다변량 적응 회귀 스플라인을 이용한 증발접시 증발량 모델링

Pan evaporation modeling using multivariate adaptive regression splines

서영민* · 김성원**

Youngmin Seo · Sungwon Kim

요 지

본 연구에서는 일 증발접시 증발량 모델링을 위한 다변량 적응 회귀 스플라인 (multivariate adaptive regression splines, MARS) 모델의 성능을 평가하였다. 모델 입력변수 집합은 부산 관측소 (기상청)로부터 수집된 기상자료를 활용하여 증발접시 증발량과의 상관성이 높은 변수들의 조합으로 구성되었으며, 일사량, 일조시간, 평균지상온도, 최대기온의 조합으로 구성된 세 가지 입력집합이 결정되었다. MARS 모델의 성능은 네 가지의 모델성능평가지표를 활용하여 정량적으로 산출되었으며, 그 결과를 인공신경망 (artificial neural network, ANN) 모델과 비교하였다. 입력변수로서 일사량 및 일조시간을 가지는 Set 1의 경우 MARS1 모델이 ANN1 모델보다 우수한 성능을 나타내었으며, Set 2 (일사량, 일조시간, 평균지상온도)의 경우 ANN2 모델, Set 3 (일사량, 일조시간, 평균지상온도, 최대기온)의 경우 MARS3 모델이 상대적으로 우수한 모델 성능을 나타내었다. 모든 분석 모델들을 비교하였을 때, MARS3, ANN2, ANN3, MARS2, MARS1, ANN1 모델의 순서로 우수한 모델 성능을 나타내었으며, 특히 MARS3 모델은 CE = 0.790, r^2 = 0.800, RMSE = 0.762, MAE = 0.587로서 가장 우수한 일 증발접시 증발량 모델링 성능을 나타내었다. 따라서 본 연구에서 적용한 MARS 모델은 지상관측 기상자료를 활용한 일 증발접시 증발량 모델링에서 효과적인 대안이 될 수 있을 것으로 판단된다.

핵심용어 : Multivariate adaptive regression splines, artificial neural network, meteorological input data, pan evaporation

1. 서 론

저수지 설계 및 관리, 용수 확보 및 공급, 가뭄 예측 및 관리 등을 포함하는 수문학적 설계 및 수자원 관리 측면에서 신뢰성 있는 증발접시 증발량 산정은 중요한 역할을 한다. 증발접시 증발량은 국내의 경우 주로 경험공식 또는 추계모델 등을 활용하여 산정되었으나, 최근 인공신경망 (artificial neural networks, ANNs), 퍼지규칙기반모델 (fuzzy rule-based models), 지지벡터머신 (support vector machines, SVMs) 등과 같은 다양한 자료기반모델 (data-driven models)의 적용을 통하여 과거에 비해 보다 우수한 성능을 가지는 증발접시 증발량 산정모델이 개발되고 있는 추세이다. 이러한 모델들 중에서 다변량 적응 회귀 스플라인 (multivariate adaptive regression splines, MARS) 모델은 변수들 간의 비선형성 및 상호작용을 자동으로 모델링하는 일종의 비매개변수적 회귀모델 (non-parametric regression model)로서 (Cheng and Cao, 2014; Kisi, 2015) 최근 증발접시 증발량 산정을 위해 그 적용이 증가하고 있다. 따라서 본 연구에서는 일 증발접시 증발량 산정을 위한 MARS 모델의 적용성을 모델성능평가지표에 근거하여 평가하였으며, 그 결과를 기존의 ANN 모델과 비교하였다.

2. 자료 및 방법

* 정회원 · 발표자 · 공학박사 · 경북대학교 과학기술대학 건설환경공학과 · E-mail: ymseo@knu.ac.kr

** 정회원 · 공학박사 · 수자원개발기술사 · 동양대학교 철도건설안전공학과 부교수 · E-mail: swkim1968@dyu.ac.kr

2.1 자료

본 연구에서는 일 증발접시 증발량 산정을 위한 MARS 및 ANN 모델을 구축하기 위하여 부산 관측소 (기상청)으로부터 2010~2015년 사이에 수집된 기상자료를 활용하였다. 여기서, 평균기온, 최소기온, 최대기온, 최대풍속, 평균풍속, 평균이슬점온도, 최소습도, 평균습도, 평균증기압, 최대해면기압, 평균해면기압, 일조시간, 일조량, 평균지상온도 및 증발접시 증발량에 대한 일자료가 수집되었다.

