Wavelet Transform 방법과 SVM 모형을 활용한 상수도 수요량 예측기법 개발

A Development of Water Demand Forecasting Model Based on Wavelet Transform and Support Vector Machine

권 현 한* / 김 민 지** / 김 윤 기***

Kwon, Hyun-Han / Kim, Min-Ji / Kim, Oon Gi

Abstract

A hybrid forecasting scheme based on wavelet decomposition coupled to a support vector machine model is presented for water demand series that exhibit nonlinear behavior. The use of wavelet transform followed by the SVM model of each leading component is explored as a model for water demand data. The proposed forecasting model yields better results than a traditional ARIMA time series forecasting model in terms of self-prediction problem as well as reproducing the properties of the observed water demand data by making use of the advantages of wavelet transform and SVM model. The proposed model can be used to substantially and significantly improve the water demand forecasting and utilized in a real operation.

Keywords : water demand, wavelet transform, SVM, nonlinear forecasting model

요 지

본 연구에서는 Wavelet Transform과 Support Vector Machine (SVM)을 결합한 Hybrid 상수도 수요량 예측 모형을 개발하였다. Wavelet Transform 방법을 활용하여 다양한 스케일이 존재하는 상수도 수요량 시계열을 분해하여 단순한 형태의 시계열로 변환하는데 이용하였으며, 비선형 예측모형인 SVM은 이들 단순화된 시계열을 예측하는데 활용하여 예측성능을 극대화시키는 방안을 수립하였다. 본 연구에서는 상수도 수요량 자료에서 내재되어 있는 주기의 특성과 비선형 예측모형의 장점을 서로 연계한 해석이 가능하였으며 시각적인 검토 및 모든 통계지표에서 개선된 예측결과를 확인할 수 있었다. 특히, 기존 ARIMA 모형 계열에서 나타나는 자기예측문제를 상당부분 개선한 결과를 보여줌으로서 실질적인 수요량 예측모형으로서 활용이 가능할 것으로 판단된다.

핵심용어 : 상수도 수요량, wavelet transform, SVM, 비선형 예측 모형

1. 서 론

최근 나타나고 있는 기상변동성 증가는 기존 물수요량

의 변동성을 증가시키는 원인으로 작용하고 있다. 과거에는 인구증가, 대규모 신도시 개발, 산업발달 등으로 인해 물수요량이 급증하였으며 이를 위해서 상수도 시설 계

* 전북대학교 공과대학 토목공학과 조교수, 방재연구센터 (e-mail: hkwon@jbnu.ac.kr)
Assistant Professor, Dept. of Civil Engrg., Chonbuk Univ., Disaster Prevention Center, Jeonju 561-756, Korea

** 전북대학교 공과대학 토목공학과 석사과정, 방재연구센터 (e-mail: kiminji0228@jbnu.ac.kr)
Graduate Student, Dept. of Civil Engrg., Chonbuk Univ., Disaster Prevention Center, Jeonju 561-756, Korea

*** 교신저자, 정읍시청 재난안전관리과 (e-mail: kog6858@korea.kr, Tel: 063-270-2426)
Corresponding Author, Disaster & Safety Administration, Jeongeup City Hall, Jeongeup 580-701, Korea

획이 이루어져 왔다. 최근에는 기존 양적 증가뿐만 아니라 생활수준의 향상으로 기인하여 질적 측면에 대해서도 상당한 높은 수준의 상수도계획 및 운영이 요구되고 있다. 이와 더불어 최근 수자원의 생산 및 분배시설의 효율성을 극대화 시키고자 하는 스마트 워터 그리드 개념이 도입되면서 정확한 상수도 수요량 예측을 통한 효율적인 상수도 공급 개념이 매우 중요한 요소로 인식되고 있다. 즉, 장단기 물 수요는 상수도 시설의 계획 및 운영의 효율성을 위해서 필수적으로 고려되어야 할 사항으로서 다양한 방법이 제안되어 왔으나 상수도 수요량 시계열의 비선형성으로 인해 정확한 예측에 어려움이 따르고 있다. 이러한 점에서 본 연구에서는 다중스케일 분석 기법인 Wavelet Transform 방법과 비선형 예측 기법인 Support Vector Machine (SVM) 기법을 통합한 Wavelet Support Vector Machine (WSVM) 모형을 개발하여 일단위 상수도 수요량을 예측하는데 활용하고자 한다.

일반적으로 상수도 수요량 예측모형은 단변량(univariate) 및 다변량(multivariate) 회귀분석 모형을 이용하거나 추계학적 자기회귀모형(autoregressive model, AR) 계열이 이용된다. 국내에서는 외부적인 요인들의 영향들을 고려하면서 상수도 수요량을 예측하기 위한 연구들이 다수 진행되었다. 기존 상수도 수요량 예측 방법으로는 물 수요량과 지역적인 특성들의 주성분 분석을 통해 물 수요량에 영향을 미치는 여러 요인들에 대한 지역특성을 파악하고 이를 활용하여 물 수요 예측 식을 수립한 사례가 있으며(최재호 등, 1998), Fourier급수와 ARIMA (autoregressive integrated moving average)모형을 합성하여 지역별 시간변동에 따른 물 수요량 예측 모형을 제안하였다(김신걸 등, 1999). 또한 시계열 자료의 주기 특성을 파악하는데 유의한 Wavelet Transform을 통해 상수도 수요량 특성을 파악한 후 ARMA모형을 통해 물 수요량을 예측하였다(조용준과 김종문, 1998). 최근에 상수도 수요량 예측 방법으로서 배수지 운영을 위한 물수요량 패턴을 분석하여 시계열 예측 모형을 개발하고 이를 활용하여 월단위, 일단위 및 시간단위 용수 사용량 예측에 활용하였다(김주환 등, 2004). 이외에도 수요량을 예측하는데 많이 활용되고 있는 AR모형과 MLP (multilayer perceptron)를 결합하여 다중신경망의 모형을 제안하고 상수도 수요량 예측에 활용하였다(최기선 등, 2009).

