A aplicação do procedimento desenvolvido por Chang (1982), permitiu identificar nove outliers ao longo de seis iterações de estimação dos parâmetros. Os resultados de detecção são aproximadamente compativeis com os obtidos nas iterações intermédias das etapas I e III deste procedimento. No entanto, os resultados finais são diferentes dos obtidos pelo actual procedimento, particularmente, a estimativa para θ_{12}, $\hat{\theta}_{12} = .89$ com o procedimento de Chang e $\tilde{\theta}_{12} = .62$ com o novo procedimento.

Aplicando o novo procedimento quando se consideram os quatro tipos de outliers (AO,IO,LC e TC), obtêm-se as seguintes estimativas conjuntas dos parâmetros do modelo e dos efeitos dos outliers detectados:

$\tilde{\theta}_{12} = .7128 \quad \tilde{\phi}_1 = -.6871 \quad \tilde{\phi}_2 = -.4617 \quad \hat{\sigma}_a = .02404$

t	Estimativa do efeito	Tipo do outlier
45	.094	TC
96	-.083	AO
112	-.176	LC

Então, admitindo um conjunto mais vasto de tipos de outliers, detecta-se um menor número de outliers, mas mais significativos. O desvio padrão residual estimado, $\hat{\sigma}_a = .02404$, também é menor do que o obtido quando só são considerados os IO e AO, $\hat{\sigma}_a = .02667$. A partir

destes resultados, é possivel incorporar os efeitos dos outliers no modelo inicial e estimar o seguinte modelo de intervenção:

$$\nabla\nabla_{12}Z_t = \frac{w_1}{1-.7B}\nabla\nabla_{12}I_t^{45} + w_2\nabla\nabla_{12}I_t^{96} + \frac{w_3}{1-B}\nabla\nabla_{12}I_t^{112} + \frac{\left(1-\theta_{12}B^{12}\right)}{\left(1-\phi_1 B-\phi_2 B^2\right)}a_t$$

As estimativas dos parâmetros deste modelo, obtidas pelo método exacto da máxima verosimilhança, são as seguintes:

$\hat{\theta}_{12} = .7123 \quad \hat{\phi}_1 = -.6877 \quad \hat{\phi}_2 = -.4622 \quad \hat{\sigma}_a = .02352$

$\hat{w}_1 = .0956 \quad \hat{w}_2 = -.0837 \quad \hat{w}_3 = 7.166$

Os autores concluem, que as estimativas obtidas pelo novo procedimento de estimação estão muito próximas das obtidas através do modelo de intervenção com a informação dos outliers incorporada.

4.4 Identificação de múltiplos outliers adjacentes em modelos ARIMA

Quando existe uma sequência de outliers aditivos, é frequente que surja o problema de "masking", porque os procedimentos usuais de identificação sequêncial de cada um dos outliers podem não identificar alguns desse grupo. Para evitar esta situação, Sánchez e Penã (1997), desenvolveram um procedimento para detecção de outliers isolados e também de múltiplos outliers adjacentes. Os autores destacam ainda os seguintes problemas nos procedimentos de tratamento de outliers existentes:

1) Quando existe um LC, este é muitas vezes identificado como um IO.

2) A estimação inicial dos parâmetros é feita sob a hipótese de inexistência de outliers, o que conduz à inicialização dos procedimentos de detecção com um conjunto enviezado de estimativas dos parâmetros.

O procedimento de detecção de múltiplos outliers, proposto pelos autores, inclui uma possível solução para o problema mencionado em 1) que se baseia em não comparar as estatísticas da máxima verosimilhança para detectar um IO, com as estatísticas para um LC. A solução para o 2° problema consiste na utilização das estatísticas $DZ(T)$ e $DL(T)$, de (3.3.4) e (3.3.5), na 1ª etapa do procedimento, para se efectuar uma remoção inicial de outliers e alterações de nível, nos dados da amostra.

O procedimento de detecção de múltiplos outliers é descrito em 4.4.1 e seguidamente apresenta-se um exemplo que ilustra o desempenho do procedimento.

4.4.1 Procedimento de detecção de múltiplos outliers em modelos ARIMA

O procedimento tem três etapas. O resultado da etapa I (estimação inicial dos parâmetros do modelo), é uma estimativa inicial robusta, obtida de uma amostra da qual foram retirados todos os pontos influentes, tanto individualmente como conjuntamente. Na etapa II (detecção de outliers), os outliers são identificados através de um algoritmo similar aos desenvolvidos por Tsay (1988) e Chen e Liu (1993), com uma modificação para evitar a confusão entre alterações de nível e outliers inovadores.

Na última etapa (estimação conjunta), o procedimento utiliza o método da máxima verosimilhança para estimar conjuntamente os parâmetros do modelo e os efeitos dos outliers. Esta etapa termina com um novo passo de detecção de outliers, utilizando as estimativas da máxima verosimilhança dos parâmetros.

