Journal of Environmental Research Institute
Analysis of the Measured Data for BOD Concentration and Runoff by the Application of Pattern Classification Method
Chonnam Nat'l. Univ. Vol. 9, pp.125-132 2009

패턴분류 기법의 적용에 의한 BOD 농도 및 유량 측정자료의 분석

박성천† · 진영훈 · 노경범* · 문병석**

동신대학교 토목공학과 · *동신대학교 공업기술연구소
**서남대학교 토목공학과

Analysis of the Measured Data for BOD Concentration and Runoff by the Application of Pattern Classification Method

Sung-Chun Park†, Young-Hoon Jin, Kyong-Bum Roh*, Byoung-Seok Moon**

Department of Civil Engineering, Dongshin University
**Institute of Industrial Research and Technology, Dongshin University*
***Department of Civil Engineering, Seonam University*

ABSTRACT

Currently, Ministry of Environment in Korea has been measuring water quality and runoff data in the outlets of each unit area for Total Maximum Daily Loads (TMDLs) with 8-day intervals. It is critical to develop various techniques for data analysis and apply for improvement of the data utilization. Therefore, in the present study, Self-Organizing Feature Map (SOFM) was applied to classify patterns and investigate the characteristics included in the classified data. BOD concentration and runoff data which have been measured in the outlet of SeomBon_D (SB_D) station were classified into three patterns by the application of SOFM. The results showed that Pattern-1 whose data occurred during flood period includes higher values of runoff data than Pattern-2 and Pattern-3 which have low runoff data and are decomposed by 1.4 *mg/L* of BOD concentration. The frequency analysis of occurrence for the data in the respective patterns on a monthly basis revealed that the classified data into Pattern-2 occurred before flood period and Pattern-3 showed high frequency of the data occurrence after flood period. Consequently, the pattern classification result could be obtained according to the amount and occurrence frequency on a monthly basis of runoff and water quality data.

Key words : BOD concentration, Runoff, Pattern Classification, Self-Organizing Feature Map (SOFM)

† 진영훈 · E-mail : nmdrjin@gmail.com

1. 서론

현재 환경부에서는 수질오염총량관리제도와 같은 환경관리 제도의 원활한 시행을 위해 수질 및 유량측정망을 통하여 자료를 측정하고 있으며, 이를 국립환경과학원의 DB 및 웹 시스템(http://smat.nier.go.kr)에 공개하고 있다[1]. 따라서 이러한 양질의 자료에 대한 활용성 제고를 위해 기초자료 분석기법의 개발이 필요하며, 그 결과물들은 상기의 수질오염총량관리제도의 원활한 시행을 위한 기초적인 자료로 활용될 수 있을 것으로 판단된다.

그러나 현재 측정되고 있는 다양한 항목의 수질자료와 유량 자료들 사이에는 자연현상에 일반적으로 내재되어 있는 강한 비선형성으로 인해 쉽게 파악되기 어려운 관계들이 존재하고 있다. 따라서 측정 자료의 전 범위에 걸친 분석보다는 각 자료에 내재된 특성을 반영한 구간별 또는 패턴별 분석이 수행될 필요가 있다. 이러한 연구의 필요성은 자료의 전체적인 범위를 고려할 경우 파악하기 어려운 자료의 특성이 분할된 자료의 패턴에 따라 각기 다른 특성관계를 나타내고 있으며, 각 패턴별 특성의 종합적 연결을 통해 전체적인 특성을 재현하고 있는 선행연구들을 통해서 제시되고 있다[2),3),4),5)].

대상자료의 패턴을 구분하고 각 패턴별 특성을 파악하기 위해 최근 가장 활발히 적용되고 있는 패턴분류 기법인 자기조직화 지도(Self-Organizing Feature Map: SOFM)의 연구결과는 하천유량자료에 대한 적용[2),3),4),6),7),8)], 수질 및 수처리 분야[9),10)] 및 기상분야[11)]를 포함하여 다양한 분야에서 그 우수성을 나타내고 있다.

