Journal of Korean Society on Water Quality, Vol. 22, No. 5, pp.952-957 (2006)

웨이블렛 변환과 인공신경망을 이용한 일 TOC 자료의 예측에 관한 연구

곽필정 · 오창열 · 진영훈† · 박성천

동신대학교 토목공학과

Study on the Prediction of Daily TOC Data by Using Wavelet Transform and Artificial Neural Networks

Pil Jeong Gwak · Chang Ryol Oh · Young Hoon Jin† · Sung Chun Park

Department of Civil Engineering, Dongshin University

(Received 26 July 2006, Accepted 28 August 2006)

Abstract

The present study applied wavelet transform and artificial neural networks (ANNs) for the prediction of daily TOC data. TOC data were transformed into denoised data by the wavelet transform and the noise-reduced data were used for the prediction model by artificial neural networks. For the application of wavelet transform, Daubechies wavelet of order 10 ('db10') was used as a basis function and decomposed the TOC data up to fifth level with five detail components and one approximation component. ANNs were calibrated with the input data of the segregated TOC data corresponding to the details from second to fifth level and the approximation. Consequently, the ANNs model for the prediction of daily TOC data showed the best result when it had seventeen hidden nodes in its layer.

keywords : Approximation component, Artificial neural networks, Detail component, Noise reduction, Total organic carbon (TOC), Wavelet transform

1. 서 론

우리나라는 1960년대 이후 산업화에 따른 인구증가, 도시화 및 공단의 집단화 등으로 인하여 오염원의 양적증가와 지역적 집중화가 이루어지고 있다. 이러한 문제를 해결하기 위하여 정부에서는 물관리종합대책을 수립하여 시행중에 있으며 수질의 상시감시 기능과 조기경보체계의 구축을 위해 전국 4대강 유역 주요 20개 지점에서 수질자동 측정망을 설치 운영 중에 있다. 이러한 수질자동측정망의 설치 목적을 달성하기 위해서는 실시간으로 관측된 수질자료에 포함된 잡음에 대한 연구 및 예측에 관한 심화연구가 필요하다.

자연현상에서 발생하는 시계열 자료 내에 존재하는 잡음은 크게 세 가지로 분류할 수 있다. 첫 번째로는 시계열자료를 수집하는 과정에서 발생하는 측정오차(Measurement error)이며, 두 번째는 시계열자료를 수집하는 측정기 자체의 오차와 측정오차를 포함한 계통오차(Systematic bias)가 있다. 마지막으로 동역학적 잡음(Dynamical noise)은 다양한 자연현상을 나타내는 변수들 사이에 존재하는 상호간섭으로 인해 임의의 변량이 갖게 되는 오차가 있다(Kaplan et al., 1997). 이러한 오차들로 인하여 원자료의 특성이 왜곡되는 현상으로 귀결된다.

수질농도 예측 및 웨이블렛 변환과 관련된 연구동향을 살펴보면 오 등(2002)은 인공신경망 이론을 이용하여 나주지점의 BOD, COD, TN, TP 수질농도 예측 모형을 개발하였다. 또한, 안 등(2004)은 평창강, 달천, 여주지점의 TOC 수질자료를 이용하여 예측모형을 개발하였으며, 이와 더불어 예측모형에 판단모형의 접합을 통해 기존의 인공신경망에 의한 수질예측모형보다 우수한 결과를 나타내었다. 조 등(1998)은 물소비특성을 분석하기 위하여 웨이블렛 변환('Coiflets5')를 사용하였으며, 진 등(2005) 목포지점의 강수자료에 웨이블렛 변환('db9')을 적용하여 강수자료가 갖고 있는 연주기성 및 장주기성을 규명하였다. 권 등(2005)은 국내 강우량 자료 및 SOI, SST의 수문기상자료에 대한 웨이블렛의 적용성 평가에서 통계적으로 유의한 결과를 제시하였다. 월별 지하수위 예측(Wang, 2003)에 웨이블렛 변환과 인공신경망의 결합으로 예측력의 우수한 입증하였고 Zanchettin 등(2005)은 석유정제 공장으로부터 추출된 냄새신호의 잡음저감을 위하여 웨이블렛 변환을 적용하였다.

