환경영향평가 제20권 제5호(2011) pp.705~716

연구논문

데이터마이닝 기법을 적용한 취수원 수질예측모형 평가

김주환* · 채수권 · 김병식*****

한국수자원공사 K-water연구원*, 을지대학교 보건환경안전과 교수**, 강원대학교 도시환경방재전공 조교수***

(2011년 8월 20일 접수, 2011년 10월 2일 승인)

Evaluation of Water Quality Prediction Models at Intake Station by Data Mining Techniques

Ju-Hwan Kim* · Soo-Kwon Chae · Byung-Sik Kim*****

K-water Research Institute, Korea Water Resources Corp., Daejeon, Korea*

Department of Environmental Health and Safety, Eulji University, Seongnam, Gyeonggi, Korea**

Disaster Prevention in Urban Environments, Kangwon National University, Samcheok, Korea***

(Manuscript received 20 August 2011; accepted 2 October 2011)

Abstract

For the efficient discovery of knowledge and information from the observed systems, data mining techniques can be an useful tool for the prediction of water quality at intake station in rivers. Deterioration of water quality can be caused at intake station in dry season due to insufficient flow. This demands additional outflow from dam since some extent of deterioration can be attenuated by dam reservoir operation to control outflow considering predicted water quality. A seasonal occurrence of high ammonia nitrogen (NH_3-N) concentrations has hampered chemical treatment processes of a water plant in Geum river. Monthly flow allocation from upstream dam is important for downstream NH_3-N control.

In this study, prediction models of water quality based on multiple regression (MR), artificial neural network and data mining methods were developed to understand water quality variation and to support dam operations through providing predicted NH_3-N concentrations at intake station. The models were calibrated with eight years of monthly data and verified with another two years of independent data. In those models, the NH_3-N concentration for next time step is dependent on dam outflow, river water quality such as alkalinity, temperature, and NH_3-N of previous time step. The model performances are compared and evaluated by error analysis and statistical characteristics like correlation and determination coefficients between the observed and the predicted water quality. It is expected that these data mining techniques

Corresponding Author: Chae, Soo-Kwon, Department of Environmental Health and Safety, Eulji University, Seongnam, Gyeonggi, Korea Tel: +82-31-740-7146 E-mail: cskwen@eulji.ac.kr

can present more efficient data-driven tools in modelling stage and it is found that those models can be applied well to predict water quality in stream river systems.

Keywords : Water Quality Model, Data Mining, Neural Network, Model Tree, Ammonia Nitrogen

I. 서 론

대부분 상 · 하류의 정수장에 원수를 공급하고 있는 댐 저수지는 유역-저수지-하천-취 · 정수장을 연계한 통합운영을 위해서는 수량과 수질을 종합적으로 관리할 필요가 있으며 이를 위해서는 각 수계별 특성에 적합한 분석 모형이 필요하다. 댐 저수지 운영은 홍수조절, 용수공급, 수력발전 그리고 하류 하천수질의 보전을 종합적으로 고려하여야 하며 특히, 갈수기 기간 중에 대부분의 댐저수지 하류 하천의 유량은 지류로부터의 유입량이 고갈되어 상류 댐 방류량에 의존하는 비율이 매우 높다할 수 있다. 최근에는 낙동강과 금강에서 갈수기 동안 하류하천의 유량 부족과 수질악화로 인해 댐으로부터의 방류량을 증가해 줄 것을 요청하는 횟수가 잦아지고 있는 실정이다. 갈수기 동안 우리나라 하천의 수질은 상류댐 방류량에 상당한 영향을 받는 것으로 보고되고 있으나, 댐방류량과 하천수질의 정량적인 상관성 분석과 이를 분석할 수 있는 모형의 구축은 매우 미흡한 실정이어서 효율적인 저수지운영의 제약조건이 되고 있다.

갈수기 동안 대청댐 하류 부여지점에서의 암모니아성 질소(NH_3-N)농도는 음용수 수질기준인 0.5 mg/L 보다 훨씬 높게 검출됨에 따라 금강수계 하류부에 위치한 S-정수장의 정수처리 공정에 많은 어려움을 주고 있다. 특히, 지난 '94~'95년 동안의 2년에 걸친 가뭄기간 중에는 암모니아성 질소농도가 음용수 수질기준 보다 약 5~10배 높게 검출되어 질소처리를 위한 염소투입량의 증가, 염소 부산물 발생 가능성 증가에 따른 정수 수질관리의 어려움, 시설물 부식, pH 및 알칼리도 저하에 따른 보조약품(소석회) 추가 투입, 잉여 슬러지 발생 등 수처리 공정의 애로뿐만 아니라 경제적으로도 많은 손실이 발생된 사례가 있다(한국수자원공사, 1993). 이와 같이 하천에서 자연유량의 감소는 하천자정능력을 저하시킴으로써 하천수질의 악화를 가속시켜 취수 이후에 정수장으로 유입되는 원수의 정수처리과정에서도 많은 지장을 초래하기도 한다. 하천 상류측에 댐이 설치되어 있는 경우 하류부에서는 댐 방류를 통한 오염물질 누출사고에 대응하기 위한 방안으로서, 일시에 많은 물을 내보냄으로써 하천 수질환경과 생태계 서식환경을 개선을 위한 플러싱(flushing) 효과를 기대하기 위한 저수지 운영방법이 플러싱 방류이다.(Chung, Kim, *et al.*, 2002) 프랑스 세느강에서 저수지 플러싱 방류가 하류수질에 미치는 영향은 Barillier(1993)에 의해 조사된 바 있으며, 여기에서는 방류초기 홍수파의 전단부가 지나가는 시기에는 하천바닥에 퇴적되어 있는 저니층이 재부상되어 영양염류와 용존 및 고형물질의 농도가 증가하여 산소가 많이 소모되는 경향을 보였으나 용존물질의 농도는 급격히 저감되었다(정 등, 2005).

국내에서도 금강을 대상으로 대청댐 하류 취수원의 수질에 미치는 영향을 예측하기 위하여 중회귀모형과 같은 통계적 모형과 신경망모형이 적용된 사례가 있다(정, 김 등, 2002).

