하천의 수질 및 유량자료의 패턴분류에 의한 특성 파악

Detection of Characteristics by Pattern Classification of Water Quality and Runoff Data in a River

박성천*, 진영훈**, 노경범***, 김용구****, 이용희*****

Sung-Chun Park, Young-Hoon Jin, Kyong-Bum Roh, Yong-Gu Kim

요 지

현재 환경부에서는 수질오염총량관리제를 위하여 각 단위유역의 말단지점에서 8일 간격으로 수질 및 유량을 측정하고 있으며, 이 자료들을 공개하고 있다. 이러한 양질의 자료의 활용성을 제고하기 위해서는 무엇보다도 자료의 분석을 위한 다양한 기법이 개발되고 제안되어야 한다. 따라서 본 연구에서는 수질 및 유량자료를 동시에 적용하여 두 자료 사이의 관계를 조사하고 특성을 파악하기 위하여 자기조직화 특성지도(Self-Organizing Feature Map: SOFM) 이론을 적용하였다. 시행착오법에 의해 적정한 SOFM 구조를 결정하였으며, 그 결과 4X4 구조의 육각형 배열을 갖는 구조를 이용하였다. SOFM에 의해 분류된 3개의 패턴 중 패턴-1은 유량자료의 크기에 의해 분류되었고, 패턴-2와 패턴-3은 BOD 농도의 크기에 따라 분류된 것으로 파악되었다. 따라서 SOFM의 적용에 의한 자료의 분류를 수행하고, 그 분류기준을 파악할 경우 SOFM의 자료분석 도구로서의 활용성이 더욱 높아질 것으로 판단된다.

핵심용어 : 수질자료, 유량자료, 패턴분류, 자기조직화 특성지도(Self-Organizing Feature Map)

1. 서론

환경부에서는 수질오염총량관리제도와 같은 환경관리 제도의 원활한 시행을 위해 수질 및 유량측정망을 통하여 관련 자료를 지속적으로 측정하고 있으며, 이를 국립환경과학원의 DB 및 웹 시스템(http://water.nier.go.kr)에 공개하고 있다. 이러한 양질의 자료의 축적에 따른 활용성 제고를 위한 자료의 분석기법 개발이 필요하며, 그 결과물들은 환경관리 제도의 원활한 시행을 위한 기초자료로 활용될 수 있을 것으로 판단된다.

현재 측정되고 있는 다양한 항목의 수질 및 유량 자료들에는 자연현상에 일반적으로 포함되어 있는 강한 비선형성으로 인해 쉽게 파악되기 어려운 관계들이 존재하고 있다. 이와 같이 파악되기 어려운 자료들 사이의 관계를 이해하기 위해서는 기존의 자료 분석 방법과는 다른 각도의 연구가 필요하다. 따라서 본 연구에서는 측정 자료의 전 범위에 걸친 분석보다는 각 자료에 내재된 특성을 반영할 수 있는 구간별 또는 패턴별 분석을 수행하여 그 관계를 파악하고자 하였다.

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 활발하게 제시되고 있다(Hsu et al., 2002; 김용구 등, 2006; 박성천 등, 2006; 진영훈 등, 2009).

* 정회원 · 동신대학교 토목공학과 교수 · E-mail : psc@dsu.ac.kr
** 정회원 · 동신대학교 공업기술연구소 연구교수 · E-mail : nmdrjin@gmail.com
*** 정회원 · 전남발전연구원 환경·생태연구팀 연구원 · E-mail : kbyj3711@jeri.re.kr
**** 정회원 · 동신대학교 토목공학과 연구원 · E-mail : kyg8987@paran.com
***** 정회원 · 동신대학교 토목공학과 박사과정 · E-mail : md49@nate.com

대상자료의 패턴을 구분하고 각 패턴별 특성을 파악하기 위해 최근 활발히 적용되고 있는 패턴분류 기법인 자기조직화 지도(Self-Organizing Feature Map: SOFM)의 연구결과는 유량자료에 대한 적용(Hsu et al., 2002; 김용구 등, 2006; 박성천 등, 2006; Jain et al., 2006; Srinivasulu et al., 2006; 김용구 등, 2008a)과 수질 및 수처리 분야(López et al., 2004; 김용구 등, 2008b), 기상분야(Nishiyama et al., 2007)를 포함하여 다양한 분야에서 그 우수성을 나타내고 있다.

