Journal of Korean Society on Water Quality, Vol. 26, No. 2, pp.297-302 (2010)

Bootstrap 기법을 이용한 BOD 평균 농도 및 신뢰구간 분석

김경섭†

국립한경대학교 환경공학과

Analysis of BOD Mean Concentration and Confidence Interval using Bootstrap Technique

Kyung Sub Kim†

Department of Environmental Engineering, Hankyong National University

(Received 21 October 2009, Revised 13 January 2010, Accepted 14 January 2010)

Abstract

It is very important to know mean and confidence interval of water-quality constituents such as BOD for water-quality control and management of rivers and reservoirs effectively. The mean and confidence interval of BOD at Anseong2 and Hwangguji3 sampling stations which are located at the border of local governments in Anseong Stream were estimated and analyzed in this paper using Bootstrap technique which is one of non-parametric statistics. The results of Bootstrap were compared with arithmetic mean, geometric mean, Biweight method mean as a point estimator and distribution mean came from the appropriate probability distribution of Log-normal. In Bootstrap technique 12 data set was randomly selected in each year and 1000 samples was produced to get parameter of population. Visual Basic for Applications (VBA) of Microsoft Excel was utilized in Bootstrap. It was revealed that the Bootstrap technique can be used to explain more rigorously and robustly the achievement or violation of BOD target concentration in Total Maximum Daily Load (TMDL).

keywords : Anseong Stream, BOD, Bootstrap, Confidence interval, Target concentration

1. 서 론

환경은 자체 정화능력이 있어 어느 정도 외부 환경압력에 견딜 수 있으나 환경압력이 자정능력을 초과하면 환경파괴 현상이 발생한다. 환경파괴로 나타나는 지표수 및 지하수의 오염은 농업용수 및 공업용수 등의 수자원으로서의 가치 상실을 유발하고 있으며, 음용수로 활용할 경우 음용수 수질기준을 만족하기 위한 수처리 비용의 과다를 요구한다.

수계 수질관리의 하나인 오염원의 농도규제(배출허용기준)는 산업화의 진전과 오염원의 집중화 등으로 발생하는 오·폐수의 과다유입을 효과적으로 통제하지 못하여 이로 인해 수체의 환경기준달성에 한계를 노출하게 되었다. 오염총량관리제는 수계가 수용할 수 있는 오염총량을 정하여 이에 따른 총량규제를 시행하는 것으로 유역의 오염원, 유량 및 수질을 종합적으로 관리하는 통합 유역관리의 하나이다. 우리나라는 「팔당호 등 한강수계 상수원 수질관리 특별종합대책('98.11.20) 수립」을 추진하면서 지역의 오염총량관리제를 공식적으로 도입하게 되었으며, 이어 「낙동강수계 물관리종합대책('99.12.30) 수립」, 「금강수계 물관리종합대책, 영산강수계 물관리종합대책('00.10.24) 수립」을 확정하면서 지방자치단체가 의무적으로 지역의 환경용량 한도 안에서 자율적으로 환경친화적 지역 개발을 추진하도록 오염총량관리제를 실시하고 있다.

하천 및 호소의 수질을 효율적으로 관리하기 위해서는 수질목표 지점에서 관리대상 수질항목의 대표성 있는 평균 농도 및 분포 형태를 파악하는 것이 매우 중요하다. 수질항목의 평균 농도는 강건성을 확보한 점추정량(point estimator)을 사용하는 방법(Kim et al., 2002), 적합한 농도분포를 파악하여 구하는 방법(김경섭과 안태진, 2009; Novotny, 2004; Tung and Hathhorn, 1988) 및 자료가 부족한 표본의 경우 자료를 생성하여 분포를 통하여 파악하는 Bootstrap 방법(전명식, 1990; Efron, 1979; Manly, 2006) 등이 있다. Bootstrap 기법은 다양한 학문분야에 이용되는데 이명우 등(2005) 및 김병식 등(2002)은 수자원 분야의 확률강우량 산정 및 하천유출량 산정에, Chang 등(2007)은 수의학 분야에서 닭 및 돼지의 수입량 위험도 불확실성 분석에, 전영두 등(2008)은 항공우주 분야에서 로켓엔진 점화로 인한 소음 및 진동 분석에, 정석근 등(2008)은 수산자원 분야에서 어획량 및 어획노력량의 신뢰구간 산정에 이 기법을 적용하였다. 수질관리 분야에서 Bootstrap 기법은 오염원관리를 위한 통계모델에 적용하기도 한다(Schwarz et al., 2006).

