Journal of the Korean Data & Information Science Society
2014, **25**(1), 119–130

http://dx.doi.org/10.7465/jkdi.2014.25.1.119
한국데이터정보과학회지

낙동강 본류 측정소들의 생물학적 산소요구량 수치에 대한 비모수적 회귀분석과 특이점분석[†]

김종태[1]

[1]대구대학교 전산통계학과

접수 2013년 12월 9일, 수정 2014년 1월 2일, 게재확정 2014년 1월 7일

요 약

본 연구는 국립환경과학원의 물환경정보시스템에서 제공한 자료를 사용하였다. 자료는 낙동강 본류 (낙본, nb)의 수질측정소 A지역에서 측정소 N지역까지 2003년 1월부터 2013년 8월까지 측정한 월별 수질데이터를 이용하였다. 생물학적 산소요구량 BOD (biological oxygen demand)의 통계학적 수질분석은 월별, 연도별, 지역별로 R 프로그래밍을 이용하여 구체화 하였다. 낙본지역 측정소들의 BOD에 대하여 탐색적 자료분석 (exploratory data analysis) 방법과 비모수 회귀분석방법 중 하나인 Lowess (locally weighted scatter plot smoother) 경향분석법을 이용하여 장기수질경향과 지역별 수질분포의 현황을 분석하였다. 그리고 특이점 (outlier)이 가장 많이 발생하는 시기와 낙본 측정지역들을 분석하였다. 그 결과 낙본하류지역인 부산광역시 강서구 명지동 측정소 nbM의 BOD 수질환경 보다 낙본중류지역인 대구광역시 달성군 구지면의 측정소 nbG와 경상남도 창원시의 측정소 nbI 지역의 수질오염이 보다 심각한 문제점들이 있는 것으로 나타난다.

주요용어: 비모수 회귀분석, 생물학적 산소요구량, 탐색적 자료분석, 특이점.

1. 서론

본 연구는 낙동강 본류 (낙본)의 최상류 지역에서부터 최남단의 하류지역까지의 14개 낙본 수질 측정소에서 측정되어진 수질측정 자료를 분석하였다. 수질측정 변수 중 하나인 생물학적 산소요구량, BOD (biological oxygen demand)에 대하여 월별, 연도별, 지역별 통계적 분석과 특이점 (outlier)을 연구하였다. 생물학적 산소요구량 (biological oxygen demand), BOD는 어떤 물속의 미생물이 산소가 존재하는 상태에서 유기물을 분해 안정시키는데 요구되는 산소량이다. 산소를 필요로 하는 박테리아가 일정시간 내 (보통 20도씨에서 5일간)에 물속의 유기물을 산화 분해시켜 정화하는데 소비되는 산소의 양을 ppm으로 나타낸 것이다. 음료수는 2ppm이하가 적당하고 농업용수는 5ppm이 좋다. 5ppm이상이 되면 하천은 자기정화 능력을 잃으며 10ppm을 넘을 때는 나쁜 냄새를 풍기며 시궁창 하천이 된다.

이명박 정부의 4대강 사업 시행결과에 대하여 많은 논란 중 하나는 수질 환경문제일 것이다. 본 연구의 목적은 4대강 사업 중 가장 많은 논란을 일으키고 있는 낙동강 본류의 수질에 대하여 수질측정 변수 중 하나인 BOD의 경향을 통계적으로 분석하는데 있다.

본 연구에 필요한 데이터는 국립환경과학원의 물환경정보시스템 (water information system)의 홈페이지 (water.nier.go.kr)에 있는 '측정자료조회'의 '수질 (총량측정망)'에서 구하였다. 이 데이터들은 수

[†] 이 논문은 2011년도 대구대학교 학술연구비 지원에 의한 연구임.
[1] (712-714) 경상북도 경산시 진량면 내리동, 대구대학교 전산통계학과, 교수. E-mail: jtkim@daegu.ac.kr

질측정소 낙본측정소 A지역에 낙본측정소 N지역까지 14개의 측정소에서 2003년 1월부터 2013년 8월까지 측정한 월별 수질데이터이다.

이 수질 데이터의 분석도구로 R 프로그래밍을 사용하였다. BOD 데이터를 월별, 연도별, 지역별로 탐색적 자료분석 EDA (exploratory data analysis) 방법과 비모수적 회귀분석방법인 Lowess (locally weighted scatter plot smoother) 경향분석법을 적용하여 장기수질경향과 수질분포를 분석한다.

수질변화의 측정에서 특이점 (outlier)의 발생은 매우 중요하게 다루어 진다. 측정에서의 특이점이 발생하는 경우는 크게 두 가지의 경우로 나누어지는데 첫 번째는 측정의 오류에서 발생하고, 두 번째는 측정 시기의 환경이나 주변의 환경오염을 발생시키는 산업체로 인한 수질오염이 주된 원인이 되어 진다. 첫 번째의 경우는 무시할 수 있지만 두 번째의 경우는 수질오염에 있어서 대단한 심각한 문제들이 발생할 수 있다. 그 예가 2013년 8월에 낙동강 본류의 중류지역에서 발생한 심각한 적조현상이다. 여기에 대한 요인으로 8월 가뭄으로 인한 강수량의 감소와 적조에 영향을 가장 많이 미치는 총질소 (total nitrogen; T_N)의 증가가 있다. 총질소는 무기성질소와 유기성질소의 질소량의 합으로 표현된다. 수질분석에서 총질소를 분석하는 이유는 질소와 인은 수질 부영양화 현상에 지대한 영향을 미치는데 물속에서 질소나 인이 다량 유입되면 그것을 먹고사는 플랑크톤이 급속히 증가하여 적조현상이 발생하고, 용존산소가 부족하여 수질오염의 원인이 되기 때문이다. 이러한 이유로 낙본 지역의 생물학적 산소요구량 BOD에 대하여 특이점 (outlier)들의 경향을 시기와 지역별로 분석 조사한다. 다른 수질측정 변수들에 대한 분석방법도 동일한 방법으로 할 수 있다.

