응용통계연구 제 9권 1호
1996년 3월 pp. 125-137.

시계열에서의 연속이상치가 예측에 미치는 영향

이 재 준[1], 편 영 숙[2]

요 약

시계열 자료는 흔히 반복되지 않는 비정상적인 사건의 영향으로 이상치를 포함한다. 시계열 자료는 관측치들 사이에 종속구조를 갖기 때문에, 이상치의 영향은 다른 통계적 분석에서 보다 더 심각할 수 있다. 본 논문에서는 연속이상치가 예측에 미치는 영향을 파악하는 데에 촛점을 두었다. 특히, l 시점 후 예측오차의 평균제곱의 증가량을 유도하고, 이 증가량으로 연속이상치가 예측에 미치는 영향을 측정하였다. 일반적으로, 연속이상치가 예측원점에서 아주 가까운 시점에서 발생하지 않았으면 그 증가량은 크지 않음을 밝히고, 실제 자료를 분석하여 확인하였다

1. 서론

시간의 흐름에 따라 관측되는 시계열자료에서 자료 입력과정의 오류에서 뿐만 아니라 기술혁신, 파업, 유류파동등 정치, 경제적 사건에 의해 이상치(outliers)가 발생할 수 있고, 특히 그러한 이상자료는 흔히 연속시점에서 발생하는 특성을 나타낸다고 알려져 있다. 일반적으로 이상치는 통계적 분석과정에서 모형의 선택, 모수의 추정, 검정, 예측과 같은 추론과정에 심각한 영향을 미칠 수 있고, 이러한 이상치의 영향을 고려하지 않은 분석방법을 적용한 경우에 잘못된 결론으로 귀결될 수 있다. 따라서 보다 정확한 추론을 위해, 이상치가 발생한 시점과 이유를 파악하고 그 영향을 추론과정에 반영할 수 있는 분석방법이 요구된다.

ARIMA모형을 이용한 시계열 자료의 분석에서 이상치는 추론과정에 심각한 영향을 끼칠 수 있고, 특히 연속시점에서 발생한 연속이상치(patchy outliers: PO)는 그 영향이 더욱 심각하다고 알려져 있다(Bruce와 Martin, 1989 ; Lee, 1990). 시계열 자료에서 이상치에 관한 연구는 Fox(1972)가 AO, IO 두 종류의 단일이상치(single outlier)를 소개한 이래, 한 시점에서 발생한 단일이상치의 탐지방법이나 이상치가 모수의 추정에 미치는 영향(Abraham등, 1979 ; Chang등, 1988 ; Tsay, 1986, 1988 ; Pena, 1990)과 연속이상치의 탐지방법(Bruce와 Martin, 1989; Lee, 1990 ; Cho등, 1993)등이 주를 이루었다.

ARIMA모형에서 이상치가 예측에 미치는 영향에 관한 연구는 단일이상치가 발생한 경우에 대하여, 이상치의 종류(type: AO, IO, TC, LS), 발생시점, 이상영향의 크기에 따른 예측 대상시점별 예측값과 예측오차 분산의 추정량 및 예측구간의 크기 등에 미치는 이상영향의 연구 결과를 들 수 있다(Ledolter, 1988 ; Chen과 Liu, 1993). 그러나, 실제 시계열 자료에서는 이상치가

1) (402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과 부교수.
2) (402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과 석사과정.

연속시점에서 발생하는 경우가 흔한 편이고(Martin과 Yohai, 1986), 연속시점에서 이상치가 발생한 자료의 분석에서 단일이상치만을 고려한 분석방법을 적용하는 것은 적절하지 않을 수 있다.(Bruce와 Martin, 1989 ; Lee, 1990). 본 논문에서는 이상치가 연속시점에서 발생한 경우에 대하여 그러한 이상자료가 예측에 미치는 영향의 크기와 특성을 밝히고, 실제 자료를 이용하여 그러한 연속이상치를 고려한 예측방법과 고려하지 않은 예측방법의 적용 결과를 비교하였다.

2. 연속이상치가 예측에 미치는 영향

참시계열(outlier-free series) Z_t는 주기가 s이고 차수가 $(p, d, q)\times(P, D, Q)_s$인 계절 ARIMA모형(Seasonal Autoregressive Integrated Moving Average)을 따른다고 하자.

$$\phi(B)(1-B)^d\Phi(B^s)(1-B^s)^D Z_t = \theta(B)\Theta(B^s)a_t, \tag{1}$$

$$\phi(B) = 1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p, \quad \theta(B) = 1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$$

