응용통계연구 제 **16**권 *1*호, *2003*년, *pp. 15-31*

생물학적 지표 자료의 탐색적 분석 : LAKE ONTARIO의 실측자료를 중심으로*

이기원[1)]

요 약

Lake Ontario에서 수년간 측정된 실제 생물학적 지표 자료의 각 변수에 대하여 관찰 시점의 불규칙성과 의존성을 고려한 탐색적 분석모형의 수립과정에 대하여 연구하였다. 이상점을 제거한 후 trend와 seasonal component를 수정한 선형모형으로부터 잔차를 계산하고 이로부터 variogram과 correlogram을 그려 보았다.

주요용어: *생물학적지표, variogram, correlogram*

1. 서론

수자원의 효율적 관리는 식수원 보호 및 농·공업용수의 원활한 공급과 관련하여 가장 시급하게 이루어져야 할 과제 중의 하나이다. 이러한 관점에서 수질오염 여부의 판단에 중요한 역할을 하는 각종 생물학적지표들을 짜임새 있게 수집하여 모니터할 필요가 있다. 이러한 자료들을 분석할 때 겪는 어려움들로는 피치 못할 사정으로 결측치가 생기거나 측정간격이 불규칙하여 기존의 분석방법을 그대로 적용할 수 없다는 점들이 대표적으로 꼽힌다.

본 연구에서는 이러한 특징을 갖고 있는 Lake Ontario의 실측 자료를 바탕으로 생물학적지표의 통계분석모형을 수립하는 전형적인 과정을 보이고 그중 Chloride 수준의 변화에 대하여 심층적인 탐색적 분석을 수행하고자 한다.

2. 자료 개요

Lake Ontario의 생물학적 지표 자료는 Station 41과 Station 81의 두 측정 장소에서 각각 445일(1981년 3월 19일부터 1995년 10월 23일까지) 및 441일(1981년 3월 16일부터 1995년 10월 20일까지) 씩 다음과 같은 변수들에 대하여 기록한 결과이다.

- Dates(in julian)
- Mixing Depth(m)

* 본 연구는 한국과학재단의 목적기초연구(R05-2001-000-00075-0)지원으로 수행되었음.

1) (200-702) 강원도 춘천시 옥천동 1번지, 한림대학교 수리정보과학부, 교수.
E-mail: kwlee@hallym.ac.kr

- Epilimnetic Temperature(℃)
- Total Phosporous
- NOx(㎎/ℓ)
- Soluble Reactive Silica(㎎/ℓ)
- Particulate Organic Carbon(㎎/ℓ)
- Chloride(㎎/ℓ)
- Chlorophyll A(㎍/ℓ)
- Total Phytoplankton Biomass(g/㎥)

이 자료를 분석하는 전형적인 과정에 대하여 S-PLUS 3.3을 이용하여 알아보도록 한다. 이 자료의 특징 중 하나는 측정일자 간격이 일정하지 않다는 점이다. 이는 다음과 같은 S-PLUS 작업을 통하여 알아볼 수 있다. 여기서 bio41.dates 와 bio81.dates 는 Dates 변수를 일-월-연도 로 나타낸 것이다.

```
> bio41.dates[1:10]
 [1] 19-Mar-1981 23-Mar-1981 31-Mar-1981 09-Apr-1981 16-Apr-1981
 [6] 22-Apr-1981 28-Apr-1981 05-May-1981 13-May-1981 21-May-1981
> bio81.dates[1:10]
 [1] 16-Mar-1981 24-Mar-1981 31-Mar-1981 08-Apr-1981 17-Apr-1981
 [6] 22-Apr-1981 28-Apr-1981 05-May-1981 12-May-1981 22-May-1981
```

매해 측정은 20-40 회 정도 이루어졌다. 이는 다음 작업으로 확인할 수 있다. 여기서 bio41.yr와 bio81.yr는 각 자료에서 연도를 추린 변수이다.

```
> table(bio41.yr)
 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
   38   36   35   31   32   29   28   26   28   27   29   27   30   24   25
> table(bio81.yr)
 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
   39   32   35   27   32   28   28   28   28   29   29   27   30   22   27
```

뿐만 아니라 호수가 얼어붙는 겨울에는 거의 측정이 이루어지지 않고 있었다. 이 또한 다음과 같은 작업으로 확인할 수 있다. 여기서 bio41.mon과 bio81.mon은 월을 추린 변수이다. 1월과 2월은 아예 측정이 이루어지지 않았고 11월, 12월, 3월의 측정회수도 매우 적은 편임을 알 수 있다.

```
> table(bio41.mon)
  3  4  5  6  7  8  9 10 11 12
 11 51 63 66 63 65 62 48 14  2
> table(bio81.mon)
 3  4  5  6  7  8  9 10 11 12
 7 51 63 67 61 64 60 53 13  2
```

