강우와 호수의 인자료에 대한 평균 및 분산추정

김형수[1], 정건희[2], 김중훈[3], 김태균[4]

1 서 론

1.1 연구배경

수계에서의 인농도는 질소와 더불어 부영양화의 지표로써 알려져 있는데 이것은 질소와 인이 조류생장의 제한요소이기 때문이다. 그러나, 이중에서도 대부분의 호수는 인을 제한인자로 가지는 것으로 알려져 있어 아주 적은 양으로도 호수의 부영양화에 있어 제한요소로 작용하게 되는데, 인과 같은 종류의 미소 데이터의 처리는 상당한 주의를 요하게 된다. 그러나 현재까지는 이와 같은 자료 처리가 잘 안되고 있는 실정이며 본 연구에서는 기준치 이하의 관측값들의 처리에 대해 언급하고자 한다.

여러가지 실험에 의해 기록되는 관측값들 가운데 가끔 DL보다 작은 관측값들이 기록곤한다. 그 값들은 수학적으로 그다지 중요하지 않아 보이므로 대체로 현재의 실무에서는 N.D.라고 처리하여 자료를 삭제하여 버리는 것을 관례화하고 있다. 일반적으로 이러한 DL 이라는 기준치를 사용하는 이유는 이렇게 산출된 관측치가 기계적인 오차로부터 분리해 내기가 어려울 정도로 작은 값을 가지기 때문이다. 이런 상태의 자료를 DL 이하의 자료(below the detection level, <DL)라고 하고, 이렇게 DL 이하의 자료를 포함하고 있는 자료를 "censored data"라고 한다.

censored data의 평균과 분산을 추정하기 위해 ADL의 관측치와 BDL 관측치의 수가 주어져 있을 경우, 자료(ADL+BDL)는 다음과 같은 가정을 가진다. 첫째로, ADL부분의 확률분포가 연속적이라는 것과 둘째로, 그 분포를 BDL부분까지 확장시킬 수 있다는 가정이다.

이 자료들을 분석하는 방법은 크게 3가지로 분류된다. 첫째는 평균이나 분산을 측정할 때 기계적인 오차나 실험오차로 간주하여 제거하거나, 혹은 DL 혹은 DL/2로 대체하여 사용하는 것이고, 둘째는 분포방법(distributional method), 셋째는 회귀방법(regression method)이다. 이중

1) 선문대학교 토목공학과 조교수
2) 고려대학교 토목환경공학과 석사과정
3) 고려대학교 토목환경공학과 부교수
4) 진주산업대학 조경학과 전임강사

에서도 이 값들은 수학적으로 그다지 중요하지 않아 보일 뿐만 아니라 대부분의 경우 통계적 분석에 많은 어려움을 초래하므로 첫번째의 방법을 사용하여 분석하는 것이 우리나라에서는 관례화 되어있다. 그러나, 이러한 방법은 모집단의 평균과 분산의 추정시 왜곡된 결과를 초래할 수가 있고, 더군다나 인처럼 미소한 농도가 부영양화의 지표로 작용하는 경우에는 더욱 그 영향이 크다고 할 수 있다. 그러므로, 본 연구에서는 기존의 삭제하거나 치환하는 방법과 분포방법(maximum likelihood estimation(MLE), one-step restricted MLE, and Bias Corrected MLE), 그리고 회귀방법을 적용하여 그 차이를 비교하고 자료계열에 가장 적합한 분포형을 찾아보고자 한다.

1.2 대상자료

본 연구의 적용대상자료로는 우선 이 방법이 우리나라에서 적용이 되지 않았었던 만큼 자료가 많이 확보되어있지 않으므로, 미국의 Florida 지방에서 측정한 강우속에 포함된 인 농도의 자료를 사용하여 분석한 후에 같은 방법을 부여의 저수지의 암모니아성질소(NH_3-N)의 농도를 사용하였다. 자료기간은 93년 1월 1일부터 98년 12월 31일까지의 2,187개의 일자료이다.