따라서 본 연구에서는 전체 자료를 사용할 경우 파악되기 어려운 자료들 사이의 관계를 이해하기 위하여 SOFM 기법을 적용하여 분류된 패턴별 특성을 파악하고자 하였다. 앞서 언급한 수질오염총량관리제의 단위유역들 중 하나인 섬본_D 지점(SB_D)에서 측정된 수질 자료인 BOD 농도와 유량자료를 대상으로 하였으며, 각 패턴의 분류기준을 파악하기 위한 분석을 수행하였다.

2. 대상지점 및 자료

본 연구에서 제안한 연구방법의 적용타당성을 검증하기 위하여 섬진강 수계의 수질오염총량관리제를 위한 단위 유역들 중 섬본_D 지점(SB_D)을 대상으로 선정하여 해당 자료를 수집하였다. 대상지점은 전라남도 구례군 구례읍 유곡나루터에 위치하고 있다. 대상지점에서 측정된 수질 항목들 중 BOD 농도와 유량자료를 환경부 국립환경과학원의 웹 시스템으로부터 수집하여 사용하였다. 자료기간은 2004년 9월 17일부터 2009년 3월 24일까지이며, 수집된 자료의 수는 179개로 8일 간격으로 측정되었다. BOD 농도 자료에 대한 최소값은 0.40 mg/L로 나타났으며, 최대값은 4.00 mg/L, 평균값은 1.38 mg/L로 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.

3. 자기조직화 특성지도

인공신경망 이론의 한 종류인 SOFM은 교사학습 방법인 오차 역전파 학습 알고리즘(Error Back-Propagation Algorithm: EBPA)과는 달리 훈련과정에 목표값을 갖지 않는 비교사 학습방법(unsupervised learning algorithm)이다. SOFM은 다차원의 입력 자료들에 근거하여 분류한 후 2차원으로 사상시킬 수 있는 특징을 가지고 있으며, 자료의 가시화가 용이하므로 자료의 특성 파악을 위한 자료 분석 도구로 활용되고 있다.

SOFM을 이루는 구조를 살펴보면, 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.

$$X = [\ x_1,\ x_2,\ \ldots,\ x_m\ ]^T \tag{1}$$

$$W_j = [\ w_{j1},\ w_{j2},\ \ldots,\ w_{jm}\ ]^T, \quad j = 1, 2, \ldots, l \tag{2}$$

$$i(X) = \arg\min_{j} \| X - W_j \| \tag{3}$$

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. 이러한 과정을 근접반경 조정과정이라 하며, 최종적으로 승자노드만이 연결강도를 조정하게 된다.

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

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

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

4. 결과

본 연구에서는 시행착오법에 의해 SOFM의 구조를 결정하였으며, 그 결과 4×4의 육각형배열을 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 López and Machón(2004)에 의해 제안된 DBI(Davies-Bouldin Index)를 각 클러스터의 수에 따라 산정하였으며, 그 결과 3개의 패턴으로 분류하는 것이 최적인 것으로 나타났다.

따라서 4×4의 육각형배열을 갖는 SOFM 구조에 의해 총 3개의 패턴으로 구분된 결과를 그림 1에 나타내었다. 이는 두 개의 변량을 갖는 179개의 입력 자료를 이용한 SOFM의 패턴분류 결과를 보여주고 있다. 각 패턴별로 구분된 자료의 수를 보면, 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.

