머신러닝 기법을 이용한 수도권 지역의 호우피해 예측함수 개발

Development of Heavy Rain Damage Prediction Functions in the Seoul Capital Area Using Machine Learning Techniques

최창현* · 김종성** · 김동현*** · 이준형**** · 김덕환***** · 김형수*****

Choi, Changhyun*, Kim, Jongsung**, Kim, Donghyun***, Lee, Junhyeong****, Kim, Deokhwan*****,
and Kim, Hung Soo*****

Abstract

In this study, we developed heavy rain damage prediction functions using three machine learning techniques (support vector machine, decision tree, and random forest) for the Seoul Capital Area, South Korea. Data on damage caused by heavy rain were used as the dependent variable for the development of the heavy rain damage prediction function, and weather observation data were used as the independent variables. When we compared the results, the best function was the support vector machines based on weather observation data of the past two days. Compared to the linear regression model used primarily in previous studies, the results showed that the functions using machine learning techniques were mostly predictable. Therefore, it was judged that the machine learning techniques could be applied to disaster management areas. Also, it is believed that using the heavy rain damage prediction function developed in this study can help reduce damage through proper disaster management before the damage occurs.

Key words : Disaster Management, Heavy Rain Damage, Machine Learning, Random Forest, Support Vector Machine

요 지

본 연구에서는 3가지의 머신러닝 기법(서포트 벡터 머신, 의사결정나무, 랜덤포레스트)을 이용하여 수도권 지역의 호우피해 예측함수를 개발하였다. 호우피해 예측함수의 종속변수로 호우피해액 자료를 사용하였고, 독립변수로 기상관측자료를 사용하였다. 분석결과 과거 2일전의 기상관측자료를 기반으로 서포트 벡터 머신을 이용하여 개발한 함수가 가장 높은 예측력을 보였다. 기존의 연구들에서 주로 사용하였던 선형회귀모형과 비교한 결과 머신러닝 기법을 이용한 함수가 대부분 예측력이 높은 것으로 나타나 재난관리 분야에 머신러닝 기법의 적용이 가능한 것으로 판단되었다. 또한, 본 연구에서 개발된 호우피해 예측함수를 활용하여 피해 발생 전에 호우피해를 예측한다면, 적절한 재난관리를 통해 피해를 저감하는데 도움이 될 수 있을 것으로 판단된다.

핵심용어 : 재난관리, 호우피해, 머신러닝, 랜덤포레스트, 서포트 벡터 머신

*정회원, 인하대학교 토목공학과 박사수료(E-mail: karesma0cch@naver.com)

Member, Ph.D. Candidate, Department of Civil Engineering, Inha University

**정회원, 인하대학교 토목공학과 박사과정

Member, Doctor's Course, Department of Civil Engineering, Inha University

***정회원, 인하대학교 토목공학과 석사과정

Member, Master's Course, Department of Civil Engineering, Inha University

****정회원, 인하대학교 토목공학과 석사과정

Member, Master's Course, Department of Civil Engineering, Inha University

*****정회원, 인하대학교 수자원시스템 연구소 박사후연구원

Member, Post-Doctoral Researcher, Institute of Water Resources System, Inha University

*****교신저자, 정회원, 인하대학교 사회인프라공학과 교수(Tel: +82-32-874-0069, Fax: +82-32-876-9787, E-mail: sookim@inha.ac.kr)

Corresponding Author, Member, Professor, Department of Civil Engineering, Inha University