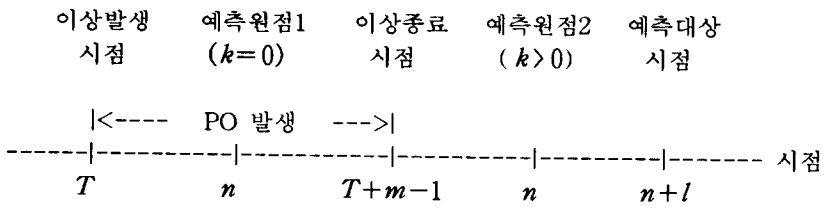
$$\Phi(B) = 1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_P B^{Ps}, \quad \Theta(B) = 1-\Theta_1 B^s-\Theta_2 B^{2s}-\cdots-\Theta_Q B^{Qs}.$$

단, B는 후향연산자($B^m X_t = X_{t-m}$), d와 D는 각각 0을 포함한 양의 정수, $\{a_t\}$는 평균이 0, 분산이 σ^2인 백색잡음과정이고, AR과 MA연산자는 각각 정상성과 가역성을 만족하며 같은 근이 없다고 가정한다. 또한, $\psi(B) = 1+\psi_1 B+\psi_2 B^2+\cdots$ 와 $\pi(B) = 1-\pi_1 B-\pi_2 B^2-\cdots$ 는 식 (2)로 부터 각각 정의된다.

$$(1-B)^d(1-B^s)^D\phi(B)\Phi(B^s)\psi(B) = \theta(B)\Theta(B^s) \tag{2}$$

$$\pi(B) = \phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D/\theta(B)\Theta(B^s)$$

본 논문에서는 시점을 나타내는 기호로서, T는 이상치 발생 시작시점, n은 예측원점(forecast origin), l 은 예측시차(lead time)이고, PO 발생시 m은 PO의 길이, k는 이상종료시점에서 예측원점까지의 시차(lags)로 정의한다. 예로, 예측원점 이전에 AO가 발생한 경우에 $n = T+k$ 가 되고, 예측원점이 PO발생이 종료된 후인 경우에는 $n = T+m-1+k$ 가 된다.



2.1 단일이상치의 영향

시점 T에서 단일이상치 발생시 이상치를 포함한 관측시계열(observed TS; Y_t)는 참시계열 Z_t에 이상 영향이 반영된 개입모형(intervention model; Box등, 1975)으로 표현될 수 있다.

$$Y_t = Z_t + \omega L(B) I_t(T), \qquad t = 1, 2, \cdots, n \tag{3}$$

단, $L(B)$는 T이후 시점들에 미치는 이상영향의 형태(type), ω는 발생시점 T에서의 이상의 크기(magnitude), $I_t(T)$는 $t = T$에서 1이고 $t \neq T$에서 0인 지시함수이다. $L(B)$는 이상치의 종류(AO,IO,LS,TC)에 따라 각각 1, $\phi(B)$, $1/(1-B)$, $1/(1-\delta B)$로 표현된다.

단일이상치가 T시점에서 발생한 경우에 이상치의 종류 $L(B)$에 따라 그 이후 시점의 관측자료에 구조화된 형태의 영향을 미치게 된다. 따라서, 예측원점 n에서 $n+l$ 시점의 이상영향의 크기를 구체적으로 표현할 수 있어 Z_{n+l} 뿐만 아니라 이상영향이 포함된 Y_{n+l}이 모두 예측대상에 해당된다. ARIMA모형에서 모형의 모수, 단일이상치의 종류와 크기를 알고 있는 경우에 l 시점 후(예측시차가 l)의 예측값과 예측오차는 각각

$$Z_n(l) = \sum_{j=1}^{\infty} \pi_j^{(l)} Z_{n-j+1} \ , \tag{4}$$

$$Y_n(l) = Z_n(l) + \omega L(B) I_{n+l}(T) \tag{5}$$

$$e_n(l) = Y_{n+l} - Y_n(l) = Z_{n+l} - Z_n(l) = \sum_{i=0}^{l-1} \psi_i a_{n+l-i} \tag{6}$$

가 되고, 식(4)에서 $\pi_j^{(l)}$은 다음의 식(7)과 같이 정의된다.

$$\pi_j^{(l)} = \pi_{j+l-1} + \sum_{h=1}^{l-1} \pi_h \pi_j^{(l-h)}, \quad \pi_j^{(1)} = \pi_j \tag{7}$$

이상치의 종류 $L(B)$는 알지만 이상치의 크기 ω와 모형의 모수 $\boldsymbol{\beta} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta})$를 모르는 경우, 자료로 부터 추정된 $\hat{\omega}$과 $\hat{\boldsymbol{\beta}}$을 (4)-(6)에 대입하여 구해지는 $n+l$ 시점의 예측치와 예측오차는 추정에 따른 오차로 인해 식(8)과 같이 표현된다.

