Journal of the Korean Data & Information Science Society
2011, **22**(3), 449–458

한국데이터정보과학회지

R을 이용한 이상점 탐지 알고리즘의 구현

송규문[1] · 문지은[2] · 박철용[3]

[123]계명대학교 통계학과

접수 2011년 4월 8일, 수정 2011년 5월 2일, 게재확정 2011년 5월 9일

요 약

불법 오물 투기는 정부가 당면한 시급한 문제들 중의 하나이다. 최근 들어 관련기관들은 실시간으로 연속적으로 수질의 상태를 감지 할 수 있는 화학적 산소요구량 자동측정기를 강과 하천 등에 설치하고 있다. 본 논문에서는 시계열 간섭모형을 이용하여 화학적 산소요구량 자동측정기로부터 발생하는 데이터를 분석하여 투기시점이라고 여겨지는 이상점을 탐지하는 알고리즘을 R언어를 이용하여 구현한다. R을 이용한 알고리즘을 통해 단계별 계산에서 수동 작업을 피할 수 있기 때문에 알고리즘의 자동화를 달성할 수 있고, 한 단계 더 나아가 모의실험에서 사용될 수 있을 것이다.

주요용어: 시계열 간섭모형, 이상점 탐지, 화학적 산소요구량.

1. 서론

급변하는 정보화 사회를 맞이하여 많은 시계열 자료들이 방대하게 쏟아져 나오고 있다. 이런 시계열 자료들은 항상 똑같은 패턴으로 움직이는 것들도 있지만, 외부의 영향으로 인해 그 패턴에서 벗어나는 양상을 보이기도 한다. 경제 분야의 예를 살펴보면 IMF 같은 경제 쇼크가 온다면 같은 패턴으로 진행되던 KOSPI 자료에 갑작스런 이상점들이 발생하게 된다. 의학 분야의 예로는 심전도 (ECG; electrocardiogram) 자료의 경우 주기적인 패턴이 일정하게 반복되다가 심장에 이상이 생기면 갑자기 주기의 패턴이 바뀌거나 이상점이 나타나는 양상을 보이게 될 것이다. 사회 분야의 예로는 택시 이용 승객의 수가 일정한 패턴을 유지하다가 택시 요금이 인상되면 일시적으로 택시 이용 승객 수가 급격히 떨어지는 양상을 보이기도 한다. 또한 환경 분야의 예로는 공업 폐기물 같은 오물 투기가 발생하면 수질을 측정하는 화학적 산소요구량 (COD; chemical oxygen demand) 혹은 생물학적 산소요구량 (BOD; biochemical oxygen demand) 값이 일상적인 패턴에서 벗어나 갑작스럽게 이상점이 나타나게 된다.

이처럼 다양한 분야에서 발생하는 많은 시계열 자료에서 급속한 큰 변화 (외부 효과)를 빠르게 탐지할 수 있다면, 신속한 대처가 가능해져 그 급격한 변화로 생길 수 있는 손실을 회피할 수 있을 것이다. 다시 말해 시계열 자료들이 같은 패턴을 지속하다가 갑자기 그 패턴에서 벗어나는 큰 변화를 보일 때 그것을 빠르게 탐지할 수 있다면, 외부의 영향 (개입)이 있는지 조사하여 즉각적인 대책을 마련할 수 있을 것이다.

이 논문에서는 자기회귀누적이동평균 (ARIMA; autoregressive integrated moving average) 모형에 외부의 영향 (개입)이 포함될 수 있는 시계열 간섭 모형으로 COD 값을 모형화하여 실시간으로 이상점

[1] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수.
[2] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 박사 과정생.
[3] 교신저자: (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수.
E-mail: cypark1@kmu.ac.kr

을 탐지하는 알고리즘을 R언어로 구현한다. 이 알고리즘은 시간의 흐름에 따라 들어오는 COD 자료의 잔차를 분석하면서 갑작스런 이상점이 발생하면 경보를 발령하게 된다.