Etapa I: Estimação inicial dos parâmetros do modelo

Passo 1- Calculam-se as estimativas dos parâmetros do modelo, utilizando a série observada, que se supõe sem outliers.

Passo 2 - Calcula-se a medida de influência $DL(t)$ para todo o t. Selecciona-se o tempo T_1, onde ocorre o máximo de $DL(t)$ e seguidamente estima-se o modelo de intervenção dado por (2.6). Então se:

a) $\hat{w}_{LC}$ é significativo, remove-se das observações Z_t o efeito do LC, utilizando a equação dada em(4.5) definindo-se deste modo a série ajustada $\{Z_t^*\}$ e repete-se o passo 2. O processo é repetido até que $\hat{w}_{LC}$ não seja significativo.

b) $\hat{w}_{LC}$ não é significativo, prossegue-se para o passo 3.

Passo 3 - Calcula-se a medida de influência individual, $DZ(t)$, com a última série ajustada $\{Z_t^*\}$. Seleccionam-se os α % valores mais influentes e removem-se os seus efeitos, utilizando a equação de (4.3) para outliers aditivos. A seguir, estimam-se novamente os parâmetros do modelo, parâmetros esses que vão ser utilizados na etapa II do procedimento.

Etapa II: Detecção de outliers

Passo 1 - Calculam-se $\hat{\sigma}_a$ e os resíduos do modelo, utilizando os parâmetros da etapa I.

Passo 2 - Calculam-se as estatísticas $\hat{\lambda}_{A,t}, \hat{\lambda}_{I,t}$ e $\hat{\lambda}_{LC,t}$, para $t = 1,...,n$.

Passo 3 - Os módulos das estatísticas para um IO e um AO, são comparados em cada t e seleccionam-se as estatísticas com maior valor absoluto, $\hat{\lambda}_{va,t}$. Seja $\hat{\lambda}_{va,T_A}$ o valor que ocorre em

$t = T_A$ e que é o máximo de $\hat{\lambda}_{va,t}$. Se $\hat{\lambda}_{va,T_A} = \left|\hat{\lambda}_{A,T_A}\right| \geq C_1$, existe possivelmente um AO em $t = T_A$ e se $\hat{\lambda}_{va,T_A} = \left|\hat{\lambda}_{I,T_A}\right| \geq C_1$ o possivel outlier é um IO.

O valor crítico C_1 é específico para outliers aditivos e inovadores, e é pré-determinado de acordo com o tamanho da amostra e a estrutura do modelo.

Passo 4 - Selecciona-se $\hat{\lambda}_{L,T_B} = \max\left|\hat{\lambda}_{LC,t}\right|$ para $t = 1,...,n$. Se $\hat{\lambda}_{L,T_B} \geq C_2$ existe possivelmente um LC em $t = T_B$.

Passo 5 - Procede-se de acordo com uma de quatro situações possíveis:

a) Se não foram encontrados outliers, nem alterações de nível, então conclui-se que a série observada está livre de efeitos de outliers e termina-se o procedimento.
b) Se somente foi detectado um outlier, obtém-se a série ajustada, removendo o efeito do AO, ou do IO.
c) Se é detectada uma alteração de nível e não é detectado nenhum outlier, remove-se o efeito do LC e obtém-se a série ajustada.
d) Se é detectado um outlier e uma alteração de nível e se ocorrem em pontos diferentes, então removem-se os efeitos de ambos. Se ocorrem no mesmo ponto, (quando existe um LC em $t = T$ é possivel que o procedimento o detecte simultâneamente com um AO ou um IO), averigua-se, se ambos os efeitos são significativos, utilizando os modelos de intervenção correspondentes.Quando as estatísticas $\left|\hat{w}_j / std(\hat{w}_j)\right|$ são maiores que os valores críticos C_1 e/ou C_2, então os efeitos do outlier e/ou da alteração de nível em $t = T$ são significativos. $std(\hat{w}_j)$ é o desvio padrão do efeito estimado do outlier em $t = T$.

O passo 5 termina quando os outliers significativos tiverem sido removidos.

Passo 6: Utilizando a série ajustada do passo 5 e os parâmetros da etapa I, retoma-se o passo 2 e repetem-se os passos 3,4 e 5. Este ciclo é repetido até não serem detectados mais outliers.