연도별 월별 측정회수를 표시하면 다음과 같다. 1981년 첫 해에만 12월에도 측정이 있었으며 1985년 이후로는 11월에도 측정을 하지 않았음을 알 수 있다.

```
> table(bio41.yr,bio41.mon)
     3 4 5 6 7 8 9 10 11 12
1981 3 4 4 5 4 4 4  4  4  2
1982 3 3 4 5 4 5 4  4  4  0
1983 1 4 4 5 4 5 4  4  4  0
1984 1 3 5 4 5 4 4  3  2  0
1985 1 5 4 4 5 4 5  4  0  0
1986 0 4 4 4 5 4 5  3  0  0
1987 0 3 4 4 5 4 5  3  0  0
1988 0 3 4 5 4 5 4  1  0  0
1989 0 3 4 5 4 5 4  3  0  0
1990 0 3 5 4 4 5 4  2  0  0
1991 0 3 5 4 4 5 3  5  0  0
1992 1 3 4 4 5 3 4  3  0  0
1993 1 4 4 4 5 4 5  3  0  0
1994 0 3 4 4 2 4 4  3  0  0
1995 0 3 4 5 3 4 3  3  0  0

> table(bio81.yr,bio81.mon)
     3 4 5 6 7 8 9 10 11 12
1981 3 4 4 5 4 4 5  4  4  2
1982 1 1 4 5 4 5 4  4  4  0
1983 1 4 4 5 4 5 4  4  4  0
1984 1 3 5 4 5 4 2  2  1  0
1985 1 5 4 4 5 4 4  5  0  0
1986 0 4 4 4 5 3 5  3  0  0
1987 0 3 4 5 4 4 5  3  0  0
1988 0 3 4 5 4 5 4  3  0  0
1989 0 3 5 4 4 5 4  3  0  0
```

```
1990 0 3 5 4 5 4 4  4  0  0
1991 0 4 4 4 5 4 3  5  0  0
1992 0 4 4 5 4 3 4  3  0  0
1993 0 4 4 5 4 5 4  4  0  0
1994 0 2 4 3 2 4 4  3  0  0
1995 0 4 4 5 2 5 4  3  0  0
```

변수별로 파악한 결측치의 개수는 다음과 같다. 여기서 bio41.na와 bio81.na는 변수별 결측치의 개수를 나타낸다.

```
> bio41.na
 [1]   0   4  29  20  17  34  35  18  38 104
> bio81.na
 [1]   0   0  22  16  14  20  22  32  30 100
```

유난히 결측치의 개수가 많이 나타나고 있는 Bioplankton Biomass 변수의 경우 1993년 이후 거의 기록이 이루어지지 않았음을 그 다음 출력물로부터 알 수 있다.

```
> table(bio41.yr[bio41.dat[,10]=="NA"])
 1981 1982 1983 1985 1986 1987 1990 1991 1993 1994 1995
    7    7    3    1    3    2    1    1   30   24   25
> table(bio81.yr[bio81.dat[,10]=="NA"])
 1981 1982 1983 1985 1986 1989 1990 1993 1994 1995
    7    5    4    2    1    1    1   30   22   27
```

3. 탐색적 자료 분석 과정

먼저 자료에 등장하는 모든 변수들로 Pairwise Scatter Plot을 그리도록 한다. 이때 Dates 변수는 UNIX의 기준 캘린더에 따라 julian으로 변환된다. 이 결과로부터 연도별 추세에서 가장 흥미가 가는 변수를 하나 골라 보다 상세한 분석을 수행토록 한다. 그림 3.1은 Station41의 관측값들에 대하여 그린 Pairwise Scatter Diagram이다.

Station 81의 관측값들에 대하여 그린 Pairwise Scatter Diagram도 비슷한 양상을 보이기 때문에 이는 생략토록 한다. 이중 우리가 관심을 가지고 분석하고자 하는 변수는 Cl로 표시되어 있는 Chloride Level 이다. 탐색적 자료 분석은 다음과 같은 단계로 수행된다.

1. 먼저 자료의 전체적인 추세를 파악하기 위하여 Chloride Level의 산점도를 Station 별로 구분하여 그린다.
2. Trend와 Seasonal Component를 고려한 선형모형을 적합한다. 이 과정에서 Outlier를 제거한다. 잔차 도표(Residual Plot)과 Box Plot을 활용한다.
3. 마지막으로 trend와 seasonal variaion을 고려한 잔차에 대하여 variogram과 correlogram의 추정값을 그린다.