본 논문에서는 안성천 유역의 경기도 수질측정지점 가운데 안성시와 평택시의 경계에 위치한 안성천2와 화성시와

† To whom correspondence should be addressed.
kskim@hknu.ac.kr

평택시의 경계에 위치한 황구지천3 지점을 선정하여 BOD의 평균 농도 및 평균 농도의 신뢰구간을 Bootstrap 기법을 사용하여 파악해 보았다. Bootstrap 기법의 결과는 점추정 방법인 산술평균, 기하평균, Biweight 방법에 의한 평균 및 BOD 농도의 분포(대수정규분포)를 통하여 파악하는 분포평균과 비교해 보았다. 본 연구의 목적은 목표수질 지점의 수질관리 대상 항목 목표수질 설정에 대표성을 확보하고 임의성을 감소하여 합리적이며 효율적인 수계관리가 가능하도록 하는 데 있다.

2. 연구방법

2.1. BOD농도 분석 지점

Bootstrap 기법을 이용한 BOD농도 분석을 위하여 대상 유역은 안성천으로 하였으며 수질분석 지점은 우리나라 오염총량관리제에서 일반적으로 지방자치단체와 단체의 행정경계에서 목표수질을 설정하므로 환경부 물환경정보시스템(Water Information System, WIS)에 나와 있는 수질측정지점 가운데 안성시와 평택시의 경계에 위치한 안성천2와 화성시와 평택시의 경계에 위치한 황구지천3 지점을 선정하였다.

2.2. Bootstrap 기법

2.2.1. 기본 개념

Bootstrap 기법은 이미 존재하는 표본자료에서 무작위로 자료를 추출하여 새로운 표본을 만들며 이와 같은 과정을 반복하여 모집단의 확률분포를 파악하는 비매개면수적 방법이다. 일반적으로 모수를 파악하기 위해서는 약 100개의 Bootstrap 표본을 이용하며, 신뢰구간을 파악하기 위해서는 약 1000개의 Bootstrap 표본을 이용한다(이명우 등, 2005; 전영두 등, 2008).

2.2.2. 적용 절차

Bootstrap 기법의 적용 절차를 요약하여 나타내면 다음과 같다(전명식, 1990; 전영두 등, 2008; Efron, 1979).

단계 1. 표본자료($X=(x_1, x_2, \cdots\cdots x_n)$)에서 임의로 자료를 추출하여 Bootstrap 표본($X_b=(x_{1b}, x_{2b}, \cdots\cdots x_{nb})$)을 생성한다. 자료 추출은 난수를 발생하여 Monte Carlo 모의를 통하여 얻는다.

단계 2. 단계 1에서 구한 Bootstrap 표본에서 구하고자 하는 파라미터($\theta_b = mean(X_b)$)를 구한다.

단계 3. 단계 1 및 2를 반복하여 파라미터군($\theta_B = \theta_{b1}, \theta_{b2}, \cdots\cdots \theta_{bB}$)을 생성한다.

단계 4. 단계 3의 파라미터군으로부터 확률밀도함수(Probability Density Function, PDF) 또는 누적분포함수(Cumulative Distribution Function, CDF)를 파악한다.

단계 5. 단계 4에서 구한 PDF 또는 CDF에서 모수 및 신뢰구간을 구한다.