$$\hat{Y}_n(l) = \hat{Z}_n(l) + \hat{\omega} \hat{L}(B) I_{n+l}(T) \tag{8}$$

$$\begin{aligned} \hat{e}_n(l) &= Y_{n+l} - \hat{Y}_n(l) \\ &= Z_{n+l} - \hat{Z}_n(l) - \{\omega L(B) I_{n+l}(T) - \hat{\omega} \hat{L}(B) I_{n+l}(T)\} \end{aligned}$$

Ledolter(1988)는 ARIMA모형에서, AO가 관측자료 Y_t로 구한 예측값, 예측오차 분산의 추

정량 및 예측구간에 미치는 영향을 유도하였다. Chen과 Liu(1993)는 단일이상치의 각 종류에 대해 그러한 이상치가 예측에 미치는 영향과 이상치의 종류를 잘못 판단했을 때 발생하는 예측 편의(bias)의 크기를 밝혔다.

2.2 연속이상치의 영향

시점 T에서 $T+m-1$ 까지 m개 시점에서 정형화되지 않은 구조를 갖는 연속이상치(PO)가 발생한 경우, 관측시계열 Y_t는 식(3)의 확장된 개입모형으로 식(9)와 같이 표현될 수 있다.

$$Y_t = Z_t + \sum_{j=0}^{m-1} \omega_j I_t(T+j), \quad t=1, 2, \cdots, n \tag{9}$$

단, ω_j는 $T+j$ 시점의 이상영향의 크기이고, $I_t(T)$는 식(3)에서 정의된 지시함수이다. PO가 예측에 미치는 영향에 관한 문제는 특정 구조를 가정하지 않는 PO의 특성에 따라, Y_{n+l}에 포함될 이상영향을 구체적으로 표현할 수 없고, 따라서 예측대상은 참시계열 Z_{n+l}에 국한된다.

식(1)을 따르는 예측대상의 $n+l$ 시점 시계열값은 식(2)에서 정의된 $\pi(B)$를 이용하여

$$\pi(B) Z_{n+l} = a_{n+l}$$

로 표현되며, 관측된 Y_t, $t=1, 2, \cdots, n$들로 식(10)과 같이 표현될 수 있다(Pyun, 1995 참조).

$$Z_{n+l} = \sum_{j=1}^{\infty} \pi_j^{(l)} Y_{n-j+1} + \sum_{j=0}^{l-1} \psi_j a_{n+l-j} - \sum_{j=0}^{m-1} \omega_j \pi_{m+k-j}^{(l)} \tag{10}$$

$$\text{단, } \pi_j^{(l)} = \sum_{i=0}^{l-1} \pi_{l-1+j-i} \psi_i \ , \ \psi_j = \sum_{i=0}^{j-1} \pi_{j-i} \psi_i \ . \tag{11}$$

식(10)의 오른쪽 세 번째항 $-\sum_{j=0}^{m-1} \omega_j \pi_{m+k-j}^{(l)}$ 는 관측시계열에서 PO의 영향을 제거하는 항이다. 참시계열 $Z_t(t=1, 2, \cdots, n)$ 또는 PO의 영향이 제거된 시계열로 구한 $n+l$ 시점의 예측값과 예측오차 $Z_n(l)$, $e_n(l)$은 식(4)와 (6)과 같고, PO의 영향을 제거하지 않고 $n+l$ 시점의 시계열값을 예측하는 경우의 예측값과 예측오차 $Z_n(l)_Y$, $e_n(l)_Y$는

$$Z_n(l)_Y = \sum_{j=1}^{\infty} \pi_j^{(l)} Y_{n-j+1} \tag{12}$$

$$e_n(l)_Y = Z_{n+l} - Z_n(l)_Y = \sum_{i=0}^{l-1} \psi_i a_{n+l-i} - \sum_{j=0}^{m-1} \omega_j \pi_{m+k-j}^{(l)}$$

가 된다. 따라서, PO의 영향을 제거하지 않은 관측시계열로 l시점 후를 예측한 경우에 예측오차의 평균제곱오차(Mean Squared Forecast Error: MSFE)와 PO에 의한 MSFE의 상대적 증가분 IMSFE는 다음과 같이 표현된다.

$$MSFE(l;k)=\sigma^2\sum_{i=0}^{l-1}\psi_i^2+\left(\sum_{j=0}^{m-1}\omega_j\pi_{m+k-j}^{(l)}\right)^2 \tag{13}$$

$$IMSFE(l;k)=\frac{\left(\sum_{j=0}^{m-1}\omega_j\pi_{m+k-j}^{(l)}\right)^2}{\sigma^2\sum_{i=0}^{l-1}\psi_i^2}$$

대표적인 ARIMA모형에 대해 PO가 예측값과 예측오차 분산에 미치는 영향을 살펴본다(단, 예측원점이 PO 발생 중인 경우에 $k=0$).