따라서 본 연구에서는 보다 정확한 TOC 수질농도를 예측하기 위한 방법으로서 웨이블렛 변환을 적용하여 TOC 수질자료에 포함되어 있는 잡음을 저감하였으며, 최종 변환된 근사성분 및 상세성분을 인공신경망의 입력자료로 이용하여 TOC 수질농도 예측모형을 개발하였다.

† To whom correspondence should be addressed.
yhjin@dsu.ac.kr

2. 이론적 배경

2.1. Wavelet 이론

웨이블렛 변환에 대한 연구는 1980년대에 들어서면서부터 본격적으로 진행되었으며 1990년 전까지 수학 분야에서 Morlet과 Grossman 등에 의해서 실증적으로 연구되었으며 공학 분야에서는 Vetterli과 Smith 등에 의해서 부대역 부호화(Filter Banks, Subband Coding)가 연구되어 왔다. 1990년 후에 Daubechies 등에 의해서 이 두 분야가 하나의 이론이라는 것이 수학적으로 유도되어 웨이블렛 이론으로 결합되었다(강 등, 2001).

이러한 웨이블렛 변환은 푸리에 변환에 비하여 계산속도가 빠르며, 주어진 신호에 대한 시간과 주파수 영역에서의 정보를 균형적으로 국소화시킬 수 있다. 또한 불규칙한 데이터를 평탄하게 해줌으로써 데이터에 대한 예측이 가능하게 해주는 평탄성(smoothness) 분석이 가능하다는 장점이 있다(Daubechies, 1994). 이 뿐만 아니라 실제 신호를 적절한 형식으로 기록하거나 전송할 때 발생하는 계통오차(Systematic bias)에 대한 잡음제거의 효율성이 입증되었다(강 등, 2001).

웨이블렛 변환의 기저 함수로 사용되는 $\Psi(t)$를 모함수(mother wavelet)라고 하며 그 수학적 표현은 식 (1)과 같다. 식 (1)은 기저 함수의 스케일링과 전이를 나타내며 여기서 j는 스케일링을 결정하는 값이고, k는 함수를 얼마나 이동시킬 것인가를 결정하는 값이며, 정수범위에서 정의된다.

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k) \tag{1}$$

웨이블렛 변환은 식 (1)의 모함수를 k만큼 이동하고 j에 의해 크기가 변화하는 웨이블렛군(wavelet family)을 형성하며, 웨이블렛 기본 함수들의 중첩으로 임의의 함수를 표현한다. 이러한 웨이블렛 기본 함수들의 중첩은 각각 다른 스케일 레벨을 가지고 임의의 함수를 만들어 내며, 각 레벨은 그 레벨에 맞는 해상도를 가지게 된다.

결국 이산형 웨이블렛 변환은 각각의 스케일과 전이항에 의해 다음 식 (2)와 같다.

$$f(t) = \sum_{k=-\infty}^{\infty} a_{J,k}\phi_{J,k}(t) + \sum_{j=1}^{J}\sum_{k=-\infty}^{\infty} d_{j,k}\,\psi_{j,k}(t) \tag{2}$$

여기서 J는 최대 분해단계를 나타내며 일반적으로 자료의 수에 따라 그 값이 결정된다. 최대 분해단계에 의해 도출된 계수인 $a_{J,k}$는 가장 낮은 주파수 성분 즉 가장 넓은 스케일 영역을 나타내며 대상자료($f(t)$)와 스케일링 함수($\phi_{j,k}(t)$)의 내적에 의해 산정된다. 반면에 각각의 이산형 스케일 j와 전이항 k에 의해 산정되는 $d_{j,k}$는 최대 분해단계 이전의 각 단계에서의 주파수 성분들에 대한 계수들이며, 원자료($f(t)$)와 웨이블렛 함수($\psi_{j,k}(t)$)의 내적값이다.

본 연구에서는 Daubechies에 의해 제안된 다양한 웨이블렛 함수 중 다음과 같은 두 가지의 기준에 의해 최적의 함수를 선택하였다. 그 기준으로는 대상자료에 대한 웨이블렛 변환 후 재현기간 45일 이상 해석이 가능한 함수 그리고, 원자료와 최종분해단계 근사성분과의 상관계수가 높은 기준으로 최종 웨이블렛 함수를 선택하였으며 그 결과 'db10' 함수를 적용하였다.