따라서 본 연구에서는 앞서 언급한 수질오염총량관리제의 단위유역들 중 하나인 섬본_D 지점(SB_D)에서 측정된 수질 자료인 BOD 농도와 유량자료를 대상으로 하여 SOFM 기법을 적용하였으며, 각 패턴의 분류기준을 파악하기 위한 분석을 수행하였다.

2. 연구방법

2.1 대상지점 및 사용자료

본 연구에서는 섬진강 수계의 수질오염총량관리제를 위한 단위 유역들 중 섬본_D 지점(SB_D)을 대상으로 하였으며(Fig 1, 점선 표시), 전라남도 구례군 구례읍 유곡나루터에 위치하고 있다.

대상지점에서 측정된 수질 항목들 중 BOD 농도와 유량자료를 앞서 언급한 환경부 국립환경부의 웹 시스템으로부터 수집하여 사용하였다. 자료기간은 2004년 9월 17일부터 2009년

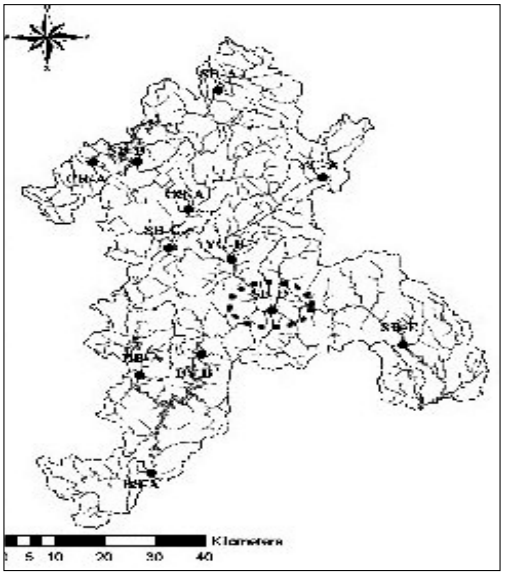


Fig 1. Location of SB_D station

3월 24일까지이며, 자료의 총 수는 179개로 8일 간격으로 측정되었다.

BOD 농도의 최소값은 0.40 *mg/L*, 최대값은 4.00 *mg/L*, 평균값은 1.38 *mg/L*로 나타났다. 또한 유량자료의 최소값, 최대값 및 평균값은 각각 5.03 m^3/s, 410.58 m^3/s 및 32.40 m^3/s로 산정되어 유량이 BOD 농도 자료보다 그 편차가 심한 것으로 나타났다.

2.2 적용방법

SOFM은 훈련과정에 목표값을 갖지 않는 비교사 학습방법(unsupervised learning algorithm)이며, 인공신경망 이론의 한 종류이다. 이러한 SOFM은 다차원의 입력 자료들을 분류하여 2차원으로 사상시킬 수 있는 특징을 가지고 있다. SOFM의 장점으로는 먼저 복잡한 다차원 자료에 대한 패턴분류를 위해 그 적용성이 뛰어나며, 자료의 가시화가 쉽고 이에 따라 자료의 특성 파악을 위한 자료 분석 도구로 활용되고 있다.

SOFM 구조는 기본적으로 입력층과 출력층을 갖게 되며, 입력층의 노드의 수는 입력 자료를 구성하는 변수의 수가 m일 때, 이에 따른 m개의 입력노드를 갖게 된다. 또한 입력된 자료를 l개의 노드로 구분하고자 할 경우, 출력층의 노드의 수는 l개를 갖게 된다. 입력층과 출력층의 모든 노드들은 서로 연결되고, 각 노드들 사이에 연결강도를 갖게 된다. 입력층의 각 노드는 입력 자료를 네트워크로 전달하며, 출력층의 노드는 입력 자료와 입·출력노드 사이의 연결강도를 이용하여 거리를 계산한다.