예 1 : AR(1) 모형

(a) $m=1$인 경우(single AO) :

Z_t가 AR(1)을 따르는 경우에 $\pi_1=\phi$, $\pi_j=0$, $j\geq 2$이고 $\psi_j=\phi^j$, $j\geq 1$이므로, 식(11)에 의해 $\pi_1^{(l)}=\phi^l$, $\pi_j^{(l)}=0, j>1$가 된다. 따라서, 식(12)로 부터 예측원점에서 이상치가 발생할 경우($k=0$)에만 $-\omega_0\phi^l$만큼의 예측오차가 발생하고, 그 외의 경우($k\neq 0$)에는 영향을 미치지 않음을 알 수 있다. 식(13)에 의해, AO가 예측 원점에서 발생하였을 때 IMSFE는 다음과 같이 표현되며, 이 결과는 Ledolter(1988)의 결과(pp. 4의 식(9))와 같다.

$$IMSFE(l;0)=\frac{\phi^{2l}(1-\phi^2)}{1-\phi^{2l}}\cdot\left(\frac{\omega_0}{\sigma}\right)^2$$

(b) $m\geq 2$인 경우(PO) :

식(12)의 오른쪽 두번째항은

$$\sum_{j=0}^{m-1}\omega_j\pi_{m+k-j}^{(l)}=\sum_{j=0}^{m-1}\left(\sum_{i=0}^{l-1}\pi_{l-1+m+k-j-i}\psi_i\right)\omega_j\ ,\ l\geq 1,\ k\geq 0$$

가 되어 예측오차는 m에 무관하게 $k=0$일 때만 $\omega_{m-1}\phi^l$이고 그 외의 경우 0이며, 오차의 크기는 예측시차(l)이 길어질 수록 급격히 줄어들게 된다. 즉, PO가 예측원점에서 발생했을 때 그 영향을 받고, 예측원점에서 발생한 것이 아니면 영향을 받지 않는다. 예측원점이 PO 종료시점인 경우에 PO가 $MSFE$에 미치는 영향은 식(13)으로 부터 $\omega_{m-1}^2\phi^{2l}$가 되고, 따라서 $MSFE(l;0)$의 상대적인 증가량은

$$IMSFE(l;0)=(\omega_{m-1}/\sigma)^2\phi^{2l}/\sum_{i=0}^{l-1}\psi_i^2$$

가 된다. 모수 ϕ, 예측시차 l, 예측원점의 이상영향의 크기 $\omega_{m-1}(k=0)$에 따른 $IMSFE$를 비교하면, ϕ의 값이 클수록 $IMSFE$는 커짐을 알 수 있다(표 1 참조).

표 1. *IMSFE* $(l\ ;\ 0)(\%)$

	$\omega_{m-1}=3\sigma$			$\omega_{m-1}=5\sigma$		
ϕ \ l	1	2	3	1	2	3
$\phi=0.1$	9	0.09	0.00	25	0.25	0.00
$\phi=0.3$	81	6.69	0.60	225	18.58	1.67
$\phi=0.5$	225	45.00	10.71	625	125.00	29.76
$\phi=0.7$	441	145.03	61.20	1225	402.85	170.00
$\phi=0.9$	729	326.24	193.95	2025	906.22	538.75

예 2 : IMA(1,1) 모형

(a) $m=1$인 경우(single AO) :

이 경우에 $\pi_j=\theta^{j-1}(1-\theta),\ j\geq 1,\ \psi_j=1-\theta,\ j\geq 1$이고, 모든 예측시차 l에 대하여 $\pi_{k+1}^{(l)}=(1-\theta)\theta^k$ 가 된다. 따라서, 식(12)의 오른쪽 두번째항은 $\omega(1-\theta)\theta^k$ 가 되고 *IMSFE*는 다음과 같이 표현된다.

$$IMSFE(\,l;k)=\frac{(1-\theta)^2\theta^{2k}}{1+(l-1)(1-\theta)^2}\cdot\left(\frac{\omega_0}{\sigma}\right)^2$$

(b) $m\geq 2$인 경우(PO) :

길이가 m인 PO가 발생한 경우에 (12)의 오른쪽 두번째항과 *IMSFE*는 다음과 같다.

$$\sum_{j=0}^{m-1}\omega_j\pi_{m+k-j}^{(l)}=(1-\theta)\sum_{j=0}^{m-1}\theta^{m+k-j-1}\omega_j \tag{14}$$

$$IMSFE(l;k)=(1-\theta)^2\{\sum_{j=0}^{m-1}\omega_j\theta^{m+k-j-1}\}^2/\{1+(l-1)(1-\theta)^2\}\sigma^2$$