1.3 개요

이번 연구의 자료로 사용되고 있는 인자료는 플로리다 지방의 대기중의 TP 농도자료로서 1980년대 초반부터 the South Florida Water Management District에서 수집되기 시작했다.

19개 관측지점에서, 대기중의 농도자료는 일주일 간격으로 매주 목요일에 wet와 dry로 나누어 수집되었고, 그 성분과 주요이온을 결정하기 위해 분석되었다. 이 자료의 DL을 3.5㎍/ℓ로, 정확도 1㎍/ℓ로 하여 저장되었다.

강우중에 TP 농도의 모집단의 평균이 DL에 가까이 접근하고 있고, 많은 TP 자료가 DL 이하에 분포하고 있으므로, 자료의 통계적 특성을 결정하기 위한 통계적 기법을 결정해야 할 필요가 있었다. censored data에 관한 초기의 연구는 전체자료의 분석을 위해 maximum likelihood(ML) methods를 사용하였다. 여기서 전체자료란, DL이하의 자료(BDL)와 DL이상의 자료(ADL)를 모두 일컫는다.

위의 ML 방법을 기초로하여, 안호성은 1996년 논문에서 BDL 통계치를 계산하기 위해 방법을 제안하고 있다. 이 방법은 censored TP 농도자료값에서 TP의 부하를 계산하는데 매우 유용하다. 그러므로 이 방법을 검증하고, 우리나라에의 적용을 꾀하기로 한다.

2 문헌연구

2.1 Analysis methods

2.1.1 Distributional methods

Distributional method는 BDL과 ADL 자료가 모두 주어진 분포형을 따른다는 가정과 대략의 통계치는 관측된 ADL 자료의 확률분포함수를 구함으로써 계산되어진다는 가정하에 평균과 분산 같은 통계치들을 구하기 위해 가정된 분포형의 특성을 사용하는 방법이다. 일반적으로 많이 쓰이는 ML method를 아래에 설명한다.

관측치의 개수를 n이라고 하면 관측된 시계열은 다음과 같이 나타내어진다.

$X = x_i, i = 1,...,n$를 고려하자.변수 X는 lognormal 분포를 따르는 $LN(\mu, \sigma^2)$이고, log 변환을 이용하여 $Y = y_i = \ln(x_i), i = 1,...,n$으로 변형하여 정규분포를 따르도록 한다 $N(\mu_y, \sigma_y^2)$. 계산상의 편의를 위해 내림차순으로 정리하고 계산을 시작하도록 한다.

그러나, 우리는 DL이하의 자료계열의 분포형에 관심이 있으므로 전체 n개의 자료들을 다음과 같이 나눌 수 있는데, log변환한 자료계열인 Y를 $d = \ln(DL)$을 기준으로하여 다음의 2개의 계열로 나눌 수 있다.

m 개의 $Y_B = \{y_i, i = 1,...m \ for\ all\ y_i < d\}$인 부분(censored observations)과 (n-m)개의 $Y_A = \{y_i, i = m+1,...n \ for\ all\ y_i \geq d\}$인 부분(uncensored observations)으로 나누어진다. 여기서, d보다 작은 값들의 계열을 BDL(Below Detection Level)이라 하고, d보다 큰 자료들을 ADL(Above Detection Level)이라고 하자. 또한, BDL자료들은 대수정규분포를 따르고, ADL자료들은 정규분포를 따른다. 모멘트법을 사용하여 구한 Y의 평균과 분산을 각각 $\bar{y}$, s라고 하고, ADL과 BDL부분의 평균과 분산을 각각 $\bar{y}_A$, s_A, $\bar{y}_B$, s_B라고 정의하였다.