하천 수질변동에 관한 연구는 인과적 분석과 추계적 분석으로 분류할 수 있으며 인과적 분석은 수질에 영향을 미치는 변수와 수질간의 인과관계를 중심으로 수질을 예측하는 것으로 하천에 유출입되는 오염물의 경계값으로 수질을 예측하는 QUAL2E, WASP5모형과 회귀분석에 의한 방법을 들 수 있으나, 이들 모형을 운영하기 위해서는 많은 양의 실측자료 및 조사연구가 필요하고 입력자료가 불충분할 경우에는 오차의 발생확률이 높아져 그 정확도를 기대하기 곤란하다. 따라서 본 연구의 목적은 금강

수계 내에 위치한 대청댐의 방류량이 하류하천의 암모니아성 질소농도에 미치는 영향을 분석하고 이를 예측할 수 있는 하천수질모형을 구축하기 위한 것으로 예측모형 개발을 위해 최근 많은 분야에서 적용되고 있는 데이터마이닝(Data mining)기법을 도입하여 그 적용성을 평가하고 예측된 하천수질을 근거로 갈수기 동안 댐하류의 수질보전을 고려한 적정 댐 방류량을 산정함으로써 보다 효율적인 댐 저수지 운영 시 도움이 되고자 하였다.

II. 데이터마이닝 기법

데이터마이닝(Data mining)은 많은 자료들 가운데 숨겨져 있는 유용한 상관관계를 발견하여 과거에는 알지 못했지만 자료들로부터 도출된 새로운 모델을 발견하여 미래에 실행 가능한 정보로부터 의사 결정에 활용하는 과정을 말한다. 즉 데이터에 숨겨진 패턴과 관계를 찾아내어 광맥을 찾아내듯이 정보를 발견해 내는 것이다. 여기에서 정보 발견이란 데이터에 고급 통계 분석과 모델링 기법을 적용하여 유용한 패턴과 관계를 찾아내는 과정이다. 데이터베이스 마케팅의 핵심 기술이라고 할 수 있다. 따라서 데이터마이닝의 필수 요소는 신뢰도가 높은 충분한 자료이다. 이것은 신뢰도 높은 충분한 자료가 정확한 예견을 가능하게 하기 때문이다. 그러나 너무 많은 자료는 오히려 데이터마이닝의 예견 능력을 떨어뜨릴 수 있으므로 최적의 결과산출이 가능한 의미 있는 자료의 확보가 필요하다.

1. 모델트리

모델트리(Model tree)는 분석용 자료를 이용하여 의사결정규칙을 나무 구조로 만들고, 관심대상을 몇 개의 하위집단으로 분류하거나 예측을 하는 데이터마이닝 기법이다. 모델트리의 분석과정은 트리구조에 의해서 표현되기 때문에, 분류 또는 예측을 목적으로 하는 다른 데이터마이닝 기법들에 비해, 분석과정의 이해와 결과의 해석이 쉽다는 장점을 가지고 있다. 특히, 데이터마이닝에서 모델트리의 활용이 많은데, 이는 의사결정트리 자체가 분류 또는 예측 모형으로 활용되어 데이터마이닝 기법으로 사용되기도 하고, 다른 데이터마이닝 기법을 적용하기 전에 자료를 처리하는 작업에도 사용할 수 있기 때문이다. 의사결정트리기법은 연속형이나 범주형 등의 예측변수를 그대로 이용하므로 자료의 변형에 필요한 시간을 줄일 수 있고, 모형을 구축하는 시간이 짧다. 이러한 특성 때문에 의사결정트리기법은 다른 예측기법을 수행하기 전에 많은 예측변수 중에서 유용한 것들만을 고르는 과정에 사용될 수 있다. 회귀모형과 모델트리는 의사결정나무의 결과로서 수치형 결과를 제공하는 면에서는 유사하나, 회귀모형은 출력결과에 대한 평균값을 제시하고 모델트리는 선형회귀모형식을 트리구조별로 제공한다는 면에서 차이가 있다. 모델트리에서 통계학적 분리기준은 자식마디의 엔트로피(entropy, 불순도(不純度)) 감소에 근거하고 있다. 즉, 유사한 성질의 자료들을 가능한 하나의 자식마디로 분류하는 방식으로 분리를 수행한다.

표 1. 의사결정트리 알고리즘 비교

구 분	CHAID	CART	QUEST	C4.5
목표변수	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형	명목형, 순서형, 연속형
예측변수	명목형, 순서형, 연속형(사전그룹화)	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형, 순서형, 연속형
분리기준	카이제곱-검정 F-검정	지니 지수 분산의 감소	카이제곱-검정 F-검정	지니 지수 분산의 감소
분리개수	다지분리(multiway)	이지분리(binary)	이지분리(binary)	다지분지(multiway)
가지치기	○	○	○	○
결손값 대체	×	○	○	○
비용함수	×	○	○	○

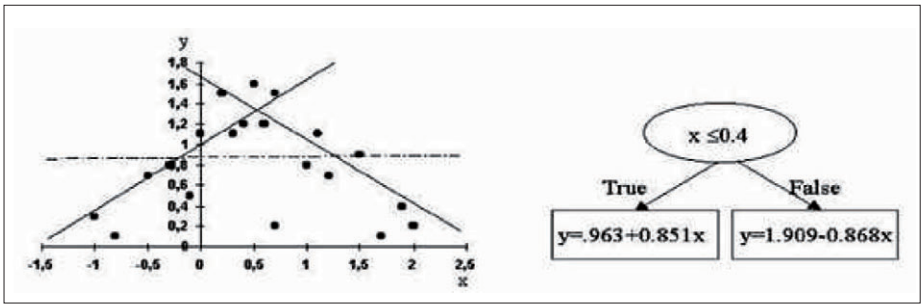


그림 1. 모델트리에 의한 모형구축

의사결정트리는 뿌리마디에서 시작해서 각 가지가 끝마디가 될 때까지 자식마디를 만들면서 형성된다. 이러한 의사결정트리를 완성하기 위해서는 마디의 분리기준(splitting rule)의 선택, 분리를 멈추기 위한 정지기준(stopping rule)의 선택, 가지치기(pruning)방법의 선택, 입력변수의 값에 결측치가 있는 경우에 결측치 대치(imputation)방법의 선택 등 여러 단계를 수행해야 한다. 이러한 과정을 수행하여 의사결정트리를 형성하는 주요알고리즘으로는 CHAID, CART, C4.5, QUEST등이 있으며 각 단계에서 서로 다른 기준을 가지고 있어 다른 의사결정트리가 만들어진다(최종후 등, 2003). Table 1은 몇가지 관점에서 이들 알고리즘을 비교한 것이다.