Lowess는 실제 데이터에 대한 평활 곡선 (smoothing curves)들을 적합 시키는데 단순하지만 매우 훌륭한 도구이다 (Jacoby, 2000). Lowess의 용어는 Cleveland (1979)에 처음 소개되어진 이후에 Lowess는 가장 대중화된 비모수적 평활 추정방법이 되었다 (Cleveland와 Devlin, 1988). Cleveland (1993)에 따르면 Lowess는 계산하기 쉽고, 사용하기 쉬운 매우 매력적인 통계적인 방법으로 소개하고 있다. Lowess는 장기 수질경향분석 도구로도 많이 사용되었다. Lee와 Lee (2006)은 Lowess 기법을 이용하여 낙동강 수계의 장기 수질경향 분석을 하였다. Kim과 Park (2004)은 Lowess 기법과 비모수 검정법을 이용하여 낙동강 수계 경향분석을 하였다. Kim 등 (2007)은 금호강 수질오염에 대한 모형을 연구하였고, Kang 등 (2011) 낙동강의 유량과 유속의 모형을 연구하였다. 이 연구에서 월별로 연도별로 각 낙본측정소에 대하여 BOD에 대한 수질의 상태의 분석과 어떤 낙본측정소 지역이 심각한 수질오염을 가지고 있는가를 분석한다.

본 연구는 다음과 같이 구성되었다. 2절에서는 Lowess 경향기법과 그 절차들을 소개하고, 3절에서는 BOD에 대한 지역별, 월별, 연도별 경향들을 분석한다. 4절에서는 BOD의 특이점 (outlier)들을 분석하고 결과를 요약한다.

2. Lowess 경향분석기법

Lowess (locally weighted scatter plot smoother)는 각 값에 대해 이동 직선 (moving line)을 구하고 이로부터 y의 평활점 (smoother)을 얻은 후 이 평활점 (smoother)들을 직선으로 연결한 것이다. 이는 1차 또는 2차 회귀모형에 대한 가정 없이 자료들을 회귀모형에 맞춤으로 유용한 경향 분석법이다. 이동선 (x_i, y_i)를 계산하는 방법으로 $x = x_i$를 중심으로 일정 간격에 걸친 수직띠 (window)를 만든다. $x = x_i$ 중심으로 $n \times f$에 가장 가까운 정수만큼의 데이터를 포함하도록 수직띠 (window)의 폭을 결정한다.

여기서 f 는 $0 < f < 1$ 인 평활상수로서 자료크기 대비 평활기의 너비 (smoother span)로 표현한다. 흔히 $1/3 < f < 2/3$가 추천된다. f의 값이 커질수록 한 평활점 (smoother)을 계산하기 위하여 많은 데이터 점들이 사용되므로 평활곡선 (smoother curve)이 더 매끄럽게 된다 (과소적합; under-fitting). 반

대로 f를 작게 하면 회귀함수의 휜 정도가 큰 곡선이 된다 (과대적합; over-fitting). 그러므로 관측자료의 적합정도를 세밀하게 관찰하여 적절한 f값을 찾아야 한다.

가중값 함수 (weighted function)을 정의하고 부근 가중값들 (neighborhood weights)을 계산한다. 가중값 함수의 바람직한 성질은 다음과 같다.

1) 중앙에 위치한 (x_i, y_i)이 가장 큰 가중값 w_0을 주고, $(x_{i\pm k}, y_{i\pm k})$에는 두 번째 큰 가중치 w_1을 주며 가장 바깥에 있는 $(x_{i\pm k}, y_{i\pm k})$에는 가장 작은 가중치 w_k를 준다.

2) 가중값 함수는 $x = x_i$에 대하여 대칭이며 중앙에서 멀어질수록 매끄러운 형태로 감소한다. 즉 $w_0 \le w_1 \le \cdots \le w_k$로 실제 존재하지 않는 관측점에는 가중치를 주지 않는다.

3) 가중값 함수는 수직띠의 양쪽 경계에서 0의 값을 가진다.

$$T(u) = \begin{cases} (1-|u|^3)^3, & |u| < 1, \\ 0, & |u| \ge 1. \end{cases} \tag{2.1}$$

따라서, x_i로부터 수직띠 (window)까지의 최대거리를 d_i라 하면 (x_k, y_k)의 가중값 w_k는 다음과 같이 계산된다.

$$w_k = T\left(\frac{x_k - x_i}{d}\right) = \left(1 - \left(\frac{d(x_i - x_k)}{\max_{l\in N_k} d(x_l - x_k)}\right)^3\right)^3 \tag{2.2}$$

수직띠 (window)안의 데이터 점들은 가중최소제곱법 (weighted least squares method)을 써서

$$\min_{\beta_0,\beta_1} \sum_{j\in N_i} w_j (y_j - (\alpha_k + \beta_k x_j))^2 \tag{2.3}$$

직선으로 적합시킨 다음 α_k와 β_k의 추정치

$$\hat{\beta_k} = \frac{\sum w_i^2 (x_i - \bar{x})(y_i - \bar{y})}{\sum w_i^2 (x_i - \bar{x})^2}, \qquad \hat{\alpha}_k = \bar{y} - \hat{\beta}_k \bar{x} \tag{2.4}$$