최근 국외에서는 물 수요량에 영향을 미치는 4가지 요인인 추세, 계절성, 기후의 상관성 및 자기 상관성을 적용하여 일단위의 물 수요량을 예측하였고(Zhou et al. 2000), 또는 도심 지역의 날씨 자료와 상수도 수요량의 자료를 활

용하여 수요량을 예측한 결과 Neuro-Fuzzy모형이 기존 Fuzzy모형보다 우수한 예측능력을 나타내었다. 이밖에도 ANN (artificial neural networks)과 ANFIS (Neuro-fuzzy inference system)모형 또한 물 수요량 예측 모형으로서 활용성을 보여주었다(Tabesh and Dini, 2009). Zhou et al. (2002)은 일수요량과 시간단위 수요량을 통합한 예측모형을 개발하였으며 시계열의 자기상관성과 기상자료를 예측인자로 활용하였다. Ghiassi et al. (2008)은 동적 인공신경망 모델(dynamic artificial neural network model)을 활용하여 다양한 시간 범위에서의 도심지역 물 수요량을 예측하고 기존 ARIMA모형과 ANN모형을 결합한 방법론과 비교하여 모형의 우수성을 평가하였다. 이외에도 주택지역의 급수영역 분배패턴을 예측에 이용한 연구사례가 있다(Gato et al., 2007).

기존 연구에서는 다양한 외부적인 요인들을 고려한 회귀분석 모형이 이용되고 있으나 모형에서의 추정매개변수 증가로 인한 불확실성 증대, 모형의 과적합(over-fitting) 문제 등으로 상수도 수요량 예측에 어려움이 따른다. 특히, 자료의 비선형성으로 인해서 기존 관측 범위를 벗어나는 외삽(extrapolation) 시에 예측 결과의 신뢰성을 저해하는 요인으로 작용한다. 기존 자기회귀형식의 모형을 이용한 일수요량 예측의 자기상관성이 너무 큰 관계로 미래를 예측하기 보다는 단순히 지체시간만큼 이동하여 예측되는 자기예측(self prediction) 문제가 발생하는 단점이 있다.

이러한 점에서 본 연구에서는 첫째, 다중스케일 분석방법의 대표적인 방법으로 이용되고 있는 Wavelet Transform 방법을 이용하여 다양한 스케일이 존재하는 시계열을 특정 스케일로 분해하여 보다 간단한 형태의 시계열로 변환하고, 둘째, 각 스케일별로 비선형 예측모형인 SVM 모형을 적용하여 예측 정도를 높일 수 있도록 모형을 구성하였다. 이를 통하여 AR 형식 모형의 자기예측 문제, 비선형성을 고려한 예측모형을 개발하였다. 논문의 구성은 다음과 같다. 2장에서는 Wavelet Transform 방법과 SVM 방법에 대해서 요약정리 하였으며 3장에서는 서울특별시 강북정수장 일상수도 생산량을 대상으로 제안된 모형의 적합성을 평가하였다. 4장에서는 연구 결론 및 향후 연구 방향에 대해서 서술하였다.

2. 연구방법

본 연구에서는 서울시 정수장 자료를 대상으로 하였으며 시간적인 검토를 통하여 판단한 결과 시계열이 급격한 변동성을 보이는 경우가 있어 강북정수장만을 연구 대상으로

Table 1. Basic Statistic

Minimum (m ³)	Maximum (m ³)	Mean (m ³ /day)	Skewness	Kurtosis	Standard deviation
532,500	881,943	717,428	-0.14625	3.42685	45,919

로 하였다. 상수도 수요량을 예측하기 위해 2001년 1월 1일 ~2010년 12월 31일 기간의 일단위 강북정수장 수요량을 활용하였다. 강북정수장의 상수도 수요량 데이터를 다중스케일 분석에서 가장 우수한 분해 능력을 가지는 Wavelet Transform을 적용하여 시계열을 분해한 후 패턴인식 기반의 비선형 예측모형인 Support Vector Machine (SVM)을 적용하고자 하며 각 방법론에 대해서 요약 정리하였다. 다음 Table 1은 기본적인 통계값을 통해 강북정수장의 상수도 수요량의 특성을 알아보았다.

2.1 Wavelet Transform

최근 다양한 시간 스케일에서 시계열을 평가할 수 있는 다중스케일(multiscale) 분석이 활용되고 있으며 기본적으로 시계열에서 존재하는 다양한 시간적 스케일의 변동성을 보다 자세히 평가하기 위한 방법으로 이용되고 있다. 특히 지구물리학(geophysics)과 수문기상학(hydrometeorology) 등 자연현상을 규명하는데 이용되고 있으며 다양한 다중스케일 분석 중에서도 Wavelet Transform가 가장 진보된 방법으로 인식되고 있다(Weng and Lau, 1994). 수공학 분야에서도 기존에 이용되었던 Fast Fourier Transform (FFT) 방법에 비해 효율적인 방법임을 제시하기도 하였다(권현환과 문영일, 2005). Wavelet Transform과 FFT의 가장 큰 차이점은 시간분해능에 있다. 즉 FFT는 시간에 따른 주기의 변동성을 파악하는데 제약이 있는 반면에 Wavelet Transform은 시간에 따른 주기의 변동성을 동시에 파악할 수 있는 장점이 있다. FFT는 주기별 스펙트럼 강도로 표현되지만 Wavelet Transform은 시간에 따른 주기의 스펙트럼의 2차원의 정보를 제공해준다. 이러한 점에서 비선형적으로 변동하는 시계열의 주기정보를 보다 자세하게 평가할 수 있으며 시계열의 분해 및 합성을 통한 시계열 예측 및 모의에도 장점을 제공한다.

