엔트로피 방법을 이용한 수질측정망의 평가

유철상[1] · ○ 권상우[2]

1. 서론

현재 우리나라는 하천수, 호소수, 상수원수, 농업용수, 공단배수, 도시관류 측정망으로 구성된 총 1,574개의 수질측정지점을 운영하고 있다(환경부, 2000). 그러나 이러한 측정망들은 미리 준비된 체계적인 측정망 설계방법이나 설계계획 등에 의한 것은 아니었고, 주로 경험적인 차원에서 측정위치를 결정했던 것으로 평가된다. 이와 같은 문제점은 관측된 자료가 수계 전체의 수질 현황을 제대로 반영하지 못하고, 수질 측정의 목적을 제대로 달성하지 못하는 결과를 초래할 가능성을 내포하고 있다. 아울러 이러한 주먹구구식 수질 측정은 경제적 측면에서도 그 효율을 크게 저하시켜 효과적인 수질관리정책의 수립을 방해하는 결과를 초래하기도 한다.

이러한 수질 측정망에 관한 국내의 연구(오경두, 1989; 최지용 등, 1996)는 최근에서야 시작되었으며, 반면에 국외의 연구(Hudak 등, 1995; Lo 등, 1996; Domagalski, 1997; Harmancioglu 등, 1999; Ozkul 등, 2000)는 국내의 경우보다 더욱 활발히 진행되고 있다.

본 연구에서는 현재까지 발표된 수질측정망의 평가 방법 중 엔트로피 방법을 이용하여 각 수질항목별로 그 적정성을 평가해 보는데 목적이 있다. 기본적으로 엔트로피 방법은 현재까지 관측된 자료에 근거하는 방법이므로 미래에 발생 가능한 수질오염 감시를 위한 측정망의 설계 목적보다는 이미 구성되어 있는 측정망을 평가하는데 장점이 있는 방법이다.따라서 엔트로피 방법은 여타 다른 방법으로 설계된 수질측정망을 추후 평가하는데 좋은 방법론을 제시한다. 본 연구는 대청댐 유역에 적용하였으며 각 수질항목별로 현재의 수질측정망이 적절한지를 판단하였다.

2. 엔트로피방법 (Entropy Method)

엔트로피방법은 수, 신호 또는 기호들로 구성된 통신신호를 분석하는 정보이론(information theory)에서 그 출처를 살펴볼 수 있다. 백색잡음(white noise)과 같이 모든 사상의 발생확률이 같을 때 엔트로피가 최대가 되며, 반대로 어떤 한 값을 나타낼 확률이 1에 가까워지는 경우 0에 접근하게 된다.

엔트로피를 근거한 4가지 기본 정보측정방법은 한계엔트로피(marginal entropy), 결합엔트로피(joint entropy), 조건엔트로피(conditional entropy) 및 정보전달(transinformation) 등이 있다. 먼저 한계엔트로피 $H(X)$는 무작위 이산변량 X에 대해 다음과 같이 정의된다.

$$H(X) = -K \sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{1}$$

1) 고려대학교 환경공학과 부교수 (envchul@tiger.korea.ac.kr)
2) 한조엔지니어링 상하수도부 사원 (ojung3@hanmail.net)

여기서, N은 확률 $p(x_i)(i=1,...,N)$를 가지는 사건들의 수를 나타낸다. 만일 위 식에서 $H(X)$가 밑이 e인 로그를 취한다면 K값은 1이 된다.

확률밀도함수 $f(x)$를 갖는 무작위 연속변량 X의 경우, X의 총 범위를 폭이 $\triangle x$가 되도록 N개의 구간으로 나누면, $i(i=1,2,...,N)$번째 구간에 있는 X값의 확률은 다음과 같이 나타난다.

$$p_i = \text{Prob}[x_i - \frac{\triangle x}{2} \leq X \leq x_i + \frac{\triangle x}{2}] = \int_{x_i-(\triangle x/2)}^{x_i+(\triangle x/2)} f(x)dx \tag{2}$$

따라서 엔트로피의 정의를 연속변량으로 확장하면 다음과 같다.