1. 서론

전 지구적인 지구온난화로 인해 기후변화가 심화됨에 따라 이상기후의 영향으로 인한 자연재난의 강도가 증가하고 있다(Han et al., 2016). 국내의 경우 지리적·기후적 특성으로 장마와 태풍의 규모가 증가하고 있으며, 최근에는 2017년 7월 청주를 비롯한 중부지역에서 발생한 기록적인 폭우로 인하여 약 547억 원의 재산피해와 2,539명의 인명피해가 발생하였다(Song et al., 2006; Bae, 2018). 자연재난으로 인한 피해를 최소화하기 위해서는 구조적 대책(배수펌프 및 제방 설치 등)과 비구조적 대책(홍수 예·경보 및 비상대처계획 공지 등) 수립이 필수적이다. 그러나 대규모 구조물의 건설 등과 같은 구조적 대책은 환경단체와 주민들의 반발로 인해 대책 실행에 어려움이 있으며, 기후변화와 급격한 도시화로 인한 자연재난의 규모와 강도가 증가함에 따라 구조적 대책의 설계빈도를 초과하는 재난의 발생 등 여러 가지 한계점을 가지고 있다(Kang et al., 2007; Kim et al., 2007). 이러한 구조적 대책의 한계점을 보완하기 위해 사전에 피해를 예측하여, 피해 발생 전에 대비 단계의 재난관리를 실시하는 비구조적 대책에 관한 연구가 활발하게 진행되고 있다. 대표적인 사례로 피해의 직접적인 원인으로 판단되는 강우량(Davis and Skaggs, 1992; Lee, Eo, et al., 2016; Choi, Kim, Lee, et al., 2017; Kim, Choi, Kim, et al., 2017)과 최대풍속(Munich Re., 2002; Murnane and Elsner, 2012; Prael et al., 2012; Zhai and Jiang, 2014) 등을 사용하여 피해를 사전에 예측할 수 있는 선형회귀모형을 개발하는 연구가 있다.

최근에는 컴퓨팅 기술과 데이터 처리 속도의 비약적인 발전으로 인해 선형회귀모형에 비해 예측력이 높다고 알려진 머신러닝 기법을 재난관리 분야에 적용하는 연구가 등장하고 있다(Sakr et al., 2010; Asim et al., 2017; Choi et al., 2018). 그러나 머신러닝을 이용하여 자연재난으로 인한 피해를 예측하는 사례는 미비하고, 주로 기상 자료를 예측하는

연구가 진행되고 있다(Radhika and Shashi, 2009; Sharma et al., 2011; Kusiak et al., 2012; Rasouli et al., 2012; Kalra et al., 2013).

선행연구들을 살펴보면 주로 피해의 직접적인 원인으로 판단되는 소수의 기상관측자료만을 기반으로 선형회귀모형을 이용한 피해예측함수를 개발하였고, 다양한 분야에서 좋은 성능을 보이고 있는 머신러닝 기법을 적용한 연구는 미비한 것으로 나타났다. 또한 대부분의 연구에서 해당 시점의 피해(종속변수)를 예측하기 위한 독립변수로 해당 시점의 기상관측자료를 사용하였다. 그러나 실제로 피해를 예측하기 위해 필요한 해당 시점의 기상관측자료를 해당 시점이 지나기 전에는 얻을 수 없는 시간적인 한계 때문에 발생하는 적용의 어려움이 있었다.

따라서 본 연구에서는 피해의 직접적인 원인으로 판단되는 소수의 기상관측자료 이외에 다양한 기상관측자료를 고려하였고, 머신러닝 기법을 적용하여 호우피해 예측함수를 개발함으로써 예측력을 극대화 하고자 하였다. 또한 과거의 기상관측자료를 활용하여 당일의 호우피해를 예측함으로써 기존 연구의 한계점을 극복하고자 하였다.

본 연구의 절차를 간략히 설명하면 다음과 같다. 대상지역을 선정하고, 행정안전부에서 발간하는 재해연보의 지역별/이벤트별 호우피해액 자료를 수집하여 종속변수로 구축하였다. 기상청에서 제공하는 기상관측자료를 각각 1일전, 1-2일전, ..., 7일전 형태로 구축하여 독립변수로 사용하였다. 선행연구들에서 주로 사용되었던 선형회귀모형과 본 연구에서 제시한 3가지의 머신러닝 기법(의사결정나무, 랜덤포레스트, 서포트 벡터 머신)을 이용하여 호우피해 예측함수를 개발하였고, 개발된 함수를 평가용 데이터에 적용하여 예측력을 평가하였다. 마지막으로, 예측력 평가를 통해 예측력이 가장 높은 함수를 수도권 지역의 최종 호우피해 예측함수로 선정하였다. Fig. 1은 연구 흐름도를 도식화한 것이다.