SOFM의 훈련과정은 경쟁과정, 근접반경 조정과정 및 연결강도 조정 과정을 포함한 3단계로 진행된다. 경쟁과정은 다음의 식 (1)과 같은 m차원의 입력자료(X)와 식 (2)와 같은 출력노드 j의 연결강도(W)에 대하여 식 (3)을 적용하며, 그 결과로 출력노드 중의 승자노드($i(X)$)를 결정한다. 즉 승자노드의 선택은 입력 자료의 패턴과 가장 유사한 연결강도를 선정하는 것이며, 유사한 정도를 측정하기 위해 유클리드 거리를 이용한다.

$$X=[\chi_1, \chi_1, \cdots, \chi_m]^T \quad (1)$$

$$Wi=[w_{i1}, w_{i2}, \cdots, w_{im}]^T, \; i=1,2,\cdots, \ell \quad (2)$$

$$i(X)=\arg\min \| X-Wi \| \quad (3)$$

여기서 T는 전치행렬을 의미하며, ℓ은 출력층의 전체 노드의 수이다.

또한 승자노드와 이에 인접한 이웃 노드들만이 제시된 입력 자료에 대한 학습이 허용된다. 인접노드를 결정하는 반경에 따라 학습이 진행되는 노드의 수가 결정되며, 이 반경은 학습이 진행됨에 따라 서서히 줄어들어 점점 적은 개수의 노드들이 학습을 하게 된다. 일반적으로 기하학적 이웃반경의 조정을 위해서 대칭성과 수렴특성을 지닌 가우시안 함수(Gaussian function)를 이용한다. 이를 근접반경 조정과정이라 하며, 최종적으로 단지 승자노드만이 그것의 연결강도를 조정하게 된다.

상기의 경쟁과정 및 근접반경 조정과정의 단계가 끝나면 마지막으로 적응학습과정에 의해 실제 연결강도의 조정이 이루어진다. 조정되기 이전의 연결강도를 , 조정된 후의 새로운 연결강도를 $Wi(n+1)$이라 할 때, 이산적인 시간간격에 대한 조정규칙은 다음 식 (4)와 같이 표현된다.

$$Wi(n+1) = W_i(n) + \eta(n) \cdot h_{j,i(X)}(n) \cdot [X - W_j(n)] \quad (4)$$

여기서 η는 시간 n이 증가함에 따라 서서히 감소하는 학습율을 나타내는 매개변수이며, $h_{j,i(X)}$는 근접반경 조정과정의 기하학적 이웃반경을 나타낸다.

3. 결과 및 고찰

3.1 패턴분류 결과

본 연구에서는 시행착오법에 의해 SOFM의 구조를 결정하였으며, 그 결과 4×4의 육각형 배열을 갖는 구조로 결정하였다. 또한 이러한 구조를 갖는 SOFM에 대하여 분류 가능한 최소 및 최대 패턴의 수를 2개에서 16개까지 적용하였다. 이에 따른 최적의 패턴 수를 결정하기 위해 Lóez and Machón(2004)에 의해 제안된 DBI(Davies-Bouldin Index)를 각 클러스터의 수에 따라 산정하여 Fig 2에 나타내었다. DBI가 낮은 값을 보일수록 적정한 패턴분류 결과를 보이는 것으로 판단되므로, 본 연구에서는 Fig 2에 도시된 바와 같이 3개의 패턴으로 분류하는 것이 최적인 것으로 나타났다.