모델트리는 자식마디에 할당된 샘플의 표준편차가 줄어드는 방향으로 분리를 수행한다. 각 마디에 할당된 자료의 표준편차는 다음 식(1)과 같은 형태의 예측오차로써 평가되고 표준편차의 감소량을 최대화 시키는 변수가 분리기준으로 선정된다(이대종 등, 2006).

$$SDR = sd(T) - \sum_{i} \frac{|T_i|}{|T|} \times sd(T_i) \qquad (1)$$

여기서, SDR(Standard Deviation Reduction)은 표준편차의 감소량, T_i는 선정된 변수에 의해 생성된 자식마디에 적용된 자료집합이다. 나무구조의 분리는 표준편차(sd) 변화가 미미하거나(약 5% 미만) 자식마디에 할당된 자료 수가 거의 없을 때 중지된다. 최종적으로 각각의 자식마디 샘플에 대하여 선형회귀모형을 구축한다(Fig. 1 참조).

2. 중회귀모형

중회귀모형은 2개이상의 독립변수를 사용하여 종속변수 y를 산정하는 방법으로 선형다중회귀모형과 지수형 다중회귀모형 등 여러 가지 형태의 관계가 가능하다. 본 연구에서는 선형중회귀모형을 적용하였다.

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n + \varepsilon \qquad (2)$$

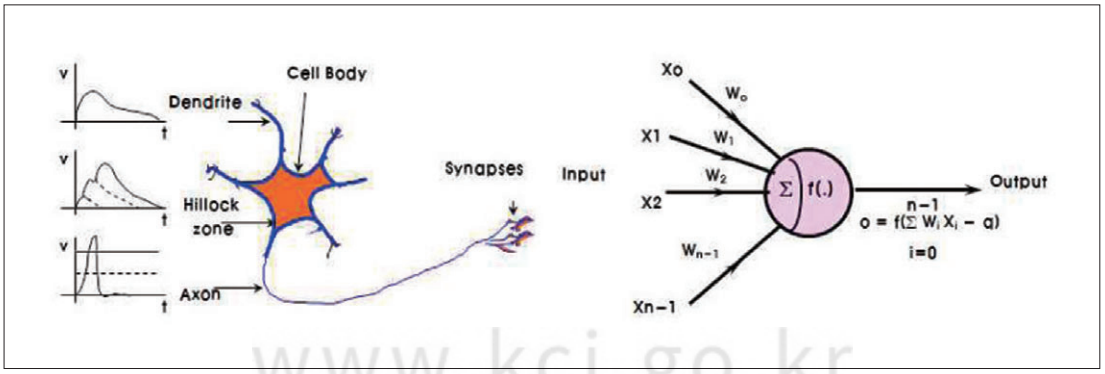


그림 2. 생물학적 뉴런과 인공 신경망 개념도

여기서, β_0, β_1,…, β_n 는 회귀분석을 통해 산정되는 매개변수이며, x_1, x_2,…, x_n는 독립변수, ε은 잔차, 그리고 y는 종속변수를 나타낸다.

3. 신경망모형

1) MLP

신경망은 인간의 두뇌를 수학적으로 모사한 모형이며, 가장 광범위하게 사용되어온 신경망 모형은 Rumelhart 등(1986)에 의해 소개된 다층퍼셉트론(Multi-Layer Perceptron; MLP)이다. 신경망의 기본소자인 뉴런은 입력된 외부자극이 일정 값 이상이면 반응하여 출력 신호를 내보내는데 이 과정을 수학적으로 모형화한 것이 인공신경망이다. 인공신경망의 뉴런은 그림 2에서는 생물학적 뉴런과 인공신경망의 개념을 나타낸 것으로 인공신경망에서는 입력된 외부자극 X_0, X_1, …, X_{n-1}과 연결강도 W_1, W_2, …, W_{n-1}를 곱한 합이 특정한 값 이상이면 반응하여 출력값을 내보낸다. 여기서, f(.)는 뉴런의 반응여부를 결정하는 활성화 함수이다.

실제 적용에 있어서는 식(3)과 같이 표시되는 시그모이드(Sigmoid) 함수가 활성화 함수로 많이 사용된다. 여기서, λ는 함수의 경사도이다.

$$f(NET) = \frac{1}{1+\exp(-\lambda NET)} \quad (3)$$

신경망이론을 적용하여 문제를 해결하려면, 먼저 설정된 신경망모형을 학습시켜야 한다. 신경망에서 학습이라 함은 특정한 응용 목적에 적합하도록 뉴런간의 연결강도를 변화시키는 과정이다. 신경망의 학습에는 적절한 입출력 학습패턴 쌍의 구성이 매우 중요하다. 신경망은 입력항목을 바탕으로 신경망모형에 의해 산출되는 결과를 목표치와 비교하여, 모형의 가중치를 오차가 작은 방향으로 변경함으로써 유의한 특성을 학습하게 된다. 실질적인 적용에 있어서 신경망의 학습에는 역전파(Backpropagation) 알고리즘이 가장 많이 사용된다.

BPA를 이용한 모형의 학습은 출력층 오차 신호를 이용하여 은닉층과 출력층간의 연결강도를 변경하고, 출력층 오차 신호를 은닉층에 역전파하여 입력층과 은닉층간의 연결강도를 변경하는 과정이다. 즉, BPA는 목표치 d_i와 모형에 의해 계산된 최종출력 y_i를 비교하여 식(4)와 같이 표시되는 학습오차 E를 줄이는 방향으로 연결강도를 변경함으로써 최적해를 찾는 기법이다.

$$E = \frac{1}{2}\sum_i (d_i - y_i)^2 \quad (4)$$

신경망 이론을 이용한 모형개발은 회귀모형과 같은 특별한 구조나 매개변수의 산정 및 자료의 변화 등이 필요치 않고 자료의 축적에 따라 모형의 능력을 향상시킬 수 있는 장점을 가지고 있다. 본 연구에서는 신경망모형의 학습을 위해 역전파알고리즘(BPA)의 학습률 및 학습속도를 개선한 모멘텀 및 적응학습률 방법을 사용하였으며 활성화함수로는 시그모이드 함수를 사용하였다.