Y는 정규분포를 따르므로 확률밀도함수(pdf)는 다음과 같은 형태를 가진다.

$$f(y) = \frac{1}{\sqrt{2\pi\bar{y}}} \exp\left[-\frac{(y_i - \bar{y})^2}{2s^2} \right]$$

위의 식에서 $\xi = \frac{d - \mu_y}{\sigma_y}$이라 나타내면, 누가확률밀도함수(cdf) 는 $F(\xi) = \int_{-\infty}^{\xi} f(y)dy$이 되고, 이 논문에서 전체 자료(BDL+ADL)의 평균과 분산을 구하는데 적용한 각 최우도법의 식을 써보면 다음과 같다.

a. ML Estimator by Cohen (1959)

log 변환한 자료의 평균과 분산을 추정하는 최우도법의 식으로 Cohen(1959, 1961)의 공식을 사용하였다. 평균과 분산을 구하는 방법은 식(1)과 같다.

$$\hat{\sigma}_y^2 = s_A^2 + \hat{\lambda}(\bar{y}_A - d)^2, \quad \hat{\mu}_y = s_A - \hat{\lambda}(\bar{y}_A - d) \qquad \text{----------(1)}$$

여기서, $\hat{\lambda} = \lambda(h,\xi)$ 으로서 $h = \frac{m}{n}$ 의 함수이다.

또한, 다른 부수적인 식들을 써보면 다음과 같다.

$$\lambda(h,\xi) = \frac{Y(h,\xi)}{Y(h,\xi) - \xi},\ Y(h,\xi) = \left[\frac{h}{1-h}\right] Z(-\xi),\ Z(\xi) = \frac{f(\xi)}{1-F(\xi)}$$

위의 식들은 Y에 대한 평균과 분산이므로, 원자료의 평균과 분산은 식(2)에 의해 역변환을 하여 구하도록 한다.

$$\hat{\mu} = \exp(\hat{\mu}_y + \frac{\hat{\sigma}_y^2}{2}),\ \hat{\sigma}^2 = \hat{\mu}^2\{\exp(\hat{\sigma}_y^2 - 1)\} \qquad \text{----------(2)}$$

b. Bias Corrected ML Estimator

n<20 인 작은 수의 자료를 가지는 시계열에서 큰 왜곡도의 영향을 줄이기 위하여 Shaw(1961)와 Schneider and Weissfeld(1985)는 다음의 식(3), (4)를 제안하였다.

$$\hat{\mu}_y = \bar{y} - \frac{s^2}{n+1}\exp\left\{2.692 - 5.439\left(\frac{n-m}{n-2m+1}\right)\right\} \qquad \text{----------(3)}$$

$$\hat{\sigma}_y = s - \frac{s^2}{n+1}\left\{0.312 - 0.859\left(\frac{n-m}{n+1}\right)\right\} \qquad \text{----------(4)}$$

역시 마찬가지로 위의 식을 이용하여 역변환을 한다.

c. One - Step Restricted ML. (persson and Rootzen, 1977)

BDL samples 가 불연속적인 이항분포(binominal distribution)를 갖는다는 가정하에 다음과 같은 계열을 발생시켜 평균과 분산을 식(5)를 이용하여 계산한다.

$$C = c_i = y_i - d, i = m+1,...,n$$

$$\hat{\mu}_y = \bar{y}_A - \alpha^*\sigma^*,\quad \sigma_y^2 = \frac{1}{k}\sum_{i=m+1}^{n} y_i^2 - \left(\frac{1}{k}\sum_{i=m+1}^{n} y_i\right)^2 - \{\alpha^*\varepsilon - (\alpha^*)^2\}(\sigma^*)^2 \qquad \text{---------}$$

(5)

여기서, $\varepsilon = \frac{d - \bar{y}}{s}$

$\alpha^* = \frac{nf(\varepsilon)}{k}$,$k = n - m$

$$\sigma^* = \frac{1}{2}\left[\varepsilon\frac{1}{k}\sum_{i=m+1}^{n} c_i + \left\{\left(\varepsilon\frac{1}{k}\sum_{i=m+1}^{n} c_i\right)^2 + \frac{4}{k}\sum_{i=m+1}^{n} c_i^2\right\}^{\frac{1}{2}}\right]$$

2.1.2 BDL Statistics

위와 같은 방법으로 전체자료(ADL+ BDL)의 평균과 분산을 계산한다. 그러나, 우리가 관심이 있는 부분은 BDL 자료의 부분이므로 위와 같은 최우도법으로 BDL자료의 평균과 표준편차를 구한다.