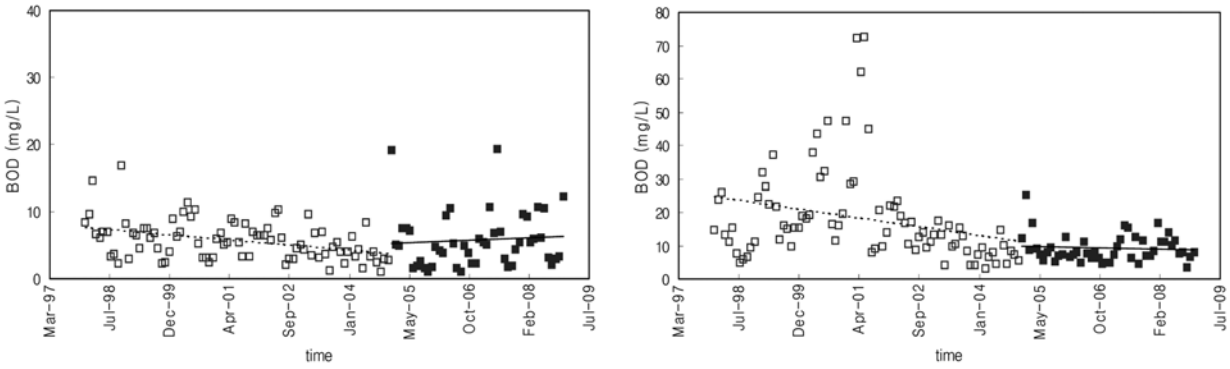
3. 결과 및 고찰

3.1. BOD농도 분석 자료

환경부에서 관리하는 수질측정지점중의 하나인 안성천2 지점의 수질자료는 1992년부터 공개되어 있으며, 1998년부터 2008년까지 과거 약 10년간의 BOD 자료를 분석한 결과 최근 4년간(2005~2008)의 변화와 과거 7년간(1998~2004)의 변화가 상이하여 최근 4년간의 BOD 자료를 근간으로 BOD농도 분석을 실시하였다. 황구지천3 지점은 1989년부터 자료가 존재하며 안성천2 지점과 같이 약 10년간(1998~2008)의 BOD 자료를 분석한 결과 안성천2 지점의 농도변화와 유사하여 최근 4년간(2005~2008)의 BOD 자료를 근간으로 BOD 농도 분석을 실시하였다. Fig. 1은 안성천2 및 황구지천3 지점의 BOD 농도를 추세선과 함께 도시한 것이다.

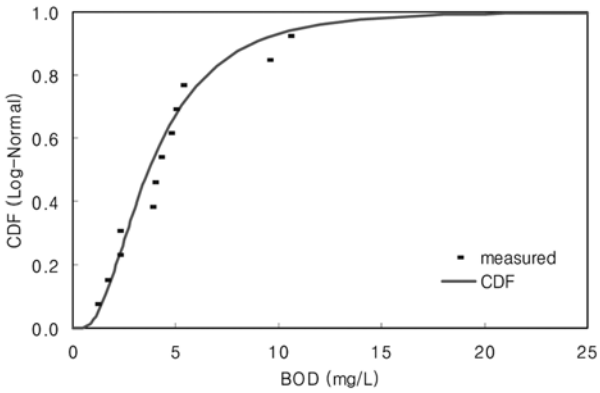
3.2. BOD농도 분포

산소요구물질인 BOD 및 DO 같은 수질항목의 농도분포는 일반적으로 대수정규분포에 잘 적합되는 것으로 알려져 있다(김경섭과 안태진, 2009; Novotny, 2004). 안성천2 및 황구지천3 지점의 BOD농도 분포도 적합도 검정을 실시하여 대수정규분포의 적용성을 파악해 보았으며, Kolmogorov-Smirnov 적합도 검정 방법에 의한 2006년 BOD농도 CDF,

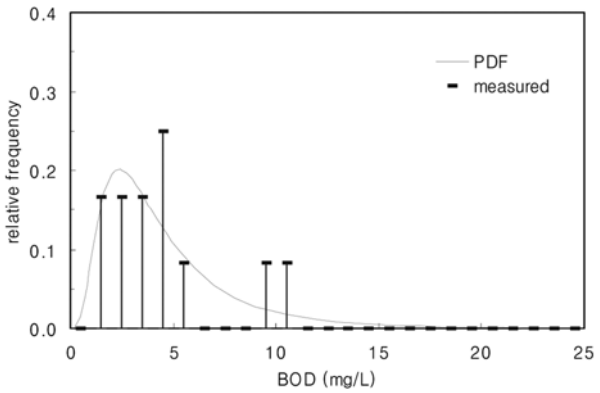


(a) Anseong2 (b) Hwangguji3

Fig. 1. BOD sampling data at Anseong2 and Hwangguji3.