$$H(X; \triangle x) = -\int_{-\infty}^{+\infty} f(x)\log f(x)dx - \log(\triangle x) \tag{3}$$

두 무작위 독립변수 X와 Y의 총 엔트로피는 각각의 한계엔트로피의 합과 같다.

$$H(X, Y) = H(X) + H(Y) \tag{4}$$

만약, X와 Y가 추계학적으로 의존적(stochastically dependent)이라면, 결합엔트로피는 식 (4)의 총 엔트로피보다 작게 나타난다. 이러한 경우 두 연속변량의 결합엔트로피는 다음과 같이 나타낼 수 있다.

$$H(X, Y; \triangle x, \triangle y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\log f(x, y)dxdy - \log(\triangle x \triangle y) \tag{5}$$

여기서, $f(x,y)$는 X와 Y의 결합확률밀도함수이다.

X와 Y의 조건엔트로피는 Y를 알고 있는 경우 X에 남아있는 불확실성을 나타낸다. 즉,

$$H(X|Y; \triangle x) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\log f(x|y)dxdy - \log(\triangle x) \tag{6}$$

여기서 $f(x|y)$는 Y에 대한 X의 조건부 확률밀도함수를 나타낸다. 반대의 경우도 동일하게 표현될 수 있다.

마지막으로 정보전달은 X와 Y사이의 여분(redundant)의 정보, 혹은 공통된(mutual) 정보를 측정하는 또 다른 엔트로피방법이다. 이는 X와 Y의 총 엔트로피와 결합엔트로피의 차로 정의된다.

$$T(X, Y) = H(X) + H(Y) - H(X|Y) \tag{7}$$

위의 정의는 M개의 변수를 가진 다변량의 경우로 확장될 수 있다(Harmancioglu와 Alpaslan, 1992). 먼저, M개의 독립변수 $X(X_1, X_2, \ldots, X_M)$에 대해 총 엔트로피는 다음과 같이 정의된다

$$H(X_1, X_2, \ldots, X_M) = \sum_{m=1}^{M} H(X_m) \tag{8}$$

만약 이들 변수들이 서로 의존적이라면, 그들의 결합엔트로피는 다음과 같이 계산된다.

$$H(X_1, X_2, \ldots, X_M) = H(X_1) + \sum_{m=2}^{M} H(X_m|X_1, \ldots, X_{m-1}) \tag{9}$$

즉, M개의 변수들에 대해 다변량 결합확률분포 $f(x_1, \ldots, x_M)$을 사용하여 결합엔트로피를 계산하는 식은 다음과 같다.

$$H(X_1, X_2, \ldots, X_M; \triangle x_1, \ldots, \triangle x_M) = -\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_1, \ldots, x_M) \log f(x_1, \ldots, x_M)dx_1dx_2\ldots dx_M - \log(\triangle x_1, \triangle x_2 \ldots \triangle x_M) \tag{10}$$

조건엔트로피는 식 (9)로 정의되는 결합엔트로피의 차로 구할 수 있다. 즉,

$$H(X_M|X_1, X_2, \ldots, X_{M-1}) = H(X_1, X_2, X_3, \ldots, X_M) - H(X_1, X_2, X_3, \ldots, X_{M-1}) \tag{11}$$

마지막으로, $f(x_1, \ldots, x_M)$을 다변량 정규분포로 가정할 수 있다면 그 결합엔트로피는 다음과 같이 나타낼 수 있다(Harmancioglu와 Alpaslan, 1992).

$$H(X) = (M/2)\ln 2\pi + (1/2)\ln|C| + M/2 - M\ln(\triangle x) \quad (12)$$

여기서 $|C|$는 공분산행렬(covariance matrix) C의 행렬식(determinant)이고, $\triangle x$는 계급구간의 크기이다. 여기서 계급구간의 크기는 모든 변수들에 대해 일정하다고 가정한다.