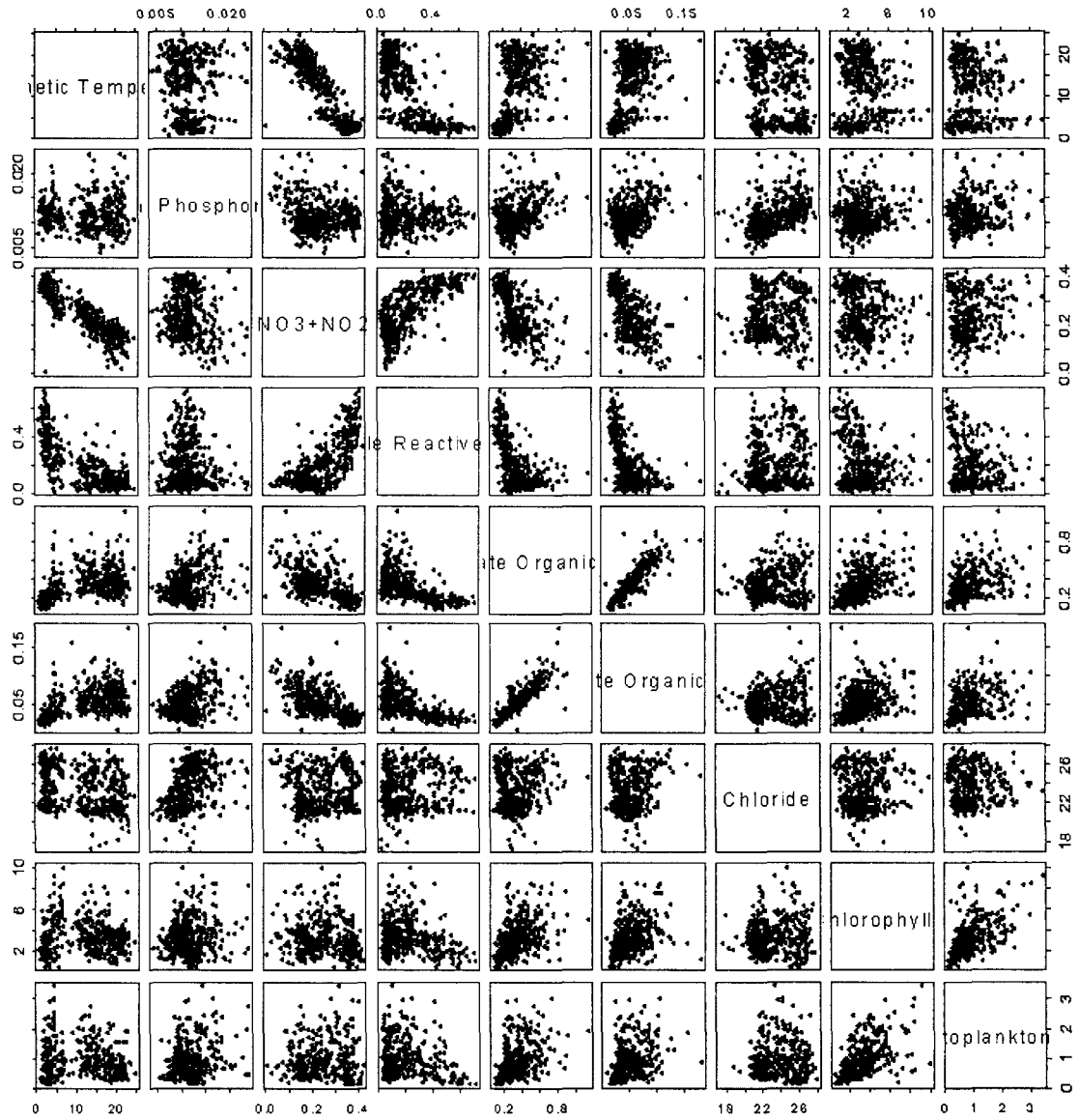

그림 3.1: Lake Ontario Bioindex Pairwise Scatter Diagram(Station41)

3.1. Chloride Level의 산점도와 추세 분석

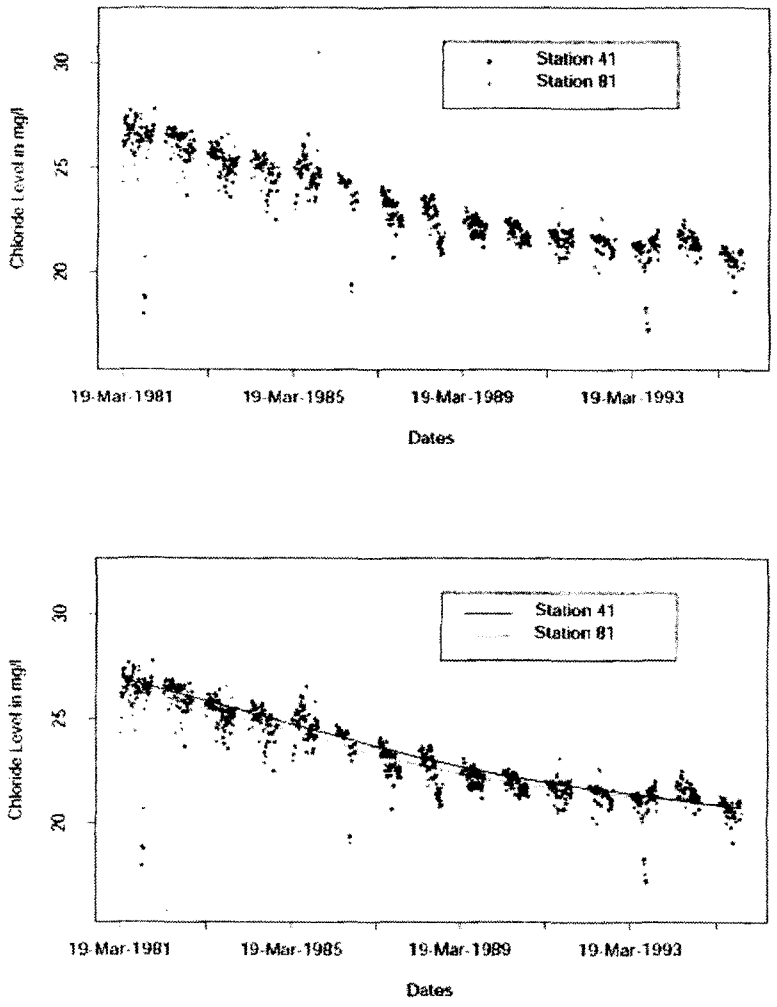

그림 3.2: Chloride Level의 산점도(상단). loess가 배경에 추가된 산점도(하단).