2.2 다변량 적응 회귀 스플라인 (Multivariate Adaptive Regression Splines, MARS)

MARS 모델은 선형모델을 확장한 일종의 자료기반모델로서 입력 및 출력변수들 간의 함수관계에 대한 어떠한 가정도 필요하지 않으며 (Friedman, 1991; Samui, 2012), 변수들 간의 비선형성 및 상호작용을 자동으로 모델링하는 비매개변수적 회귀모델 (non-parametric regression model)이다. MARS 모델에서 해공간은 예측변수들에 대하여 여러 구간으로 분할되고, 각 구간에 대해 스플라인 적합 (spline fitting)이 이루어진다. 각 스플라인 함수는 주어진 각 구간과 매듭점 (knot)라고 불리는 구간의 끝점들에 대하여 정의된다. MARS 모델은 기저함수 (basis function)의 가중합의 형태로 식 (1)과 같이 나타낼 수 있다 (Friedman, 1991).

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} a_m B_m(\mathbf{x}) \tag{1}$$

여기서, $\mathbf{x}$는 예측변수, $B_m(\mathbf{x})$는 기저함수, a_m는 계수이다. 각 기저함수는 상수 1, 경첩함수 (hinge function), 둘 이상의 경첩함수의 곱 중 하나를 취하게 된다. 경첩함수는 MARS 모델의 핵심부분으로서 $\max(0, x-c)$ 또는 $\max(0, c-x)$ (여기서, c는 매듭점)의 형태를 가진다. MARS 모델링은 순방향 및 후방향 단계로 구성되는 2단계 절차로 이루어진다. 순방향 단계에서는 매우 많은 매듭점들을 가지는 과적합모델 (over-fitted model)이 생성되고, 다음으로 후방향 단계에서는 불필요한 매듭점들을 제거하는 가지치기 (pruning) 절차가 진행된다 (Kisi, 2015).

2.3 인공신경망 (Artificial Neural Networks, ANNs)

ANN은 생물학적 신경망 시스템과 유사하게 뉴런 (neuron)이라고 불리는 다중노드를 가지는 여러 층으로 구성된 인공지능연산시스템이다. 다층퍼셉트론 (multilayer perceptron, MLP)은 은닉층 (hidden layer)라고 불리는 중간층을 가지는 순방향 ANN 모델 (feedforward ANN model)로서 분류 및 회귀문제와 같은 비선형 연산을 위해 널리 적용되고 있다. 일반적으로 MLP는 입력층, 은닉층 및 출력층과 같은 세 개의 층으로 구성되며, 식 (2)와 같이 나타낼 수 있다 (Lima et al., 2016).

$$\hat{\mathbf{y}}_j = \sum_{i=1}^{L} \boldsymbol{\beta}_i h(\mathbf{w}_i \mathbf{x}_j + b_i) + \boldsymbol{\beta}_0, \quad j = 1, 2, \cdots, N \tag{2}$$

여기서, $\mathbf{x}_j$는 입력벡터, $\hat{\mathbf{y}}_j$는 출력벡터, L은 은닉뉴런의 개수, h는 활성화함수 (activation function), $\mathbf{w}_i$는 은닉층의 가중치, $\boldsymbol{\beta}_i$는 출력층의 가중치, b_i는 은닉층의 편의 (bias), $\boldsymbol{\beta}_0$는 출력층의 편의, N은 자료크기이다. MLP의 매개변수는 역전파 알고리즘 (backpropagation algorithm)과 같은 학습 알고리즘을 이용하여 반복적으로 조정된다 (Alpaydin, 2010).

2.4 모델성능평가지표

본 연구에서 MARS 및 ANN 모델의 성능은 모델성능평가지표에 근거하여 정량적으로 평가되었으며, 평가지표로서 효율성계수 (coefficient of efficiency, CE), 결정계수 (coefficient of determination, r^2), 평균제곱근오차 (root-mean-square error, RMSE) 및 평균절대오차 (mean absolute error, MAE)가 적용되었다. 여기서, CE 및 r^2은 값이 클수록, RMSE 및 MAE는 값이 작을수록 모델성능이 우수함을 나타낸다. 모델성능평가지표에 대한 자세한 내용은 Dawson et al. (2007)을 참조할 수 있다.