2) RBFNN

RBFN(Radial Basis Function Neural Network)은 1998년 Broomhead & Lowe와 1989년 Moody & Darken등에 의해 방사함수(Radial function)에 기반을 둔 신경망을 제시함으로써 이에 대한 연구가 활발히 진행되어 왔다. MLP는 일반적으로 높은 분류능력을 가지고 있는 반면 학습시간 및 국부 최소값의 문제점을 가지고 있다. 이에 반해 RBFNN은 학습속도가 빠르고 구성이 간단하며 분류능력이 우수한 장점이 있다. RBF 신경망에서 활성화 함수로 사용되는 h는 중심 c_i에서 멀어질수록 단조증가나 또는 단조감소하는 특징을 갖는 함수를 사용한다. 은닉층에서 MLP가 시그모이드 함수인 것에 반해 RBF신경망에서는 다음 (5)식과 같은 가우시안(Gaussian) 함수가 주로 사용되며, 이외에도 Multi-Quadratic 함수나 코우시(Cauchy) 함수 등이 사용된다.

$$f(\vec{x_i}) = \sum_{j=1}^{m} w_j h(\|\vec{x_i} - \vec{c_i}\|) \quad (5)$$

여기서, h(은닉층 뉴런의 활성화 함수)는 RBF함수를, $\vec{c_i}$는 입력값들의 중심을, w_j는 연결강도를 의미한다. RBF는 $\vec{c_i}$와 입력값 $\vec{x_i}$사이의 유클리드 기

하학에 근거를 두고 있다. 따라서, 일반적으로 RBF h는 거리가 0일 경우에 최대값을 갖게 된다. 가우시안 함수의 경우 중심값 $\vec{c_i}$와 $\vec{x_i}$값이 같을 때 이 함수는 1.0의 값을 산출하고 반면에, $\vec{t_i}$와 $\vec{x_i}$의 값의 차이가 한계범위에 가까워질수록 0에 가까운 값을 산출하게 된다. RBFNN의 학습 알고리즘은 여러 가지가 있는데 대부분이 학습 알고리즘을 두 단계로 나누어 학습을 한다. 즉, 은닉층에서의 학습과 출력층에서의 학습 알고리즘으로 나누어진다. 입력층은 단지 은닉층으로 입력값을 전달하는 역할만을 하기 때문에 모든 연결강도 값은 1로 고정되어 있다. 은닉층의 활성화함수는 방사함수를 이용하고 은닉층 뉴런의 활성화 값과 연결강도 w_j를 곱하여 선형적으로 합한 값을 출력층 뉴런이 출력한다. 은닉층에서의 학습은 클러스터링 알고리즘을 이용한 자율학습(unsupervised learning)을 한다. 여러 가지 클러스터링 알고리즘이 있지만 그 중 구현이 쉽고 간단하며 여러 응용 분야에서 많이 사용되는 알고리즘으로 k-means 클러스터링 알고리즘이 있다. k-means클러스터링 알고리즘은 데이터의 집단화와 분석의 연구에 많이 이용되는 방법으로 각 집단의 중심으로부터 그 집단 내에 포함된 데이터 사이의 오차를 최소화 하도록 집단의 중심점을 결정하게 된다.

RBFNN 신경망은 입력층과 은닉층이 선형으로 연결되어 은닉층과 출력층만으로 구성된 단층 신경망 형태를 갖는 단층으로 구성된 모형이기 때문에, 수학적 표현이 명료하고 기존 신경망의 학습알고리즘으로 사용되고 있는 수치해 탐색알고리즘으로 최급강하법(Gradient descent method)과 같은 반복적이고 시간이 걸리는 알고리즘 대신에 역행렬을 사용하여 연산속도가 매우 빠른 강점을 갖고 있다.

III. 적용 대상유역

금강유역의 수원은 전라북도 장수군 장수읍 수분리 신무산(EL.896.8m)에서 발원하여 북쪽으로 흐르면서 남대천, 봉황천, 초강천, 보청천, 등과 차례로 합류한 후 대청댐에 유입된다. 대청댐은 상류지역의 용담댐과 함께 금강수계의 다목적댐으로서 생공 및 농업용수를 공급하고 있으며 홍수기에 홍수를 댐에 저류 하였다가 갈수나 평수기에 발전용수로 이용하고 대청댐 하류에 위치한 조정지댐에 의해 하류지역에 조절방류로 각종 용수를 공급하고 있다.

다목적댐의 주요한 기능은 하절기에 홍수를 조절 · 저장하고, 평갈수기 동안에는 확보된 저수량으로 용수공급, 수력발전, 그리고 하류수질보전을 위한 방류를 수행하는 것이다. 즉, 다목적댐은 홍수기 동안 하류 하천의 홍수의 크기와 빈도는 줄여주고 갈수기 하천유량을 증가시켜 주는 역할을 하는 것이다. 갈수기 동안 하천의 유하량 크기는 하천의 자

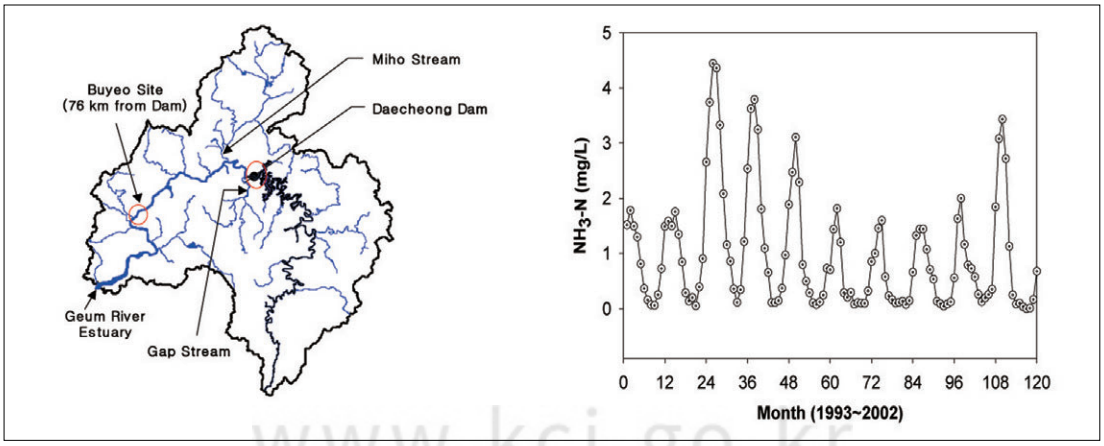


그림 3. 금강유역도 및 대상 취수지점의 NH_3-N 농도의 계절변동

정능력에 큰 영향을 미치므로 하천 수질에 직접적인 영향을 미치는 중요한 요소 중의 하나이다.

그림 3에서는 대상유역인 금강유역과 '93~'02년 동안의 부여지점에서 관측된 NH_3-N의 월단위농도를 도시한 것이다. 대청댐 하류 부여지점에서의 암모니아성 질소(NH_3-N)농도는 음용수 수질기준인 0.5 mg/L 보다 높게 나타나고 있으며 특히, 지난 '94~'95년 동안의 2년에 걸친 가뭄기간 중에는 암모니아성 질소농도가 음용수 수질기준 보다 약 5~10배 가량 높게 검출되어 질소처리를 위한 염소투입량의 증가, 염소 부산물 발생 가능성 증가에 따른 정수 수질관리의 어려움을 겪은 사례가 조시된 바 있다.