그림 3.2의 상단부는 Chloride Level에 대하여 그린 산점도이다. 검은 색으로 표현된 점들은 Station 41, 붉은 색으로 표현된 점들은 Stations 81에서 관찰된 값이다. Outlier들이 같은 시기에 각 Station에서 관찰됨을 알 수 있다. 하단부에는 loess를 이용하여 요약한 전반적 추세를 삽입하였다. Station 41이 전반적으로 Station 81에 비하여 높은 수준을 유지하다가 점차 그 차이가 소멸되어 감을 알 수 있다. 또한 그 추세는 선형에 약간의 2차적 요소가 반영되어 있는 것으로 판단할 수 있다. 다음 단계에서는 Outlier의 제거방법에 대하여 알아보도록 한다.

3.1.1. Trend의 산정과 Outlier의 제거

3.1에서 관찰된 Trend를 설명하기 위하여 Dates 변수를 설명변수로 하는 2차 선형모형을 적합한다. 적합한 결과는 다음과 같다. 여기서 cl41.narm과 cl81.narm은 각Station의 Chloride Level 수준 관찰값에서 결측치를 제외한 것이다. S-PLUS의 glm()을 이용하여 적합하였다.

```
> cl41.glm.fit
Call: glm(formula = cl41.narm ~ poly(cl41.jul, 2))

Coefficients:
 (Intercept) poly(cl41.jul, 2)1 poly(cl41.jul, 2)2
    23.38056          -38.95415            6.36241
Degrees of Freedom: 427 Total; 424 Residual Residual
Deviance:341.9826

> cl81.glm.fit
Call: glm(formula = cl81.narm ~ poly(cl81.jul, 2))

Coefficients:
 (Intercept) poly(cl81.jul, 2)1 poly(cl81.jul, 2)2
    23.10782          -36.82542           5.641723
Degrees of Freedom: 409 Total; 406 Residual Residual
Deviance:316.6941
```

다음으로 적합된 모형에 대하여 잔차를 계산하고 Dates 변수에 대하여 그리면 그림 3.3의 상단과 같이 나타난다. 여기서 환경통계학에서 흔히 하는 방법대로 표준 정규곡선의 .9995th quantile에 해당하는 2.58을 기준으로 이를 벗어나는 관찰값들을 제거하였다(Gilbert(1987) 참조). 이를 도표의 수직축에 점선으로 나타내었으며 그림 3.3의 하단에 이와 같이 Outlier들을 제거한 후의 Residual Plot을 그렸다. Outlier를 제거하고 다시 2차 선형모형을 적합하면 다음과 같은 결과를 얻게 된다. Residual Deviance가 주목할 만큼 감소되었음을 알 수 있다.

```
> cl41.glm.fit.1
Call: glm(formula = cl41.narm.2 ~ poly(cl41.jul.2, 2))

Coefficients:
 (Intercept) poly(cl41.jul.2, 2)1 poly(cl41.jul.2, 2)2
    23.45392           -39.10645             7.528427
Degrees of Freedom: 421 Total; 418 Residual Residual
Deviance:149.6545
```

```
> cl81.glm.fit.1
Call: glm(formula = cl81.narm.2 ~ poly(cl81.jul.2, 2))

Coefficients:
 (Intercept) poly(cl81.jul.2, 2)1 poly(cl81.jul.2, 2)2
    23.15672           -36.61465            6.753313
Degrees of Freedom: 402 Total; 399 Residual Residual
Deviance:132.9074
```

그림 3.3의 하단으로부터 계절의 변화에 규칙적으로 대응하고 있는 상당한 seasonal component가 존재함을 알 수 있다.

3.1.2. Seasonal component의 수정

3.1.1에서 파악한 seasonal component를 수정하기 위하여 먼저 Chloride Level 잔차의 월별 Box Plot을 작성한다(그림 3.4). 그림 3.4로부터 Station 41과 Station 81 공히 4월에서 7월까지는 높은 수준을 보이다가 8월부터는 낮아지는 경향이 있음을 알 수 있다.

Seasonal component를 수정하는 방법으로는 월별로 관찰값에서 월별 median을 차감하는 방법과 mean을 차감하는 방법이 있으나 그 결과에 큰 차이가 없었기 때문에 여기서는 mean을 차감한 결과만 소개한다(그림 3.5). Station 81의 3월 경우 1986년 이후로는 관찰이 이루어지지 않았다. Seasonal component를 수정하기 전보다 훨씬 안정적으로 바뀌었음을 알 수 있다.