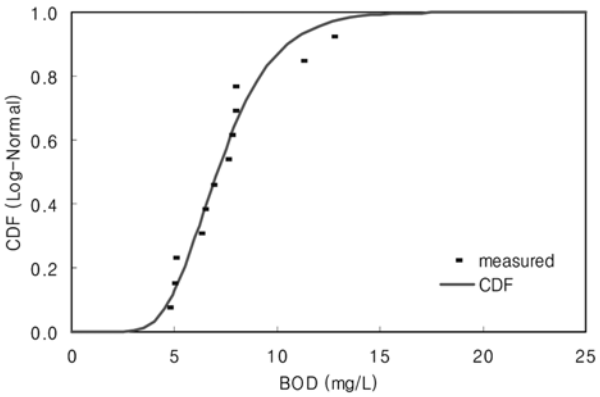


(a) CDF

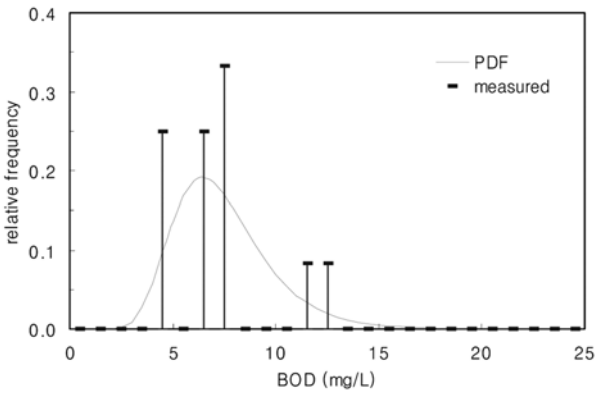


(b) PDF and relative frequency

Fig. 2. Anseong2 in 2006.



(a) CDF



(b) PDF and relative frequency

Fig. 3. Hwangguji3 in 2006.

PDF 및 상대도수가 Figs. 2 및 3에 나타나 있다. 안성천2 지점에서 2006년 실측자료의 누적확률분포 한계편차(D_n^{α})는 $D_{12}^{0.2}$의 경우 0.300로 주어지며, 가정된 이론 확률분포의 누적확률분포 최대편차($D_{\max}$)는 0.209로 나타나 최대편차가 한계편차보다 작아 대수정규분포를 채택할 수 있는 것으로 나타났다. 황구지천3 지점에서도 $D_{\max}$ = 0.163이 $D_{12}^{0.2}$ = 0.300보다 작아 안성천2 지점과 같이 대수정규분포가 적합한 것으로 파악되었다. 2004, 2005 및 2007년도 BOD 농도분포도 Figs. 2 및 3에 나타나 있는 2006년도 결과와 같이 대수정규분포로 잘 적합되는 것으로 파악되었다.