Etapa 3: Estimação conjunta

Calculam-se os resíduos , $\hat{e}_t$, do modelo utilizando os parâmetros estimados depois da correcção dos k outliers detectados em $T_1,...,T_k$. A seguir estima-se conjuntamente o efeito dos outliers utilizando,

$$\hat{e}_t = \sum_{j=1}^{k} w_j \hat{\phi}(B) v_j(B) I_t^{T_j} + a_t \qquad (4.8)$$

Averigua-se se os efeitos dos outliers são significativos, utilizando um valor critico menor, do que os utilizados na etapa 2. Se o efeito de alguns não é significativo, então remove-se do conjunto dos outliers detectados, o que tem menor efeito e estima-se novamente o efeito dos k-1 outliers. Este procedimento é repetido até que sejam significativos todos os outliers do conjunto final de outliers. Removem-se então os os seus efeitos da série observada e obtém-se a série ajustada. Com a última série, estimam-se os parâmetros e retomam-se as etapas 2 e 3 do procedimento.Quando não são detectados outliers adicionais na etapa 2, removem-se todos os efeitos dos outliers detectados e obtém-se a série ajustada. Finalmente, identifica-se o modelo para a série ajustada e estimam-se conjuntamente os efeitos dos outliers e os parâmetros do modelo.

Exemplo 4.3: Os autores utilizam uma simulação de uma série gerada por um modelo AR(1), com parâmetro $\phi = 0.7$ e com $a_t \sim N(0,1)$. Foram geradas 150 observações e as primeiras 50 foram eliminadas. O tamanho da amostra é então, n = 100 e o modelo estimado é $(1-0.78B)Z_t = a_t$, com $\hat{s}_R = 0.82$. Os autores contaminaram esta série com os seguintes outliers: um AO em t =13 de magnitude 5, um AO em t = 15 de magnitude -5, um LC em t = 91 de magnitude 4 e um AO em t= 92 de magnitude 3.

A figura 4.2 mostra a série observada.

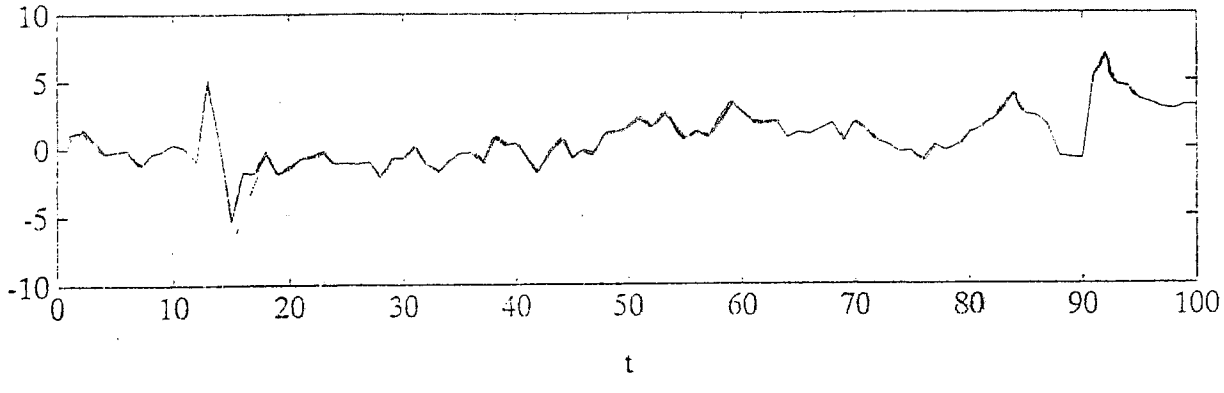

Figura 4.2

O procedimento proposto é descrito do seguinte modo:

Na etapa I, o valor máximo da medida de influência para um LC, ocorre em t =91. Estima-se o modelo de intervenção, resultando que o efeito (4.1714) é significativo. Então remove-se este efeito, definindo-se a série ajustada $\{Z_t^*\}$ e recalcula-se a medida $DL(t)$. Como, o efeito estimado da observação seleccionada não é significativo, calcula-se a medida de influência individual DZ^* em todos os pontos t. Seleccionam-se os 10% maiores valores de DZ^* e removem-se os efeitos das observações correspondentes, tratando-as como outliers aditivos.

A tabela 4.8 mostra os valores seleccionados e os tempos de ocorrência dessas observações.

T	12	13	15	16	18	28	37	42	84	92
DZ^*	11.3	27.6	20.3	1.7	2.1	1.7	1.7	1.7	2.2	4.0

Tabela 4.8

No fim desta etapa a estimativa do parâmetro autoregressivo é $\hat{\phi} = 0.75$.

A etapa II tem início com a série observada e com a estimativa do parâmetro, obtida na etapa I. Calculam-se os resíduos e as estatísticas $\hat{\lambda}_{A,t}, \hat{\lambda}_{I,t}$ e $\hat{\lambda}_{L,t}$ para t = 1,...,n.

As estatísticas $|\hat{\lambda}_{A,t}|$ e $|\hat{\lambda}_{I,t}|$ são comparadas em todos os pontos t e de entre todas, selecciona-se a que tem valor máximo. A seguir compara-se este valor com o valor $C_1 = 3.0$. Seguidamente compara-se o valor máximo de $|\hat{\lambda}_{L,t}|$ com o valor crítico $C_2 = 3.0$.