각각 구한다. 여기서 $\bar{x}$와 $\bar{y}$는 가중 평균이다. $\hat{\alpha_k}$와 $\hat{\beta_k}$을 이용하여 $x = x_i$에서의 y의 적합값 $\hat{y}$들을 계산한다.

$$\hat{y} = \hat{\alpha}_k + \hat{\beta}_k x \tag{2.5}$$

각 관측점별로 독자적인 회귀적합을 하기 때문에 따라서 많은 계산이 필요하다. 마지막으로 Lowess는 적합점들 $(x_1, \hat{y_1}), (x_2, \hat{y_2}), \cdots, (x_n, \hat{y_n})$을 연결하여 최종적인 회귀곡선 (regression curve)을 산출한다.

참고로, 특이점 (outlier)들에 의한 영향을 제한하기 위해 추가로 겹제곱 가중값 함수 (bisquare weight function)를 사용하여 가중값 (weighted values)들을 새로 계산하기도 하는데 이는 로버스트 단계 (robust step)라 불린다. 로버스트 Lowess 방법은 다음과 같다.

1) 식 (2.5)의 추정된 회귀직선을 이용하여 $\hat{y_i}$ 들을 구한다.

2) 초기 잔차 $\hat{\epsilon}_k = y_k - \hat{y}_k$를 구한다. 그리고 다음과 같이 로버스트 가충치 (robust weight) δ_k을 다음과 같이 계산한다.

$$\delta_k = B\left(\frac{\epsilon_k}{6s}\right), \quad s = \text{median of } \left|\hat{\epsilon}_1\right|, \left|\hat{\epsilon}_2\right|, \cdots, \left|\hat{\epsilon}_n\right|.$$

여기서 $B(z)$는 겹제곱 함수 (bisquare function)로서 다음과 같이 정의된다.

$$B(z) = \begin{cases} (1-|z|^2)^2, & |z| < 1 \\ 0, & |u| \ge 1. \end{cases}$$

3) 가중치들 $\delta_k w_k$을 사용하여, x의 값들에 대한 y들의 가중 단순 선형 회귀 (weighted simple linear regression)를 적합 시킨다. 이 과정에서 α_k와 β_k을 추정하고, 또한 추정값 $\hat{y}_k = \hat{\alpha}_k + \hat{\beta}_k x_k$을 구한다.

4) 잔차 $\hat{\epsilon}_k = y_k - \hat{y}_k$들을 다시 구한 다음에 로버스트 가중치 δ_k들을 다시 계산한다.

5) 가중치들 $\delta_k w_k$을 사용하여, x의 값들에 y대한 들의 가중 단순 선형 회귀 (weighted simple linear regression)를 적합 시킨다. 이 과정에서 마지막 추정치들 α_k, β_k, $\hat{y}$와 $\hat{\epsilon}_j$들을 계산한다.

$$y = g(x) + \epsilon$$

의 모형을 가정하면, Lowess를 하기 위한 R의 함수는 lowess()로서 다음과 같은 형식으로 사용된다.

$$\text{lowess}(x, y, f = 2/3, iter = 3)$$

여기서 f는 평활 폭 (smoother span)으로 각 관찰값에서 평활의 영향력을 정하는점들의 비율로서 값이 클수록 평활의 정도가 심해진다. R에서의 디폴트 값은 2/3이다. $iter$은 로버스트 회귀를 수행하기 위한 반복수를 나타낸다. 값이 클수록 수행속도가 느려진다.

3. BOD 경향분석

낙동강 물 환경 연구소에서 담당하고 있는 각 낙본측정소의 위치는 Table 3.1에 나타나 있다. 측정망들의 위치는 낙본상류지역에서 낙본하류지역까지 차례대로 낙본의 측정소의 위치를 나타낸 것이다.

Table 3.1 The place of measurement in Nakdong main stream river

Measuring Station	Adress
nbA	Kyungbuk Bonghwa Sekpo Daehyen
nbB	Kyungbuk Bonghwa Mtengho Ganchang
nbC	Kyungbuk Andong Pungchen Sinsang
nbD	Kyungbuk Sangju Nakdong Nakdong
nbE	Kyungbuk Gumi OtaeDong
nbF	Kyungbuk Seongju Yongam Donglak
nbG	Daegu Dalseong Guji Daeam
nbH	Kyungnam Yryeong Jijeong Seongsan
nbI	Kyungnam Changweong Ychang Bukmyeon
nbJ	Kyungnam Jimhae Saengrim Masa
nbK	Kyungnam Yaqngsan Mulgum Mulgum
nbL	Busan BukGu GumgokDong
nbM	Busan GangseoGu MyongjiDong
nbL	Busan GangseoGu NoksanDong

3.1. 지역별 탐색적 자료분석

탐색적 자료 분석 (exploratory data analysis; EDA)에서는 최소값 (min), 아래 사분위수 (제1사분위수; H_L), 중앙값 (median; M), 위 사분위수 (제3사분위수; H_U), 최대값 (max)을 다섯 수치 요약 (five number summary)이라 하는데 이는 (min, H_L, M, H_U, max) 표현된다. Table 3.2는 다섯 수치 요약에서 평균 (mean)과 표준편차 (sd), 그리고 사분위수 산포 ($spr(H)$)를 포함하여 지역별 BOD를 분석하였다. 가장 하류 낙본지역인 낙동강 하구언지역은 본류가 두 줄기로 나누어지는데 하나는 부산광역시 강서구 명지동의 nbM지역과 또 다른 하나는 부산광역시 강서구 녹산동의 nbN 지역이다. 이들 두 낙동강 하구언지역의 수질오염을 비교하면 nbN 지역의 측정치들은 nbM 낙본측정 지역뿐만 아니라