2.2. 인공신경망 이론

인공신경망 모형에서 학습이란 입력층, 은닉층, 출력층으로 구성된 다층신경망의 각 층 노드들 간의 연결강도를 최적의 상태로 적응시키는 과정을 말하며, 입력 자료의 형태에 따른 학습방법은 Fig. 1과 같다.

일반적으로 인공신경망 모형의 학습을 위해 Fig. 1의 역전파학습알고리즘을 이용하며, 이에 대한 기본 방법은 최급강하법이다. 이러한 최급강하법은 목적함수의 1차 도함수를 이용한 기울기에 비례하는 조정량을 산출하여 목적함수의 값이 개선될 수 있도록 매개변수의 최적화를 위해 반복적으로 탐색하는 방법이다. 본 연구에서는 최급강하법의 보다 효율적인 훈련과 더 나은 결과의 도출을 위해 모멘텀 상수와 적응식 학습율을 사용하였으며 그 식은 다음과 같다.

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha[w_{ji}^{(i)}(n-1)] + \eta\delta_j^{(l)}y_i^{(l-1)}(n) \tag{3}$$

여기서, η는 학습율, α는 모멘텀 상수, $w_{ji}^{(l)}(n)$는 각 층으로 연결되는 가중치, δ는 역방향계산시 출력층 노드로 연결될 때에는 오차값과 활성화함수의 곱이며, 은닉측으로 연결될 때에는 가중치와 활성화함수의 곱이다.

본 연구에 사용된 인공 신경망의 종류는 다층(multi-layer) 신경망이며, 입력층과 은닉층, 그리고 출력층으로 구성된다. 또한 매개변수 최적화를 위하여 사용된 학습 알고리즘은 모멘텀 상수(momentum constant)와 적응식 학습율(adaptive learning rate)이 적용된 최급 강하법을 이용한 오차 역전파

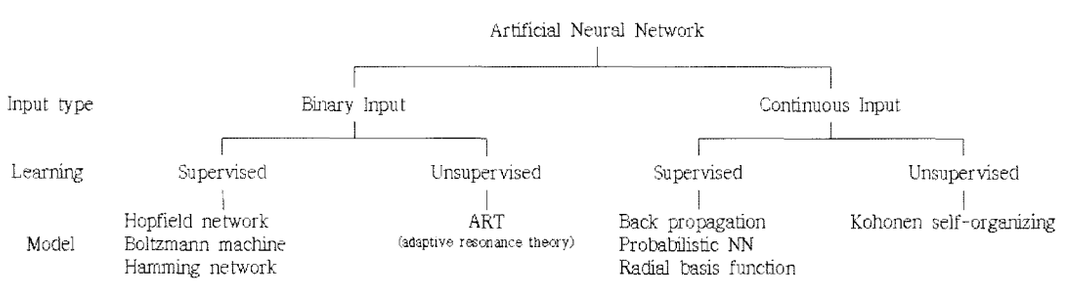


Fig. 1. Training methods of artificial neural networks according to input data.

학습 알고리즘이며, 이에 포함된 활성화 함수는 은닉층에서는 탄젠트 시그모이드(tangent sigmoid) 함수와 출력층에서 선형(linear) 함수를 사용하였다.

3. 대상지점 및 자료

본 연구의 대상지점인 주암호(Fig. 2)는 유역면적 1,010 km^2, 저수량 4억 5700만 ton으로 보성군 · 순천시 · 화순군의 3개 시군에 걸쳐 있으며, 광주광역시 · 나주시 · 목포시 · 화순군 등 전라남도 서부권에 하루 80만 ton의 생활용수를 공급하고 있다.

주암호1 수질관측망은 전라남도 순천시 대광면 주암리에 위치해 있으며, 본 연구를 위해 주암호 지점으로부터 2002년 1월 1일 ~ 2004년 12월 31일까지 총 1,096개의 일평균 TOC 수질농도 자료를 이용하였다.

TOC 수질자료에 대한 기술통계 분석 결과는 Table 1과 같으며, 평균농도는 1.46 mg/L로 비교적 양호한 수질농도임을 알 수 있었다. 또한 TOC 수질농도의 분포는 0.98~2.93 mg/L로서 최저농도는 2003년 3월 26일의 0.98 mg/L, 최고농도는 2004년 7월 11일의 2.93 mg/L로 나타났다. 이에 따른 TOC 수질농도 곡선 및 월별 박스플롯은 Fig. 3과 같다.