또한, 수문시계열이 갖는 시간적인 특성을 Wavelet Transform을 활용하여 강수량, 기온 등의 수문기상자료의 비선형 동역학적 성분을 추출하기 위해서 활용하기도 하였고(진영훈과 박성천, 2006) 수문시계열 자료인 강수량 자료에 Wavelet Transform을 적용하여 대상 자료의 장·단기 주기성분을 파악하였다. 비선형 시계열의 저빈도 주기특성을 효과적으로 재현할 수 있는 방법으로 Wavelet Transform을

이용한 시계열모의기법을 개발하였다(Kwon et al., 2007). 이러한 점에서 본 연구에서는 다중스케일 방법으로서 Wavelet 방법을 활용하였다.

Wavelet Transform에 대한 이론적 배경은 권현환과 문영일(2005)의 논문을 중심으로 정리 요약 하였다. Wavelet Transform은 기저함수(basis function)의 스케일항과 천이항(translation)이라는 두 변수로 표현되고 시계열의 서로 다른 스케일 성분들로 분해 가능하다. 여기서 스케일은 주파수영역에서의 우리가 평가하고자 하는 주파수를 의미한다. 즉, 스케일 항을 이용하여 기저함수의 폭을 축소 또는 확대하여 변환시킴으로서 고빈도(high frequency)에서 저빈도(low frequency)에 해당하는 다양한 주파수영역에서 시계열의 주기를 평가하게 된다. 천이항은 시간에 따른 각 주파수가 가지는 스펙트럼의 강도를 평가하기 위해서 도입된 항으로서 천이항을 우리가 원하는 시간으로 이동시키면서 시간에 따른 주기의 강도를 평가할 수 있다. 이러한 점에서 기존 FFT와는 다르게 시간에 따라 변동하는 주기의 강도를 용이하게 평가할 수 있으며 시간-주파수영역에서 전체 또는 지역적인 특징도 분석 가능한 특성을 지닌다. 이러한 Wavelet Transform의 특성들은 2차원 영역의 복잡성을 효과적으로 분석할 수 있으며 시계열분석에 있어 또한 많이 활용된다.

Wavelet Transform은 기저함수의 형식이 특별히 정해진 것이 아니며 기저함수가 될 수 있는 조건만이 정해져 있기 때문에 다양한 기저함수가 존재한다. Eq. (1)은 Wavelet Transform의 기저 함수를 나타내며 여기서 sc 는 스케일을 결정하는 값이고, sh 는 함수를 얼마나 이동시킬 것인가를 천이항을 의미한다.

$$\Psi_{sc,sh}(t) = \frac{1}{\sqrt{|sc|}} \Psi\left(\frac{t-sh}{sc}\right) \quad sc, sh \in R \quad (1)$$

Eq. (1)과 같이 기저 함수가 될 수 있는 함수 $\Psi(t)$ 의 스케일과 천이를 통해서 Wavelet Transform을 수행한다. Wavelet Transform의 기저 함수로 사용되는 $\Psi(t)$ 를 모(mother) Wavelet 함수라고 하며 다음의 Eqs. (2) and (3)의 두 가지 조건을 만족 시키면 모 Wavelet 함수가 될 수 있다.

$$\int_{-\infty}^{\infty} \Psi(t) dt = 0 \quad (2)$$

$$\int_{-\infty}^{\infty} |\Psi(t)|^2 dt < \infty \quad (3)$$

Wavelet Transform은 Eq. (1)에서 모 Wavelet을 sh 만 큼 이동하고 sc 에 의해 크기를 변화시켜가는 기저 함수를 사용한다. 이는 고주파로 갈수록 Wavelet은 함수의 폭이 좁아지고, 저주파로 갈수록 함수의 폭이 넓어지는 것을 나타낸다. Wavelet Transform은 Wavelet 기본 함수들의 중첩으로 임의의 함수를 표현하는 것인데 이러한 Wavelet 기본 함수들의 중첩은 각각 다른 스케일 레벨을 가지고 임의의 함수를 만들어 내며, 각 레벨은 그 레벨에 맞는 해상도를 가지게 된다. 연속적인 신호의 Wavelet Transform (Continuous Wavelet Transform, CWT)은 Eq. (4)와 같이 정의되며 그것의 역 변환은 Eq. (5)와 같이 정의 된다. 여기서 $X(t)$ 는 원자료를 나타내며 $W(sc, sh)$ 는 에너지 스펙트럼을 의미한다. Ψ^* 는 Ψ 의 켈레복소수를 의미한다. 여기서 $\Psi(w)$ 는 Fourier Transform $\Psi(t)$ 의 값을 의미하며, w 는 주파수를 나타낸다.

$$W(sc, sh) = \int_{-\infty}^{\infty} X(t) \Psi_{sc, sh}^*(t) dt \quad (4)$$

$$X(t) = \frac{1}{C} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{|sc|^2} W(sc, sh) \Psi_{sc, sh}(t) dsh dsc \quad (5)$$

$$C \equiv \int_{-\infty}^{\infty} \frac{|\Psi(w)|^2}{|w|} dw$$

일반적으로 CWT에서 스펙트럼 추정 시에 효율적인 방법은 이산(discrete) Fourier Transform (DWT)을 이용해서 Fourier Transform 영역에서 스펙트럼을 추정하는 것이다. DFT의 Convolution 이론에 근거하여, Wavelet Transform은 Fourier Transform의 역변환과 같으며 다음 Eq. (6)과 같이 나타낼 수 있다.

$$W_t(s) = \sum_{k=0}^T \hat{X}_k \hat{\Psi}^*(sw_k) e^{i w_k n \delta_t} \quad (6)$$

여기서 $\hat{X}_k$ 은 원시계열 X 의 Fourier Transform을, k 는 $(0, \dots, T)$ 의 주파수를 의미하며 $\hat{\Psi}^*(sw_k)$ 는 Wavelet Transform의 Fourier Transform을 나타낸다. Wavelet 스펙트럼은 다음 Eq. (7)의 이산 스케일을 이용하여 계산된다.

$$s_j = s_0 2^{j\delta_j} \quad (7)$$

여기서, s_0 는 추출 가능한 가장 작은 스케일을 나타내며 δ_j 는 추출간격을 의미하고 모 Wavelet함수의 특성에 따

라 다른 값을 나타낸다. 본 연구에서는 Morlet 모함수를 이용한 CWT를 적용하여 연구를 진행하였으며 최종적으로 강북정수장의 상수도 수요량 시계열이 가지는 주기성 및 경향성 성분을 추출하여 SVM을 이용한 시계열 예측에 이용하였다.