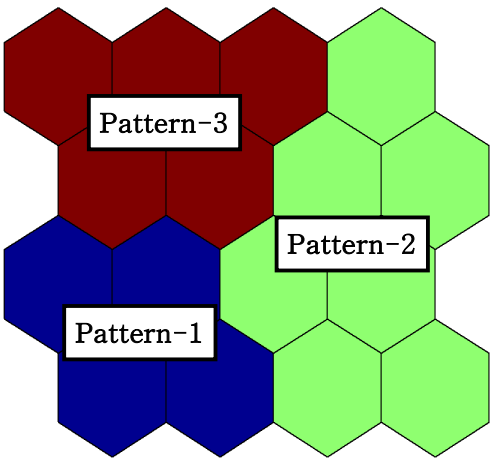

그림 1. SOFM에 의한 패턴분류 결과

분류된 자료의 분포를 살펴보기 위하여 표 1에 각 패턴별 BOD 농도 및 유량자료에 대한 최소값, 평균값, 중앙값 및 최대값을 나타내었다. 패턴-1의 경우 49.88 m^3/s 이상의 유량에 해당하는 자료가 분류된 것으로 나타났으며, 패턴-2와 패턴-3에 비하여 상대적으로 큰 값의 유량자료가 분류되었다. 패턴-1에 비하여 상대적으로 저유량에 해당하는 패턴-2와 패턴-3은 BOD 농도 1.40 mg/L의 값을 기준으로 분류되었다. 상대적으로 낮은 값의 BOD 농도자료가 패턴-2로 분류되었으며, 고농도 자료는 패

턴-3으로 분류되었다.

표 1. 각 패턴별 BOD 농도 및 유량 분류자료에 대한 최소, 평균, 중앙, 최대값

	Pattern-1 (Data No. = 20)		Pattern-2 (Data No. = 90)		Pattern-3 (Data No. = 69)	
	Runoff (m^3/s)	BOD (mg/L)	Runoff (m^3/s)	BOD (mg/L)	Runoff (m^3/s)	BOD (mg/L)
Minimum	49.8800	0.5000	7.5500	0.4000	5.0300	1.4000
Mean	139.6015	1.2000	19.3439	0.9444	17.5684	1.9884
Median	106.9800	1.2000	16.3550	0.9000	13.2800	1.8000
Maximum	410.5800	2.5000	55.0000	1.4000	60.5900	4.0000

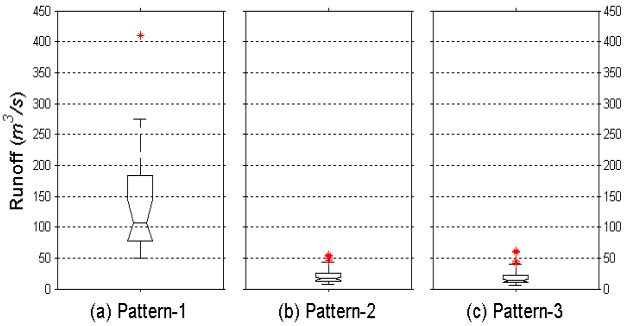

그림 2. 각 패턴별 유량자료에 대한 박스플롯

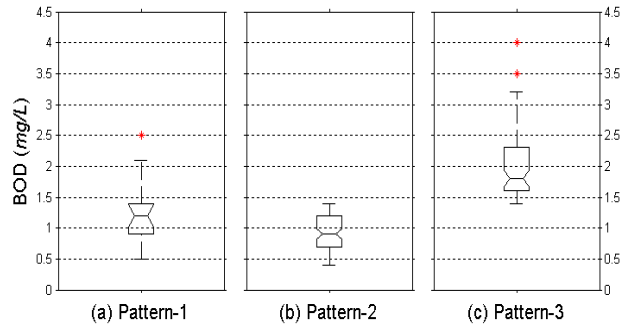

그림 3. 각 패턴별 BOD 농도자료에 대한 박스플롯

또한 SOFM을 적용하여 분류된 패턴의 분류기준을 보다 가시적으로 나타내기 위해 유량자료 및 BOD 농도에 대한 최소, 1분위값, 중앙값, 3분위값, 최대값 및 이상치를 박스플롯을 이용하여 그림 2와 그림3에 도시하였다. 그림 2에서 보는 바와 같이 패턴-1의 경우 패턴-2 및 패턴-3에 비해 상대적으로 고유량 자료를 포함하고 있으며, 이에 의해 독립적으로 분류된 것으로 파악된다. 특히 표 1에 나타낸 바와 같이 패턴-1의 최소값과 패턴-2 및 패턴-3의 최대값의 범위가 다소 겹친 것으로 나타났으나, 그림 2에 도시한 바와 같이 각 패턴별 자료의 이상치를 배제할 경우 유량자료에 대한 경계값이 49.88 m^3/s로 판단된다.