SAS를 이용하여 자료 분석을 할 수 있는데도 불구하고 R언어를 통해 실시간 이상점 탐지 알고리즘을 구현하려고 하는 것은 알고리즘의 자동화가 쉽기 때문이다. 다시 말해 R을 이용한 알고리즘을 통해 단계별 계산에서 필요로 하는 수동 작업을 쉽게 회피할 수 있기 때문에 알고리즘의 자동화를 달성할 수 있다. 알고리즘의 자동화를 달성한다면 한 단계 더 나아가 궁극적으로 필요한 모의실험을 할 수 있는 기초를 마련할 수 있다. 또한 R은 공개소프트웨어이기 때문에 R로 구현된 알고리즘은 누구나 추가비용 없이 접근할 수 있는 커다란 장점도 존재한다.

본 논문에서는 2005년 6월 1일부터 6월 30일까지의 매 2시간 간격으로 서울 근교 A하천의 COD 자동 측정 장치로부터 관찰된 240개의 시계열 자료를 분석 대상으로 한다. 이 COD 자료를 대상으로 시계열 간섭 모형을 이용하여 이상점이라 여겨지는 변화시점을 탐지 할 수 있는 알고리즘을 R언어를 통해 구현한다.

본 논문의 구성은 다음과 같다. 2절에서는 SAS를 이용하여 실시간 이상점 탐지 알고리즘의 과정을 단계별로 설명한 후, 3절에서는 R을 이용하여 알고리즘을 구현하는 과정을 단계별로 설명한다. 마지막으로 4절에서는 이 연구의 간단한 요약과 함께 결론을 맺는다.

2. SAS를 이용한 자료 분석

본 논문에서는 2005년 6월 1일부터 6월 30일까지의 매 2시간 간격으로 서울 근교 A하천의 COD 자동 측정 장치로부터 관찰된 240개의 시계열 자료를 분석 대상으로 하였다. 이 240개의 자료 $Z_1, \cdots, Z_{240}$ 중 처음 200개의 자료 $Z_1, \cdots, Z_{200}$을 사용하여 시계열 모형을 적합 시키고, 남아있는 $Z_{201}, \cdots, Z_{240}$ 까지 40개 자료는 최종 모형 테스트를 위한 자료로 남겨둔다.

표 2.1 COD자료의 사용

훈련용 자료	테스트용 자료
$Z_1, \cdots, Z_{240}$	$Z_{201}, \cdots, Z_{240}$

이 COD 자료의 시도표는 그림 2.1과 같다.

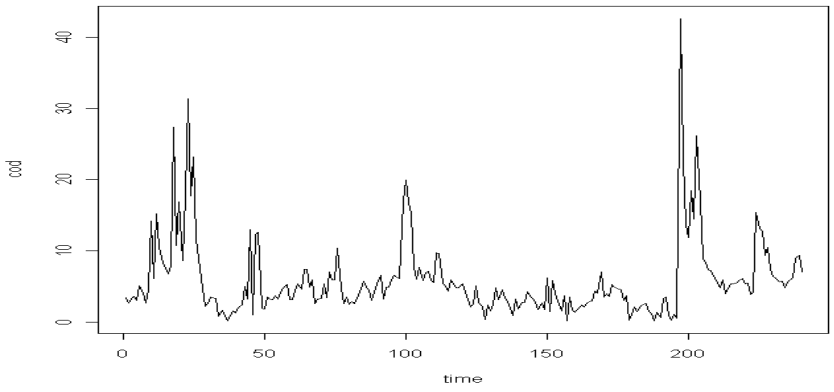

그림 2.1 240개 COD 자료의 시도표

먼저 COD 자료 Z의 모형을 살펴보기로 한다. $\{Z_t : i = 1, 2, 3, \cdots\}$를 COD 관측값들의 수열이라고 나타낼 때, 이것은 다음과 같은 시계열 간섭모형을 따른다고 가정할 수 있다 (Choi 등, 2007; 문지은 등, 2010).

$$\begin{aligned} Z_t &= N_t + f_1(I_{R_{t_{11}}}(t), \cdots I_{R_{t_{1p}}}(t)) + f_2(I_{D_{t_{21}}}(t), \cdots I_{D_{t_{2q}}}(t)) + \epsilon_t \\ &= N_t + f_1(t) + f_2(t) + \epsilon_t. \end{aligned}$$