이와 같은 과정을 거쳐 적합된 모형은 다음과 같이 요약된다.

```
> cl41.glm.fit.2
Call: glm(formula = cl41.narm.2 ~ poly(cl41.jul.2, 2) +
cl41.mon.2)

Coefficients:
 (Intercept) poly(cl41.jul.2, 2)1 poly(cl41.jul.2, 2)2
    23.51231           -38.82414            7.427725
 cl41.mon.21 cl41.mon.22 cl41.mon.23  cl41.mon.24 cl41.mon.25
   0.1990121  0.06814715  0.04228211 -0.006067543 -0.08616248
 cl41.mon.26 cl41.mon.27  cl41.mon.28 cl41.mon.29
 -0.07114464 -0.02857532 -0.002692709  0.08443004
Degrees of Freedom: 421 Total; 409 Residual Residual
Deviance:113.8416

> cl81.glm.fit.2
```

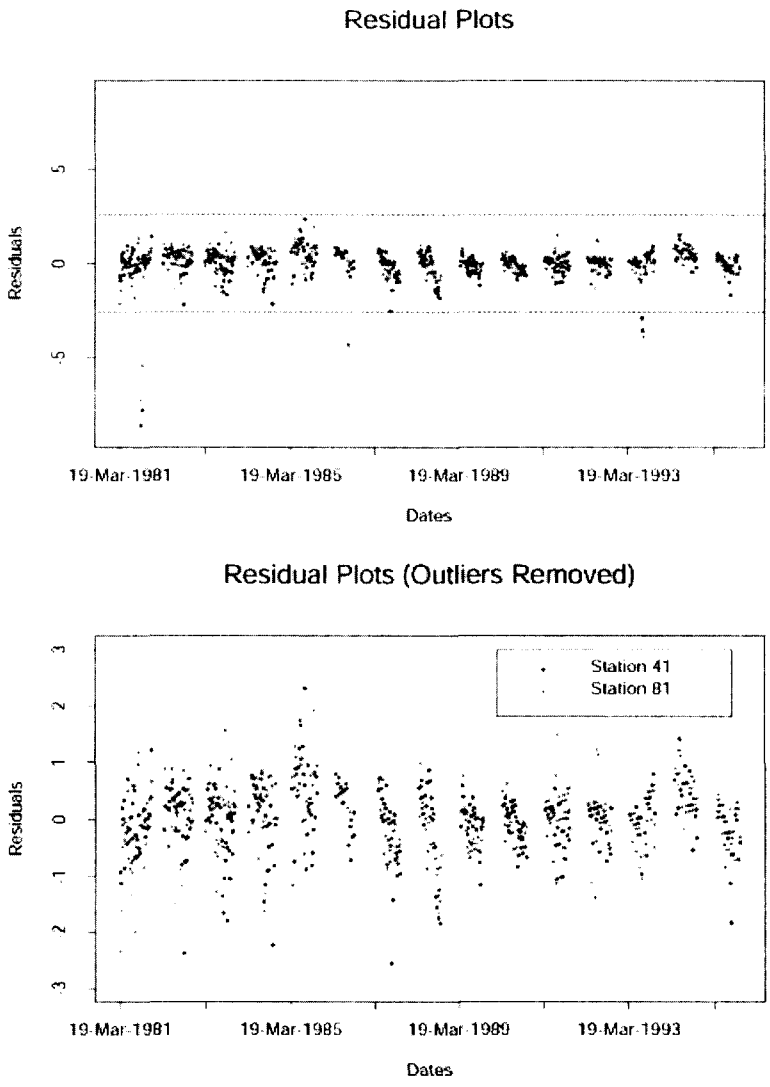

그림 3.3: Outlier를 제거하기 전(상단)과 후(하단)의 Residual Plot

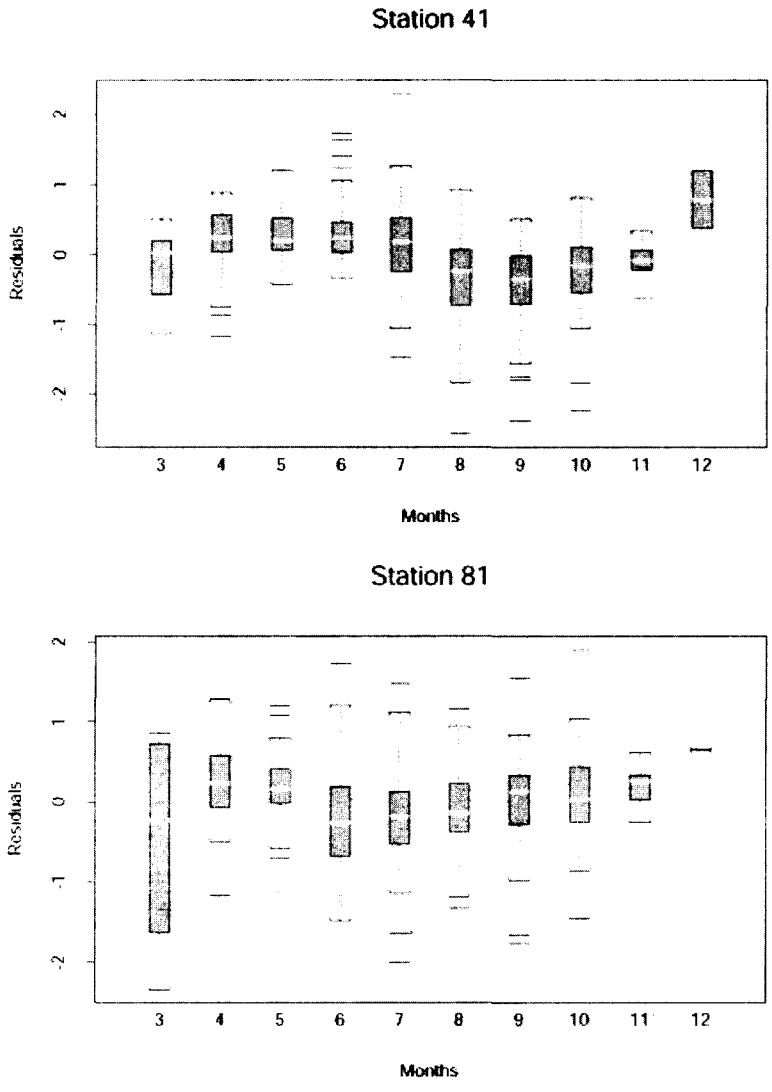

그림 3.4: Seasonal Component 제거 전 월별 Chloride Level 잔차의 Boxplot

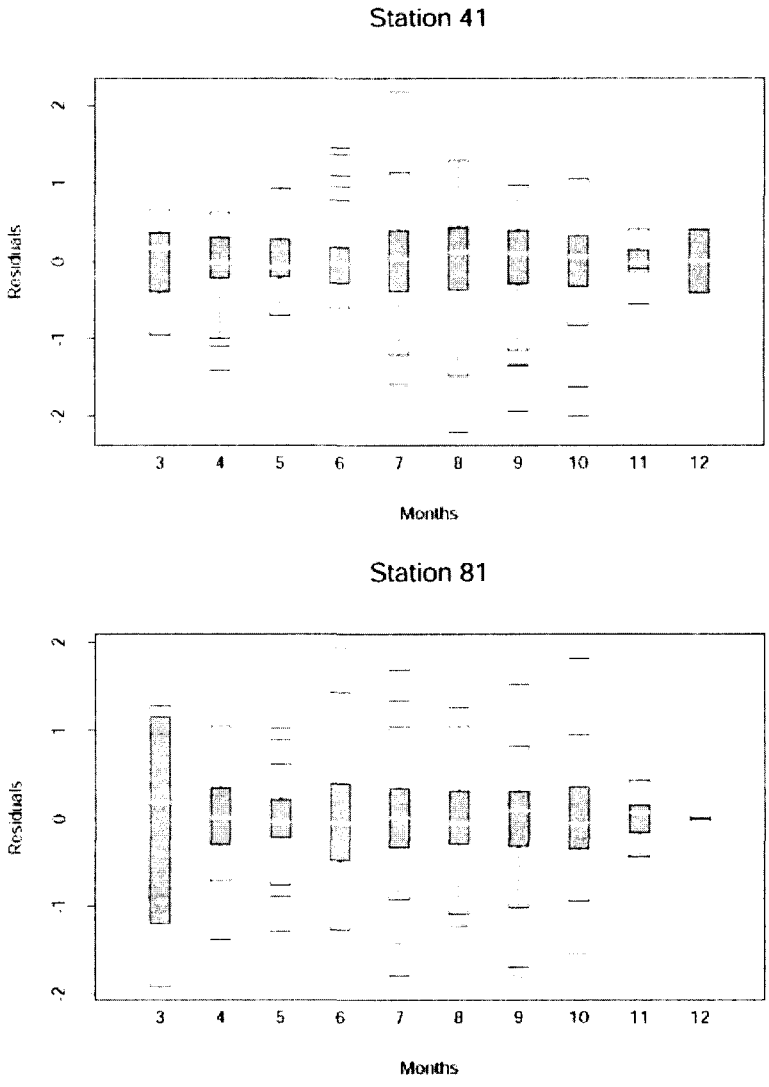