3. 결과

MARS 및 ANN 모델을 이용한 일 증발접시 증발량을 모델링하기 위하여 부산 관측소 (기상청)에서 측정된 2010~2014년의 기상자료를 수집하였으며, 증발접시 증발량과 상관성이 높은 변수들을 조합하여 세 가지 입력집합 (Set 1~3)에 대한 각 모델들을 구축하였다. Table 1은 본 연구에서 적용된 입력집합을 나타낸다. 또한 모델구축 및 성능평가를 위하여 입력자료들은 학습자료 (2010~2014년)와 테스트 자료 (2015년)로 분할되었다.

Table 2와 Fig. 1은 각각 모델성능평가지표 산정결과와 산점도를 나타낸다. 입력변수로서 일사량 및 일조시간을 가지는 Set 1의 경우, MARS1 모델이 ANN1 모델보다 CE 및 r^2 값은 크고 RMSE 및 MAE는 작은 결과를 산출하였으며, 상대적으로 우수한 모델성능을 보여주었다. 일사량, 일조시간 및 평균지상온도를 입력변수로 가지는 Set 2의 경우, ANN2 모델이 MARS2 모델보다 우수한 모델성능을 나타내었으며, 입력변수로서 일사량, 일조시간, 평균지상온도 및 최대기온을 가지는 Set 3의 경우, MARS3 모델이 ANN3 모델보다 우수한 모델성능을 보여주었다. MARS 모델의 경우, Set 3가 가장 우수한 모델성능을 산출하는 최적입력집합인 반면, ANN 모델의 경우, Set 2가 최적입력집합인 것으로 분석되었다. 모든 분석모델들을 비교하였을 때, MARS3, ANN2, ANN3, MARS2, MARS1, ANN1 모델의 순서로 우수한 모델성능을 나타내었다. 이러한 결과로부터 MARS 및 ANN 모델의 성능은 입력변수의 구성에 큰 영향을 받으며, 최적입력변수 구성은 적용모델에 따라 다르게 나타남을 알 수 있다. 또한 MARS3 모델은 비교모델들 중에서 가장 우수한 모델성능을 나타내었으나 입력변수 구성에 따라서는 ANN 모델보다 모델성능이 떨어질 수도 있음을 알 수 있다. 따라서 MARS 모델을 이용한 증발접시 증발량 모델링의 경우 입력변수 구성을 최적화하는 것이 최적의 모델성능을 얻는데 중요한 역할을 함을 알 수 있다.

4. 결론

본 연구에서는 일 증발접시 증발량 산정을 위한 MARS 모델을 구축하고 ANN 모델과의 비교를 통해 MARS 모델의 적용성을 평가하였다. 본 연구에서는 지상관측 기상자료 (부산 관측소)의 수집, 상관성 분석을 통한 입력집합의 구축, MARS 및 ANN 모델의 구축, 모델성능평가지표의 산정 및 비교가 수행되었으며, 이를 통해 MARS 및 ANN 모델의 성능을 비교·평가하였다. 그 결과, MARS3, ANN2, ANN3, MARS2, MARS1, ANN1 모델의 순서로 우수한 모델성능을 나타내었으며, 특히 모든 비교모델 중에서 MARS3 모델은 가장 우수한 증발접시 증발량 모델링 성능을 나타내었다. 또한 MARS 및 ANN 모델의 성능은 입력변수의 구성에 큰 영향을 받으며, 최적입력변수 구성은 적용 모델에 따라 다르게 나타남을 알 수 있었다. 따라서 MARS 모델을 이용한 증발접시 증발량 모델링에 있어서 최적 입력변수 구성은 최적 모델성능을 얻는데 중요한 역할을 함을 알 수 있었으며, 입력변수로서 일사량, 일조시간, 평균지상온도 및 최대기온를 가지는 MARS 모델은 일 증발접시 증발량을 모델링하는데 효과적인 대안으로 적용될 수 있을 것으로 판단된다.