여기에서 N_t는 (0이 아닌 평균을 포함할 수 있는) 정상 시계열이며, ϵ_t는 무상관 오차이다. 또한 $f_1(t)$은 자연적인 변화인 강우효과를 나타내는 함수로서 $t_{11}, \cdots, t_{1p}$ 시점에서 강우가 내린 것을 의미하며, $f_2(t)$는 인위적인 변화인 투기효과를 나타내는 함수로서 $t_{21}, \cdots, t_{2q}$ 시점에서 투기가 일어난 것을 의미한다. 구체적으로 지시함수 (indicator function)는 다음과 같이 정의된다.

$$I_{R_{t_{1i}}}(t) = \begin{cases} 1, & \text{만약 } t = t_{1i} \\ 0, & \text{그 외}, \end{cases}$$

$$I_{D_{t_{2i}}}(t) = \begin{cases} 1, & \text{만약 } t = t_{2i} \\ 0, & \text{그 외}. \end{cases}$$

강우효과를 나타내는 함수 $f_1(t)$를 더욱 단순화 시켜 강우 시에 동일한 효과를 미친다고 가정하고 다음과 같이 나타내기로 한다.

$$f_1(I_{R_{t_{11}}}(t), \cdots I_{R_{t_{1p}}}(t)) = \alpha I_R(t),$$

$$I_R(t) = \begin{cases} 1, & \text{만약 } t = t_{1i},\ i = 1, 2, \cdots, p \\ 0, & \text{그 외}. \end{cases}$$

이제 SAS를 통해 200개 학습용 COD 자료를 이용하여 앞의 시계열 간섭 모형에 적합하는 과정을 설명한다. 우선 학습용 COD 자료의 자기상관함수 (ACF; autocorrelation function)와 편자기상관함수 (PACF; partial autocorrelation function)의 값을 살펴보면 각각 그림 2.2와 그림 2.3과 같다.

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	29.683387	1.00000	|********************|	0
1	16.909480	0.56966	|***********. |	0.070711
2	13.605928	0.45837	|*********. |	0.090803
3	9.493294	0.31982	|******. |	0.101716
4	6.102507	0.20559	|****. |	0.106626
5	7.005214	0.23600	|*****. |	0.108589
6	5.368416	0.18086	|****. |	0.111124
7	3.664710	0.12346	|**. . |	0.112586
8	3.172797	0.10689	|**. . |	0.113261
9	1.444721	0.04867	|*. . |	0.113765
10	1.141851	0.03847	|*. . |	0.113869
11	0.448079	0.01510	|. . |	0.113934
12	0.064582	0.00218	|. . |	0.113944
13	0.173445	0.00584	|. . |	0.113944
14	-1.508057	-.05080	|. *. |	0.113945
15	-1.998959	-.06734	|. *. |	0.114059
16	-2.946303	-.09926	|. **. |	0.114257
17	-3.167713	-.10672	|. **. |	0.114688
18	-3.572996	-.12037	|. **. |	0.115183
19	-2.855705	-.09621	|. **. |	0.115810
20	-1.762790	-.05939	|. *. |	0.116209
21	-1.917314	-.06459	|. *. |	0.116361
22	-0.383329	-.01291	|. . |	0.116540
23	-0.049767	-.00168	|. . |	0.116547
24	0.284757	0.00959	|. . |	0.116547

그림 2.2 COD의 자기상관함수 (ACF)

그림 2.2의 자기상관함수가 지수적으로 감소하는 모습을 보이고 그림 2.3의 편자기상관함수에서 lag3인 지점에서 절단이 된 모양을 나타내므로, 이 자료는 AR(2)모형을 따르리라는 것을 짐작할 수 있다.

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.56966	. |********** .
2	0.19816	. |**** .
3	-0.00375	. | . .
4	-0.04656	. * | . .
5	0.14447	. |*** .
6	0.00352	. | . .
7	-0.05972	. * | . .
8	0.01387	. | . .
9	-0.02479	. | . .
10	-0.00719	. | . .
11	-0.02190	. | . .
12	0.00183	. | . .
13	0.00749	. | . .
14	-0.07847	.** | . .
15	-0.03138	. * | . .
16	-0.03888	. * | . .
17	-0.01324	. | . .
18	-0.04966	. * | . .
19	0.02950	. | * . .
20	0.05098	. | * . .
21	-0.02729	. * | . .
22	0.05892	. | * . .
23	0.02748	. | * . .
24	0.00889	. | . .