2.2 Support Vector Machine (SVM)

Support Vector Machine (SVM)은 상수도 수요량 예측을 위해서 사용되어지고 있는 비선형 회귀분석 모형 중 하나이다. 또 SVM은 서로 다른 Class에 속하는 자료들의 Margin을 최대화하는 초평면(hyperplane)을 구함으로써 최적의 일반화 성능을 지니는 모델을 구현하는 것이며, 본 연구에서는 상수도 수요량의 Pattern을 분류(classification)의 문제로 해결하기 위한 수단으로 응용되었다. Fig. 1은 간단한 2개의 Class를 갖는 자료에 대한 SVM의 Pattern 인식 개념 즉, 초평면의 결정 과정을 나타낸다. Fig. 1과 같이 기하학적으로 w 는 평면 h 의 방향을 나타내며 다양한 b 의 변화는 평면을 움직이는 역할을 한다. 즉 w 와 b 를 적절히 조정하여 최적의 평면을 찾음으로써 2개의 다른 특성을 가지는 집단(class)을 효과적으로 구분할 수 있다. 여기서 최적의 경계면이란 Fig. 1과 같이 마진(margin)이 가장 큰 경계면을 찾는 것이다. Fig. 1에서 h 는 마진이 최대가 되는 마진의 중심을 이등분하는 초평면을 최적 분류 초평면이라 하며, 마진을 이루는 영역을 떠받치고 있는 w 와 b 거리에 있는 데이터 점들을 Support Vector이라고 한다.

이를 SVM 회귀분석에 응용한다면, $x \in R$ 에서 추출된

Fig. 1. Hyperplane Concept for an SVM Trained with Samples from Two Classes

입력항 $\{x_1, x_2, \dots, x_l\}$ 과 L차원의 $y \in \mathbb{R}$ 인 $\{y_1, y_2, \dots, y_l\}$ 출력항의 함수적인 의존관계 $f(x)$ 를 추정할 수 있다. 여기서 함수 $f(x)$ 의 추정은 다음 위험도 함수를 최소화 하는 문제로 귀결될 수 있다. 즉, $f(x)$ 는 기하학적으로 다수의 입력 벡터와 출력항 중에서 이들 관계를 규정할 수 있는 최적의 회귀함수를 의미한다.

$$R[f] = \int \Gamma(x, y, f(X)) dP(X, y) \quad (8)$$

$$f(X) = \langle W, X \rangle + b \quad \text{with} \quad W \in X, b \in \mathbb{R} \quad (9)$$

$$f(X) = \sum_{j=1}^K w_j x_j + b \quad (10)$$

SVM 회귀분석에서 b 는 bias를 w 는 Basis 함수를 나타낸다. $\langle, \rangle$ 는 x 의 벡터내적(dot product)을 의미하며 K 는 입력자료의 차원을 나타낸다. $\Gamma(x, y, f(X))$ 는 관측치 y 와 예측치 $f(x)$ 와의 차이 즉, Loss 함수를 나타낸다. 즉, Eq. (8)의 오차를 최소화 하는 최적화 과정을 통해서 매개 변수들이 결정된다.

결국 SVM 회귀분석은 입력항과 출력항의 비선형 함수 관계를 선형의 조합으로 분류함으로써 가능하며 결국 이러한 문제는 Pattern 분류와 밀접한 관계가 있다. SVM에서 비선형성은 커널함수(Kernel Function)를 이용하여 훈련자료를 고차원으로 보냄으로써 고려할 수 있으며 훈련

자료는 항상 표본들의 내적(inner product)의 형태로 항상 표현된다. 따라서 내적항을 적절하게 선택된 커널함수로 대체함으로써, 무수적인 매개변수의 증가 없이 고차원의 함수로 Mapping할 수 있으며 이를 소위 “커널트릭(Kernel Trick)”이라고 한다. 즉 Kernel 함수를 이용하여 입력항과 출력항의 상호관계를 효과적으로 규명할 수 있다.

Fig. 2는 Mapping함수를 이용한 비선형 SVM의 기본 개념도를 나타낸다. 그림에서 보는 바와 같이 Fig. 2(a)와 2(b)는 각각 1차원 데이터로 선형 분리 가능한 경우와 선형 분리 불가능한 경우를 나타내며 2(c)는 2(b)의 선형 분리 불가능한 1차원 데이터를 Mapping 함수 x^2 을 도입함으로써 특성이 다른 두 집단을 효과적으로 구분할 수 있다. 즉, Fig. 2(a)와 같이 성격이 다른 2개의 집단에 대해서 분리 작업을 수행한다면 2개의 집단의 마진이 최대가 되는 선을 가지고 쉽게 구분이 가능하다. 반면 Fig. 2(b)의 경우에는 하나의 직선으로 성격이 다른 2개의 집단을 분리할 수 없으며 이런 경우 1차원의 자료를 2차원의 자료로 투영함으로서 자료를 분리하게 된다. 그림에서 점선으로 나타낸 부분은 Epsilon Tube(ϵ)라고 하며 오차의 허용 범위를 나타낸다. 즉, ϵ 를 너무 작게 하게 되면 집단간의 분리를 위해서 과적합(overfitting)이 이루어지게 되어 다수의 Support Vector들이 생성되게 된다. 따라서 SVM에서는 ϵ 를 통하여 모형의 일반성을 확보하는 과정을 수행하며 이를 통해 모형의 과적합을 방지하게 된다. 모형의