참 고 문 헌

1. Alpaydin E (2010). Introduction to machine learning. 2nd edn, The MIT Press, Cambridge, MA, USA.
2. Cheng MY, Cao MT (2014). Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. Applied Soft Computing, 22: 178-188.
3. Dawson CW, Abrahart RJ, See LM (2007). HydroTest: A web-based toolbox of evaluation metrics for the standardized assessment of hydrological forecasts. Environmental Modelling and Software, 22(7): 1034-1052.
4. Friedman JH (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19(1): 1-67.
5. Kisi O (2015). Pan evaporation modeling using least squares support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology, 528: 312-320.
6. Lima AR, Cannon AJ, Hsieh WW (2016). Forecasting daily streamflow using online sequential extreme learning machines. Journal of Hydrology, 537: 431-443.
7. Samui P (2012). Determination of ultimate capacity of driven piles in cohesionless soil: a multivariate

adaptive regression spline approach. International Journal for Numerical and Analytical Methods in Geomechanics, 36(11): 1434-1439.

8. Seo Y, Kim S (2017). Pan evaporation modeling using deep learning theory. Proceedings of KWRA Annual Conference, May 25-26, 2017, Changwon, Republic of Korea, pp. 392-395.

Table 1. Combination of input and output variables (Seo et al., 2017)

Input sets	Input variables	Output variables
Set 1	R_s, D	E_p
Set 2	R_s, D, GT_{mean}	E_p
Set 3	R_s, D, GT_{mean}, T_{max}	E_p

R_s: solar radiation, D: sunshine duration, GT_{mean}: mean ground temperature, T_{max}: maximum air temperature, E_p: evaporation

Table 2. Evaluation of model performance

Input sets	Models	CE	r^2	RMSE (mm)	MAE (mm)
Set 1	MARS1	0.779	0.781	0.783	0.613
	ANN1	0.774	0.778	0.791	0.619
Set 2	MARS2	0.786	0.796	0.770	0.595
	ANN2	0.788	0.794	0.767	0.585
Set 3	MARS3	0.790	0.800	0.762	0.587
	ANN3	0.786	0.793	0.770	0.590

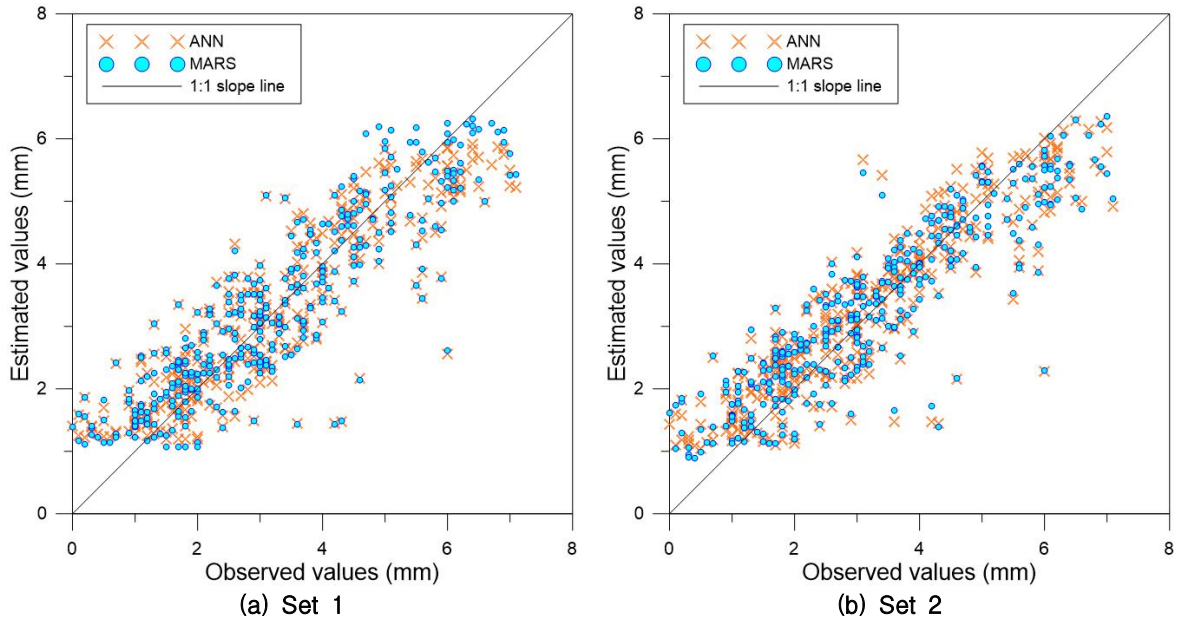

Fig. 1. Scatter plots for observed and estimated pan evaporation values