그림 2.3 COD의 편자기상관함수 (PACF)

다음으로 COD 자료 중에서 비가 온 시점 (t=43, 44, 45, 46, 49)은 이미 알고 있기 때문에 비온 시점을 지시변수를 이용하여 비가 온 시점에서는 1, 비가 오지 않은 시점에서는 0으로 나타낸다 (이정형과 조신섭, 1997).

$$I_R(t) = \begin{cases} 1, & \text{만약 } t = 43, 44, 45, 46, 49 \\ 0, & \text{그 외.} \end{cases}$$

다음으로 강우효과를 포함하고 정상 시계열 부분을 AR(2)로 상정한 후 이상점이라고 판단되는 시점을 탐지하였다. SAS에 의한 이상점으로 탐지된 시점은 다음과 같다.

표 2.2 이상점으로 탐지된 시점

시점	타입	추정값	χ^2	$p > \chi^2$
197.0000	Additive	32.9521	451.50	<0.0001
18.0000	Additive	17.6010	1280.81	<0.0001
23.0000	Additive	15.2004	96.07	<0.0001
198.0000	Additive	13.3559	74.17	<0.0001

이상점이라고 판단되는 시점은 t=18, 23, 197, 198로서 4군데 나왔으나 t=198시점은 앞의 t=197시점의 그 후의 영향이라고 판단하여 t=18, 23, 197시점을 이상점으로 간주한다. 다시 말해 $t_{21} = 18$, $t_{22} = 23$, $t_{23} = 197$가 되며 투기 시점이라고 판단되는 세 시점을 강우효과와 마찬가지로 지시함수 $D_{t_{21}}(t), D_{t_{22}}(t), D_{t_{23}}(t)$로 나타낼 수 있다.

강우 시점과 투기 시점을 고려하여 최종 선택된 시계열 간섭 모형의 모수 추정값을 구한 것이 표 2.3에 주어져 있다. 따라서 최종적으로 선택된 시계열 간섭 모형의 구체적인 모양은 다음과 같다.

$$Z_t = 5.12 + 32.44 I_{D_{197}}(t) + 14.49 I_{D_{18}}(t) + 17.53 I_{D_{23}}(t) - 3.94 I_R(t) + \frac{e_t}{I - 0.58B - 0.19B^2}.$$

여기서 e_t는 ϵ_t의 실현값인 잔차이며, B는 시계열 Z_t에 대해 $BZ_t = Z_{t-1}$를 만족하는 후방이동 연산자 (back-shift operator)이다.

표 2.3 최종 선택된 시계열 간섭 모형의 모수 추정값

모수	추정값	표준편차	t 값	$p > \|t\|$	lag	변수
평균	5.1208	0.91054	5.17	<0.0001	0	COD
AR1,1	0.5853	0.07262	7.21	<0.0001	1	COD
AR1,2	0.1912	0.07309	3.83	0.0002	2	COD
NUM1	32.4444	2.58280	15.68	<0.0001	0	AO197
NUM2	14.4861	0.35869	6.22	<0.0001	0	AO23
NUM3	17.5346	0.30713	7.44	<0.0001	0	AO18
NUM4	-3.9366	1.50150	-2.50	0.0131	0	RAIN

마지막으로 모형의 검증을 위하여 40개의 테스트용 자료를 이용한다. 정상 시계열과 강우효과만으로 이루어진 간섭모형을 이용하여 SAS프로그램에서 40개의 예측값을 계산한다. 실제 COD자료에서 이 예측값을 뺀 잔차를 계산한다. 이 잔차를 표준화시킨 뒤 40개의 잔차가 표준정규분포의 95%분위수를 넘어가는 시점을 이상점으로 간주하여 투기가 있었다고 경고한다.

SAS에 의한 결과는 다음과 같다.