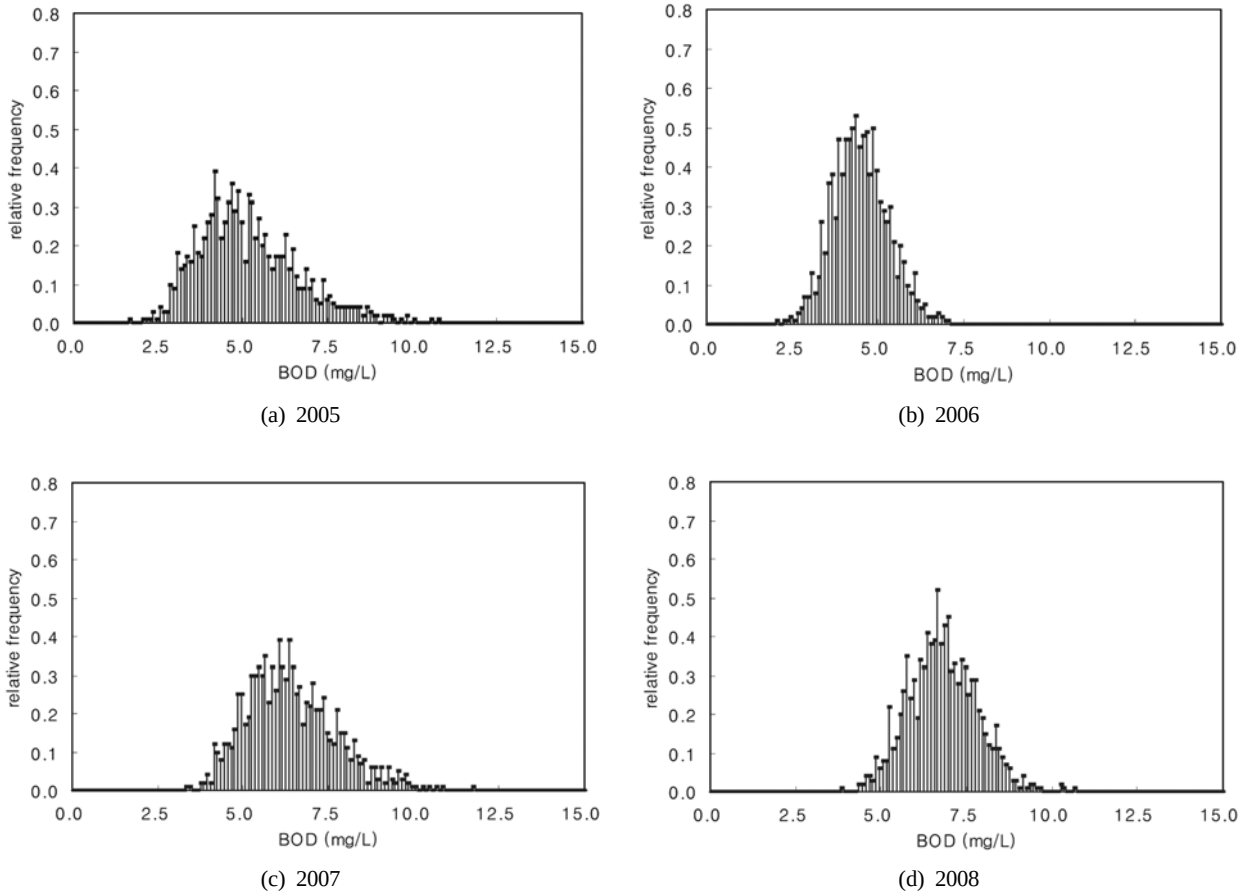
3.3. Bootstrap 기법

3.3.1. 안성천2

Bootstrap 기법을 이용한 안성천2 지점의 년도별 BOD 평균 농도분포가 Fig. 4에 나타나 있다. 자료추출은 1년에 12개의 월 자료가 가용하므로 12개를 기본으로 하였으며, Bootstrap 표본은 신뢰구간 파악을 위하여 1000개로 하였다. 모의수행은 Microsoft Excel에서 Visual Basic for Applications(VBA)를 이용하여 실시하였다. Fig. 4로부터 최빈값(mode)은 2008년도, 첨도(kurtosis)는 2006년도에 크게 나타나는 것으로 파악되었다. BOD 평균 농도 분포를 파악하기 위하여 Bootstrap 표본의 신뢰구간을 산정하였으며 95% 신뢰구간이 Table 1 및 Fig. 5에 나타나 있다. 2005년도 BOD 95% 신뢰구간은 2.867~8.558 mg/L로 가장 넓게 나타났으며, 2006년도 95% 신뢰구간은 2.950~6.150 mg/L로 제일 좁게 파악되었다. 2005년부터 2008년까지 년평균 BOD 농도는 2006년도에 약간 감소하다가 이후 다시 증가하는 양상을 나타냈다.

3.3.2. 황구지천3

Bootstrap 기법에 의한 황구지천3 지점의 BOD농도 분석도 안성천2 지점과 같은 과정으로 실시하였으며 결과가 Fig. 6에 나타나 있다. Fig. 6에 보이듯이 2006년도 BOD 평균 농도 첨도가 제일 높았으며 2005년도 BOD 평균 농도가 다른 해에 비하여 상당히 퍼져있는 것으로 나타났다. 황구지천3 지점의 BOD 95% 신뢰구간이 Table 2 및 Fig. 7에 나타나 있으며, 2005년도 95% 신뢰구간이 7.500~13.800 mg/L로 제일 넓게 분포하는 것으로 파악되었다. 황구지천3 지점의 년도별 BOD 평균 농도는 2005년도가 제일 높았으며 Fig. 7에 보이듯이 해를 지나면서 감소하는 추세를 나타냈다.