2.2 Statistics for Combined Population of the BDL and ADL Portions

위에서 구한 BDL자료들의 최우도법 결과치와 모멘트법을 사용하여 구한 ADL자료의 평균과 분산을 결합하여 각 시계열의 특성을 고려한 분포형의 평균과 분산을 구한다. 식은 다음과 같다.

$$\hat{\mu} = h\hat{\mu}_B + (1-h)\hat{\mu}_A \quad \text{----------(6)}$$

$$\hat{\sigma}^2 = h\hat{\sigma}_B^2 + (1-h)\hat{\sigma}_A^2 + h(1-h)(\hat{\mu}_B - \hat{\mu}_A)^2 \quad \text{----------(7)}$$

여기서, $\hat{\mu}_A, \hat{\sigma}_A$는 ADL자료의 평균과 표준편차이고, $\hat{\mu}_B, \hat{\sigma}_B$는 BDL자료의 평균과 표준편차를 나타낸다.

3 연구결과

3.1 Florida지방의 강우 속 총인 자료의 분석

이 연구에 사용된 자료들은 Florida지방의 Okeechobee(위도 26°43', 경도 80°43') 호수의 동서부 늪지역의 자료로서 총 9개의 측정지점 중 BG1WET과 BG2WET지점의 대기중 인 농도 함유량이다. 이 지역의 동쪽 1Km정도는 농업지역으로 이루어져 있다. 자료기간은 1996년에서 1993년의 WET TP자료로서 이 자료는 DL인 3.5㎍/ℓ 이하의 자료들을 포함하고 있었다.

이 자료들의 기초 통계량을 구하여 정리하여 보면 < 표. 1 >과 같다. 원자료 계열이 많이 왜곡되어 있어 log변환이 필요하였다.

<그림 1>에서 알 수 있듯이 DL이하의 값이 많은 부분을 차지하고 있으므로 단일한 분포형으로 분석을 시도하였을 때 왼쪽 꼬리부분이 많이 왜곡된 결과를 줄 것으로 예상된다. 자료계열의 왜곡도가 크므로 대수변환을 하면 왜곡도가 3.179, 3.791에서 0.803, 0.709로 줄어들어 정규분포의 형태가 됨을 알 수 있다. log변환한 자료들의 기초통계량에서 첨도는 BG2의 ADL자료만을 제외하고는 유의수준 0.01에 맞도록 줄어들었지만 표준편차는 유의수준 0.01에 해당하는 0.567보다 약간씩 큰 값

을 보이고 있다. 그러나 이런 왜곡도의 효과는 제안된 결합분석식 (16)과 (17)에 의해 줄어들므로 대수정규분포(lognormal distribution)의 가정은 타당하다.

표 1 플로리다 지방의 강우속 총인자료의 기초통계량

	전체 자료(ADL, BDL)		BDL 자료		ADL 자료	
	BG1WET	BG2WET	BG1WET	BG2WET	BG1WET	BG2WET
개수	116개	109개	31개(27%)	25개(25%)	85개	84개
평균(㎍/ℓ)	9.732	13.664	2.074	1.957	12.165	16.869
표준편차(㎍/ℓ)	12.034	20.247	0.917	1.022	12.895	21.792
왜곡도	3.179	3.791	-0.478	-0.467	2.919	3.493
왜곡도($Y_1=\ln X_1$)	0.803	0.709	-0.900	-0.895	1.149	1.067
첨도	12.178	17.212	-0.917	-0.988	9.868	14.189
첨도($Y_1=\ln X_1$)	0.577	0.482	0.343	0.039	0.719	0.039