표 2와 3은 $m=2,\ \omega_j=3\sigma$인 경우와 $m=3,\ \omega_j=3\sigma$인 경우에 대하여 각각 *IMSFE*(%)를 요약한 것으로, 다음과 같은 사실을 확인할 수 있다.

i) k가 증가할 수록 *IMSFE*는 급격하게 감소한다. 즉, PO가 예측원점에서 멀리 떨어져 있을수록 PO의 영향은 급격하게 감소한다(표 2, 3 참조).

ii) l 이 증가할 수록 *IMSFE*값은 감소한다.

iii) θ가 0.1인 경우에는 $k=0$일때는 매우 큰 *IMSFE*값을 가지지만 k가 증가할 수록 급격하게 감소하여 k가 5이상이 되면 PO는 영향을 미치지 않는다.

iv) θ가 0.9인 경우에는 θ가 0.1인 경우와 비교하여 *IMSFE*가 상당히 작다. 그러나, k가 증가할수록 *IMSFE*의 값은 감소하지만 감소하는 정도가 θ가 0.1인 경우에 비해 느리다.

표 2 $m=2$, $\omega_j=3\sigma$일때 $IMSFE(\%)$

k	l \ θ	θ = 0.1	θ = 0.3	θ = 0.5	θ = 0.7	θ = 0.9
0	1	882.09	745.29	506.25	234.09	32.49
	2	487.34	500.19	405.00	214.76	32.17
	3	336.68	376.41	337.50	198.38	31.85
1	1	8.82	67.08	126.56	114.70	26.32
	2	4.87	45.02	101.25	105.23	26.06
	3	3.37	33.88	84.38	97.21	25.81
2	1	0.09	6.04	31.64	56.21	21.32
	2	0.05	4.05	25.31	51.56	21.11
	3	0.03	3.05	21.09	47.63	20.90
5	1	0.00	0.00	0.49	6.61	11.33
	2	0.00	0.00	0.40	6.07	11.22
	3	0.00	0.00	0.33	5.60	11.11
10	1	0.00	0.00	0.00	0.19	3.95
	2	0.00	0.00	0.00	0.17	3.91
	3	0.00	0.00	0.00	0.16	3.87

표 3 $m=3$, $\omega_j=3\sigma$일때 $IMSFE(\%)$

k	l \ θ	θ = 0.1	θ = 0.3	θ = 0.5	θ = 0.7	θ = 0.9
0	1	898.20	852.06	689.06	388.48	66.10
	2	496.24	571.85	551.25	356.41	65.44
	3	342.82	430.33	459.38	329.22	64.80
1	1	8.98	76.69	172.27	190.36	53.54
	2	4.96	51.47	137.81	174.64	53.01
	3	3.43	38.73	114.84	161.32	52.49
2	1	0.09	6.90	43.07	93.28	43.37
	2	0.05	4.63	34.45	85.57	42.94
	3	0.03	3.49	28.71	79.05	42.52
5	1	0.00	0.01	0.67	10.97	23.05
	2	0.00	0.00	0.54	10.07	22.82
	3	0.00	0.00	0.45	9.30	22.59
10	1	0.00	0.00	0.00	0.31	8.04
	2	0.00	0.00	0.00	0.28	7.96
	3	0.00	0.00	0.00	0.27	7.88

3. 사례분석

3.1 분석 방법

이 자료는 80년 1월 부터 95년 1월 까지 통계월보에 수록된 181개 월별 건축허가면적 자료로서, 분석과정에서는 제곱근 변환자료(그림 1)이 이용되었고, 모수의 추정과 예측에는 Windows용 SCA V4.2 소프트웨어가 사용되었다. 그림 1에서 보여지듯이, 89년 2월 부터 91년 8월 사이(110-140 번째)의 자료는 주택 200만호 건설로 인해 다른 기간의 자료와 차이가 남을 알 수 있다. 이 기간 동안의 자료를 PO라 보고 PO가 예측에 미치는 영향을 분석하였다.