3. 대상유역 및 자료의 특성

우리나라 중부내륙에서 서쪽의 황해로 흐르는 금강은 그 유역면적이 9,810 ㎢이며, 유로연장은 401㎞로써 미호천을 제외하고는 큰 지류가 없으며 본류는 소백산맥에서 발원하여 대청댐의 상류구간에서 노령산맥을 지나 하행하고 있다. 따라서 금강 상류부에는 협곡을 이룬 곳이 많으며 하천 구배는 상류부에서 1/500~1/1,000, 중류부에서는 1/1,000~1/3,000로 급한 구배를 보이며 대청댐 이하 하류구간은 1/5,000~ 1/8,000의 완만한 구배를 보이고 있다.

금강유역은 환경부의 "수질측정망 운영 계획"에 의해 선정된 수질측정망에 대해 금강환경관리청이 중심이 되어 시·도 보건환경연구원(대전광역시, 충청남·북도)에서는 하천수, 상수원수 및 도시관류하수에 대한 측정망을, 한국수자원공사는 하천수 및 호소수와 관련된 측정망을, 농어촌진흥공사에서는 농업용수에 관련된 측정망을 각각 운영하고 있다.

본 연구의 적용유역인 대청댐 상류부에는 총 17개의 수질 측정지점이 설치되어 있다. 대청댐 유역의 수질자료는 환경부를 통해 구할 수 있으며, 1994년 이전의 자료는 일부 지점의 자료가 누락되어있어 본 연구에도 1994년 1월 이후의 자료(BOD, COD, pH, SS, T-N, T-P)를 사용하였다(1994년 1월~1999년 12월).

4. 적용 예

본 연구에서는 각 수질변수가 모두 대수정규분포를 따른다고 가정하여 χ^2 검정법 및 Kolmogorov-Smirnov 검정법을 통해 그 적정성을 판단하였다. 검정 결과 대부분의 경우 각 수질항목에 대해 대수정규분포를 가정하는 것에 문제가 없음을 나타내고 있으나 일부 지점 일부 수질항목의 경우에는 대수정규분포가 적절하지 않은 것으로 나타나기도 하였다. 그러나 본 연구에서 사용하는 엔트로피방법의 적용을 위해서는 모든 지점에 동일한 확률밀도함수를 사용해야 하므로 모든 지점 모든 수질항목에 대해 대수정규분포를 적용하였다.

각 수질항목에 대해 현재의 수질 측정망이 적절한지를 판단하기 위해 정보 행렬(information matrix)을 구성하였다. 이 정보 행렬은 각 수질항목별로 계산되며 각 측정지점에서의 한계 엔트로피, 총 엔트로피, 결합엔트로피 및 정보전달 정도를 쉽게 파악할 수 있도록 구성된다. 예를 들어 BOD 및 COD에 대한 정보 행렬은 표 3과 같이 구성된다. 수질측정지점 1번의 한계엔트로피는 정보 행렬 (1,1)에 위치한 1.997이 되며 나머지 값은 각각 수질측정지점 2,...,17과의 정보전달을 의미한다. 따라서 총 17개 수질측정지점 중 1번 측정지점을 유지할 경우 얻을 수 있는 총 정보의 양은 4.084가 된다.