(a) 2005 (b) 2006

(c) 2007 (d) 2008

Fig. 4. BOD mean distribution at Anseong2.

Table 1. Confidence interval of each year at Anseong2

(unit : mg/L)

Confidence interval	2005	2006	2007	2008
Lower limit (2.5%)	2.867	2.950	4.208	4.875
Upper limit (97.5%)	8.558	6.150	9.333	8.742

3.4. BOD농도 분석

안성천2 지점의 년도별 BOD 평균 농도를 파악해 보았으며, BOD 평균 농도는 산술평균, 기하평균, 실측치에 가중

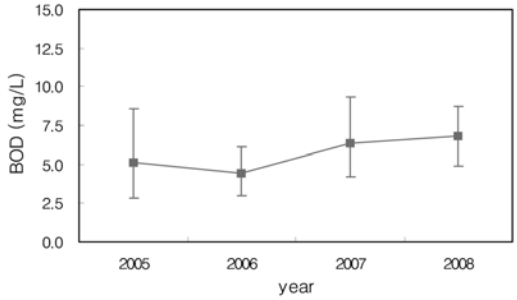
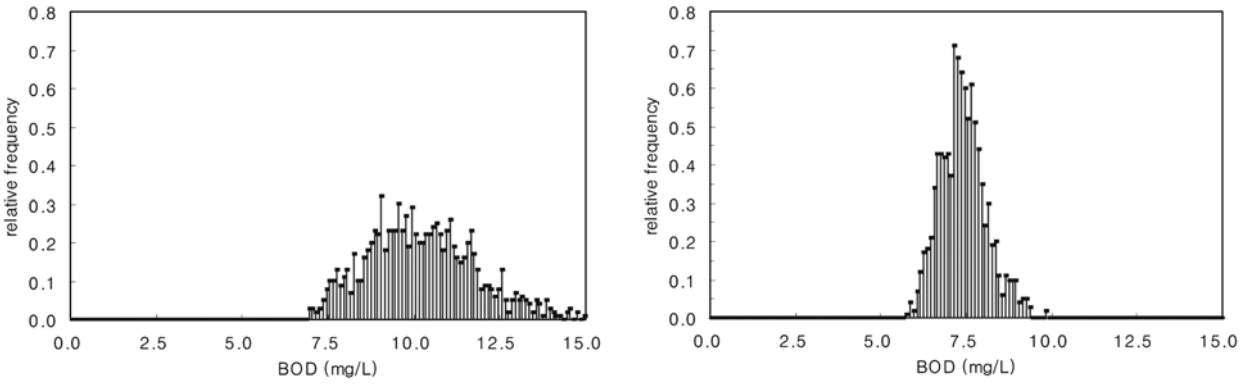
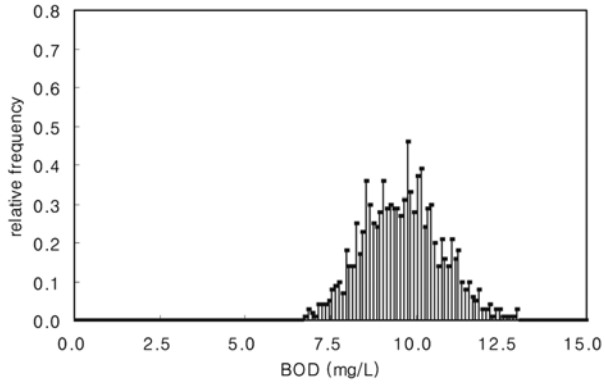


Fig. 5. Mean and 95% confidence interval at Anseong2.

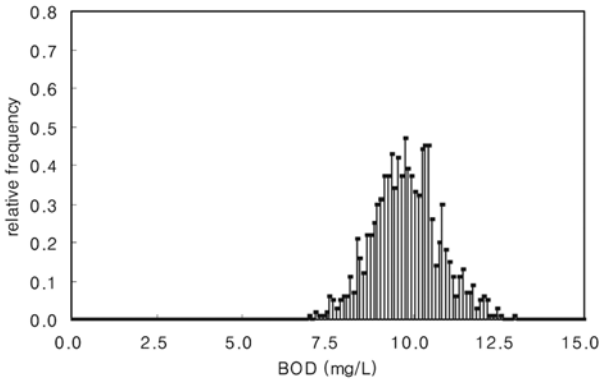


(a) 2005 (b) 2006

Fig. 6. BOD mean distribution at Hwangguji3.