Table 1. Descriptive statistics of data used

	Max	Min	Mean	Stdev
TOC	2.93	0.98	1.46	0.26

Fig. 2. Location of the study area.

4. 모의결과 및 고찰

4.1. 웨이블렛 변환 결과

웨이블렛 변환 함수 선택기준에 대하여 2절에서 밝힌 바와 같이 Daubechies의 'db10'를 적용하였다.

일 TOC 자료에 대하여 'db10'를 적용한 결과 약 46.7일의 강한 주기성을 갖는 것으로 나타났으며 단계별 주기성 및 상세 · 근사성분의 에너지 값을 Table 2에 제시하였다. 또한, TOC의 원자료 및 각 단계별 웨이블렛 변환 결과는 Fig. 4에 나타내었다. Table 2에 나타난 바와 같이 전체에너지 중 근사성분(A5)의 에너지는 99.37%로 나타났으며 이는 근사성분이 원자료에 대하여 99.37% 설명해 주고 있다. 그리고 상세성분(D5~D1)의 에너지는 극히 낮은 0.63%의 에너지 값을 갖고 있다. 여기서 에너지 정보는 웨이블렛 변환에 의한 각 단계별로 추출된 상세 및 근사성분이 원자료를 설명하고 있는 정도를 나타내는 정량적인 지표로서 다음의 식과 같다.

$$\text{Energy(P(A, Dj) ; \%)} = \frac{\sum_{n=1}^{k} P_n^2}{\sum_{n=1}^{k} R_n^2} \times 100 \quad (4)$$

여기서, A는 근사성분, Dj는 각 단계별 상세성분, R은 원자료, P는 웨이블렛 변환된 근사 또는 상세성분 자료, k는 자료의 수이다.

4.2. TOC 수질농도 예측모형의 개발

주암호1지점에 대한 1일 후 TOC 수질농도를 예측하기 위하여 각 단계별로 최종 변환된 근사성분 및 상세성분 자료를 이용하여 다음의 식 (5)~(10)까지 인공신경망의 예측모형을 구성하였다.

최적의 인공신경망 모형 선별 조건으로는 일차적으로 인공신경망 모형의 매개변수인 모멘텀 상수와 초기 학습율은 모든 모형에서 각각 0.1과 0.7, 은닉층의 수를 10개로 일괄적으로 적용하여 모형을 선별하였으며, 선별된 모형에 대하여 은닉층의 수를 10개부터 20개까지 순차적으로 1개씩 증가하여 최적의 TOC 수질농도 예측모형을 선별하였다.

일차적인 모형 선택의 결과는 Table 3에서 잘 보여주고 있는데 TOC_M5 모형이 가장 좋은 예측 결과를 보였다.

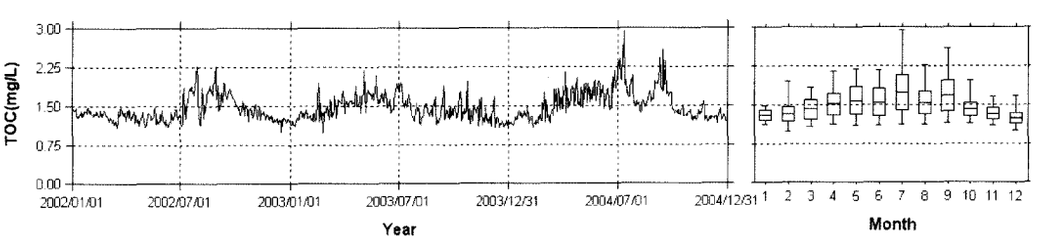


Fig. 3. Time series plot and box-whisker plot of daily TOC on a monthly basis.