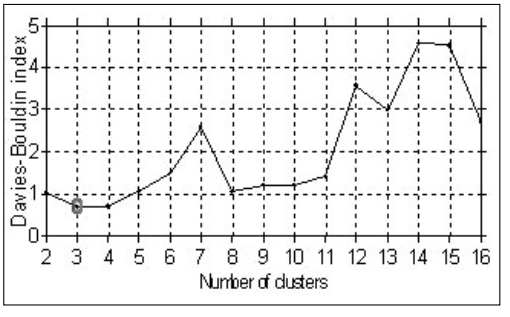


Fig 2. DBI values according to the number of clustersÄÄ

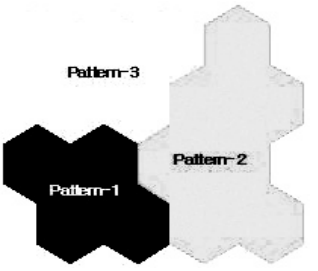


Fig 3. Pattern classification result

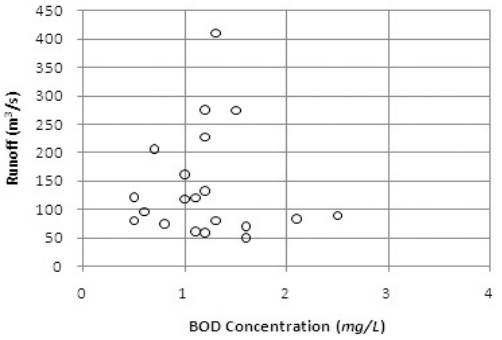
따라서 4×4의 육각형배열을 갖는 SOFM 구조에 의해 총 3개의 패턴으로 구분된 결과를 Fig 3에 나타내었다. 이는 2차원(BOD 농도 및 유량자료)을 갖는 179개의 입력 자료를 이용한 SOM의 패턴분류 결과를 보여주고 있다. 첫 번째 패턴(Pattern-1)에는 20개의 자료가 분류되었으며, 두 번째 패턴(Pattern-2)에는 90개, 마지막으로 세 번째 패턴(Pattern-3)에는 69개의 자료가 분류되었다.

분류된 자료의 분포를 살펴보기 위하여 Fig 4에 BOD 농도에 대한 유량자료를 각 패턴별로 도시하였다. Fig 4. (a)에 도시한 바와 같이 패턴-1은 49.88 m^3/s 이상의 유량에 해당하는 자료가 분류된 것으로 나타났으며, 패턴-2와 패턴-3에 비하여 상대적으로 고유량 자료가 분류되었다. 패턴-1에 비하여 상대적으로 저유량에 해당하는 패턴-2(Fig 4. (b))와 패턴-3(Fig 4. (c))은 BOD 농도 1.40 *mg/L*의 값을 기준으로 분류되었다.

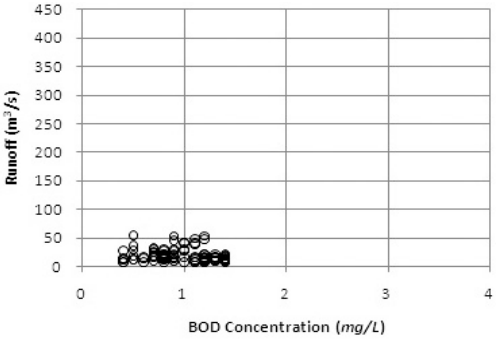
이러한 패턴분류 결과에 대한 분류기준을 각 자료별로 나누어 분석하였으며, 그 분석결과를 아래에 기술하였다.

3.2 유량 자료에 의한 패턴분류

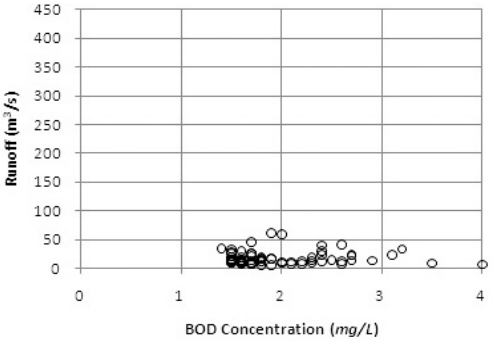
SOFM을 적용하여 총 3개로 분류된 패턴의



(a) Pattern-1



(b) Pattern-2



(c) Pattern-3

Fig 4. Data distribution plot according to the respective patterns classified by SOFM (4x4 hexagonal array)