2(a) Datasets that are linearly separable with some noise work out great

2(b) Datasets that are not linearly separable

2(c) By employing x^2 , Mapping data to a higher-dimensional space and data set are separable

Fig. 2. A Concept of Nonlinear SVM Model using Mapping Function

훈련단계에서의 적합성 향상뿐만 아니라 실제 예측능력을 향상시킬 수 있도록 모형의 일반화에 주안점을 두고 모형이 구축된다. 그림에서 나타내어지는 Support Vector 들은 두 개의 집단을 분류시키는 기준점이 되며 두 개의 집단을 분류하는 초평면에 마진이 최대가 되도록 최적화 과정을 통하여 도출된다. 논문의 지면 제약상 SVM에 대한 좀 더 구체적인 이론적 배경은 기존 연구 등을 참조할 수 있다(Boser et al., 1992; Guyon et al., 1993; Cortes and Vapnik, 1995; Vapnik, 1995; Schölkopf et al., 1995; 권현한과 문영일, 2005; 황석환 등, 2007).

2.3 Multiscale 분석 및 비선형 예측 모형 적용

본 연구에서는 상수도 수요량을 예측하는데 있어서 다중스케일 기반 비선형 예측 모형을 개발하고자 한다. 여기서 다중스케일이라는 의미는 상수도 수요량 시계열에서 나타나는 다양한 시간적 스케일을 이용한다는 의미로서 복잡한 구조를 가지는 시계열을 분해하여 단순한 시계열로 만든 후 시계열 예측모형을 적용하는 복합예측 모의 기법이라 할 수 있다. 이에 본 연구에서는 앞에서 설명하

였던 Wavelet Transform을 적용하여 2001년 1월 1일~2010년 12월 31일 기간의 일단위별 강북정수장 수요량 시계열을 분해한 후 패턴인식 기반의 비선형 예측모형인 SVM을 적용하여 다음 Fig. 3과 같은 과정을 통해 상수도 수요량을 예측하였다. 모형 적용과정을 구체적으로 설명하면 다음과 같다.

첫째, 상수도 수요량 자료를 Wavelet Transform 기법을 통하여 단순화시킨다. Wavelet Transform을 적용하는 목적은 앞에서 언급했듯이 시계열이 가지는 주기특성을 효과적으로 파악하는데 있다.

둘째, Wavelet Transform을 통하여 통계적으로 의미 있는 주기만을 선별할 필요가 있으며 본 연구에서 Global Wavelet Spectrum을 통하여 통계적으로 의미 있는 성분만을 추출하고 이를 해석 대상으로 한다. 통계적으로 유의한 독립된 주기성분만을 독립적으로 모의 대상으로 하며 유의하지 않은 성분은 하나의 시계열로 통합하여 하나의 모형을 적용한다.

셋째, 특정 주기를 갖는 유의한 독립성분들에 대해서 ACF (autocorrelation function), PACF (partial autocor-

Fig. 3. A Proposed Forecasting Concept for Water Demand

relation function), MI (mutual information)을 이용하여 최적 지체시간을 결정한 후 SVM 모델을 통해 예측 모델을 구축한다.

넷째, 노이즈 성분에 대해서도 SVM 모델을 적용하여 예측을 실시한 후 앞서 예측된 성분과 모두 결합하여 최종적으로 예측시계열을 구성한다.

본 연구에서는 10년의 자료 중 9년의 자료를 대상으로 모형의 학습을 실시하였으며 1년에 대해서 모형 검증을 실시하였다. 최종적으로 C 개의 주요 주기성분에 대해서 최적지체시간이 p 가 결정되면 입력벡터 $\mathbf{x} = \{x_{t-1}, \dots, x_{t-p}\}$ 가 SVM 모형의 자료로 활용되며 다음 Eq. (11)과 같이 나타낼 수 있다.

$$x_t = \sum_{i=1}^C \left(\sum_{j=1}^K w_{i,j} x_{i,j} + b_i \right) \quad (11)$$

3. 결과 및 고찰

본 절에서는 앞서 제시된 Fig. 3의 예측 과정을 토대로 시계열 예측모형을 구축하였다. 본 연구에서는 모형의 적합성을 평가하고자 서울특별시 강북정수장을 연구 대상으로 선정하였다. 서울특별시의 6개 정수장을 고려하였으나 시각적인 검토를 통하여 판단한 결과 시계열이 급격한

변동성을 보이는 경우가 있어 강북정수장만을 연구 대상으로 하였다. 모형을 구축하는데 있어서 첫 번째 과정으로 강북정수장을 대상으로 Wavelet Transform을 적용하였다. Fig. 4는 강북정수장의 상수도 수요량에 대해서 Wavelet Transform을 수행한 결과를 나타낸다.