표 2.4 40개 테스트용 자료 중 이상점으로 판단되는 시점

시점	잔차	표준화된 잔차
201	9.5209	3.6140
203	13.0853	4.9589
224	11.0627	4.1958

다시 말해 t=201에서 t=240까지의 테스트 자료 중 투기가 있었다고 여겨지는 이상점은 t=201, 203, 224이다. 실제 COD 자료의 시도표와 비교해 보면 시점 t=203과 t=224에서 COD 값이 갑자기 증가하는 것을 알 수 있다. 하지만 시점 t=201에서 경고를 준 이유는 시점 t=197에서 투기된 효과가 그 후에도 영향을 미친 결과가 아닐까 생각된다.

3. R을 이용한 알고리즘 구축

2절에서 사용한 동일한 COD 자료를 훈련용 자료와 테스트용 자료로 각각 200개, 40개씩 분리하여 R을 사용하여 분석한다.

구현하고자 하는 R 프로그램의 단계는 다음과 같다.

[단계1] 원자료의 시도표 확인 후 240개의 자료에 이미 알고 있는 강우 시점을 지시함수로 지정한다. 처음 200개의 자료에 대한 강우효과를 포함한 정상 시계열 모형의 차수를 추정한 후 이상점을 찾아낸다. 정상 시계열의 AR 모형 차수와 이상점을 자동적으로 찾아주는 R 함수를 이용하여 모형 선택의 자동으로 선택할 수 있도록 한다.

[단계2] 이상점으로탐지된 투기 시점을 지시변수로 두고 최종 모형의 모수를 추정한다.

[단계3] 정상 시계열과 강우효과만을 고려한 모형을 이용하여 40개의 잔차를 계산한다.

[단계4] 잔차가 표준정규분포의 95%분위수를 넘어가는 시점을 이상점으로 판단한다.

자료가 들어오면 가장 먼저 시도표를 이용하여 개략적인 자료의 구조를 파악한다. 자료를 읽어 들이고 COD 자료의 시계열 도표를 그리는 R코드는 다음과 같다.

```
library("TSA")
```

```
y<-scan("C:/thesis/cod.txt")
plot(y, pch="*", type="l", xlab="time", ylab="cod")
```

첫 번째는 이상점 탐지에 필요한 프로그램을 포함하는 패키지를 불러오는 명령어이다. 여기서는 패키지인 "TSA"를 이용한다. 두 번째는 경로를 지정하여 자료를 읽어 들여 y에 두는 명령어로서 240개의 자료가 y에 저장된다. 세 번째는 시도표를 그리는 명령어이다. plot 명령어를 이용하여 옵션으로 점의 모양, 시도표의 각 축의 이름 등을 넣을 수 있다.

그 다음 단계로 강우 시점을 이미 알기 때문에 강우 시점 (t=43, 44, 45, 46, 49)을 지시함수로서 정해주는데 그 R코드는 다음과 같다.

```
rain<-rep(0,240)
rain[c(43,44,45,46,49)]<-1
```

첫 번째는 rain에 0이라는 숫자를 240개만 생성하라는 명령어이다. 두 번째는 1부터 200개의 자료 중 43, 44, 45, 46, 49번째 시점에는 1로 바꾸라는 명령어이다. 이렇게 함으로써 43, 44, 45, 46, 49번째 시점에는 1, 나머지 시점에는 0이 들어가게 되어 강우 시점에 대한 지시함수가 된다.

총 240개의 자료 중 훈련용 자료인 200개의 자료만 모형 설정에 사용하고 남은 40개는 테스트를 하기 위하여 남겨둔다. 이것을 위해 240개의 자료가 다 들어가 있는 y에서 1에서 200까지의 자료는 y200에 두고, 201번째부터 240번째 자료는 y40에 둔다. 마찬가지로 rain 자료도 분리한다. 이 작업을 수행하는 R코드는 다음과 같다.

```
y200<-y[1:200]
y40<-y[201:240]
rain200<-rain[1:200]
rain40<-rain[201:240]
```

다음으로 훈련용 자료 y200을 이용하여 적절한 모형을 선택하여 선택된 모형의 모수를 추정한다. 먼저 자기상관함수와 편자기상관함수를 그리는 R코드이다.

```
acf(y200, lag.max=36)
pacf(y200,lag.max=36)
```

그림 2.2와 그림 2.3을 살펴보면 자기상관함수가 지수적으로 감소하고 편자기상관함수가 lag3에서 절단된 양상을 보이므로 자료가 차수가 2인 자기상관 모형 AR(2)를 따르리라는 것을 짐작할 수 있다. (구체적인 그림은 그림 2.2에서 그림 2.3에서 살펴보았기 때문에 여기서는 생략한다.)