표 1. BOD에 대한 정보 행렬

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.997	0 327	0 192	0 124	0.119	0.164	0.086	0 236	0 131	0 158	0 105	0 117	0 106	0 091	0 000	0 010	0 121
2	0.327	2 175	0.323	0.075	0.105	0.083	0.168	0.064	0 046	0.089	0 108	0.068	0 034	0 036	0 000	0 061	0 076
3	0 192	0.323	2 311	0.173	0.111	0.095	0.087	0 119	0 135	0 071	0.126	0 071	0 041	0 065	0 023	0 042	0 153
4	0 124	0 075	0.173	1.594	0.035	0.045	0.034	0.070	0 152	0.126	0.048	0.109	0 071	0 092	0.006	0 027	0 083
5	0 119	0 105	0 111	0 035	2.086	0.028	0 006	0.082	0.090	0 113	0 018	0.047	0 025	0 066	0.000	0 034	0 083
6	0 164	0 083	0 095	0.045	0.028	2 349	0 051	0.060	0.054	0 029	0 289	0.033	0 049	0 072	0 006	0.075	0 131
7	0.086	0.168	0 087	0 034	0.006	0 051	2.912	0 018	0.015	0 024	0 128	0 007	0 001	0 003	0 000	0 016	0 018
8	0.236	0.064	0.119	0.070	0 082	0 060	0.018	1 967	0.209	0.125	0.044	0 048	0 031	0.056	0 002	0 001	0 008
9	0 131	0 046	0.135	0.152	0 090	0.054	0.015	0 209	2.251	0 025	0 055	0 019	0 006	0.025	0 001	0 000	0 045
10	0 158	0 089	0.071	0.126	0.113	0.029	0.024	0 125	0 025	3 006	0.022	0.262	0.339	0.207	0 011	0 051	0 055
11	0.106	0.108	0.126	0.048	0.018	0.289	0.128	0.044	0 055	0 022	2.738	0 023	0 014	0 014	0 036	0 066	0 086
12	0.117	0 068	0.071	0.109	0.047	0 033	0.007	0.048	0.019	0 262	0 023	1 957	0 602	0 413	0 017	0 066	0 186
13	0.106	0.034	0 041	0 071	0.025	0 049	0 001	0.031	0.006	0 339	0 014	0.602	2 565	0.527	0 043	0 086	0 157
14	0 091	0 036	0 065	0 092	0 066	0 072	0 003	0.056	0.025	0 207	0 014	0 413	0 527	2 004	0 053	0 068	0 176
15	0 000	0.000	0 023	0.006	0.000	0 006	0 000	0.002	0.001	0 011	0 036	0.017	0 043	0 053	1 302	0 021	0 026
16	0 010	0.061	0 042	0 027	0 034	0.075	0.016	0.001	0.000	0 051	0 066	0.066	0 086	0.068	0.021	2 990	0 125
17	0 121	0 076	0.153	0.083	0 083	0 131	0 018	0 008	0.045	0 055	0 086	0.186	0 157	0 176	0 026	0 125	2 012
SUM	4 084	3 837	4.137	2.865	3 048	3.611	3 572	3 139	3.260	4 712	3 919	4.042	4 697	3 965	1 547	3 737	3 542

만일 한 개의 수질 측정지점만을 선택해야한다면 총 엔트로피가 가장 큰 지점 10을 선택해야 한다. 그러나 만일 2개 이상의 지점을 유지시킨다면 그 지점들은 단순히 각 지점에서 얻을 수 있는 총 엔트로피의 비교로 결정될 수 없다. 이는 지점에 따라 한계엔트로피와 각 지점사이의 정보 전달이 다르게 나타나기 때문이다. 따라서, 임의의 m개 지점을 선정하는 경우 결정되어야 할 지점은 다음의 조건식에 의하여 결정된다.

$$MaxT(X_1, X_2, \ldots, X_m; X_k, X_l, \ldots, X_s) = Max\left[H(X_k)+H(X_l)\cdots+H(X_s)+\sum_{i=1}^{m-p}\sum_{j=1}^{p}T(X_i;X_j)\right] \quad (13)$$

여기서, p는 총 m개의 수질측정지점에서 선택된 수질측정지점의 수를 나타낸다. 이와 같은 방법으로 각 수질지점의 중요도에 따라 순위를 매길 수 있으며 이를 각 수질항목별로 정리하여 표 5에 나타내었다.

표 5에서 살펴볼 수 있듯이 각 수질항목별로 결정된 중요 지점의 순위는 서로 다르게 나타났으며, 경우에 따라서는 아주 극단적인 결과를 보여주기도 했다. 예를들어 항건천 지점의 경우 BOD, pH, SS는 상위순위를, COD, T-N, T-P의 경우는 최하위의 순위를 보여주고 있다. 이는 앞에서도 언급한 것처럼 각 수질항목별로 그 영향 인자가 서로 다를 수 있다는 점과, 아울러 엔트로피 방법의 적용은 관측치에만 의존하므로 그 결과도 관측자료의 양 및 질에 의해 결정된다는 점을 이유로 들 수 있다.