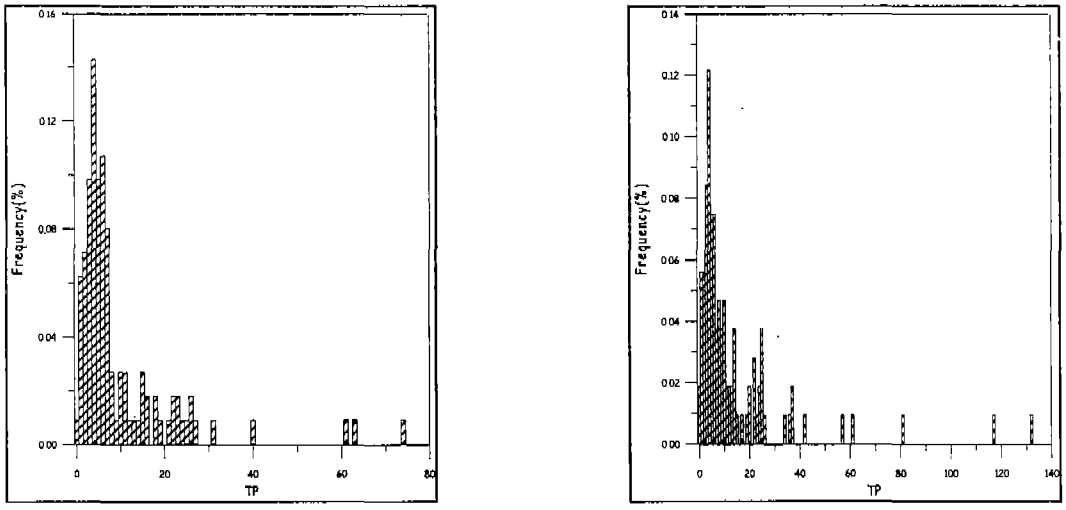

그림 1 BG1WET(n=117)과 BG2WET(n=109)의 빈도분석결과

<그림 1>은 전체자료와 DL을 기준으로 하여 BDL과 ADL로 나누어 분석을 시도하여 결합한 결과를 나타내면 다음과 같다. <그림 1>에서 보듯이 기존의 방법으로 분석한 전체 자료에 대한 평균과 표준편차 추정치는 관측치와 많은 차이를 보인다. 이것은 BDL부분의 영향이라고 생각된다. 특히 회귀분석에 대한 추정치는 많은 차이를 보였다. 그러므로 BDL부분의 통계량은 따로 적용하였다(column 5, 6). BDL통계량 추정은 역시 같은 방법으로 적용하였고, 이때의 오차를 추정하여 보니 one-step restricted ML method가 가장 작은 오차를 보였다. 또한 회귀분석의 오차는 여전히 크게 남아 있는 것은 볼 수 있다. 이것은 식(13)에 의해 계산된 p와 자료의 수에 의해 계산된 h를 비교하여 보면 알 수 있다. BDL부분의 p는 0.305, 0.261인데 반해 h는 0.267, 0.229의 값을 보인다. 이것은 분포의 왼쪽 꼬리 부분이 과장되었음을 나타낸다. case C는 모멘트법을 이용하여 계산한 ADL부분과 BDL통계치들을 결합시킨 통계치들을 구해보았다. 여기서 우리는 회귀분석에 의한 추정치들이 현저히 개선되었음을 알 수 있다. 그러므로 이 때의 회귀식이 자료의 특성을 가장 잘 나타내어줌을 알 수 있다. 또한 관측치들이나 전체자료에 대한 추정치들과도 다른 값

들을 보임을 알 수 있다. 이것은 BDL자료들의 분포형을 고려하여 개선된 결과라 할 수 있다. 결과의 오차를 비교해 보면 One-Step Restricted ML.의 오차가 가장 작은 것을 알 수 있다.

표 2 자료의 평균과 표준편차의 추정

방법		전체 자료 (ADL+BDL)		BDL 자료 (<3.5㎍/ℓ)		BDL과 ADL 자료를 결합시킨 분포형	
		BG1WET	BG2WET	BG1WET	BG2WET	BG1WET	BG2WET
ML by Cohen	평균	9.222	12.867	2.004	1.891	9.450	13.434
	표준편차	10.151	16.757	0.959	1.104	11.929	20.147
Bias Corrected ML	평균	9.260	12.738	2.089	1.978	9.472	13.454
	표준편차	9.259	15.206	1.038	1.189	11.917	20.137
One-Step Restricted ML	평균	9.124	12.764	2.082	1.966	9.470	13.451
	표준편차	10.652	17.405	1.009	1.141	11.917	20.138