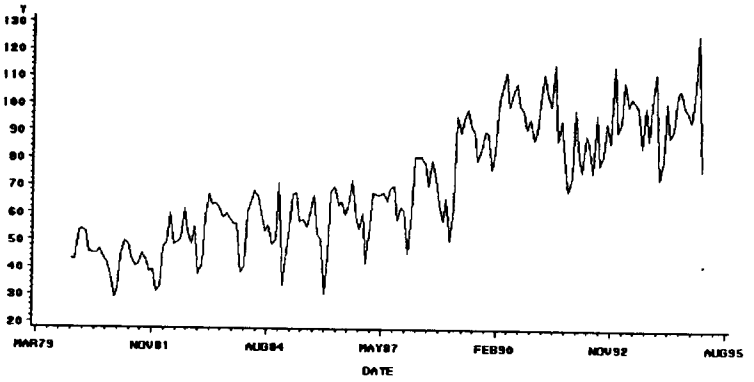


그림 1 제곱근 변환된 건축허가면적자료의 시계열 도표

본 분석에서는 1) 예측시점(T), 2) 예측구간 ($n+1$ 에서 $n+l$), 3) 예측방법, 4) 적용모형 등 비교 요인들과 적용 자료를 다음과 같이 구분하여 비교한다. 단, 적용 모형은 ① 모형화 과정에 이용된 자료의 시구간, ② 이용된 자료에서 PO의 제거 여부에 따라(원자료, 수정자료) 다르게 선택되고 적합되었으며, 그 결과는 부록 1과 같다.

1) 예측 시점	① T=130(PO발생 중)	② T=140(PO종료 직후)	③ T=144(PO종료 후)
2) 예측 구간	① 향후 12 개월(n+1 에서 n+12)	② 자료 보유기간 까지(n+1 에서 181)	
3) 예측 방법	① 한시점앞 예측(l=1)	② 여러 시점앞 예측	
4) 적용 모형	모형 1 PO발생 직전까지(1-109 번째)의 원자료를 이용한 모형선택과 모수추정	모형 2 예측원점까지(1에서 n=130, 140, 또는 144)의 원자료를 이용한 모형선택과 모수추정	모형 3 모형 1의 결과를 이용한 PO 탐지, 자료에서 PO를 제거한 수정자료를 이용해 모형선택과 모수추정

3.2 분석 결과

본 분석에서는 대상 모형과 적용 자료중의 조합에서 다음의 3 경우를 비교하였다.

① 모형 1의 결과를 원자료를 적용하여 예측
② 모형 2의 결과를 원자료를 적용하여 예측
③ 모형 3의 결과를 수정자료를 적용하여 예측

또한, 예측의 정확도를 측정하는 기준으로 *MSE*, *MAPE*, *MAE*를 고려하였고, 다음의 3 경우의 예측 구간과 방법에 대해 각 예측시점별로 비교하였다(표 4-6 참조).

① 예측원점에서 자료 보유 최종 시점(181)까지 한 시점앞(one-step-ahead) 예측
② 예측원점에서 향후 12개월을 한 시점앞(one-step-ahead) 예측
③ 예측원점에서 향후 12개월을 여러 시점앞(multi-step-ahead) 예측

표 4-6의 분석 결과로 부터 다음과 같은 사실을 확인할 수 있다.

i) 예측시점별 분석에서, 예측시점이 PO발생 중인 경우($T=130$) 수정된 자료의 예측결과가 모든 기준과 예측방법·예측구간의 조합에서 가장 나쁘다. 예측시점이 PO종료 직후인 경우($T=140$)에는 모든 기준에서 수정된 자료로 선택한 모형과 수정된 자료를 이용한 예측이 가장 좋은 결과를 보이며, 예측시점이 PO 종료 후인 경우($T=144$)에도 수정된 자료로 선택한 모형과 수정된 자료를 이용한 예측이 가장 좋은 결과를 보이지만, 그 정도가 PO종료 직후 보다 약하다(표 7의 여러 시점앞 예측결과 제외시).

표 4 예측원점에서 95년 1월까지 한 시점앞 예측결과

경우	T=130			T=140			T=144		
	MSE	MAPE	MAE	MSE	MAPE	MAE	MSE	MAPE	MAE
모형1,원자료	150.44	10.84	9.92	154.20	11.14	10.06	162.05	11.13	10.20
모형2,원자료	162.57	11.02	10.04	175.52	11.59	10.45	185.74	11.48	10.56
모형3,수정자료	157.04	11.00	10.17	153.77	100.39	9.70	161.17	10.36	9.84

표 5 예측원점에서 향후 12개월의 한 시점앞 예측결과

경우	T=130			T=140			T=144		
	MSE	MAPE	MAE	MSE	MAPE	MAE	MSE	MAPE	MAE
모형1,원자료	128.43	9.91	9.40	216.55	15.00	12.37	264.78	14.85	13.19
모형2,원자료	105.33	8.78	8.20	229.30	14.72	12.08	305.12	15.01	13.65
모형3,수정자료	165.09	11.70	11.25	204.70	13.09	11.21	257.60	13.36	12.27