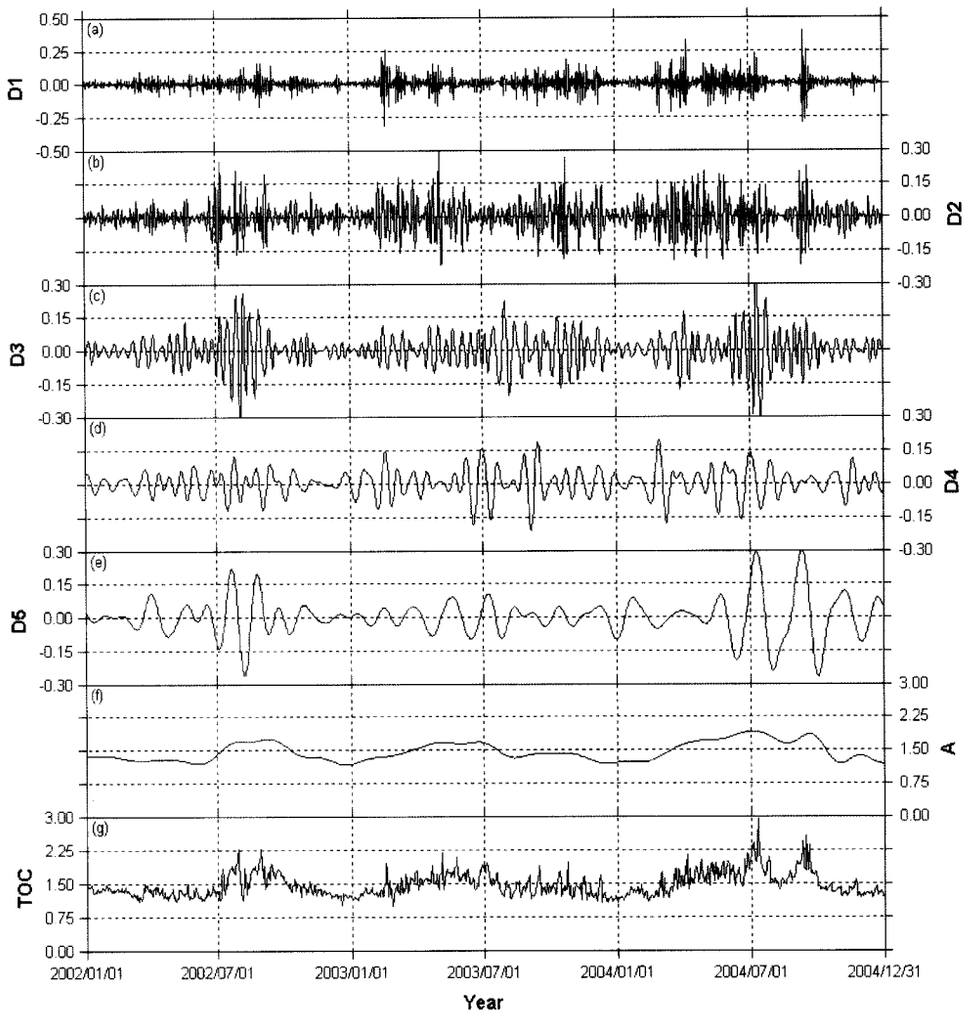


Fig. 4. Comparison between daily TOC data and decomposed data by wavelet transform. (A is stands for qpproximation and D for datail)

Table 2. Scale, frequency and energy values at the respcetive levels of discrete wavelet transform for daily TOC data

	Level (j)	Scale(a=2j)	Frequency (1/Days)	Period (Days)	Energy(%) of details	Energy(%) of approximation
TOC	1	2	0.3421	2.9230	0.08	
	2	4	0.1712	5.8462	0.12	
	3	8	0.0855	11.6918	0.16	
	4	16	0.0428	23.3863	0.12	
	5	32	0.0214	46.7727	0.15	**99.37**

또한, 근사성분으로 구성된 모형(TOC_M1)은 검정, 검증Ⅰ, 검증Ⅱ과정에서 가장 낮은 예측력을 보였으나 근사성분 및 상세성분을 포함한 모형에서는 순차적으로 예측력이 증가함을 보였다. 반면에 D1 상세성분을 포함한 TOC_M6는 상세성분이 포함된 5개의 모형 중 가장 낮은 예측력을 보이고 있는데 이는 웨이블렛 변환함수 'db10'에 의하여 최종 분해된 상세성분 D1은 TOC 원자료에 포함되어 있는 잡음성분임에 기인한 것으로 판단된다.