표 3 BDL자료의 오차(=100|Obs.-Est.|/Obs.)의 추정

	평균	표준편차
cohen methods	3.359	6.349
bias corrected	0.902	14.833
one-step restricted	0.431	10.854

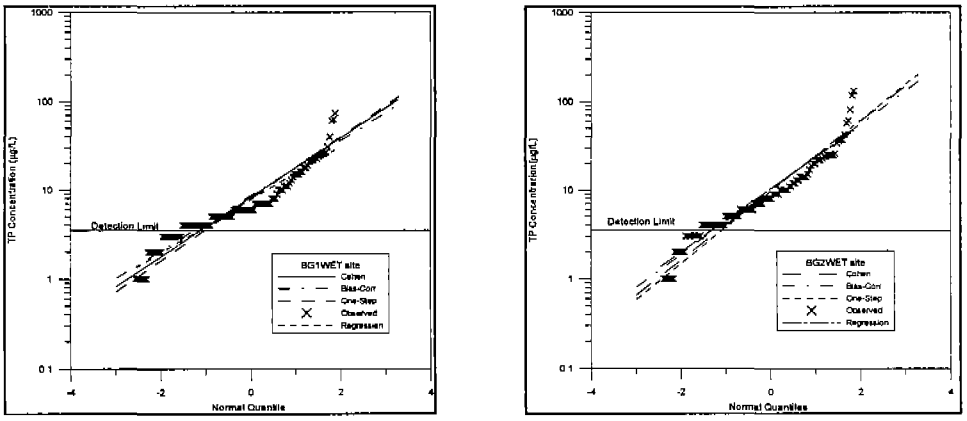

그림 2 각 방법의 결과비교

위의 결과그래프에서 알 수 있듯이 BG1, BG2지점 모두 중간에 분포되어 있는 값들은 모든 방법들이 잘 표현하고 있는 반면, 양극값들은 방법에 따라 약간씩 그 정확도가 다름을 알 수 있다. 특히, 우리가 관심있는 부분인 DL이하의 값들은 두 지점 모두 One-Step Maximum Likelihood방법이 관측값을 잘 표현하는 것으로 증명되었다. 그러므로, 플로리다의 강우속에 포함되어 있는 인농도의 분포를 나타내기 위해서는 One-Step ML방법으로 분석하여 BDL부분과 ADL부분의 결합을 시도하는 것이 가장 적합한 것으로 분석된다.

3.2 질소자료에의 적용

이 방법을 우리나라에 적용하기 위해 Florida와 같은 형태의 자료를 찾아보았으나 미소한 값들은 모두 N.D.라고 처리하여 무시해 버리는 것이 현재 우리나라의 상황이어서 Detection Level을 가지는 자료를 구할 수가 없었다. 그러므로 부득이하게 Detection Level을 가지지는 않지만 통계적 특성이 Florida의 자료와 대체적으로 비슷하고, 대수정규분포를 이루고 있으며 자료의 수도 풍부한 저수지의 암모니아성질소(NH_3-N)자료를 사용하게 되었다. 이 자료를 선정한 이유는 우선 Censored Distribution이라는 분포형의 적용에 초점을 맞추었기 때문이다. 같은 이유로 반응기작에는 주의를 기울이지 않고 통계적 기법의 적용에만 관심을 기울였다. 자료의 특성을 살펴보면, 대수정규분포를 따르고 DL(=500㎍/ℓ) 이하에서 꼬리가 매우 두터운 형태를 띄고 있어, 이부분이 모집단과 다른 특성을 나타내므로 적용에 무리가 없다고 판단된다. 기초통계량 분석은 <표 4>와 같고, 적용한 결과는 <표 5>와 같다.