Através destas comparações, o procedimento detecta iterativamente, um AO em t = 13, um LC em t = 91, um AO em t = 92 e um AO em t = 15. Os outliers aditivos e o LC são corrigidos iterativamente, e a detecção é reeniciada com a série ajustada e com os parâmetros obtidos na etapa I. Como não se detectam mais outliers, prossegue-se para a etapa III.

Na etapa III, estima-se o parâmetro autoregressivo do modelo, utilizando a última série ajustada e estimam-se conjuntamente os efeitos dos três outliers aditivos e do LC. O modelo final é:

$$(1-\phi_1 B)\left(\tilde{z}_t - w_{13} I_t^{13} - w_{15} I_t^{15} - w_{91} I_t^{91} - w_{92} I_t^{92}\right) = a_t$$

A tabela 4.9 mostra os parâmetros estimados e o erro padrão residual.

Tipo de parâmetro	valor	$\hat{s}_R$
Autoregressivo	0.81	
Efeito do AO em t = 13	5.51	
Efeito do AO em t = 15	-4.66	
Efeito do LC em t = 91	5.15	
Efeito do AO em t = 92	2.02	
		0.81

Tabela 4.9

Seguidamente, removem-se os efeitos destes outliers das observações obtendo-se a série ajustada. A etapa II é reeniciada, com a série ajustada e com o parâmetro estimado na etapa III, mas como não são detectados mais outliers, o procedimento é concluído.

5. Conclusão

Os procedimentos apresentados neste trabalho foram desenvolvidos pelos diversos autores tendo como objectivo a modelação linear de uma série temporal na presença de outliers. Estes procedimentos contribuem para aumentar a eficácia da análise estatística de situações reais, pois permitem identificar as observações responsáveis por alterações da estrutura da série e que necessitam de uma análise mais aprofundada.

Existem abordagens alternativas ao problema dos outliers em séries temporais. Por exemplo, a formulação de Harvey (1981), que se baseia em modelos estruturais utilizando a representação de espaço de estados, com componentes não observados. Denby e Martin (1979), Bustos e Yohai (1986), consideraram métodos robustos de estimação de parâmetros.

Um outro domínio bastante importante, é a previsão em séries temporais na presença de outliers, sendo ainda reduzidos os estudos nesta área. Por exemplo, Chen e Liu (1993), utilizaram o procedimento iterativo de estimação conjunta dos parâmetros do modelo e dos efeitos dos outliers, na etapa de previsão.

Em situações práticas de análise estatística de uma série temporal, tem vindo a impor-se o recurso cada vez mais frequente a procedimentos específicos de tratamento de outliers. A intencionalidade subjacente à análise, os resultados teóricos e o software disponíveis e a própria educação estatística do analista, justificam a preferência de determinada abordagem em detrimento de outras. Por outro lado, a multiplicidade de abordagens possíveis ao estudo do problema dos outliers em séries temporais, o que explica em parte a inexistência de uma estrutura unificadora dos procedimentos e resultados teóricos associados, mantém a necessidade de contribuições futuras e motiva a continuidade do estudo deste tema.

Importa ainda reconhecer, que o software de apoio à maior parte dos procedimentos disponíveis é insuficiente e que o seu desenvolvimento também será neste caso, deveras útil como complemento ao progresso do estudo de outliers em séries temporais.

Bibliografia

Balke (1993) - "Detecting level shifts in time series", Journal of Business & Economic Statistics, 11, 81-92.

Barnett e Lewis (1980) - "Outliers in Statistical Data", Wiley.

Bento J. F. Murteira (1993) – "Análise de Sucessões Cronológicas", McGraw-Hill.

Bruce e Martin (1989) - "Leave-k-out diagnostics for time series", Journal of the Royal Statistical Society, Ser.B, 51, 363-424.

Chen e Liu (1993) - "Joint estimation of model parameters and outliers efects in time series", Journal of the American Statistical Association, 88, 284-297.

Chen e Liu (1993) - "Forecasting time series with outliers", Journal of Forecasting,12, 13-35.

Fox (1972) - "Outliers in time series", Journal of the Royal Statistical Society, Ser. B, 34, 350-363.

Peña (1990) - "Influential observations in time series", Journal of Business & Economic Statistics, 8, 235-241.

Sánchez e Peña (1997) - "The identification of multiple outliers in ARIMA models", Universidade Carlos III de Madrid, Statistics and Econometrics Ser. 27, 1-27.

Tsay (1986) – "Time Series models specification in the presence of outliers", Journal of the American Statistical Association, 81, 132-141.

Tsay (1988) – "Outliers, level shifts, and variance changes in time series", Journal of Forecasting, 7, 1-20.