자료를 강우효과를 포함한 AR(2) 모형에 적합하는 R코드는 다음과 같다

```
y.rain<-arima(y200, order=c(2,0,0), xreg=data.frame(rain200))
```

그런데 앞의 절차는 수동적인 방법으로 차수를 결정하는 것이기 때문에 자동화된 알고리즘을 위해서는 AR 모형의 차수를 자동적으로 결정해 줄 필요가 있다. AR 모형의 차수를 자동 탐지하기 위해서는 다음과 같은 두 단계의 R코드를 사용할 수 있을 것이다. (다른 패키지에서 AR 모형이 아닌 ARIMA 모형에서 여러 정보기준 (information criterion)에 의해 최적의 차수를 찾아주는 함수가 존재하나, 여기서는 TSA 패키지 내에서 사용할 수 있는 함수로 한정하기로 한다.)

```
ar.order <- ar(y200, aic=TRUE)$order
y.rain <- arima(y200, order(ar.order,0,0), xreg=data.frame(rain200)
```

그 다음 단계로 강우효과를 포함한 AR(2) 모형에서 벗어나는 이상점을 자동으로 탐지해 내는 R코드이다.

```
detAO <- detectAO(y.rain)
detIO <- detectIO(y.rain)
```

여기서 detectAO와 detectIO는 각각 가법적 이상점 (AO; additive outlier)과 혁신적 이상점 (IO; innovative outlier)을 탐지하는 명령어이다. 구체적으로 가법적 이상점과 혁신적 이상점은 다음과 같이 정의된다 (Chang 등, 1988; Cryer와 Chan, 2006).

$$AO : z_t = \phi^{-1}(B)\theta(B)\epsilon_t + wI_t^T,$$
$$IO : z_t = \phi^{-1}(B)\theta(B)(\epsilon_t + wI_t^T).$$

여기서 $\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$, $\theta(B) = (1 - \theta_1 B - \cdots - \theta_q B^q)$, 그리고

$$I_t^T = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}$$

이다. 만약 $w > 0$이면 시점 T에 특이치가 있다는 것을 의미한다. 그러므로 이상점의 문제는 이상점의 발생 시기 T와 그 효과 w의 결정 문제가 된다 (박유성과 김기환, 2002).

함수 detectAO와 detectIO에 의해 탐지된 이상점은 각각 표 3.1과 표 3.2와 같다.

표 3.1 가법적 이상점 (AO)의 시점

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
IND	18	21	23	45	196	197
LAMBDA2	7.2244	-3.6699	6.2391	-3.8077	-6.7689	13.5217

표 3.2 혁신적 이상점 (IO)의 시점

	[,1]	[,2]	[,3]	[,4]
IND	18	23	99	197
LAMBDA2	7.508259	7.13866	4.038404	14.71888

그런데 잔차값이 음으로 크게 나타나는 것은 관심이 없기 때문에 이것을 제외하고 해당 시점을 정수값으로 저장하는 R코드는 다음과 같다.

```
indAO<-as.integer(detAO$ind[detAO$lam>0])
indIO<-as.integer(detIO$ind[detIO$lam>0])
```

자료에 따라서 혁신적 이상점과 가법적 이상점의 시점과 개수가 다르게 나타날 것이기 때문에, 어떤 자료가 들어와도 이상점을 자동으로 지정할 수 있는 방법이 필요하다. 먼저 이상점들의 시점을 묶어서 행렬로 만들어주는 R코드는 다음과 같다.

```
AO=matrix(0,200,length(indAO))
for(i in 1:length(indAO))AO[,i]<-1*(seq(200)==indAO[i])
IO=matrix(0,200,length(indIO))
for(i in 1:length(indIO))IO[,i]<-1*(seq(200)==indIO[i])
```

위 명령어는 AO와 IO 시점들을 묶어서 각각 하나의 행렬 형태로 만들어 주는 명령어이다. 이상점의 시점과 개수는 자료에 따라 항상 바뀌게 되는데 for 문장을 이용하면 자료의 길이나 개수에 상관없이 적용가능 하도록 만든 것이다.