분류기준을 분석하기 위해 먼저 유량자료의 분석을 수행하였다. Fig 5에 도시한 바와 같이 각 패턴별 유량자료의 최소, 1분위값, 중앙값, 3분위값 및 최대값을 박스플롯을 이용하여 나타내었다. 패턴-1로 분류된 유량자료의 분포를 Fig 5. (a)에 나타내었으며, Fig 5. (b)의 패턴-2 및 Fig 5. (c)의 패턴-3에 대한 박스플롯과 비교할 때 49.88 m^3/s 이상의 상대적으로 큰 자료들이 패턴-1로 분류되었다.

그러나 패턴-2와 패턴-3은 유량자료의 분포를 고려할 때 유사한 범위를 나타내고 있어 유량자료에 의해 분류된 것으로 판단하기 어렵다. 따라서 유량자료의 크기에 따라 고유량 자료를 포함하고 있는 패턴-1이 나머지 패턴들과 상이성을 나타냄으로써 독립된 패턴으로 분류된 것으로 판단된다.

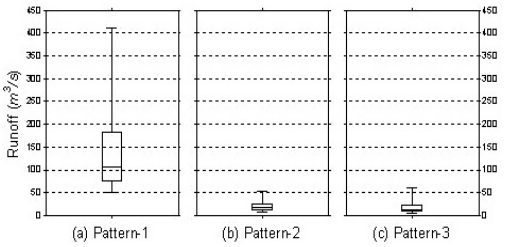


Fig 5. Box-whisker plot for the classified runoff data patterns

3.3 BOD 농도자료에 의한 패턴분류

앞서 언급한 바와 같이 패턴-1은 고유량 자료를 포함하고 있어 저유량 자료를 포함하고 있는 패턴-2와 패턴-3으로부터 독립적으로 분류되었으며, 나머지 분할결과인 패턴-2와 패턴-3의 분류기준을 파악하기 위하여 BOD 농도자료에 대한 분석을 수행하였다

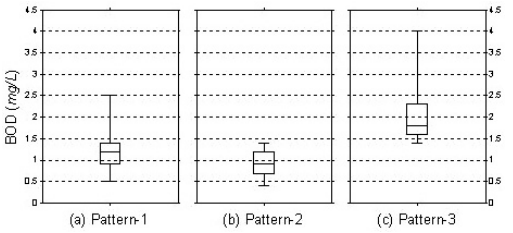
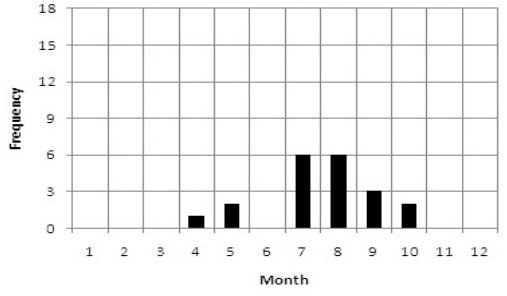
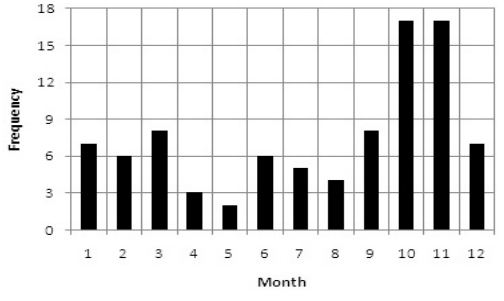


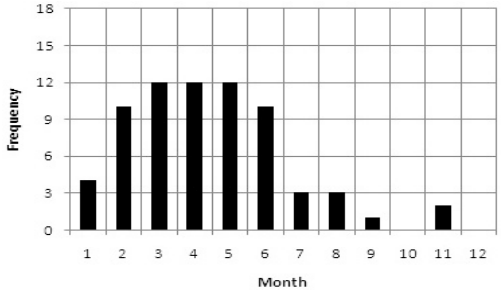
Fig 6. Box-whisker plot of BOD concentration for the classified data