다른 어떤 낙본측정지역의 측정치들보다 매우 심하게 수질오염을 나타내며, 큰 산포를 보인다. Table 3.2에서 나타난 모든 통계적 수치들에서 nbN 지역은 다른 낙본 측정지역들의 모든 통계적 수치들에 비해 월등히 높은 것으로 나타난다. 이런 이유로 본 연구에서는 nbN 측정지역과 다른 지역들 간의 통계적인 비교를 제외한다. 그러나 nbN 지역이 왜 다른 지역들보다 심한 수질오염의 측정치를 나타내는지는 기회가 된다면 다음 연구에서 분석하고자 한다. 낙동강 본류인 nbN과는 다른 줄기지만, 동일한 위치에 있는 nbM 지역을 볼 때, 단순히 낙동강 최하류지역에 있다는 이유만으로 수질이 나쁘다는 결론을 내릴 수 없기 때문이다.

Table 3.2에서 2003년 1월부터 2013년 8월까지의 BOD 데이터의 평균 (mean)과 중앙값 (M)을 비교해보면 nbI > nbG > nbL > nbJ > nbH > nbK > nbM > nbF > nbE > nbA > nbD > nbC > nbB의 순으로 나타난다. 다음과 사실들이 보여진다 (단, nbN 지역은 비교에서 제외한다).

1) 낙동강 본류 (낙본)의 중류지역인 nbG, nbH, nbI 지역의 BOD의 평균수치들은 하류지역인 nbJ, nbK, nbL, nbM 지역의 평균들보다 높게 나타난다. 특히 중류의 도시지역인 대구 달성군 nbG지역과 경남 창원시 nbI 지역이 경남 양산의 물금 nbK, 부산 북구 금곡동 nbL, 부산 강서구 명지동 nbM지역보다 높은 것은 더 깊이 연구해 볼 가치가 있다. BOD의 수치가 대체로 높은 nbG 지점은 낙동강수계의 종류를 대표할 수 있는 지점으로, 산업화와 공업화로 발전된 대구광역시의 수많은 공단과 폐수 및 생활오폐수의 영향을 받는 금호강이 유입한 직후의 지점으로 낙동강 수계의 수질 오염도가 급격히 변하는 지점이다. nbI지역은 낙동강 수계구간 중 남강 합류점 후부터 밀양시 청도천 합류점 전까지 지점으로, 낙동강 전체 유역면적의 14.1%를 차지하는 남강의 영향을 크게 받을 것으로 예상되는 지역이다. 남강은 진주시를 제외하고 대부분 군 지역으로 주 오염원은 농경지와 소규모 축산농가로부터 유출되는 축산 폐수 등이다.

2) 낙동강 본류의 가장 상류에 있는 경북 봉화군 석포면 nbA측정소가 nbB, nbC, nbD 지역의 BOD 수치의 평균보다 높게 나타날 뿐 아니라 다른 모든 통계수치들에서도 높게 나타나고 있다. 상식적으로 nbA 지역이 가장 낮은 BOD 수치를 가질 것으로 예측되었지만, 상류지역에서 가장 높은 BOD 평균 수치를 나타낸다.

3) 위의 1)과 2)의 결과를 두고 볼 때, BOD의 수치는 측정소의 순위에 따른 지리적 위치도 영향을 받겠지만 그 보다는 주변의 수질오염에 영향을 미치는 인프라에 더 큰 영향을 받는다는 것을 알 수 있다. 또한 강은 상류에서 하류로 흘러가면서 어느 정도 자체 정화 능력이 있다는 사실을 보여주고 있다.

Table 3.2 BOD data summary by place in Nakdong main stream river

place	n	min	H_L	mean	M	H_U	max	$spr(H)$	sd
nbA	396	0.2	1.13	1.20	1.1	1.27	6.5	0.14	0.69
nbB	410	0.1	0.75	0.79	0.7	0.83	4.2	0.08	0.43
nbC	408	0.1	0.93	0.98	0.9	1.03	4.4	0.10	0.49
nbD	451	0.2	1.03	1.08	1.00	1.13	3.3	0.10	0.55
nbE	408	0.4	1.55	1.63	1.5	1.70	4.8	0.14	0.74
nbF	401	0.3	1.90	2.00	1.7	2.10	8.6	0.20	1.02
nbG	406	0.8	2.51	2.63	2.3	2.75	7.9	0.24	1.22
nbH	406	0.6	2.29	2.40	2.1	2.52	7.8	0.23	1.17
nbI	407	0.2	2.51	2.64	2.3	2.77	8.7	0.26	1.32
nbJ	408	0.5	2.37	2.48	2.2	2.59	6.9	0.23	1.17
nbK	451	0.5	2.27	2.37	2.1	2.48	6.1	0.21	1.12
nbL	404	0.3	2.41	2.54	2.2	2.66	8.8	0.25	1.26
nbM	403	0.3	2.07	2.20	1.8	2.32	7.8	0.25	1.29
nbN	403	1	3.62	3.80	3.5	3.99	12.1	0.37	1.89

Table 3.3은 BOD에 대한 Tukey의 다중비교 분석의 결과를 나타낸 것이다. Tukey의 다중비교 분석 결과는 (nbA, nbB, nbC, nbD, nbE)의 낙본상류지역들로 하나의 그룹과 (nbG, nbH, nbI, nbJ, nbK, nbL, nbM)인 낙동강 본류의 중하류지역을 동일한 그룹으로 나타내고 있다. 그리고 김해시 쪽에 가깝게 있는 낙본 최하단의 낙동강 하구언지역인 nbN은 낙본지역의 다른 어떤 지역들과도 독립된 그룹으로 나타난다.