(c) 2007



(d) 2008

Fig. 6. BOD mean distribution at Hwangguji3 (continued).

Table 2. Confidence interval of each year at Hwangguji3

(unit : mg/L)

Confidence interval	2005	2006	2007	2008
Lower limit (2.5%)	7.500	6.167	7.508	7.842
Upper limit (97.5%)	13.800	8.942	11.892	11.858

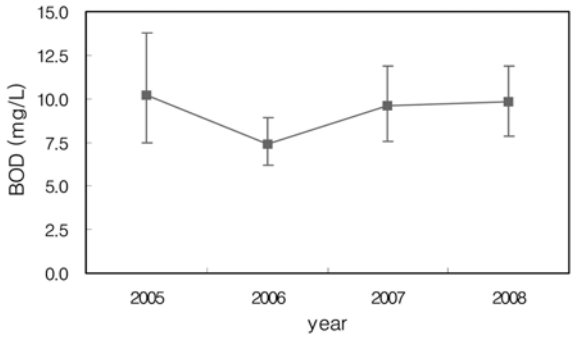


Fig. 7. Mean and 95% confidence interval at Hwangguji3.

치를 부여하여 추정하는 Biweight 방법에 의한 평균, 분포평균 및 Bootstrap 평균으로 구분하여 구하였다(Table 3). Table 3에 나타나 있듯이 각 년도별 Bootstrap 평균은 산술평균보다 다소 작게 파악되었으며 기하평균 및 Biweight 평균보다는 크게 나타났다. Bootstrap 평균과 분포평균을 비교해 보면 전반적으로 Bootstrap 평균이 분포평균보다 작게 나타났으며, 기하평균이 최저의 값을 보이는 것으로 파악되었다. 2007년 Biweight 평균(5.054 mg/L)이 기하평균(5.203 mg/L)보다 작은 것은 큰 수치를 보이는 5월 농도값(19.3 mg/L)의 가중치가 고려되지 않은 결과로 분석된다. 일년에 12개의 실측자료를 통하여 얻은 산술평균, 기하평균 및 Biweight 평균보다는 평균 농도의 분포를 구하여 얻은 분포평균이나 부족한 자료를 생성하여 모수를 구하는 Bootstrap 평균이 대표성을 더욱 나타낼 것으로 판단한다. 황구지천3 지점의 BOD 평균 농도는 Table 4와 같이 주어지며 안성천2 지점과 유사하게 분석되었다.

4. 결 론

본 연구에서는 Bootstrap 기법을 이용한 BOD 농도를 안성천 수계에 속한 안성천2 및 황구지천3 지점에 적용하여 분석해 보았으며, 이를 통하여 얻어진 연구결과 및 장래연구방향을 서술하면 다음과 같다.

1) 안성천2 및 황구지천3 지점의 BOD농도는 하수도 보급률의 꾸준한 상승으로 2000년 중반(약 2005년)부터 감소하는 것으로 나타났다.
2) 2005년부터 2008년까지 매해 BOD 평균 농도를 산정하기 위하여 산술 평균법, 기하 평균법, 실측치에 가중치를 적용하는 Biweight 방법, BOD 농도의 분포를 파악

Table 3. Mean values of Anseong2

(unit : mg/L)

Year	Arithmetic	Geometric	Biweight	Distribution	Bootstrap
2005	5.150	3.569	3.888	5.255	5.119
2006	4.492	3.694	3.757	4.632	4.472
2007	6.400	5.203	5.054	6.503	6.377
2008	6.808	5.873	6.785	7.016	6.778

Table 4. Mean values of Hwangguji3

(unit : mg/L)

Year	Arithmetic	Geometric	Biweight	Distribution	Bootstrap
2005	10.183	9.166	8.113	10.138	10.223
2006	7.408	7.082	6.759	7.425	7.394
2007	9.617	8.874	9.437	9.705	9.608
2008	9.808	9.177	9.561	9.941	9.793

하여 구하는 방법 및 Bootstrap을 이용하여 구하는 방법 등이 적용되었으며, 안성천2 및 황구지천3 지점의 경우 BOD 평균 농도는 전반적으로 Bootstrap 평균이 산술평균보다 다소 작았고 분포평균이 산술, 기하, Biweight 및 Bootstrap 평균보다 높게 나타났다.