본 연구에서 선별한 TOC 수질농도 예측 모형은 전체적으로 주암호1 지점의 TOC 수질농도에 대한 특성을 가장 잘 반영하고있는 TOC_M5_17 모형을 선택하였다. 본 모형의 결과를 살펴보면 검정 및 보정, 검증과정에서의 상관계수는 0.953, 0.968, 0.943이며 RMSE는 각각 0.067, 0.086, 0.090으로 가장 우수한 예측력을 보였으며 그 결과를 Table 4에 나타내었다. Fig. 5는 TOC_M5_17 모형의 검정, 검증Ⅰ, 검증Ⅱ과정에 대한 관측값과 모의값 및 산포도를 도시한 그림이다. Fig. 5(c)를 살펴보면 TOC 수질농도의 최고값을 갖는 2004년 7월 28일의 2.93 mg/L에 대한 예측농도는 2.45 mg/L로 관측값에 비하여 저평가된 것으로 나타났다. 이는 검정과정에서의 최대값은 2.26 mg/L이나 검증Ⅱ과정에서의 최대값은 2.93 mg/L임에 따라 최대 TOC 수질농도를 갖는 입력패턴에 대한 적절한 훈련이 이루어지지 않음에 기인한 것으로 판단된다.

$$TOC\_M1(t) = f\left[\ A5(t-1,\ 2)\ \right] \quad (5)$$

$$TOC\_M2(t) = f\left[\ A5(t-1,\ 2), D5(t-1,\ 2)\ \right] \quad (6)$$

$$TOC\_M3(t) = f\left[\ A5(t-1,\ 2), D5(t-1,\ 2), D4(t-1,\ 2)\ \right] \quad (7)$$

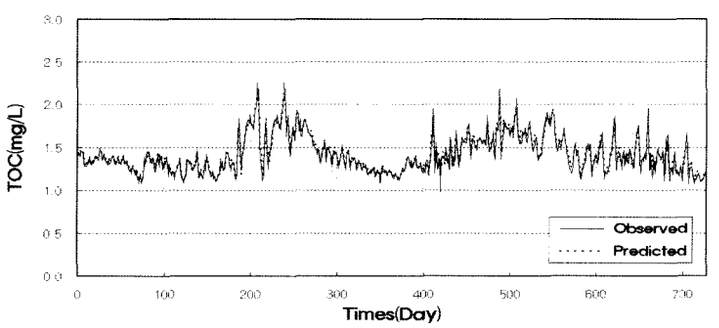
$$TOC\_M4(t) = f\left[\ A5(t-1,\ 2), D5(t-1,\ 2), D4(t-1,\ 2), D3(t-1,\ 2)\ \right] \quad (8)$$

$$TOC\_M5(t) = f\left[\ A5(t-1,\ 2), D5(t-1,\ 2), D4(t-1,\ 2), D3(t-1,\ 2), D2(t-1,\ 2)\ \right] \quad (9)$$

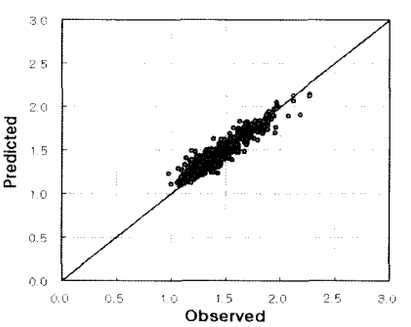
$$TOC\_M6(t) = f\left[\ A5(t-1,\ 2), D5(t-1,\ 2), D4(t-1,\ 2), D3(t-1,\ 2), D2(t-1,\ 2),\ D1(t-1,\ 2)\ \right] \quad (10)$$

Tabel 3. Correlation coefficient and RMSE according to the respective models

Model	CC			RMSE			Model	CC			RMSE		
	Calibration	Validation	Verification	Calibration	Validation	Verification		Calibration	Validation	Verification	Calibration	Validation	Verification
TOC_M1	0.760	0.750	0.785	0.149	0.223	0.176	TOC_M4	0.870	0.915	0.883	0.112	0.146	0.128
TOC_M2	0.817	0.899	0.832	0.134	0.163	0.159	TOC_M5	0.932	0.950	0.932	0.083	0.117	0.107
TOC_M3	0.854	0.909	0.865	0.115	0.148	0.139	TOC_M6	0.784	0.831	0.810	0.227	0.397	0.330