Table 3.3 Tukey's multiple test for BOD by place in Nakdong main stream river

	nbB	nbC	nbD	nbE	nbF	nbG	nbH	nbI	nbJ	nbK	nbL	nbM	nbN
nbA	0.95	1.00	1.00	0.98	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
nbB		1.00	1.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nbC			1.00	0.57	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nbD				0.81	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nbE					1.00	0.03	0.22	0.02	0.08	0.20	0.03	0.610	0.00
nbF						0.48	0.92	0.33	0.70	0.91	0.47	1.00	0.00
nbG							1.00	1.00	1.00	1.00	1.00	0.99	0.00
nbH								1.00	1.00	1.00	1.00	1.00	0.00
nbI									1.00	1.00	1.00	0.95	0.00
nbJ										1.00	1.00	1.00	0.00
nbK											1.00	1.00	0.00
nbL												0.98	0.00
nbM													0.00

3.2. 월 기준 탐색적 자료분석

Table 3.4은 각 수치들은 각 지역별로 BOD 월 (month) 평균들을 구한 다음에, 다시 월을 기준으로 최소값 (min), 아래 사분위수 (제1사분위수; H_L), 평균 (mean), 중앙값 (median; M), 위 사분위수 (제3사분위수; H_U), 최대값 (max), 사분위수 산포 ($spr(H)$), 표준편차 (sd), 구한 것이다. BOD의 총 평균 (mean)과 중앙값 (M)은 3월 > 2월 > 4월 > 1월 > 6월 > 5월 > 12월 > 7월 > 8월 > 11월 > 10월 > 9월의 순으로 2월, 3월, 4월을 중심으로 BOD수치가 높게 나타나고, 가을인 9월, 10월, 11월에는 상대적으로 BOD 평균 수치에 대한 측도들의 값들이 낮게 나타난다. 이는 강우량과 상관성이 있어 보인다. 비가 적게 오는 계절일수록 BOD의 평균 수치가 올라가고, 비가 많이 오는 계절일수록 상대적으로 BOD 수치가 낮아지는 경향을 보인다. 흥미로운 사실은 산포의 측도인 사분위수 범위 ($spr(H)$)와 표준편차 (sd)의 순위 역시 평균과 중앙값들의 순위들과 매우 비슷한 경향을 띄고 있다.

Table 3.4 BOD data summary by month in Nakdong main stream river

month	min	H_L	mean	M	H_U	max	$spr(H)$	sd
1	0.60	1.35	2.03	2.26	2.84	3.53	1.49	1.01
2	0.62	1.48	2.57	3.06	3.89	4.29	2.42	1.38
3	0.86	1.42	2.67	13.33	3.73	4.02	2.31	1.26
4	0.92	1.38	2.43	2.83	3.01	3.70	1.63	0.93
5	1.00	1.44	1.88	2.05	2.30	2.59	0.87	0.52
6	1.11	1.40	1.91	2.07	2.23	2.51	0.83	0.50
7	0.80	1.16	1.61	1.89	1.96	2.14	0.80	0.48
8	0.85	1.29	1.57	1.72	1.86	2.07	0.57	0.43
9	0.77	1.02	1.42	1.58	1.68	2.02	0.66	0.40
10	0.64	0.93	1.52	1.79	1.90	2.35	0.97	0.58
11	0.49	1.08	1.54	1.46	2.07	2.54	0.99	0.66
12	0.41	1.18	1.74	1.93	2.57	2.85	1.40	0.88

3.3. 월 기준 지역별 경향분석

Figure 3.1은 BOD 수치에 대한 Table3.4의 월별 결과들을 낙본지역의 측정소별로 EDA 기법과 Lowess 경향들을 표현한 것이다. 여기서 Lowess 함수추정 기법에 사용된 평활 폭 (smoother span) f의 값은 2/3로 주었고, 반복수 $iter$은 2로 주었다. 사실 비모수 함수의 추정기법에서 평활폭의 선택은 평활의 정도를 결정하는 중요한 값으로, 여러 가지 다양한 값을 이용해서 가장 적합한 값을 결정하는 것이 중요한 이슈 중 하나이다. 그러므로 각 연도별 데이터마다 평활의 적절한 폭을 선택하고 결정해야 하는 번거로운을 가져야한다. 특히 커널함수나 스플라인 평활함수를 사용할 때는 더욱 많은 신경을 쓰야한다 (Hub, 2012). 그러나 다행히도 평활 폭을 2/3로 했을 때에, Figure 3.1과 Figure 3.2에서의 Lowess 함수추정은 무리가 없어 보인다. 이런 이유로 평활 폭을 2/3를 동일하게 사용하였지만 Lowess 한수추정에서 좀더 세밀히 적절한 평활폭을 찾는 알고리즘의 개발이 필요하다.