Fig. 4(a)는 대상 시계열을 나타내며 Fig. 4(b)는 상수도 수요량을 이용하여 스펙트럼 분석을 실시하여 얻어진 Wavelet Power Spectrum을 의미한다. 여기서 y축은 주기를 나타내며 단위는 연으로 정리하였다. Fig. 4(c)는 Fig. 4(b) y축의 시간의 방향으로 Wavelet Power Spectrum을 평균한 값으로서 GWP라 하며 시계열의 통계적으로 유의한 주기를 판단하는데 유용한 정보를 제공한다. 여기서 y축은 Fig. 4(b)와 동일하게 주기를 나타내며 x축은 스펙트럼의 강도를 나타낸다. 마지막으로 Fig. 4(d)는 Scaled Average Wavelet Power (SAWP)라 하며 GWP와는 반대로 Scale의 방향으로 Wavelet Power Spectrum을 평균한 값으로서 시계열의 주기의 시간에 따른 특정 주기에 해당하는 진폭(amplitude)의 변동성을 평가하는데 유리하다. Fig. 4(c)의 GWP와 중첩되어 도시된 실선은 유의수준에 해당하는 Wavelet Power Spectrum으로서 이 유의수준을 넘는 주기만이 통계적으로 유의한 성분이라 판단할 수 있다. 따라서 Fig. 4(c)의 결과를 대상으로 평가해보면 신뢰구간 95%에서 7일 주변의 주기 및 1년 주변의 주

Fig. 4. Wavelet Transform Analysis for Water Demand Data

기가 통계적으로 유의한 특성을 가지는 것으로 판단되었다. 또한 2년 이상의 경년주기 및 저빈도의 성분의 경우에도 강한 스펙트럼 강도를 보여주고 있으나 사용된 자료가 10년인 점을 고려하면 유의성이 상대적으로 작다고 할 수 있다. 여기서 4년 이상의 주기는 주로 증가 경향성을 나타내는 성분으로서 자료의 연환을 고려할 때 통계적으로 유의한 성분으로 판단하기는 어렵다. Fig. 4에서 상수도 수요량의 주요 주기로 평가된 7일, 1년, 2년, 4년 이상의 주기 등을 Wavelet Transform을 통해 재구성하여 추출하여 나타내면 Fig. 5와 같다. 즉, Fig. 5에서 나타난 시계열

은 특정 주기만을 가지는 성분을 시계열로부터 추출한 것으로서 첫 번째 그림은 10년 자료부터 추출한 7일주기의 시계열을 나타내며 두 번째 그림은 1년 주기의 해당하는 시계열을 나타낸다. 이런 방식으로 주요 주기에 대한 시계열을 나타내며 이를 모두 합하면 강북정수장의 상수도 수요량 자료가 된다.

모형의 적용과정을 설명하기 위해서 Fig. 5 왼쪽 첫 번째와 같이 Wavelet Transform을 통해서 추출된 첫 번째 Weekly 성분에 대해서 SVM 모델을 적용하여 예측 결과를 Figs. 6 and 7에 나타내었다. Fig. 6의 상단 그림은 모

Fig. 5. Dominant Wavelet Reconstructed Component Derived from Wavelet Analysis

Fig. 6. Forecasting Results of Calibration and Verification for Weekly Cycle of Wavelet Component

형 구축단계에서의 시각적 비교를 나타내며 하단의 그림은 모형 구축단계에서 한 구간인 1년에 대해서 1-Day Ahead 예측 결과를 나타내었다. Fig. 7은 1년에 대한 모형 구축단계에서의 상수도 수요량 관측값과 1-Day Ahead 예측 결과값의 결과를 비교 분석하여 나타내었으며, 단주기 (high frequency)에 속한다. Fig. 7의 산포도에서는 x, y축이 동일한 축척을 가지며 직선상에 자료들이 위치한다면 예측력이 있는 것으로 판단할 수 있다.

첫 번째 성분인 7일 주기 성분에 대해서 모형을 적용한

결과 모형 구축 및 검증 구간에서 모두 0.9 이상의 상관계수를 확보할 수 있었다. 두 번째 성분인 연주기 성분에 대해서도 동일한 방법으로 예측을 실시하여 나타내면 Fig. 8과 같으며 장주기(low frequency)에 속한다. 동일하게 모든 성분에 대해서도 SVM 모형을 적용하여 예측시계열을 구성하여 Figs. 9 and 10에서 비교 하였다. 본 연구에서 제시된 방법론의 적합성을 평가하기 위해서 시계열을 구분하여 검정과 검증 단계로 예측을 수행하여 각각 Figs. 10(a) and 10(b)로 나타내었다. 전체시계열 중 90%의 자

Fig. 7. A Scatter Plot of 1-day Ahead Forecasting Result for Weekly Cycle (high frequency)

Fig. 8. A Scatter Plot of 1-day Ahead Forecasting Result for Inter Annual Cycle

Fig. 9. A Combined Results by Summing up the Predicted Wavelet Components

Fig. 10. A Scatter Plot of Predicted Results for Calibration and Verification Period

료를 활용해 모형의 검정을 수행하였으며 검정 과정을 통해 추정된 매개변수를 토대로 나머지 10%의 자료의 대해서 모형의 검증과정을 수행하였다. 모형의 검증작업과 별개로 모형의 검정과정에서는 매개변수의 신뢰성을 더하기 위해서 Cross Validation 과정을 수행하여 매개변수를 추정하였다. Cross Validation은 검정기간의 자료기간 중 5개의 구간을 임의로 선택하여 매개변수를 추적하고 나머지를 검증에 활용하는 과정을 수행하였다.