그림 3.5: Seasonal Component 제거 후 월별 Chloride Level 잔차의 Boxplot

```
Call: glm(formula = cl81.narm.2 ~ poly(cl81.jul.2, 2) +
cl81.mon.2)

Coefficients:
 (Intercept) poly(cl81.jul.2, 2)1 poly(cl81.jul.2, 2)2
    23.20434            -36.68096             6.686211
 cl81.mon.21 cl81.mon.22 cl81.mon.23 cl81.mon.24 cl81.mon.25
   0.3235833  0.09517363 -0.05090353 -0.02956033 -0.00137921
 cl81.mon.26 cl81.mon.27 cl81.mon.28 cl81.mon.29
   0.0160649  0.01957397  0.02532219  0.06751848
Degrees of Freedom: 402 Total; 390 Residual Residual
Deviance:120.4059
```

따라서 Chloride Level은 Dates와 Month의 선형모형으로 다음과 같이 요약된다.

$$Cl_{41} = 23.5123 - 38.8241 \times Dates + 7.4277 \times Dates^2 + Month_k,$$
$$Cl_{81} = 23.2043 - 36.6810 \times Dates + 6.6862 \times Dates^2 + Month_k,$$

단 여기서 Cl41 과 Cl81 은 Station 41과 Station 81의 Chloride Level을 나타내고 Month는 월별 효과인 요인이다.