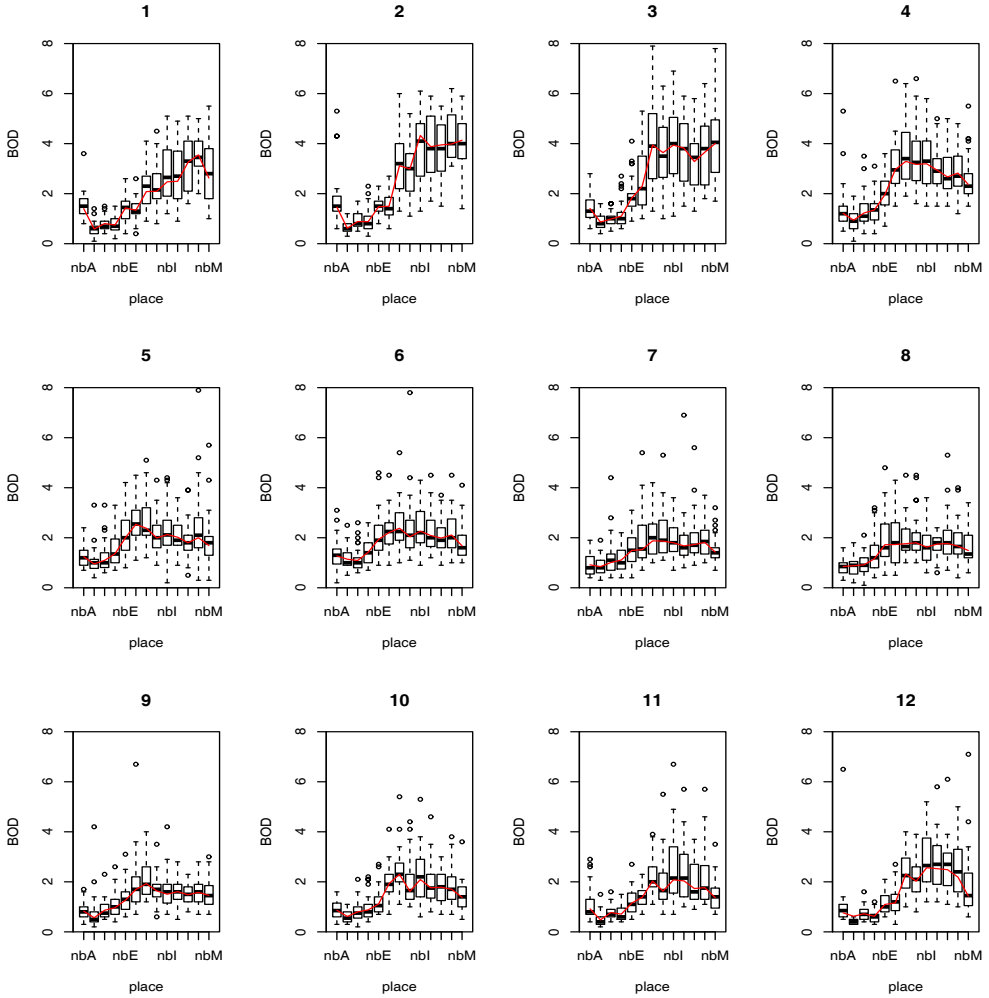

Figure 3.1 EDA and Lowess anaysis of BOD by month in Nakdong main stream river

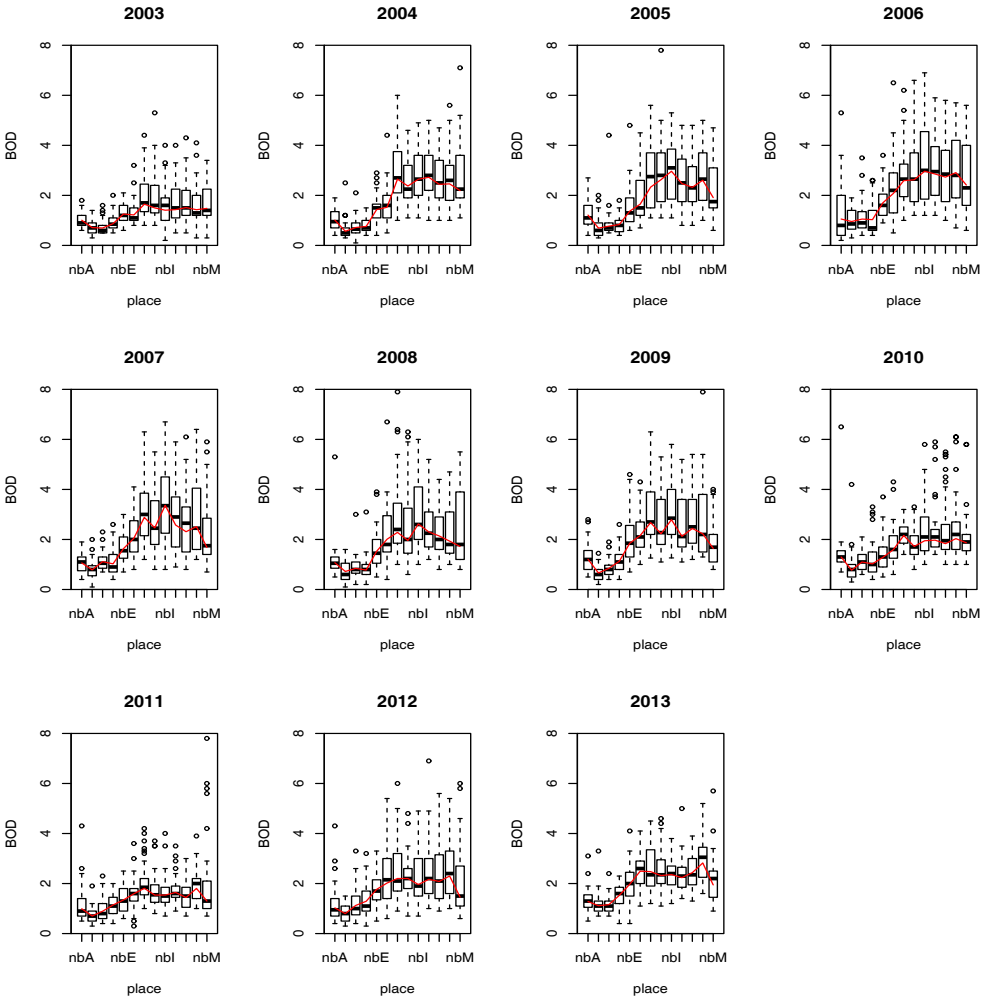

Figure 3.2 EDA and Lowess anaysis of BOD by year in Nakdong main stream river

Figure 3.1 안의 상자그림들에서 위 · 아래 분위수간 차이인 산포 ($spr(H)$) 경향은 1월에서 3월로 갈수록 $spr(H)$의 값들은 점차 증가하여 3월에 이르러 최고점에 달하고, 4월부터는 $spr(H)$의 값들이 점차 낮아지다가, 11월을 저점으로 다시 $spr(H)$의 값들이 커짐을 보인다. 7월, 8월 9월의 BOD 수치들의 산포 $spr(H)$들이 다른 달들보다 대체로 낮게 나타난다. 10월부터 BOD 수치들은 점차 증가하여 3월에 이르러서 최고에 도달한다.

Lowess의 경향을 살펴보면 1월에는 nbL인 낙본하류지역인 부산광역시 북구 금곡동지역이 BOD 수치의 최대값을 보이고, 2월에는 낙본중류지역인 nbI, 3월에는 낙본중류지역인 nbG와 nbI, 4월에는 nbG, 5월에는 경상북도 성주군 용암면 nbF, 6에는 nbG, 7월에는 nbG, 8월에 nbI, 9월에 nbG, 10월에 nbG, 11월에 nbI, 12월에 nbI 순으로 BOD의 최대값들을 나타낸다.

가장 최대값들을 많이 보이는 지역은 낙본중류지역인 nbG의 대구광역시 달성군 구지면 지역과 낙본

중류지역인 nbI인 경상남도 창원시 의창구 북면 지역이 BOD 수치가 높이 나타남을 보인다. 낙본상류지역은 낙동강 제일 상류지역 측정소 nbA의 경우, 8월에 가장 낮은 BOD 수치를 나타내지만, 9월부터 점차 nbA의 BOD 수치가 증가하기 시작하여 2월에 이르러서는 nbA보다 하류지역들인 nbB, nbC, nbD 지역의 BOD 수치들보다 높은 수치를 나타낸다.

3.4. 연도 기준 지역별 경향분석

Figure 3.2는 2003년부터 2013년까지 연도별 낙본상류지역에서 낙본하류지역까지의 BOD 수치에 대한 상자그림과 Lowess추정을 한 것이다. 2003년의 낮은 BOD의 수치들은 2004년부터 점차 낙본중류지역을 중심으로 높아지는 경향을 나타낸다. 2006년에 이르러 낙본중하류지역의 $spr(H)$의 폭들이 매우 넓게 나타나고 있다. 2007년부터는 낙본중류지역이 낙본하류지역보다 더 높은 수치를 보인다. 이는 2009년까지도 같은 분포 모양을 나타내고 있다. 2010년에는 낙본중류지역의 BOD 수치가 낮아져서 낙본하류지역의 수치와 비슷한 경향을 보이고, 2011년에는 좀 더 낮은 수치를 보인다. 그러나 2012과 2013년에는 BOD의 수치가 다시 높아지는 현상을 보이고 있다.