표 6은 상위 5개 지점. 상위 10개 지점 및 상위 15개 지점을 유지시키는 경우와 현재의 측정망을 유지하는 경우를 비교하여 얻을 수 있는 정보의 양을 상대적으로 나타낸 것이다. 표 6에서 살펴 볼 수 있듯이 상위 5개 지점을 유지시킬 경우 얻을 수 있는 정보의 양은 현재상태와 비교하여 약 40-50%, 상위 10개 지점을 유지시키는 경우는 약 70-80%, 마지막으로 상위 15개 지점을 유지시키는 경우는 약 93-99%가 된다. 특히, 상위 몇 개 지점에서 얻을 수 있는 정보의 양이 측정지점의 상대적 비율에 비해 크게 나타나고 있음에 주목할 만하다. 이러한 결과는 주요 수질측정지점 선정의 중요성을 나타내는 결과로 이해될 수 있다. 아울러, 각 수질항목별로 얻을 수 있는 정보의 양이 지점 수의 증가에 따라 다른 형태로 나타나는 것을 발견할 수 있는데, 이는 각 수질항목별로 그 영향인자가 다르기 때문이다. 즉, 일부 수질항목의 경우는 일정 수 이상의 지점에서 총 정보의 양이 정체하고 있는 형태를 보여 주고 있으나 (BOD, pH, T-N), 다른 수질항목의 경우는 (COD, SS, T-P) 계속 선형적으로 증가하고 있는 형태를 나타내고 있어 주어진 수질관측망이 이들 수질

을 관측하기에 충분한 상태가 아님을 알 수 있다. 따라서, 관측목적에 따라 (또는 수질항목에 따라) 현재의 수질관측망이 적절할 수도 또는 부적절할(불충분할) 수도 있다는 결론이 된다.

표 5. 각 수질항목별로 엔트로피방법을 적용하여 얻은 지점별 중요순위

지점번호	지점명	수질항목					
		BOD	COD	pH	SS	T-N	T-P
1	옥천천	8	6	17	8	9	6
2	영동	15	1	2	16	12	1
3	황간천	3	17	1	1	17	17
4	우산	16	16	13	9	14	16
5	보청천1	10	9	14	3	3	9
6	보청천2	11	13	8	14	13	13
7	초상2	6	10	15	2	8	10
8	봉황천	13	14	16	5	15	14
9	보청천4	7	5	10	17	7	5
10	용포	1	4	5	11	4	4
11	영동천2	4	8	3	6	1	8
12	보청천3	17	15	4	12	11	15
13	옥천	2	11	12	10	5	11
14	영동천1	12	2	6	7	2	2
15	제원	14	12	9	15	16	12
16	무주남대천	5	3	7	4	6	3
17	초강1	9	7	11	13	10	7

표 6. 각 수질항목별 총 엔트로피 및 상위 5, 10, 15개 지점의 엔트로피 비교 (괄호 안 : 총 엔트로피에 대한 상대 %비).

지점번호	총 엔트로피	상위 5	상위 10	상위 15
BOD	38.2(100%)	19.5(51.0%)	31.7(83.0%)	37.8(98.9%)
COD	46.5(100%)	17.8(38.2%)	32.0(68.9%)	43.2(93.0%)
Ph	40.5(100%)	19.9(49.1%)	32.5(80.2%)	40.3(99.5%)
SS	74.8(100%)	29.0(38.8%)	51.5(68.8%)	69.9(93.4%)
T-N	51.7(100%)	25.6(49.5%)	42.2(81.5%)	51.5(99.6%)
T-P	62.2(100%)	26.5(42.5%)	45.3(72.8%)	59.7(96.0%)

5. 결론

본 연구에서는 대청댐 유역을 적용유역으로 하여 각 수질항목별로 현재의 수질측정망이 적절한지를 판단하였으며 연구의 결과를 요약하면 다음과 같다.

(1) 각 수질항목별로 결정된 중요 지점의 순위가 각각 다르게 나타났다. 이는 각 수질항목별로 그 영향 인자가 서로 다를 수 있다는 점과, 아울러 엔트로피 방법의 적용은 관측치에만 의존하므로 그 결과도 관측자료의 양 및 질에 의해 결정된다는 점을 이유로 들 수 있다.