3) BOD 평균 농도 및 신뢰구간을 파악하기 위하여 적용한 Bootstrap 기법은 매해 12개의 자료를 임의로 추출하여 1000개의 Bootstrap 표본을 생성하여 모집단의 확률분포를 파악하였다. Bootstrap 기법을 이용하여 구한 BOD 평균 농도는 부족한 자료를 생성하는 단계를 거치므로 단순히 주어진 자료를 활용하여 구하는 산술 평균, 기하 평균 및 Biweight 평균보다 더욱 대표성을 확보한다.
4) Bootstrap 기법을 이용하여 구한 대표성을 갖춘 BOD농도 분포는 평균 농도, 신뢰구간, 자료의 비대칭도 및 첨도 등을 설명하는데 적절한 정보를 제공할 것이다.
5) Bootstrap 기법을 이용하여 파악한 수질항목 분포는 의무 또는 임의로 추진되는 오염총량관리제에서 지방자치단체의 경계에서 제시되는 목표수질 설정, 목표수질의 위험도 분석 및 안전부하량 설정에 객관적 타당성을 제공할 것이다.

사 사

본 연구의 일부는 건설교통부 및 한국건설교통기술평가원 건설핵심기술연구개발사업의 연구비지원(09건설핵심B01)에 의해 수행되었습니다.

참고문헌

김경섭, 안태진(2009). 안성천 유역의 BOD농도 확률분포 특성. *수질보전 한국물환경학회지*, **25**(3), pp. 425-431.

김병식, 김형수, 서병하(2002). Bootstrap 방법에 의한 하천 유출량 모의와 왜곡도. *한국수자원학회논문집*, **35**(3), pp. 275-284.

이명우, 이충성, 김형수, 심명필(2005). Bootstrap방법과 SIR 알고리즘을 이용한 확률강우량 결정과 위험도 분석. *대한토목학회논문집*, **25**(5B), pp. 365-373.

전명식(1990). 통계적 데이터 분석방법을 위한 컴퓨터의 활용 I : 붓스트랩 이론과 응용. *응용통계연구*, **3**(1), pp. 121-141.

전영두, 박종천, 정의승(2008). 부트스트랩 기법을 이용한 소음진동 스펙트럼 분석법 소개. *2008년 춘계학술대회논문집*, 한국소음진동공학회, pp. 185-188.

정석근, 최일수, 장대수(2008). 부트스트랩과 베이지안 방법으로 추정한 수자원관리에서의 생물학적 기준점의 신뢰구간. *한국수산학회지*, **41**(2), pp. 107-112.

환경부(2009). 물환경정보시스템. http://water.nier.go.kr/.

Chang, K. Y., Hong, K. O., and Pak, S. I. (2007). Bootstrap simulation for quantification of uncertainty in risk assessment. *Korean J. Vet. Res.*, **47**(2), pp. 259-263.

Efron, B. (1979). Bootstrap Method: Another Look at the Jackknife. The Annuals of Statistics, *Institute of Mathematical Statistics*, **7**(1), pp. 1-26.

Kim, K. S., Kim, B., and Kim, J. H. (2002). Robust measures of location in water-quality data. *Water Engineering Research*, **3**(3), pp. 195-202.

Manly, B F. J. (2006). *Randomization, Bootstrap and Monte Carlo methods in Biology*, CRC, Boca Ranton, FL.

Novotny, V. (2004). Simplified Databased Total Maximum Daily Loads, or the World is Log-Normal. *J. Environ. Eng.*, **130**(6), pp. 674-683.

Schwarz, G. E., Hoos, A. B., Alexander, R. B., and Smith, R. A. (2006). *The SPARROW Surface Water-Quality Model: Theory, Application and User Documentation*. U.S. Geological Survey, Reston, Virginia.

Tung, Y. K. and Hathhorn, W. E. (1988). Assessment of probability distribution of dissolved oxygen deficit. *Journal of Environmental Engineering, ASCE*, **114**(6), pp. 1421-1435.