표 6 예측원점에서 향후 12개월의 여러 시점앞 예측결과

경우	T=130			T=140			T=144		
	MSE	MAPE	MAE	MSE	MAPE	MAE	MSE	MAPE	MAE
모형1,원자료	172.38	11.73	11.38	237.44	16.62	13.53	310.47	17.13	15.10
모형2,원자료	246.23	13.62	12.40	417.28	23.39	18.68	323.41	15.05	13.89
모형3,수정자료	333.01	16.22	16.02	174.59	12.07	10.31	263.50	13.70	12.59

ii) PO발생 중에($T=130$) 한 시점앞 예측에서, 대부분의 기준에서 수정자료를 이용한 예측결과가 원자료를 이용한 결과 보다 부정확하고, 특히, 예측구간이 짧을수록 PO의 영향이 더 심각하다(표 4와 5). 또한, 모형 2와 원자료를 사용한 예측시 예측구간이 짧은 경우(표 5)에 모든 기준에서 가장 좋은 결과를 보이지만, 예측구간이 길어지면(표 4) 대부분의 기준에서 가장 나쁜 결과가 나타남을 알 수 있다.

iii) 예측시점이 PO발생 직후($T=140$)와 PO종료 4개월 후($T=144$)인 경우의 한 시점앞 예측에서 모형 3과 수정자료를 이용한 예측결과가 모든 기준과 예측시점에서 우월하나 그 정도는 ii)에 비하여 미약한 편이며(표 4, 5), 특히 예측시점이 PO종료 시점으로 부터 멀어질 수록 그 정도는 약해진다. 또한, 모형 2와 원자료의 사용시 대부분의 경우에서 가장 부정확한 예측결과를 보이고 있다. 즉, 모형선택 과정과 모수추정 과정에서 PO의 영향을 무시한 경우에 예측의 정확도는 상대적으로 낮아짐을 알 수 있다.

iv) 표 6의 여러 시점앞 예측결과를 비교하면, 예측시점이 PO발생 중($T=130$)인 때에는 수정자료와 모형 3을 이용한 예측결과가 한시점앞 예측에서 보다 훨씬 PO의 영향을 많이 받지만 예측시점이 PO발생 후인 경우에는 수정자료와 모형 3을 이용한 예측결과가 모형 1과 원자료를 이용한 예측결과보다 정확함을 알 수 있다.

4. 결론과 토의

본 논문은 시계열 자료에서 PO가 예측에 미치는 영향을 연구한 결과로서, Ledolter(1988)와 Chen과 Lieu(1993)에서와 같이 모형, $\boldsymbol{\beta}$, 그리고 $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_{m-1})'$ 를 안다고 가정하였다. PO가 모형의 선택과정에서 영향을 미칠 수 있다는 사실을 고려할 때, 최초 이상치 이전의 자료를 이용하여 모형을 선택하거나, 적절한 모형과 이상치 탐지방법을 적용하여 그러한 이상치의 영향을 제거한 수정자료를 이용하는 방법이 고려될 수 있을 것이다(Tsay, 1988 ; Lee, 1990).

제한된 범위에서 다음과 같은 결론을 생각할 수 있다.

i) 모형의 선택과 모수의 추정과정에서 PO의 영향이 고려된 방법이 적용될 때, 보다 정확한 예측결과를 기대할 수 있다.

ii) 예측시점이 PO발생 중인 경우에, 원자료를 이용한 예측이 수정자료를 이용한 예측보다 더 정확한 예측결과를 제공한다.

iii) 예측시점이 PO종료 후인 경우에, PO를 고려한 모형선택과 수정자료 이용시 보다 정확한 예측을 기대할 수 있다.

iv) 한 시점앞 예측시 예측원점이 PO종료 시점에서 멀어질수록 PO가 예측에 미치는 영향은 급격히 작아지고, 원자료와 수정자료를 이용한 결과의 차이는 급격히 줄어든다.

v) 여러 시점앞 예측시에는 PO의 영향이 더 심각하며, 예측시점이 PO종료 시점으로 부터 멀어지드라도 한 시점앞 예측에서 보다 PO의 영향이 더 크다고 할 수 있다.

참고문헌

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

[2] Box, G. E. P. and Tiao, G. C.(1975). Intervention Analysis with Applications to Economic and Environmental Problems, *Journal of American Statistical Association,* Vol. 70, 70-79.

[3] Box, G. E. P. and Jenkins, G.(1976). *Time Series Analysis Forecasting and Control.* 2nd ed., Holden-Day, San Francisco.

[4] Bruce, A. G. and Martin, R. D.(1989). Leave-k-out Diagnostics for Time Series(with discussion), *Journal of the Royal Statistical Society Series B,* Vol. 51, 363-424.

[5] Chang, I., G. C. Tiao, and Chen, C.(1988). Estimation of Time Series Parameters in the Presence of Outliers, *Technometrics,* Vol. 30, 193-204.

[6] Chen, C. and Liu, L. M.(1993). Forecasting Time Series with Outliers, *Journal of Forecasting,* Vol. 12, 13-35.