(2) 대부분의 경우 상위 순위의 지점들에 의한 정보의 양은 측정지점 수의 상대적인 비율에 비해 아주 크게 나타났으며, 이는 상위 순위의 지점들에 의한 정보의 양이 총 정보의 양에 큰 영향을 미치고 있음을 의미하는 결과이다. 이러한 결과는 주요 수질측정지점 선정의 중요성을 나타내는 것으로 이해될 수 있다.

(3) 각 수질항목별로 얻을 수 있는 정보의 양이 지점 수에 따라 다르게 나타났다. 즉, 일부 수질항목의 경우는 일정 수 이상의 지점에서 총 정보의 양이 정체하는 형태를 (본 연구에서는 BOD, pH, T-N의 경우), 다른 수질항목의 경우는 (COD, SS, T-P) 계속 선형적으로 증가하고 있는 형태를 나타내고 있어 주어진 수질측정망이 수질항목에 따라 적절하지(또는 충분하지) 않은 상

태임을 알 수 있다.

서론에서도 언급한 바와 같이 엔트로피 방법은 새로운 수질 측정망의 설계보다는 이미 설치된 측정망의 평가에 유리한 방법이라 평가할 수 있다. 그러나 이 방법은 과거 관측자료의 양 및 질에 전적으로 의존하므로 관측자료에 문제가 있는 경우 그 결과의 신뢰도도 당연히 떨어지게 된다는 점을 밝혀둔다. 아울러 우리나라의 경우 기존의 측정망은 수질측정의 측면에서 포화된 상태라고 판단하기는 어려우므로 수질측정망의 평가 결과 부정적인 결과가 도출되었다고 하더라도 현 측정지점의 폐지 또는 재선정 보다 측정지점의 추가를 통한 기존 측정망의 보완이 더 효율적인 방법이라는 판단을 하게된다. 새로운 측정지점을 추가하는 경우에도 엔트로피 방법의 적용은 가능하며, 이러한 경우 추가하고자 하는 지점의 결정은 가상의 지점에 대해 주변의 기존 측정지점의 확률분포를 고려하여 확률분포형을 가정하고 본 연구에서의 적용방법에 따라 수질측정망의 재평가를 수행하면 된다. 새로운 최적 측정지점의 결정은 위 과정을 여러 지점에 대해 수행한 후 비교평가를 통해 이루어지게 된다.

6. 참고문헌

오경두 (1989). "수질측정망 최적설계기법에 관하여, 미워싱턴주의 수질측정망 설계기법을 중심으로." **한국수문학회지**, 한국수문학회, 제22권, 제4호, pp. 354-361.

최지용, 박원규, 이상일 (1996). "하천 및 호소수 수질관리를 위한 자동측정망의 설계." **한국수자원학회논문집**, 한국수자원학회, 제29권, 제2호, pp. 167-178.

환경부 (2000). 수질측정망 운영 계획, 환경부.

Domagalski, J. (1997). "Results of a prototype surface water network design for pesticides developed for the San Joaquin River Basin, California." *Journal of Hydrology*, Vol. 192, pp. 33-50.

Harmancioglu, N. B. and Alpaslan, N. (1992). "Water quality monitoring network design: A problem of multi-objective decision making." *Water Resources Bulletin*, Vol. 28, No. 1, pp. 179-192.

Harmancioglu, N. B., Ozkul, S. D., and Icaga, Y. (1999). "Comparison of optimization and entropy methods in assessment of water quality sampling sites." *Proceedings of the International Conference on WEESHE*, Water resources publications, LLC.

Hudak, P. F., Loaiciga, H. A., and Marino, M. A. (1995). "Regional-scale ground water quality monitoring via integer programming." *Journal of Hydrology*, Vol. 164, pp. 153-170.

Lo, S. L., Kuo, J. T., and Wang, S. M. (1996). "Water quality monitoring network design of Keelung River, Northern Taiwan." *Water Science and Technology*, Vol. 34, No. 12, pp. 49-57.

Ozkul, S., Harmancioglu, N. B., and Singh, V. P. (2000). "Entropy-based assessment of water quality monitoring networks." *Journal of Hydrologic Engineering*, ASCE, Vol. 5, No. 1, pp. 90-100.