4. 특이점 분석과 결론

다음의 표들은 월별 연도별 지역별 특이점 (outlier)들을 분석한 것이다. EDA를 이용한 특이점의 계산은 다음과 같다. 상자그림에서 위 · 아래 분위수간 차이를 산포 (spread; $spr(H)$)이라 할 때, 양쪽 안울타리 (inner fence)값 IF_L과 IF_U은 다음과 같이 정의된다.

$$IF_L = H_L - 1.5 \cdot spr(H)$$

$$IF_U = H_U + 1.5 \cdot spr(H)$$

Table 4.1 Outlier of BOD data by month in Nakdong main stream river

month	n	nooutiier	outlier
1	295	7	2.37 %
2	387	7	1.81 %
3	516	9	1.74 %
4	530	14	2.64 %
5	536	18	3.36 %
6	534	19	3.56 %
7	458	18	3.93 %
8	386	19	4.92 %
9	430	13	3.02 %
10	467	21	4.50 %
11	487	14	2.87 %
12	333	14	2.70 %

Table 4.2 Outlier of BOD data by year in Nakdong main stream river

year	n	nooutiier	outlier
2003	444	17	3.83 %
2004	494	10	2.02 %
2005	465	9	1.94 %
2006	466	10	2.15 %
2007	486	11	2.35 %
2008	489	13	2.34 %
2009	555	13	2.34 %
2010	524	31	5.92 %
2011	559	28	5.01 %
2012	521	15	2.88 %
2013	374	12	3.21 %

Table 4.3 Outlier of BOD data by place in Nakdong main stream river

place	n	nooutiier	outlier
nbA	396	14	3.54 %
nbB	410	12	2.93 %
nbC	408	17	4.17 %
nbD	451	12	2.66 %
nbE	408	11	2.70 %
nbF	401	12	2.99 %
nbG	406	16	3.94 %
nbH	406	15	3.69 %
nbI	407	7	1.72 %
nbJ	408	13	3.19 %
nbK	451	10	2.22 %
nbL	404	11	2.72 %
nbM	403	19	4.71 %

양쪽 안 울타리의 바깥에 있는 자료점들을 특이점 (outlier)으로 간주한다 (Huh, 2012). BOD의 특이점은 8월 > 10월 > 7월 > 6월 > 5월 > 9월 > 11월 > 12월 > 4월 > 1월 > 2월 > 3월의 순으로 가장 많이 나타난 달인 8월은 가장 적게 나타난 달인 3월 보다는 2.8배 이상의 특이점이 발생하는 것으로 분석된다.