마지막으로 구성된 예측 시계열의 적합성을 평가하기 위해서 다음 Table 2와 같은 지표들을 이용하였다. 여기서 y_t 는 관측치를 $\hat{y}_t$ 은 예측치를 나타내며 $E(y)$ 는 평균을 나타낸다. 평균제곱오차(root mean square error, RMSE)는 모형의 수행결과가 평균적으로 어느 정도 오차가 발생하는지를 나타내는 지표로 일종의 평균치를 나타낸다. 상관 계수(correlation coefficients)는 관측치 y_t 과 예측치 $\hat{y}_t$ 의 비례 또는 반비례 관계를 나타내는 계수이며, 1에 가까울

Table 2. Different Efficiency Measures Used This Study

Statistics	Formula
RMSE	$Q_1 = \sqrt{n^{-1} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$
Index of agreement	$Q_2 = 1 - \frac{\sum_{t=1}^n y_t - \hat{y}_t }{\sum_{t=1}^n y_t - E(y_t) + \hat{y}_t - E(\hat{y}_t) }$
Coefficient of efficiency	$Q_3 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (\hat{y}_t - E(\hat{y}_t))^2}$
Correlation coefficient	$Q_4 = \frac{\sum_{t=1}^n (y_t - E(y_t))(\hat{y}_t - E(\hat{y}_t))}{\sqrt{\sum_{t=1}^n (y_t - E(y_t))^2} \sqrt{\sum_{t=1}^n (\hat{y}_t - E(\hat{y}_t))^2}}$

수록 실측값에 가까운 거동을 보인다고 판단할 수 있다. 그러나 상관계수는 상대적으로 극치값에 민감한 단점이 있기 때문에 상대적으로 극치값에 영향을 상쇄시키는 통계치인 모형일치계수(index of agreement, IoA)를 활용하였다. 이와 함께 기존 모형의 예측능력을 평가하는데 많이 활용된 모형효율성계수(coefficient of efficiency)도 통계적 지표로 이용하였다. 모형효율성계수는 편의를 줄일 수 있는 무차원계수이며, 1에 가까워질수록 관측치 y_t 와 예측치 $\hat{y}_t$ 가 일치하다고 평가할 수 있다(Legates et al., 1999).

Table 3(a)는 본 연구에서 제안하는 WSVM 모형의 적합성을 평가한 결과로서 모형 구축과정 단계인 검정 단계와 구축된 모형을 바탕으로 검증하는 2가지 단계로 평가하였다. 이와 더불어 WSVM의 적합성을 평가하기 위해서

기존 수량 예측모형으로 주로 이용되고 있는 ARIMA 모형을 비교대상으로 하였다. Fig. 11은 ARIMA 모형의 자기예측문제점을 확인하기 위해서 2009년 1월 1일~2009년 12월 31일까지의 특정 기간을 확대하여 나타낸 그림으로서 기존 ARIMA 모형은 모형에서 사용된 차수만큼 오른쪽으로 이동하는 자기예측문제점이 나타나고 있다. 그러나 본 연구에서 제시된 모형의 경우 이러한 자기예측문제점이 상당히 개선된 것을 확인할 수 있다. ARIMA 모형의 최적 차수 및 매개변수 추정에는 SPSS를 통해서 수행하였으며 본 연구에서 활용된 데이터 특성상 ARIMA (1,2,1) 모형을 따르는 것으로 확인되어 최적 모형으로 적용하였다. Table 3(b)는 ARIMA (1,2,1) 모형을 적용한 결과로서 본 연구에서 제안된 WSVM에 비해서 예측력이 낮은 결과를 나타내고 있다. 특히 WSVM은 상대적으로 RMSE

Table 3. A Comparison of Modeling Results between WSVM and ARIMA Model

Statistics \ Model	WSVM (a)		ARIMA(1,2,1) (b)	
	Calibration Period	Verification Period	Calibration Period	Verification Period
RMSE (m ³ /day)	16202	18965	23658	30538
Index of agreement	0.9419	0.9555	0.8843	0.8751
Coefficient of efficiency	0.8080	0.8331	0.5906	0.5673
Correlation coefficient	0.9007	0.9148	0.7868	0.7741

Fig. 11. Improvement of Self Prediction Problem using the Proposed WSVM Model. The upper Panel Represents the WSVM Prediction Result and the Lower Panel Indicates the ARIMA Prediction Result

의 개선 효과가 크게 나타나고 있으며 이밖에도 다른 통계지표에서도 기존 ARIMA모형보다 개선된 예측력을 보여주고 있다.

4. 결 론

본 연구에서는 시계열에서 나타내는 비선형성, 주기성 및 경향성을 효과적으로 고려한 예측모형을 개발하기 위해서 Hybrid 상수도 수요량 예측모형을 개발하였다. 본 연구에서는 Wavelet Transform 방법을 활용하여 다양한 주기특성이 존재하는 상수도 수요량 시계열을 분해하여 단순한 형태의 시계열로 변환하는데 이용하였으며, 비선형 예측모형인 SVM은 이들 단순화된 시계열을 예측하는데 활용하여 예측성능을 극대화시키는 방안을 수립하였다. 본 연구를 통해서 얻은 결론은 다음과 같다.

- 1) 강북정수장의 상수도 수요량에 대해 Wavelet Transform을 활용하여 통계적으로 의미 있는 주기성분만을 추출하였다. 그 결과 강북정수장의 상수도 수요량은 7일 주기, 1년 주기, 2년 주기에서 통계적 유의성을 가지는 주기성분을 추출할 수 있었으며 이밖에도 증가경향을 저빈도(low frequency)의 형태로 효과적으로 분리하는 것을 확인할 수 있었다. 기존의 연구에서 요일별로 Weekly 주기 등을 예측에 활용한 사례가 있는데 본 연구에서 제안한 방법은 이러한 기존의 방법론을 포괄하는 방안으로서 다양한 스케일을 고려할 수 있는 장점이 존재하는 확인할 수 있었다.
- 2) 비선형 예측 모형으로서 우수한 성능을 나타내는 SVM 모형을 Wavelet Transform 방안과 연계하여 상수도 수요량 예측모형으로 개발하였다. 앞서 언급되었듯이 자료의 내재되어 있는 주기 특성과 비선형 예측모형의 장점을 서로 연계한 해석이 가능함을 확인할 수 있었다.
- 3) 관측치와 예측치 통계지표를 통해 ARIMA 모형과 WSVM 모형의 예측력을 확인할 수 있었다. 앞서 보여주었듯이 상관계수 값은 ARIMA 모형의 경우 대략 0.8의 값을 보였으며, WSVM 모형의 경우 0.9의 값을 보였다. 모형일치계수(IoA)는 각각 0.5, 0.8의 통계치를 모형효율성계수는 0.7, 0.9의 통계치를 보였다. 이러한 결과를 살펴보면 ARIMA 모형 보다 WSVM 모형의 예측력이 좋은 것을 확인할 수 있었다.
- 4) 시각적인 검토 과정에서는 특히, 기존 ARIMA 모형 계열에서 나타나는 자기예측문제를 상당부분 개선

한 결과를 보여줌으로서 실질적인 수요량 예측모형으로서 활용이 가능할 것으로 판단된다.