3.2. Variogram 및 Correlogram의 작성

$y_{t_1}, \cdots .y_{t_n}$ 을 Chloride Level의 $t_1, \cdots .t_n$ 시점(julian)에서 관찰값이라 하면 trend와 seasonal component를 고려한 통계모형은 다음과 같이 기술할 수 있다.

$$\hat{y}_{t_k} = \hat{p}(t_k) + s_j, \qquad k = 1, \ldots, n, \quad j = 1, \ldots, 12,$$

단, 여기서 $\hat{p}(t_k)$는 추정된 계수로 표현된 t_k의 다항식이고 s_j는 j번째 달의 효과를 뜻하는 요인(factor)이다. 이제 r_k를 $y_{t_k} - \hat{y}_{t_k}$, 즉 t_k시점에서의 잔차라 하면 variogram의 추정량은 다음과 같이 주어진다(Cressie(1991) 참조).

$$2\hat{\gamma}(h) = \frac{1}{\mid N(h) \mid} \sum_{N(h)} (r_i - r_j)^2.$$

단 여기서 $N(h) \equiv \{(i,j) : t_i - t_j = h\}$ 즉 관측 시점이 서로 h일만큼 떨어져 있는 모든 관측값들의 모임을 의미하고 $\mid N(h) \mid$는 그러한 관측시점 쌍들의 총 개수를 나타낸다. 보통 $\mid N(h) \mid$가 5이상인 경우만 계산에 포함시킨다.

한편 시계열자료의 분석에서 자기상관함수와 같은 역할을 하는 correlogram의 추정량은

$$\hat{\rho}(h) = 1 - \frac{\hat{\gamma}(h)}{2 \times var(\text{Observed Values})}$$

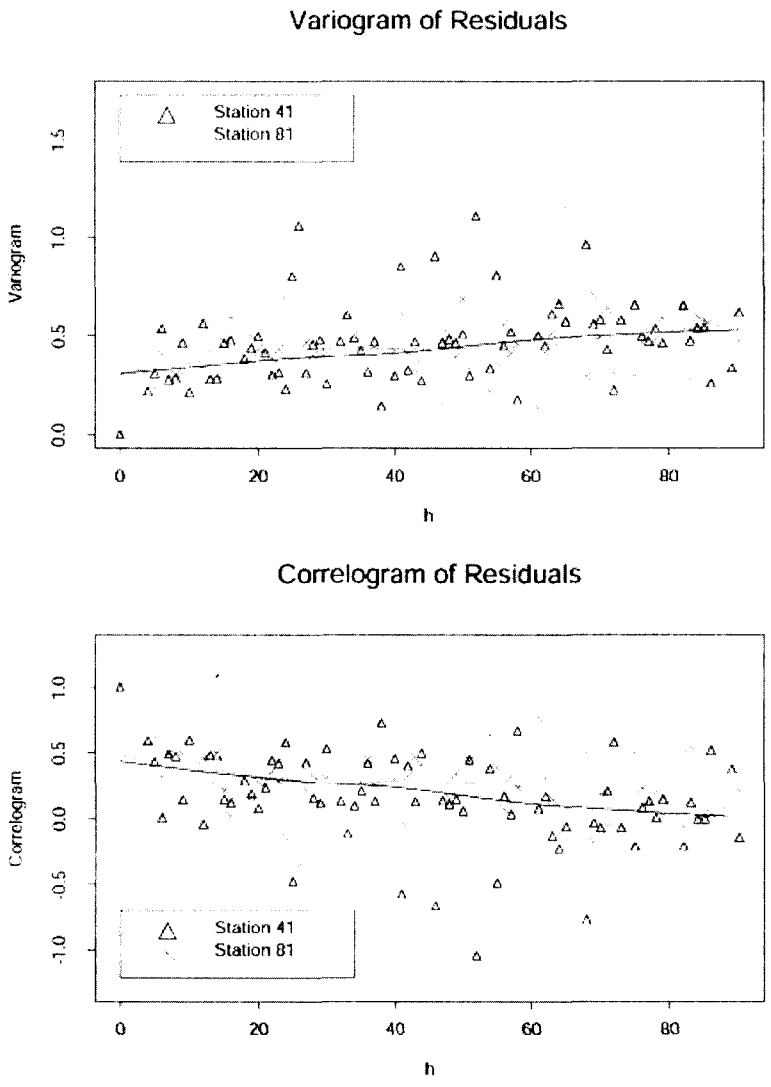

그림 3.6: Chloride Level 잔차의 Variogram(상단)과 Correlogram(하단)

로 주어진다. 3.1과 3.2의 작업으로 trend와 seasonal component를 제거한 잔차에 대하여 variogram(그림 3.6의 상단) 및 correlogram(그림 3.6의 하단)을 작성한다.