IV. 적용 및 결과

본 연구에서 제시한 모형의 개발을 위하여 사용된 변수는 표 2에 수록하였으며 자료는 1993년 1월부터 2000년 12월까지 월자료를 사용하였으며 개발된 모델의 검증을 위해서는 2001년 1월~2002년 12월 자료를 사용하였다. 여기서, 사용된 변수는 취수지점에서의 하천수질에 영향을 미치는 인자로서 각각 temp는 온도, Turb는 탁도, Alk는 알카리도, Q는 댐방류량을 그리고 NH_3-N는 암모니아성질소를 의미하며 t는 시간을 나타낸다.

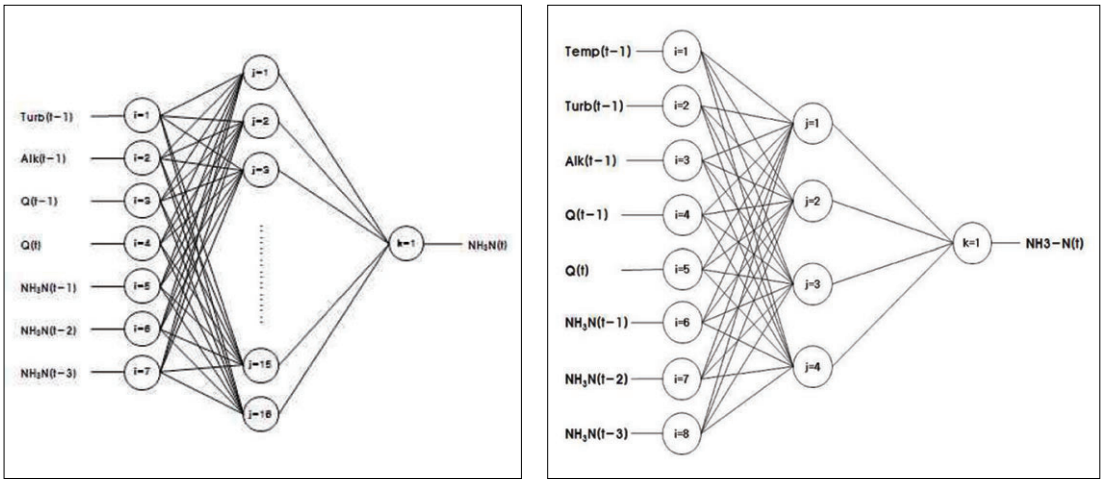
MLP 신경망모형의 경우, MLP-02와 MLP-03 모형의 구조를 그림 4에 도시하였다. MLP-02는 수질에 영향을 미치는 인자로서 7개의 입력변수에 의해 출력변수로서 예측하고자하는 암모니아성질소를 출력하는 구조를 가지며, 마찬가지로 MLP-03에서는 8개의 입력변수에 의해 출력변수로서 예측하고자하는 암모니아성질소를 출력하는 구조를 갖는다.

또한 그림 5에서는 모델트리를 적용한 모형 MT-01과 MT-03의 구조를 보여주고 있다. 여기에서 각각 뿌리 형태로 분기되는 선상에 구분되어 있는 숫자는 각 변수의 해당값에서 각기 다른 자료특성을 보여주는 것으로 예를 들어 그림 5에서, 댐 방류량을 나타내는 Q(t)와 Q(t-1)의 경우 Q(t)는 21.95cms를 기준으로 각각 다른 자료특성을 보이며 한달 전 방류량인 Q(t-1)은 55cms를 기준으로 자료상의 다른 특성을 보이고 있다는 것이다. 즉, 해당 값으로 자료를 분류하여 모형화 시킬 경우 모형의 표준편차가 가장 작게 나타남을 의미한다.

표 3 에서와 같이 3가지 형태의 다중회귀분석모형, 3가지 형태의 MLP 신경망, 3가지 형태의 모델트리, 3가지 형태의 RBF신경망 모형에 대하여 결정계수, 절대평균오차, 평균제곱근오차, 상대오차 및 제곱근상대제곱오차를 비교 · 검토하였다. 그 결과 중회귀모형에서는 8개의 입력변수로 구축된 MR-03가, MLP 신경망모형 중에서는 MLP-01, 모델트리에서는 MT-02 그리고, RBF신경망모형

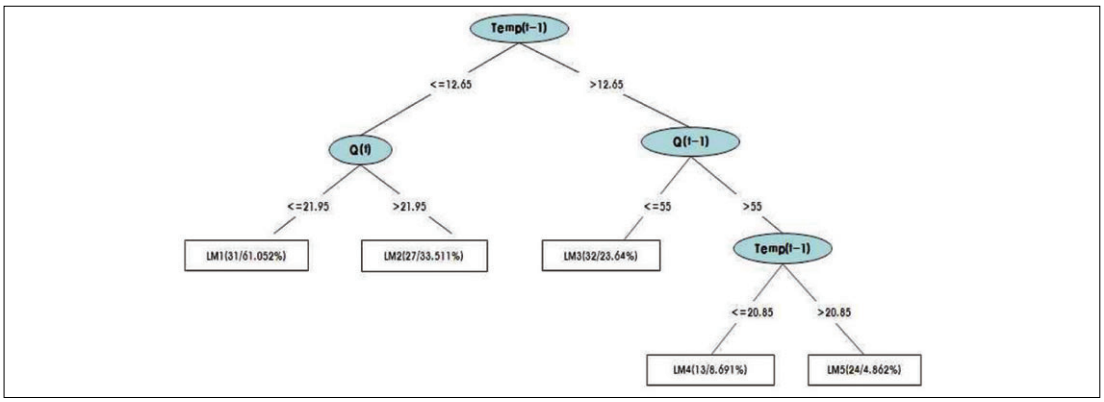
표 2. 적용모형 및 모형별 매개변수

구 분	모델명	매개변수
Multiple Linear Regression	MR-01	$NH_3N(t)$ = f[temp(t-1), Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$]
	MR-02	$NH_3N(t)$ = f[Turb(t-1), Alk(t-1) ,Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
	MR-03	$NH_3N(t)$ = f[Temp(t-1), Turb(t-1), Alk(t-1),Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
Multi Layer Perceptron Neural Network	MLP-01	$NH_3N(t)$ = f[temp(t-1), Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$]
	MLP-02	$NH_3N(t)$ = f[Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
	MLP-03	$NH_3N(t)$ = f[Temp(t-1), Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
Model Tree	MT-01	$NH_3N(t)$ = f[temp(t-1),Turb(t-1),Alk(t-1),Q(t-1),Q(t), $NH_3N(t-1)$]
	MT-02	$NH_3N(t)$ = f[Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
	MT-03	$NH_3N(t)$ = f[Temp(t-1), Turb(t-1), Alk(t-1), Q(t-1),Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
Radial Basis Function Neural Network	RBF_NN1	$NH_3N(t)$ = f[temp(t-1), Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$]
	RBF_NN2	$NH_3N(t)$ = f[Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]
	RBF_NN3	$NH_3N(t)$ = f[Temp(t-1), Turb(t-1), Alk(t-1), Q(t-1), Q(t), $NH_3N(t-1)$, $NH_3N(t-2)$, $NH_3N(t-3)$]