[7] Cho, Sinsup, Ryu, Gui Yeol, Park, Byeong Uk, and Lee, Jae June(1993). Outlier Detection Diagnostic based on Interpolation Method in Autoregressive Models, Journal of the Korean Statistical Society, Vol. 22, 283-306.

[8] Fox, A. J. (1972). Outliers in Time Series, *Journal of the Royal Statistical Society Series B,* Vol. 32, 337-645.

[9] Ledolter, J. (1988). The effect of Additive Outliers on the Forecasts from ARIMA Models, *International Journal of Forecasting,* Vol. 5, 231-240.

[10] Lee, J. J. (1990). A Study on Influential Observations in Linear Regression and Time Series, *Unpublished Ph. D. dissertation, University of Wisconsin, Dept. of Statistics.*

[11] Martin, R. D. and Yohai, V. J. (1986). Influential Functionals for Time Series(with discussion), *Annals of Statistics,* Vol. 14, 781-818.

[12] Pena, D.(1990). Influential Observations in Time Series, Journal of Business & Economic Statistics, Vol. 8, 235-241.

[13] Pyun, Y. S. (1995). The Effect of Patchy Outliers on the Time Series Forecasting, *Master Thesis, Inha University, Dept. of Statistics.*

[14] Tsay, R.S. (1986). Time Series Model Specification in the Presence of Outliers, *Journal of American Statistical Association,* Vol. 81, 131-141.

[15] ———— (1988). Outliers, Level Shifts, and Variance Changes in Time Series, *Journal of Forecasting,* Vol. 7, 1-20.

[16] Wegman, E. J.(1986). Another Look at Box-Jenkins Forecasting Procedure, *Communication in Statistics(B) Simulation and Computation,* Vol. 15, 523-530.

[17] Wei, W. S.(1990). *Time Series Analysis Univariate and Multivariate Methods.* Addison Wesley.

부 록 1

(1) 모형 1 : $ARIMA(1,0,0)\times(0,1,1)_{12}$

$$(1-\underset{(0.06843)}{0.51014B})(1-B^{12})\sqrt{Z_t}=\underset{(0.46595)}{1.42294}+(1-\underset{(0.11242)}{0.69954B^{12}})a_t$$

(2) 모형 2 : $ARIMA(0,1,1)\times(0,1,1)_{12}$

① 예측원점(130) : $(1-B)(1-B)^{12}\sqrt{Z_t}=(1-\underset{(0.07695)}{0.58915B})(1-\underset{(0.09093)}{0.63907B^{12}})a_t$

② 예측원점(140) : $(1-B)(1-B)^{12}\sqrt{Z_t}=(1-\underset{(0.07267)}{0.62410B})(1-\underset{(0.08723)}{0.71168B^{12}})a_t$

③ 예측원점(144) : $(1-B)(1-B)^{12}\sqrt{Z_t}=(1-\underset{(0.07803)}{0.52957B})(1-\underset{(0.09161)}{0.69448B^{12}})a_t$

(3) 모형 3 : $ARIMA(1,0,0)\times(0,1,1)_{12}$

① 예측원점(130) : $(1-\underset{(0.07869)}{0.52989B})(1-B^{12})\sqrt{Z_t}=\underset{(0.41002)}{1.58918}+(1-\underset{(0.09342)}{0.72532B^{12}})a_t$

② 예측원점(140) : $(1-\underset{(0.08066)}{0.43554B})(1-B^{12})\sqrt{Z_t}=\underset{(0.28512)}{1.93614}+(1-\underset{(0.10326)}{0.82075B^{12}})a_t$

③ 예측원점(144) : $(1-\underset{(0.07823)}{0.43733B})(1-B^{12})\sqrt{Z_t}=\underset{(0.26307)}{1.90325}+(1-\underset{(0.10630)}{0.84179B^{12}})a_t$

The Effect of Patchy Outliers in Time Series Forecasting

Jae June Lee[1], Young Sook Pyun[2]

Abstract

Time series data are often contaminated with outliers due to the influence of unusal and non-repetitive events. The effect of the outliers is larger in the time series analysis than in the other statistical analysis, because the time series data have dependent structure over time. This paper focuses on the effect of patchy outliers on forecasting. Especially, the increase of the mean square of the l -step-ahead forecast error is derived and used to evaluate the impact of those outliers on the forecast. We find, in general, that this increase is rather small, provided that the patchy outliers does not occur too close to the forecast origin.

1) Associate Professor, Department of Statistics, INHA University, # 253 Yonghyun-Dong, Nam-Ku, Incheon, 402-751, Korea.

2) M.S. Student, Department of Statistics, INHA University, # 253 Yonghyun-Dong, Nam-Ku, Incheon, 402-751, Korea.