(a) Pattern-1



(b) Pattern-2



(c) Pattern-3

Fig 7. Histograms according to respective patterns classified by SOFM

유량자료의 분석과정과 마찬가지로 각 패턴별 박스플롯을 통하여 BOD 농도자료의 분포를 파악하였다(Fig 6). 패턴-1(Fig 6. (a))의 BOD 농도자료의 분포는 패턴-2(Fig 6. (b)) 및 패턴-3(Fig 6. (c))으로 분류된 자료의 분포와 중첩되어 BOD 농도자료에 의해 패턴-1이 분류되었다고 판단하기 어렵다.

그러나 패턴-2와 패턴-3은 BOD 농도 1.4 *mg/L*를 기준으로 하여 저농도의 자료가 패턴-2로 분류되었으며, 고농도의 자료가 패턴-3으로 포함되었다. 따라서 BOD 농도자료의 분류결과는 1.4 *mg/L*보다 낮은 값들의 자료가 패턴-2로 분류되고, 그보다 높은 값들의 자료가 패턴-3으로 분류되었다고 판단할 수 있다.

이에 대한 추가적인 분석을 위해 각 자료가 측정된 시기를 파악하여 각 월별 자료발생 빈도를 Fig 7에 히스토그램으로 나타내었다. 고유량의 자료가 포함된 패턴-1의 자료들은 7월에서 9월 사이에 높은 빈도를 보여 주었으며(Fig 7. (a)), 이는 패턴-1이 유량자료의 크기에 따라 독립적으로 분류된 결과의 근거를 제공해 주고 있다.

패턴-2에 포함된 자료들의 각 월별 발생빈도를 보면(Fig 7. (b)), 10월과 11월에 가장 높은 빈도로 발생한 것을 알 수 있으며, 패턴-3(Fig 7. (c))의 경우는 2월에서 6월 사이에 집중적으로 발생하였다. 이러한 자료의 각 월별 발생빈도에 따라 저농도의 BOD 농도자료가 포함된 패턴-2의 경우 홍수기 이후에 측정된 자료가 분류되었으며, 고농도의 자료가 포함된 패턴-3은 홍수기 이전의 BOD 농도자료가 분류되었음을 알 수 있다.

4. 결론

본 연구에서는 수질오염총량관리제를 위한 단위유역 중 섬본_D 지점을 대상으로 하였으며, 현재 환경부 국립환경과학원에서 8일 간격으로 측정하고 있는 항목들 중 BOD 농도자료 및 유량자료를 이용하여 패턴분류 분석을 수행하였다. 패턴분류를 위해 최근 가장 널리 사용되고 있는 방법들 중 하나인 SOFM 기법을 적용하였으며, SOFM의 구조를 4×4의 육각형 배열로 결정하였다.

SOFM의 훈련결과 총 3개의 패턴으로 분류한 것이 최적의 결과로 나타났으며, 이에 따라 각 패턴별 분류기준을 파악하기 위해 유량자료 및 BOD 농도자료의 분석을 수행하였다. 분석을 위해 유량자료에 대한 각 패턴별 박스플롯을 도시하여 패턴-1이 49.88 m^3/s 이상의 고유량 자료를 포함하고 있음을 파악하였다. 따라서 패턴-1은 유량자료에 의해 저유량 자료를 포함하고 있는 패턴-2와 패턴-3로부터 독립적으로 분류되었음을 알 수 있다.

또한 패턴-2와 패턴-3의 분류기준은 BOD 농도자료의 박스플롯 및 자료 발생빈도를 나타낸 히스토그램의 분석을 통해 파악할 수 있었다. 박스플롯의 도시 결과로부터 패턴-2와 패턴-3은 BOD 농도 1.4 mg/L를 기준으로 하여 분류된 것으로 파악되다. 이에 따라 저농도의 자료가 패턴-2로 분류되었으며, 고농도의 자료가 패턴-3에 포함되었다. 또한 자료의 각 월별 발생빈도를 도시한 히스토그램을 통해서 여름철 홍수기 이후의 자료가 패턴-2로 분류되었고, 홍수기 이전의 자료가 패턴-3으로 분류됨을 알 수 있었다.