4. 결론

수자원관리에 활용되는 각종 생물학적지표들의 실제 측정값들로부터 관찰되는 장기적 추세와 계절적 변화를 통계적으로 분석하는 데는 여건상 어쩔 수 없이 나타나는 결측치와 관측 시기의 불규칙성 등으로 인하여 동일 주기 가정 하에 개발된 통계분석 기법들을 그대로 적용시킬 수 없는 문제점들이 있다.

따라서 탐색적 자료분석의 관점에서 시간 변수에 대한 각 지표들의 변화를 도표로 요약한 후 전통적 방법으로 장기적 추세(trend)와 계절적 변인(seasonal variation)을 수정한 잔차들에 대하여 variogram과 correlogram을 작성한 후 적절한 모형을 찾을 수밖에 없다. 본 연구에서는 이러한 작업의 탐색 과정을 Lake Ontario의 실측자료를 이용하여 보여준다.

먼저 그림 3.2에서 관찰되는 trend를 분석하기 위하여 결측치를 제거한 자료에 Dates변수를 설명변수로 설정한 선형모형을 적합시키고 잔차를 계산하여 이상점을 제거한 후 그림 3.3 하단의 잔차 도표를 얻는다. 이 잔차 도표로부터 상당한 seasonal component를 관찰할 수 있으므로 Station별로 월별 잔차들을 그림 3.4에 Boxplot으로 표현하였다. 여기서 각 잔차들로부터 월별평균을 차감하여 그림 3.5의 Boxplot을 얻는다. 이와 같은 작업을 거쳐 수정된 잔차들에 대하여 variogram과 correlogram을 그려 그림 3.6을 얻는다.

그림 3.6의 하단으로부터 향후 모형 구축작업에 필요한 몇 가지 방향을 잡을 수 있다. 먼저 Station 41의 경우 대부분의 lag에서 양의 자기상관을 보이다가 60일 이후 0을 중심으로 진동하면서 절대값이 작아져가는 양상을 보여주고 있으므로 AR(1) 모형으로의 적합가능성을 시사해 준다. Station 81의 경우 45일 이후 값이 커지는 등 상당히 불규칙적인 양상을 보이고 있어서 향후 모형 구축 작업에서 상당한 난관이 예상된다.

부록: VARIOGRAM과 CORRELOGRAM을 작성하는 S-PLUS 소스 코드

Variogram

```
> variogram.ts
function(x, y, l = 15, hmax = 90, plotit = T, ...) {
        n <- length(x)
        h <- NULL
        diff <- NULL
        for(i in 1:l) {
                h <- append(h,x[-(1:i)]-x[-((n-i+1):n)])
                diff <- append(diff,y[-(1:i)]-y[-((n-i+1):n)])
        }
        h <- h[h <= hmax]
```

```
        cnt <- sapply(split(diff, h), length)
        yp <- sapply(split(diff^2, h), mean)
        xp <- as.numeric(names(cnt))[cnt > 5]
        yp <- yp[cnt > 5]
        if(xp[1] > 0) {
                xp <- c(0, xp)
                yp <- c(0, yp)
        }
        z <- list(cnt = cnt, x = xp, y = as.vector(yp))
        if(plotit)
                if(exists(".Device")) {
                        plot(xp, yp, type = "p", ...)
                        invisible(z)
                }
                else {
                        warning("Device not active")
                        return(z)
                }
        z
}
>
```

Correlogram

```
> correlogram.ts
function(x, y, l = 15, hmax = 90, plotit = T, ...) {
        v <- variogram.ts(x, y, l, hmax, plotit = F)
        xp <- v$x
        yp <- v$y
        variance <- var(y)
        yp <- 1 - yp/(2 * variance)
        z <- list(x = xp, y = yp, variance = variance)
        if(plotit)
                if(exists(".Device")) {
                        plot(xp, yp, type = "p", ...)
                        invisible(z)
                }
                else {
                        warning("Device not active")
```

```
                    return(z)
                }
        z
}
```

참고문헌

[1] Cressie, N. A. C. (1991), Statistics for Spatial Data, Wiley.

[2] Gilbert, R. O. (1987), Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold.

[2002년 10월 접수, 2002년 12월 채택]

Exploratory Analysis of Bioindex Data : Based on a Data Set from Lake Ontario

Kee-Won Lee [1)]

ABSTRACT

In this study, we will construct a statistical model which considered the irregularity of observed time sequence in order to analyze sets of bioindex data gathered from stations in Lake Ontario for a number of years. We fit a linear model to account for the trend and seasonal component in an exploratory way and draw variogram and correlogram for further confirmatory studies.

Keywords: Bioindex; Variogram; Correlogram

1) Professor, Department of Statistics, Hallym University, Chunchon, Kangwon-Do, 200-702, Korea
E-mail: kwlee@hallym.ac.kr