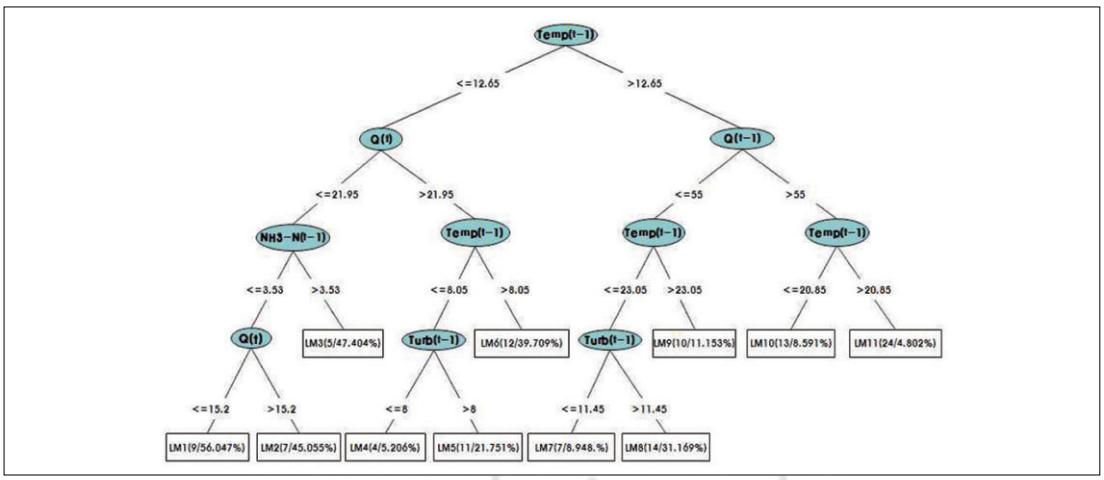


(a) (b)

그림 4. 다층퍼셉트론 신경망모형 MLP-02(a)와 MLP-03의 구조(b)



(a)



(b)

그림 5. 모델트리 MT-01(a)과 MT-03(b)의 구조

표 3. 모형별 예측성능 평가

Categories	Model	R^2	Mean absolute error	Root mean squared error	Relative absolute error(%)	Root relative squared error(%)
Multiple Linear Regression	MR-01	0.806	0.4650	0.6423	45.5548	47.6935
	MR-02	0.894	0.3696	0.4521	36.2090	33.5650
	MR-03	0.898	0.3599	0.4480	35.2580	33.2620
Multi Layer Perceptron Neural Network	MLP-01	0.935	0.3664	0.4915	35.8958	36.4931
	MLP-02	0.876	0.3655	0.5465	35.8043	40.5756
	MLP-03	0.928	0.3429	0.4414	33.5971	32.7774
Model Tree	MT-01	0.847	0.4185	0.6394	41.0004	47.4776
	MT-02	0.877	0.4171	0.4965	40.8654	36.8652
	MT-03	0.851	0.4154	0.6314	40.6998	46.8828
Radial Basis Function Neural Network	RBF_NN1	0.883	0.3639	0.5073	35.6549	37.6636
	RBF_NN2	0.845	0.4697	0.6384	46.0208	47.4020
	RBF_NN3	0.835	0.3986	0.6756	39.0553	50.1617

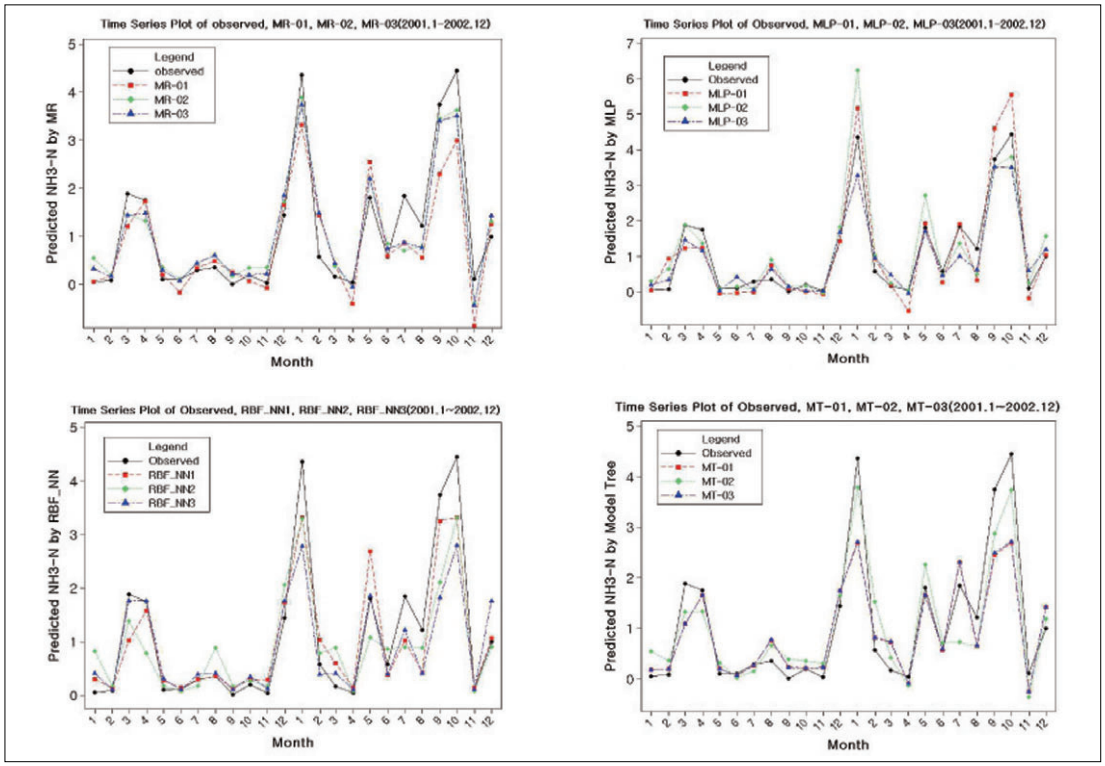


그림 6. 모형 검증결과에 비교를 위한 관측치와 예측치 시계열도

중에서는 RBF_NN1가 가장 양호한 결과를 보이고 있음을 알 수 있었다.