결론적으로, 패턴분류를 위한 SOFM의 적용결과, 총 3개의 패턴으로 구분할 수 있었으며, 패턴-1은 유량자료에 의한 분류되었고, 패턴-2와 패턴-3은 BOD 농도자료에 의해 분류됨을 알 수 있었다. 또한 자료의 각 월별 발생빈도의 분석결과, 패턴-1은 여름철 홍수기에 측정된 자료이며, 패턴-2는 홍수기 이후의 자료로 파악되었고, 패턴-3은 홍수기 이전의 자료임을 알 수 있었다. 이는 홍수기를 전 · 후로 하여 대상지점의 수질이 현저히 변화하고 있음을 반영하고 있다.

따라서 향후 역전파 학습 알고리즘을 이용한 인공신경망 모형과 같은 이론적 모형에 의한 수질예측 모형 구축 시 SOFM 기법을 전처리 과정으로 활용하여 패턴분류를 수행한 후 각 패턴별 예측모형을 구축하여 예측결과의 정도(accuracy)를 향상시킬 수 있을 것으로 기대된다. 또한 본 연구의 결과 중 유량자료에 의한 분류기준인 49.88 m^3/s의 유량과 BOD 농도자료의 분류기준인 1.4 mg/L에 대한 물리적 의미를 파악하기 위한 추가적인 분석이 필요한 것으로 판단된다.

사 사

본 연구는 환경부지정 전남지역환경기술개발센터의 연구비지원에 의해 수행되었으며 이에 감사드립니다.

참고 문헌

1) 이호열, “영산강유역 수질측정망 운영 및 활용 현황”, 한국수자원학회지, 42(3), 17-

23(2009).

2) Hsu, K.L., Gupta, H.V., Gao, X., Sorooshian, S. and Imam, B., "Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis", *Water Resources Research*, 38(12), 1302 (doi:10.1029/2001WR000795)(2002).

3) 김용구, 진영훈, 박성천, "강우-유출특성 분석을 위한 자기조직화방법의 적용", 대한토목학회논문집, 26(1B), 61-67(2006).

4) 박성천, 진영훈, 김용구, "강우-유출 예측모형 개발을 위한 자기조직화 이론의 적용", 대한토목학회논문집, 26(4B), 389-398(2006).

5) 진영훈, 김용구, 노경범, 박성천, "수질 및 유량자료의 기초통계량 분석에 따른 공간분포 파악을 위한 SOM의 적용", 한국물환경학회논문집, 25(5), 735-741(2009).

6) Jain, A. and Srinivasulu, S., "Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques", *Journal of Hydrology*, 317, 291-306(2006).

7) Srinivasulu, S. and Jain, A., "A comparative analysis of training methods for artificial neural network rainfall-runoff models", *Applied Soft Computing*, 6, 295-306(2006).

8) 김용구, 진영훈, 박성천, 정천리, "나주지점의 강우-유출 해석을 위한 최적의 SOM 구조 결정", 한국수자원학회논문집, 41(10), 995-1007(2008).

9) Lóez, H. and Machón I., "Self-organizing map and clustering for wastewater treatment monitoring", *Engineering Applications of Artificial Intelligence*, 17, 215-225 (2004).

10) 김용구, 진영훈, 정우철, 박성천, "호소수의 강우, 저류량 및 TOC변동 특성분석을 위한 자기조직화 방법의 적용", 한국물환경학회논문집, 24(5), 611-617(2008).

11) Nishiyama, K., Endo, S., Jinno, K., Uvo, C.B., Olsson, J. and Berndtsson, R., "Indentification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a Self-Organizing Map", *Atmospheric Research*, 83, 185-200(2007).