투기 시점과 강우 시점을 고려한 모형의 추정값을 구하기 위해서 다음과 같은 R코드를 사용한다.

```
IO.list<-vector("list",length(indIO))
for(i in 1:length(indAO))IO.list[[i]]<-c(1,0)
y.dump200<-arimax(y200, order=c(ar.order,0,0), method="ML",
  xtransf=data.frame(IO), transfer=IO.list, xreg=data.frame(AO,rain200))
```

첫 두 줄의 명령어는 혁신적 이상점에 대해 일반적으로 모수 추정을 위해서 필요하다. IO.list가 c(1,0)이 length(indIO)개 반복되는 list(c(1,0), $\cdots$,c(1,0))와 동일하게 인식되지 못하는 문제가 발생하고 있는데, 이 경우를 위해 부득이 미리 여러 개의 IO.list.k (k=1, 2, $\cdots$)를 만들어 놓고 k = length(indIO)를 만족하는 IO.list.k를 선택하여 arimax 함수의 transfer= 옵션 뒤에다 지정하는 방법을 사용할 수 있을 것이다.

정상 시계열과 강우효과만 고려된 모형에서의 모수를 추정하기 위해서 다음과 같은 R코드를 이용할 수 있다.

```
y.rain200<-arima(y200, order=c(ar.order,0,0), method="ML", xreg=data.frame(rain200))
```

그 결과는 다음의 표 3.3에 주어져 있다.

표 3.3 정상 시계열과 강우효과만 고려된 모형의 모수 추정

	ar1	ar2	intercept	rain
	0.4639	0.2037	5.5000	-3.5311
s.e	0.0692	0.0693	0.9159	2.4303

마지막으로 정상 시계열과 강우효과만 고려된 추정식을 이용하여 40개의 잔차를 계산해 낸다.

```
e40<-arima(y40, order=c(ar.order,0,0), fixed=y.rain200$coef, xreg=data.frame(rain40))$resid
```

여기에서 fixed=y.rain200$coef라는 옵션을 사용함으로서 y.rain200의 계수 (coefficients)를 이용하여 모수 추정값을 고정 시킨 후 잔차를 계산하게 된다. 알고리즘의 성능을 평가하기 위하여 표준화된 잔차를 이용하여 표준정규분포의 95%분위수를 상회하면 이상점으로 판단하는 R코드는 다음과 같이 작성할 수 있다.

```
outlier<-((e40-mean(e40))/sd(y.rain200$resid)>qnorm(0.95))
```

결과적으로 투기가 발생했다고 탐색된 시점은 t=201, 203, 224로 나타났다. (e40에서 mean(e40)를 빼주었지만 개개의 시계열 값이 들어오고 이것에 대한 이상점 여부를 판단할 때는 mean(e40) 부분을 생략하면 될 것이다.) 이것은 2절에서 나온 SAS의 결과물과 동일한 결과이다.

4. 결론

본 논문의 알고리즘 원리를 종합적으로 정리해 보면 다음과 같다. COD 자료의 모형으로는 (0이 아닌 평균이 포함될 수 있는) 정상적인 수질자료와 강우효과 및 투기효과가 더해진 다음과 같은 가법모형이라고 가정한다.

$$Z_t = N_t + f_1(t) + f_2(t) + \epsilon_t.$$

먼저 수질 자료의 시도표를 이용하여 자료의 추세를 확인한다. 강우 시점은 이미 알기 때문에 그 시점을 지시함수로 포함시켜 준다. 그 다음 처음 200개로 구성된 훈련용 자료의 자기상관함수와 편자기상관함수를 통해 정상 시계열의 차수를 파악한다. 이렇게 파악된 차수를 가지는 강우효과가 포함된 정상 시계열 모형을 이용하여 이상점을 찾아낸다. 정상 시계열의 AR 차수를 결정하기 위해서 차수 자동선택 함수를 이용할 수 있다. 이상점이라고 판단되는 시점을 투기 시점으로 두고 강우효과를 포함한 최종 시계열 간섭 모형을 적합시킨다. 모형의 평가를 위해서 강우효과만이 포함된 최종 모형의 모수 추정값을 이용하여 40개 테스트용 자료의 잔차를 계산한다. 잔차를 표준화시킨 후 표준정규분포의 95%분위수를 넘어가는 시점에 경고를 발령한다.