그러나 패턴-2와 패턴-3은 유량자료의 분포만을 고려할 경우 서로 유사한 범위를 나타내고 있으므로 유량자료에 의해 분류된 것으로 판단하기 어렵다. BOD 농도자료에 대한 각 패턴별 박스플롯(그림 3)을 통하여 볼 수 있듯이 패턴-2와 패턴-3은 BOD 농도에 의해 분류된 것으로 판단되며, 앞서 언급한 바와 같이 그 경계값이 BOD 농도 1.4 mg/L을 나타내고 있다.

5. 결론

본 연구에서는 현재 환경부 국립환경과학원에서 8일 간격으로 측정하고 있는 수질 및 유량자료에 대한 활용성 제고를 위하여 자료 분석 기법을 제안하고 그 적용성을 검토하였다. 측정되고 있는 다양한 수질 항목들 중 BOD 농도자료와 함께 유량자료를 이용하여 패턴분류 분석을 수행하기 위하여 수질오염총량관리제의 단위유역들 중 섬본_D 지점을 대상지점으로 선정하였다. 연구를 위한 적용 방법으로는 패턴분류를 위해 최근 널리 사용되고 있는 방법인 SOFM 기법을 사용하였

으며, 시행착오법에 의해 SOFM의 구조를 4×4의 육각형 배열로 결정하였다.

SOFM의 적용에 따른 결과는 BOD 농도와 유량자료를 동시에 이용할 경우 총 3개의 패턴으로 분류하는 것이 최적의 분류 결과인 것으로 나타났다. 패턴분류 결과에 따른 각 패턴별 분류기준을 파악하기 위한 분석을 수행하였다. 각 자료에 대한 패턴별 최소값, 평균값, 중앙값 및 최대값을 산정하였으며, 이와 더불어 박스플롯을 도시하여 그 분류 양상을 파악하였다. 그 결과 패턴-1이 49.88 m^3/s 이상의 고유량 자료를 포함하고 있음을 파악할 수 있었으며, 패턴-2와 패턴-3의 분류기준은 BOD 농도 1.4 mg/L를 기준으로 하여 분류된 것으로 파악되었다. 이에 따라 저농도의 자료가 패턴-2로 분류되었으며, 고농도의 자료가 패턴-3로 분류됨을 파악할 수 있었다.

본 연구의 결과로부터 다양한 자료의 특성 파악을 위하여 SOFM의 적용성이 높은 것으로 판단되며, 적용 자료에 근거한 패턴분류 기준을 파악할 수 있다는 점에서 인위적인 개입이 배제될 수 있는 것으로 판단된다.

감 사 의 글

본 연구과제는 환경부지정 전남지역환경기술개발센터의 연구비지원에 의해 수행한 연구과제입니다.

참 고 문 헌

1. 김용구, 진영훈, 박성천, “강우-유출특성 분석을 위한 자기조직화방법의 적용”, 대한토목학회논문집, 26(1B), 61-67 (2006).
2. 김용구, 진영훈, 박성천, 정천리, “나주지점의 강우-유출 해석을 위한 최적의 SOM 구조 결정”, 한국수자원학회논문집, 41(10), 995-1007 (2008a).
3. 김용구, 진영훈, 정우철, 박성천, “호소수의 강우, 저류량 및 TOC변동 특성분석을 위한 자기조직화 방법의 적용”, 한국물환경학회논문집, 24(5), 611-617 (2008b).
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. 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).
7. 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).
8. López, H. and Machón I., “Self-organizing map and clustering for wastewater treatment monitoring”, Engineering Applications of Artificial Intelligence, 17, 215-225 (2004).
9. Nishiyama, K., Endo, S., Jinno, K., Uvo, C.B., Olsson, J. and Berndtsson, R., “Identification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a Self-Organizing Map”, Atmospheric Research, 83, 185-200 (2007).
10. 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).