표 4 질소자료의 기초 통계량

		전체자료(㎍/ℓ)	BDL (≤ 500㎍/ℓ)	ADL (≥ 500㎍/ℓ)
원자료 X_i	평균	1137.447	177.199	1861.290
	표준편차	1182.330	136.516	1103.839
$Y_i=\ln(X_i)$	평균	6.273	4.829	7.357
	표준편차	1.458	0.916	0.593

표 5 질소자료에의 적용

	결 과 (㎍/ℓ)		BDL자료의 오차 (%) (=100|Obs.−Est.|/Obs.)	
	평균	표준편차	평균	표준편차
observed value	1137.448	1182.330	-	-
ML Estimator by Cohen	1130.305	1189.465	10.347	-15.515
Bias Corrected ML Estimator	1143.179	1182.956	-6.999	-37.534
One-Step Restricted ML.	1151.137	1182.098	-15.236	-51.157

결과를 보면, 미국의 폴로리다 자료와는 달리 Bias Corrected ML 방법이 가장 잘 맞는 것으로 나타나 있다. 이것이 자료의 두꺼운 꼬리부분의 영향을 고려한 결과라 할 수 있겠다.

4. 결론

우선, 이번 논문에서는 플로리다와 우리나라의 각각 자료에 대하여 같은 방법을 적용하여 보았다. 대상이 되는 자료들은 Detection Level을 갖는 자료들로써 기준이 되는 값을 경계로하여 아래쪽의 자료(BDL)와 위쪽의 자료(ADL)가 각기 다른 통계적특징을 가질 때 적합한 방법이라 하겠다. 플로리다 자료는 ADL부분은 모멘트법을 이용하여 통계량을 구하고, BDL부분은 대수정규분포를 사

용하여 One-Step Restricted ML를 적용한 값을 결합한 것이 가장 적합한 분포형으로 결정지어졌다. 그러나, 같은 방법을 적용하였을 때, 우리나라의 질소자료는 One-Step restricted ML가 아닌 Bias Corrected ML를 최적의 모형이라고 선택하였다. 이것은 기본이 되는 자료가 다른 종류였기 때문이고, 원래 논문에서 의도한 DL을 가지는 자료가 아니고, 임의로 DL을 정하여 주었기 때문에 플로리다의 자료와는 다소 차이를 보이는 것으로 분석된다. 그러나 역시 정확히 같은 성질을 가지는 자료는 아니므로 그 결과에 대한 절대적인 비교를 하는 것은 무리가 있다고 생각되나, 앞에서 언급한 것처럼 분포형의 적용에 그 의미를 두고 있으므로 각 분포형에 따라 다른 결과를 주고, 적절한 분포형을 찾을 수 있다는 데에서 결론을 지을 수 있겠다.

이와 같은 분포형의 적용이 일반화되면 환경자료 같은 미소농도를 가지고 있는 자료에 대한 더욱 더 정밀한 적용이 이루어질 수 있고, 미소농도의 기여도 역시 더욱 중요하게 다루어질 것으로 생각된다.

5. 참고문헌

• Hosung Ahn, 1998, Estimating the Mean and Variance of Censored Phosphorus Concentrations in Florida Rainfall, Journal of the American Water Resource Association Vol. 34 No. 3, pp. 583-593.

• Aichison, J. and J. A. C. Brown, 1957. The Lognormal Distribution, Cambridge University Press, Cambridge, Massachusets.

• Cohen, A. C., Jr.,1959. Simplified Estimates for the Normal Distribution When Samples Are Singly Censored or Truncated. Technometrics 1(3):217-237.

• Newman, M. C., P. M. Dixon, B. B. Looney, and J. E. Pinder, III, 1989. Estimating Mean and Variance for Environmental Samples With Below Detection Limit Observations. Water Resources Bulletin 25(4):905-916.

• Persson T. and H Rootzen, 1977. Simple and Highly Efficient Estimators for a Type I Censored Normal Sanple. Biometrika 64(1):123-128.

• Cohen, A. C., Jr., 1961. Tables for Maximum Likelihood Estimates:Singly Truncated and Singly Censored Samples. Technometrics 3(4):535-541.

• 윤용남, 1994, 공업수문학, 청문각.

• 최병선, 1992, 회귀분석(上), 세경사.