향후 연구계획으로서 본 연구에서 제안된 WSVM기반의 일수요량 예측모형과 연계할 수 있는 시간단위 수요량 예측모형을 개발하는 것이 필요하며 이를 위해 시간단위 일수요량의 패턴분석, 일단위안에서의 특이일 처리 등에 대한 연구가 선행되어야 할 것으로 판단된다.

감사의 글

본 연구는 “환경부 - 차세대 에코이노베이션 기술개발사업(GT-11-G-02-001-6)”에 의해 지원되었습니다. 저자들은 전북대학교 방재연구센터에 소속되어 연구를 수행하였습니다.

참고문헌

- 권현한, 문영일 (2005). “Wavelet Transform을 이용한 수문시계열 분석.” **한국수자원학회논문집**, 한국수자원학회, 제38권, 제6호, pp. 439-448.
- 김신걸, 주충남, 구자용, 유명진 (1999). “지역별 시간변동에 따른 물수요량 예측.” **대한환경공학회 추계학술연구발표회논문집**, 대한환경공학회, pp. 517-518.
- 김주환, 조임영, 우형민, 이두진 (2004). “배수지 운영을 위한 용수공급패턴분석 및 예측 모형.” **대한상하수도학회·한국물환경학회 학술발표회논문집**, 대한상하수도학회·한국물환경학회, pp. 297.
- 조용준, 김종문 (1998). “Wavelet Transform을 이용한 물수요량의 특성분석 및 다원 ARMA모형을 통한 물수요량 예측.” **한국수자원학회논문집**, 한국수자원학회, 제31권, 제3호, pp. 317-326.
- 진영훈, 박성천 (2006). “수문기상자료의 웨이블릿 변환에 의한 비선형 동역학적 성분의 추출.” **대한토목학회논문집**, 대한토목학회, 제26권, 제5호, pp. 439-446.
- 최기선, 유철, 지력민, 유성근, 전명근 (2009). “AR모형과 MLP를 이용한 단기 물 수요 예측 알고리즘 개발.” **한국지능시스템학회논문집**, 한국지능시스템학회, 제19권, 제5호, pp. 713-719.
- 최재호, 주충남, 구자용, 유명진 (1998). “도시의 지역특성과 장기 물 수요량 추정방법에 관한 연구.” **대한환경공학회 추계학술대회논문집**, 대한환경공학회, pp. 343-344.
- 황석환, 김중훈, 정성원 (2007). “SVM을 이용한 수문 시계열 자료의 예측.” **한국수자원학회 학술대회논문집**, 한

- 국수자원학회, pp. 1972-1976.
- Boser, B.E., Guyon, I., and Vapnik, V. (1992). "A training algorithm for optimal margin classifiers." *5th Annual ACM Workshop on Computational Learning Theory*, pp. 144-152.
- Cortes, C., and Vapnik, V. (1995). "Support vector networks." *Machine Learnings*, 20, pp. 273-297.
- Gato, S., Jayasuriya, N., and Raberts, P. (2007). "Forecasting Residential Water Demand: Case Study." *Journal of Water Resources Planning and Management*, Vol. 133, No. 4, pp. 309-319.
- Ghiassi, M., Zimbra, D.K., and Saidane, H. (2008). "Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model." *Journal of Water Resources Planning and Management*, Vol. 134, No. 2, pp. 138-146.
- Guyon, I., Boser, B., and Vapnik, V. (1993). "Automatic capacity tuning of very large VC-dimension classifiers." *Advances in Neural Information Processing Systems*, Vol. 5, pp. 147-155.
- Kwon, H.H., Lall, U., and Khalil, A.F. (2007). "Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: Applications to rainfall and temperature." *Water Resources Research*, Vol. 43, W05407.
- Legates, D.R., and McCabe, G.J. (1999). "Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation." *Water Resources Research*, Vol. 35, pp. 233-241.
- Schölkopf, B., Burges, C., and Vapnik, V. (1995). "Extracting support data for a given task." First International Conference on Knowledge Discovery & Data Mining, Menlo Park, AAAI Press.
- Tabesh, M., and Dini, M. (2009). "Fuzzy and Neuro-Fuzzy Models for Short-Term Water Demand forecasting in Tehran." *Iranian Journal of Science & Technology*, Vol. 33, No. B1, pp. 61-77.
- Vapnik, V. (1995). *The nature of statistical learning theory*. Springer, New York.
- Weng, H., and Lau, K.M. (1994). "Wavelets, period doubling, and time-frequency localization with application to organization of convection over the tropical western Pacific." *J. Atmos. Sci.*, Vol. 51, No. 17, pp. 2523-2541.
- Zhou, S.L., McMahon, T.A., Walton, A., and Lewis, J. (2000). "Forecasting daily urban water demand: a case study of Melbourne." *Journal of Hydrology*, Vol. 236, No. 3-4, pp. 153-164.
- Zhou, S.L., McMahon, T.A., Walton, A., and Lewis, J. (2002). "Forecasting operational demand for an urban water supply zone" *Journal of Hydrology*, Vol. 259, No. 1-4, pp. 189-202.

논문번호: 12-087	접수: 2012.08.02
수정일자: 2012.09.04/09.12	심사완료: 2012.09.12