그림 6은 각 기법별로 개발된 모형의 검증결과를 시계열로 도시한 결과이며 그림 7에서는 각 모델에 의해 예측된 값과 관측값과의 선형관계를 보여주기 위한 것으로 예측구간과 신뢰구간을 함께 도시한 것이다. 그림 7에서는 예측치와 관측치간의 선형회귀선을 중심으로 도시된 것은 95% 예측구간(Prediction level)과 95% 신뢰구간(Confidence level)을 함께 도시한 것으로 길게 표현된 점선

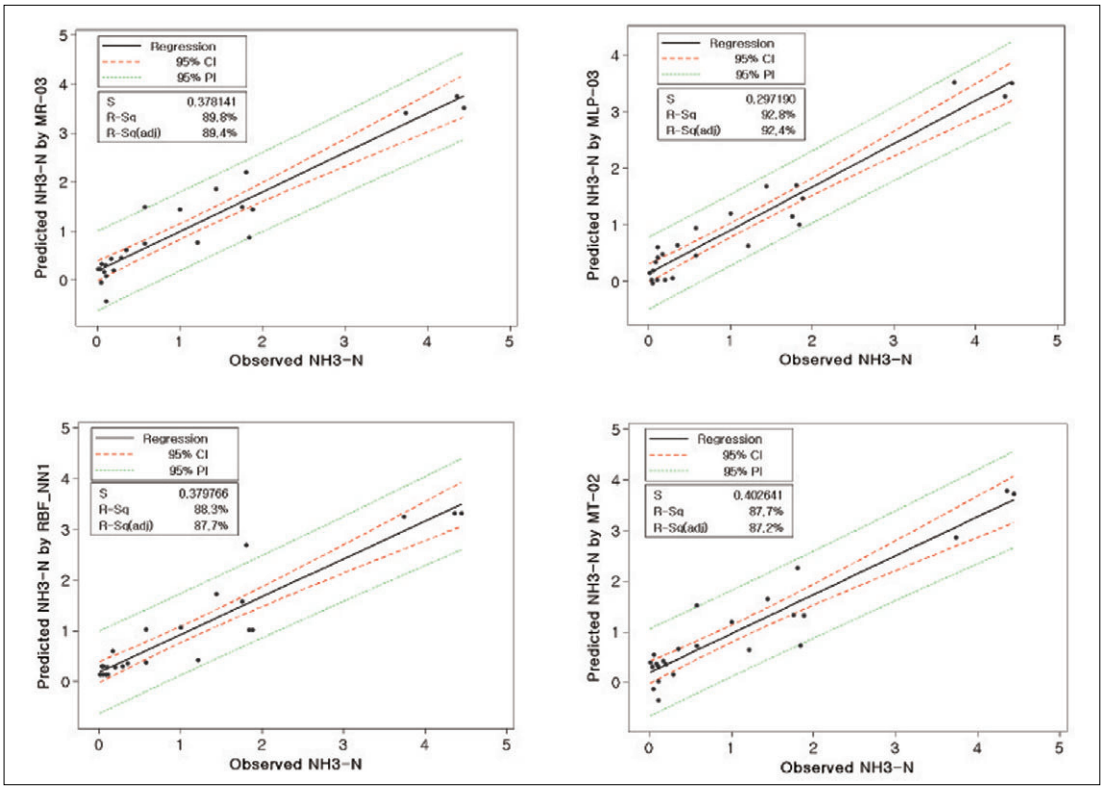


그림 7. 각 모형별 관측치와 예측치간 선형성 검토결과

(Long-dashed)구간은 예측값에 대한 95% 신뢰구간을 의미하며 짧게 표현된 점선(Short-dashed)구간은 예측치에 대한 95% 신뢰구간을 나타낸다. 여기에서는 중회귀, 모델트리, MLP 신경망 및 RBF 신경망 모형이 적용된 것 중 대표적으로 하나의 모형결과를 도시한 것으로 예측구간과 신뢰구간 범위가 좁게 나타나는 모형이 가장 예측력이 뛰어난 것으로 판단할 수 있으며 이를 근거로 판단해 볼 때 4가지 모형 중 MLP 신경망 모형의 예측성능 가장 우수하게 하천수질을 재현해 주는 것으로 나타났다.

표 3에서는 각 모형별로 검정 및 검증기간 동안의 월별 관측값과 예측값 사이의 결정계수와 오차를 나타낸 것이다. 여기서, 결정계수를 토대로 평가해 볼 때 중회귀모형은 MR-03, MLP 신경망모형은 MLP-01, 모델트리모형에서는 MT-02, RBF 신경망모형에서는 RBF_NN1모형이 가장 좋은 예측성능을 보여주었다, 이중에서 MLP 신경망모형이 다른 모형보다 다소 양호한 결과를 보여주었으며, 대체로 암모니아성 질소농도의 변화 추세를 잘 반영하는 것으로 나타났다. 그러나 오차분석 측면에서는 MLP-03모형이 MLP-01모형보다 작은 오차를 보였다.

본 연구에서 하천수질 예측에 적용된 중회귀모형, 모델트리 및 신경망 모형 중에서는 다층퍼셉트론 신경망 모형이 가장 예측력이 뛰어난 것으로 보인다.

V. 결 론

정수장의 취수지점에서의 수질은 처리시설에서 약품투입, 전력 등 운영에 영향을 미치며 특히 갈수기에는 수량부족에 의하여 수질이 악화됨으로써 이에 대한 대비를 위한 하류부 취수지점에서의 수질예측이 필요하다. 본 연구에서는 댐 하류 취수지점에서 수질예측을 위한 모형의 개발을 위해 최근 다양한 적용이 시도되고 있는 데이터마이닝 기법을 도입

함으로써 각 모형별 특성을 파악하고 예측모형의 수립에 적용하여 그들의 성능을 평가하고자 하였다.