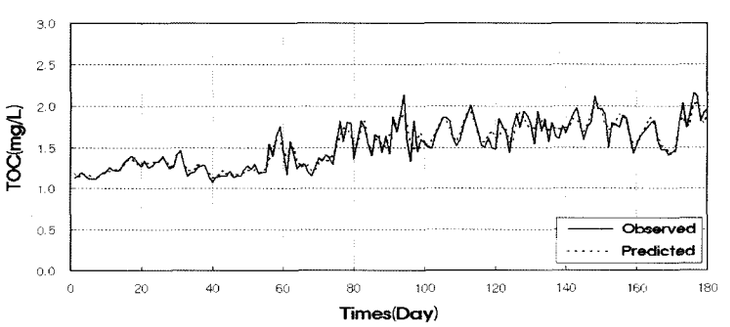


ⓐ Time series plots between raw and predicted water quality for calibration

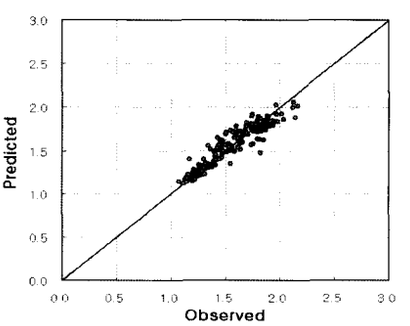


ⓑ Scatter plot of the calibration

(a) The process of calibration

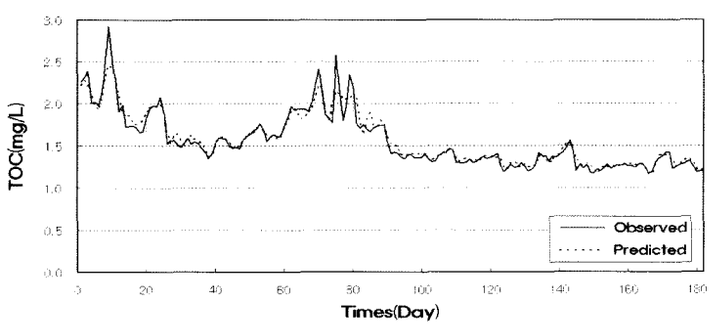


ⓐ Time series plots between raw and predicted water quality for validation

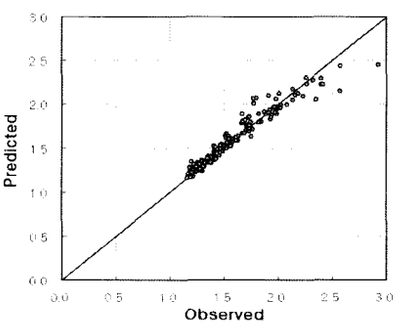


ⓑ Scatter plot of the validation

(b) The process of validation



ⓐ Time series plots between raw and predicted water quality for verification



ⓑ Scatter plot of the verification

(c) The process of verification

Fig. 5. Time series plots between raw and predicted data by TOC_M5_17.

Table 4. Correlation coefficient and RMSE according to the number of hidden node for the models

Index	CC			RMSE			Index	CC			RMSE		
	Training	Validation	Verification	Training	Validation	Verification		Training	Validation	Verification	Training	Validation	Verification
TOC_M5_6	0.945	0.960	0.940	0.074	0.098	0.097	TOC_M5_14	0.856	0.859	0.874	0.118	0.202	0.147
TOC_M5_7	0.950	0.969	0.939	0.073	0.089	0.095	TOC_M5_15	0.900	0.933	0.895	0.098	0.139	0.128
TOC_M5_8	0.945	0.958	0.942	0.076	0.101	0.098	TOC_M5_16	0.881	0.905	0.880	0.108	0.156	0.129
TOC_M5_9	0.913	0.938	0.921	0.095	0.122	0.108	TOC_M5_17	0.953	0.968	0.943	0.067	0.086	0.090
TOC_M5_10	0.932	0.950	0.932	0.083	0.117	0.107	TOC_M5_18	0.896	0.923	0.905	0.097	0.137	0.115
TOC_M5_11	0.926	0.932	0.930	0.089	0.139	0.110	TOC_M5_19	0.935	0.956	0.929	0.078	0.107	0.101
TOC_M5_12	0.844	0.839	0.858	0.118	0.194	0.140	TOC_M5_20	0.920	0.936	0.927	0.086	0.124	0.102
TOC_M5_13	0.938	0.959	0.933	0.079	0.105	0.104							

5. 결 론

본 연구에서는 수질자료에 내재되어 있는 잡음 성분의 저감을 위하여 웨이블렛 변환을 적용하였다. 또한 최종 변환된 상세성분 및 근사성분을 이용하여 비선형 시계열자료에 대한 예측력이 우수한 인공신경망을 적용하여 예측모형을 개발하였다.