테스트용 자료에 대한 이상점 탐지 결과를 살펴보면 SAS와 R의 결과가 동일하다. 하지만 SAS에서는 정상 시계열의 차수와 이상점 선택을 위한 자동화 작업이 상당히 어렵지만, R에서는 쉽게 자동화 작업이 가능하다. 따라서 R로 제대로 구현된 알고리즘을 이용하면 다양한 형태의 자료에 대해 쉽게 자동화된 모형 적합과 이상점 탐지를 행할 수 있다. 이 알고리즘은 한 단계 더 나아가 모의실험에서도 당연히 사용될 수 있으리라 생각된다.

본 논문에서는 COD 자료를 이용하여 알고리즘을 제공했지만 다양한 분야의 시계열 자료들에서도 이와 유사한 알고리즘이 적용될 수 있으리라 생각된다 (박인찬 등, 2009; 박철용과 김현일, 2009). 특히, 대구·경북 의료복합 단지조성으로 다양하게 수집되는 의학 시계열 자료의 해석과 이상점 탐지에도 이용될 수 있으리라 생각된다. 이 알고리즘의 성능을 좀 더 구체적으로 알아보기 위하여 다양한 상황 속에서의 모의실험이 필요하다. 구체적으로 어느 정도의 투기 양을 탐지하는지 알아보는 모의실험이 필요할 것이다. 또한 강우 양에 따라 COD 값이 영향을 받을 것이라 생각되는데, 이 부분을 분석 모형에 포함시킬 필요가 있을 것이다. 이 부분에 대한 연구는 추후연구과제로 남기기로 한다.

참고문헌

문지은, 송규문, 김태윤 (2010). 시계열간섭모형을 이용한 불법 오물 투기 실시간 탐지 알고리즘 구축. <한국데이터정보과학회지>, **21**, 883-890

박유성, 김기환 (2002). <SAS/ETS를 이용한 시계열 자료 분석 I>, 자유아카데미, 서울.

박인찬, 권오진, 김태윤 (2009), 시계열 모형을 이용한 주가지수 방향성 예측. <한국데이터정보과학회지>, **20**, 991-998.

박철용, 김현일 (2009). 최적 시계열 모형에 기초한 오존주의보 날짜 예측. <한국데이터정보과학회지>, **20**, 293-299.

이정형, 조신섭 (1997). <SAS/ETS를 이용한 경제 시계열 분석>, 자유아카데미, 서울.

Chang, I., Tiao, G. C. and Chen, C. (2008), Estimation of time series parameters in the presence of outliers. *Technometrics*, **30**, 193-204.

Choi, H. S., Song, G. M. and Kim, T. Y. (2007), A study on error detection algorithm of COD measurement machine. *Journal of the Korean & Data Information Science Society*, **18**, 847-857.

Cryer, J. D. and Chan, K. S. (2006). *Time series analysis with applications in R*, 2nd Edition, Springer, New York.

Journal of the Korean Data & Information Science Society
2011, **22**(3), 449–458

한국데이터정보과학회지

Realization of an outlier detection algorithm using R

Gyu Moon Song[1] · Ji Eun Moon[2] · Cheolyong Park[3]

[123]Department of Statistics, Keimyung University

Received 8 April 2011, revised 2 May 2011, accepted 9 May 2011

Abstract

Illegal waste dumping is one of the major problems that the government agency monitoring water quality has to face. Recently government agency installed COD (chemical oxygen demand) auto-monitering machines in river. In this article we provide an outlier detection algorithm using R based on the time series intervention model that detects some outlier values among those COD time series values generated from an auto-monitering machine. Through this algorithm using R, we can achieve an automatic algorithm that does not need manual intervention in each step, and that can further be used in simulation study.

Keywords: Chemical oxygen demand, outlier detection, time series intervention model.

[1] Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
[2] Ph. D. student, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
[3] Corresponding author: Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea. E-mail: cypark1@kmu.ac.kr