취수지점 수질특성 분석결과, 갈수기 기간 중 대청댐 방류량은 하류 하천의 암모니아성 질소 농도에 영향을 미치는 것으로 나타고 있으며 댐저수지 운영계획 수립시 방류조건별로 하류지점의 암모니아성 질소농도의 파악과 예측이 필요한 것으로 나타났다.

데이터마이닝기법을 적용한 모형의 개발은 취수지점 수질자료로서 과거 10년간의 해당월의 암모니아성 질소농도를 댐방류량, 하천의 수온과 알칼리도, 그리고 질소농도의 자기상관성을 고려하여 모형 매개변수를 선정하였다. 모형개발을 위해 10년간의 기록자료 중 8년간 관측된 자료를 사용하였고 나머지 2년간의 자료를 사용하여 검증하였다.

모형의 예측성능평가 결과에서는 중회귀모형중 MR-01을 제외한 모델트리, 다층퍼셉트론 및 방사함수 신경망모형 등 모형의 예측성능을 평가하는 결정계수값이 0.87이상으로 나타나 하천수질의 예측에 양호한 성능을 보임으로써 실적용이 가능할 것으로 평가되었다. 이중 검증과정에서의 예측성능이 가장 뛰어나 모형을 다층퍼셉트론 신경망 모형으로 분석되었다. 따라서 갈수기 정수장 취수지점의 수질을 고려한 댐 운영을 위해서는 하류부의 수질을 예측하여 이에 대한 적정 방류량의 월별 배분이 검토되어야 할 것으로 판단되며, 저수지의 용수공급능력을 고려하여 이와 연계된 최적화모형 등의 적용을 통한 하천유역 통합관리 및 평가가 이루어져야 할 것이다.

참고문헌

김상단, 유철상, 시계열 모형의 적용을 통한 댐 방류의 수질개선 효과검토, 한국물환경학회, V20(6), (2004)

김주환, 신경회로망을 이용한 하천유출량의 수문학적 예측에 관한 연구, 박사학위논문, 인하대학교, (1993)

류병로, 한양수, ARIMA 모형에 의한 하천수질 예측, 한국환경과학회지, V7(4), (1998)

이대종, 박진일, 박상영, 정남정, 전명근, 클러스터 기반 퍼지 모델트리를 이용한 데이터 모델링, 퍼지 및 지능시스템학회 논문지 Vol. 16, No. 5, pp. 608-615, (2006)

정세웅, 김유경, 상류댐 플러싱 방류가 금강의 겨울철 암모니아성 질소농도 저감에 미치는 효과분석, 한국물환경학회, V21(6), (2005)

최종후, AnswerTree 3.0을 이용한 데이터마이닝 예측 및 활용, SPSS 아카데미, (2003)

한태환, (1998), 다변수 시스템 인공지능 모델링에 의한 정수장 약품 주입공정 자동화 시스템의 구현, 공학박사학위논문, 충북대학교.

한국수자원공사, 댐방류량이 하천 수질에 미치는 영향에 관한 연구, (1993)

Ambrose R.B., Wool T.A. Jr. and Martin J. L., The Water Quality Analysis Simulation Program, WASP5: Model Documentation. Environmental Research Laboratory, Athens, GA, (1993)

B. Bhattacharya, D.P. Solomatine, Application of artificial neural networks and M5 model trees to modelling stage-discharge relationship, in: B.S. Wu, Z.Y. Wang, G.Q. Wang, G.H. Huang, H.W.Fang, J.C. Huang (Eds.), Proceedings of the Second International Symposium on Flood Defence, Beijing, China, Science Press, New York Ltd., New York, pp. 1029-1036, (2002)

B. Bhattacharya, D.P. Solomatine, Neural networks and M5 model trees in modelling water leveld ischarge relationship for an Indian river, in: M. Verleysen (Ed.), Proceedings of the 11th European Symposium on Artificial Neural Network, Bruges, Belgium, d-side, Evere Belgium, pp. 407-412,

(2003)

Brown L. and Barnwell T., The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS: Documentation and User's Manual, EPA/600/3-87/007, USEPA, Georgia, USA, (1987)

Chung, F., Sandhu N., Wilson, D. and Finch, R., Modeling flow-salinity relationships in the Sacramento-San Joaquin Delta using artificial neural networks, Report OSP-99-1, Department of Water Resources, California, USA, (1999)

Chung S. W. and Kim J. H., Development of artificial neural network models supporting reservoir operation for the control of downstream water quality. Wat. Engr. Res., Korea Water Resources Association, 3(2), 143-153, (2002)

D.P. Solomatine, M.B. Siek, Flexible and optimal M5 model trees with applications to flow predictions, Proceedings of the Sixth International Conference on Hydroinformatics, World Scientific, Singapore, (2004)

Dimitri P. Solomatine and Yupeng Xue, "M5 Model Tree and Neural Network: Application to Flood Forecasting in the Upper Reach of the Huai River in China', Journal of Hydrologic Engineering, ASCE/November/December 2004, pp.491-501, (2004)

Fujita M. and Zhu M. L. Runoff prediction using Fuzzy reasoning method and Neural Network method. Proc., Conf. on International Conference on Environmentally Sound Water Resources Utilization, Bangkok, Thailand, 2, pp. 221-228, (1993)

L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees, Wadsworth, Belmont, CA, (1984)

Karunanithi N., Grenney W.J., Whitley D. and Bovee K., Neural networks for river flow prediction. J. Comp. in Civ. Engrg., ASCE, 8(2), 201-220, (1994)

Kim J.H., Kang K.W. and Park C. Y., Nonlinear forecasting of streamflows by pattern recognition method. Korean J. of Hydroscience, Korean Ed., 25(3), 105-113, (1992)

Lisboa P.G.J., Neural Networks. Chapman & hall, London, (1992)

Quinlan, J. R., "Learning with continuous classes." Proc., 5th Australian Joint Conf. on Artificial Intelligence, Adams & Sterling, eds., World Scientific, Singapore, pp.343-348, (1992)

Rodriguez M. J., Serodes J. B. and Cote P. A. Advanced chlorination control in drinking water systems using Artificial Neural Networks. Water Supply, 15(2), pp. 159-168, (1997)

S.E. Darby, C.R. Throne, Predicting stage-discharge curves in channels with bank vegetation, ASCE, J. Hydraulic Eng. 122 (10) pp.583-586, (1996)

Witten, I. H. and Eibe Frank, Data Mining, Morgan Kaufmann Publishers, (1999)

Zou R., Lung W. S., and Guo H., Neural network embedded Monte Carlo approach for water quality modeling under input information uncertainty. J. of Comput. in Civ. Eng., 16(2),135-142, (2002)