연도별로는 2010년 5.92%, 2011년 5.01%가 발생한 것으로 분석된다. 그러나 2012년과 2013년에도 2.88% 3.21%으로 2009년 이전에 비하면 BOD 수치의 특이점이 높게 나타나고 있는 것으로 분석된다. 이것은 4대강 유역 사업과 연관성이 있어 보인다.

지역에 따른 BOD 수치의 특이점의 발생은 가장 BOD의 수치가 높게 분석되었던 nbI 지역에서 특이점의 발생 비율이 1.72%로 가장 적게 나온 것은 흥미 있는 현상이다. 특이점 비율이 가장 높게 나온 낙본하류지역의 가장 끝자리인 nbM은 4.71%으로 가장 높게 나타나고, nbC 4.17%, nbG 3.94%, nbH 3.69%, nbA 3.54%, nbF 2.99%, nbJ 3.19% 낙본중류지역과 낙본하류지역에서 산발적으로 높게 나타나고 있다. nbM과 nbC 지역이 특이점이 가장 많이 발생하고 있고, 이에 따른 대책이 필요해 보인다. nbI 지역은 특이점의 발생은 낮은 비율을 보이지만 BOD의 수치가 높은 지역으로 주의 깊게 관찰할 필요성이 있다.

본 연구에서의 분석방법으로 '공분산 함수에 기초한 공간모형'을 기법을 고려하여 선형모형의 시계열적 방법을 시도하짐지만, 거리가 가까운 유역에서는 적용이 잘되지만 낙동강본류의 전역을 고려할 때에 이모형이 잘 적용되지 못했다. 이러한 이유는, 이 논문의 결과에서 보듯이, 강물의 흐름의 영향보다는 탐사 지역이 샛강을 끼고 있는가 하는 것이 굉장히 큰 변수로 작용한다. nbG 유역은 금호강을 샛강으로 nbI 유역은 남강을 샛강으로 nbN은 김해시의 하수천을 샛강으로 끼고 있다. 만약 기회가 된다면 샛강들과 강물의 흐름에 영향 연계한 'Matern 공분산 함수에 기초한 공간모형을 고려한 연구'를 해볼 생각이다.

References

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistician Assocociation*, **74**, 829-836.

Cleveland, W. S. (1993). *Visualizing data*, Hobart Press, New York.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of American Statistician Assocociation*, **83**, 596-610.

Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, **19**, 577-613.

Lee, H. S. and Lee, S. U. (2006). A study on the long-term trend analyses of water qualities in Nakdong river. *Proceedings of the Autumn Conference of Korean Society of Environmental Engineers*, 1144-1149.

Kang, H., Jang, J. H., Ahn, J. H. and Kim, I. J. (2011). Numerical estimations of Nakdong river flows through linking of watershed and river flow models. *Journal of Korea Water Resource Association*, **44**, 577-590.

Kim, J. H. and Park, S. S. (2004). Long-term trend analysis of water qualities in Nakdong river based on non-parametric statistical methods. *Journal of Korean Society Water Quality*, **20**, 63-71.

Kim, J. T., Lee, B. J. and Kim, J. Y. (2007). Trend analysis of distribution of stream qualities in Gumho river. *Journal of the Korean Data & Information Science Society*, **18**, 713-719.

Huh, J. (2012). Bandwidth selection for discontinuity point estimation in density. *Journal of the Korean Data & Information Science Society*, **23**, 79-87.

Journal of the Korean Data & Information Science Society
2014, **25**(1), 119–130

http://dx.doi.org/10.7465/jkdi.2014.25.1.119
한국데이터정보과학회지

Lowess and outlier analysis of biological oxygen demand on Nakdong main stream river[†]

Jong Tae Kim[1]

[1]Department of Computing and Statistics, Daegu University

Received 9 December 2013, revised 2 January 2014, accepted 7 January 2014

Abstract

This paper is based on water information system of NIE, National Institute of Environmental Research. We used monthly data of water quality from January, 2013 to August, 2013 starting from measuring point A (nbA) to measuring point N (nbN) located along the Nakdong river main stream. Statistical water quality analysis of BOD (biological oxygen demand) is specified by R programming depending on month, year, and points. Based on BOD measured from Nakdong river's measuring points, we used exploratory data analysis and locally weighted scatter plot smoother (Lowess) trend analysis, which is a method of non-parametic regression analysis, to analyze long-term water tendency and water quality distribution depending on points. Also, we analyzed the period and the measuring point of which the outliers are abundant. As a result, compared to BOD measured in nbM located in Busan along the downstream, BOD measured in nbG located in Daegu and nbI located in Changwon along the midstream showed higher rate of water pollution at a severe level.

Keywords: Biological oxygen demand, exploratory data analysis, locally weighted scatter plot smoother, outlier.

[†] This research was supported by the Daegu University Research Grant 2011.
[1] Professor, Department of Computing and Statistics, Daegu University, Kyungsan 712-714, Korea. E-mail: jtkim@daegu.ac.kr