TOC 자료의 잡음 저감을 위하여 웨이블렛 변환함수 'db10'를 적용한 결과 5단계까지 분해되었으며 전체에너지 중 근사성분의 에너지는 매우 높은 99.37%로 나타났다.

최적의 TOC 수질농도 예측 모형은 웨이블렛 변환에 의한 5단계 근사성분과 2~5단계까지의 상세성분을 입력자료로 갖으며 은닉층의 노드의 수가 17개인 TOC_M5_17 모형으로 나타났다. 본 모형은 검정, 검증Ⅰ, 검증Ⅱ과정에서의 상관계수는 0.953, 0.968, 0.943으로 가장 우수한 예측력을 보였으며 RMSE 값 역시 0.067, 0.086, 0.090으로 가장 작은 오차값을 보임으로써 TOC_M5_17 모형이 주암호1지점의 TOC 수질농도에 특성을 가장 잘 반영한 모형으로 판단된다.

특히 웨이블렛 변환에 의해 분해된 상세성분들(D1~D5) 중에서 1단계의 상세성분(D1)을 인공신경망 모형의 입력자료로 활용할 경우 구축된 모형의 예측력이 현저하게 감소하였다. 이는 D1 성분이 TOC 원자료에 포함되어 있는 잡음 성분으로 판단된다.

따라서 본 연구에서 적용된 방법론은 연속적으로 측정되고 있는 수질 시계열 자료의 잡음 성분 제거를 위해 활용될 수 있을 것이다. 또한 1일 후 TOC 수질농도 예측모형의 개발은 호소수 수질관리 정책 입안의 기초자료로 활용될 수 있을 뿐만 아니라 수질의 상시감시 기능과 조기경보체계의 구축에 이용될 수 있을 것으로 기대된다.

참고문헌

강현배, 김대경, 서진근, *웨이블릿 이론과 응용*, 아카넷, pp. 1~8 (2001).

권현한, 문영일, Wavelet Transform을 이용한 수문시계열 분석, *한국수자원학회논문집*, **38**(6), pp. 439-448 (2005).

안상진, 연인성, 실시간 자동측정망 자료를 이용한 수질관리, *대한토목학회 논문집*, **24**(3B), pp. 221-228 (2004).

오창열, 박성천, 이한민, 표영평, 신경망을 이용한 영산강의 수질예측, *대한토목학회 논문집*, **22**(3B), pp. 371-382 (2002).

조용준, 김종문, Wavelet Transform을 이용한 물수요량 특성분석 및 다원 ARMA 모형을 통한 물수요량 예측, *한국수자원학회 논문집*, **31**(3), pp. 317-326 (1998).

진영훈, 박성천, 이연길, 수문시계의 장·단기 성분 추출을 위한 웨이블렛 변환의 적용, *대한토목학회 논문집*, **25**(6B), pp. 493-499 (2005).

Daubechies, I., Ten Lectures on Wavelets, *SIAM*, **61**, pp. 258-261 (1994).

Kaplan, D. and Glass, L., Understanding Nonlinear Dynamics, *Springer*, pp. 280-286 (1997).

Wang, W. and Ding, J., Wavelet Network Model and Its Application to the Prediction of Hydrology, *Nature and Science*, **1**(1), pp. 67-71 (2003).

Yuehui, C., Bo, Y. and Jiwen, D., Time-series Prediction using a Local Lnear Wavelet Neural Network, *Neurocomputing*, **69**, pp. 449-465 (2006).

Zanchettin, C. and Ludermir, T. B., Wavelet Filter for Noise Reduction and Signal Compression in an Artificial Noise, *Applied Soft Computing*, in press (2005).