Fig. 1. Flow Chart

2. 선형회귀모형과 머신러닝 기법

2.1 선형회귀모형

선형회귀분석(Linear regression analysis)은 종속변수 y_i 에 유의한 영향을 미칠 것으로 기대되는 독립변수($x_1, x_2, \dots, x_i$)를 고려하여 종속변수를 통계적 방법에 의해 추정하는 방법을 의미하며, 선형회귀모형(Linear regression model)은 선형회귀분석을 통해 종속변수와 독립변수간의 관계를 Eq. (1)과 같이 선형식으로 나타낸 모형이다(Kim, Hong, et al., 2015; Choi, Kim, Kim, et al., 2017).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_i x_i + \epsilon_i \quad (1)$$

여기서, β_0 은 상수항, $\beta_1 \dots \beta_i$ 은 독립변수의 회귀계수, ϵ_i 는 종속변수를 추정할 때 발생하는 오차를 말한다(Kim, Choi, Lee, et al., 2017). 여러 독립변수를 이용하여 타당성 있고 합리적인 회귀식을 유도하기 위해서는 전진선택법(Forward selection method), 후진제거법(Backward elimination method), 단계선택법(Stepwise selection method) 등의 변수선택법을 고려할 수 있다(Kim et al., 2016). 본 연구에서는 독립변수가 회귀식에 존재할 때 종속변수에 영향력이 있는 독립변수들만을 회귀식에 포함시키는 단계선택법을 사용하였고, 다중공선성(multicollinearity) 문제를 해결하기 위해 주성분 분석(Principal Component Analysis, PCA)을 이용하였다(Lee and Lim, 2005; Kim, Choi, Kim, et al., 2017).

2.2 머신러닝 기법

머신러닝(기계학습)은 인공지능의 한 분야로서 기계가 주어진 데이터로부터 자동으로 학습할 수 있도록 하는 수리적 모델과 알고리즘을 연구하고 개발하는 학문분야이다(Kim, Hwang, et al., 2015). 예를 들어 (x, y)의 집합에서 (1, 7), (2, 14), (3, 21), (5, 35)의 자료가 있다고 할 때, x와 y에 관한 함수를 기계가 알지 못하더라도, 앞의 자료들을 입력하여 기계가 학습한 후에 (7, ?), (10, ?) 등의 질문을 던지면 y에 대한 답을 기계가 할 수 있도록 만드는 것이 머신러닝이다(Choi et al., 2018). 본 연구에서는 대표적인 머신러닝 기법인 의사결정나무, 랜덤포레스트, 서포트 벡터 머신을 사용하였다.

2.2.1 의사결정나무

의사결정나무(Decision tree) 모형은 변수들의 규칙 혹은 조건문을 토대로 자료를 나무(tree) 형태의 그래프로 표현하며, 분리규칙에 의해 서로 유사한 데이터들로 세분화하고 최종 분류 기준을 만족할 때까지 분류하여 전체 나무를 구성한다(Breiman et al., 1984; Bae, 2014). 의사결정나무는 반복 이분화(binary recursive partitioning) 과정을 통해 각

단계에서 평균제곱오차(Mean Squared Error, MSE)를 최소화하는 분기변수 및 분기점을 찾아 데이터를 이분화하는 과정이 반복적으로 실행되면서 전체 나무의 모습을 이루게 되며, 교차검증(cross validation)을 이용하여 평균제곱오차를 가장 최소화하는 나무의 크기를 결정하도록 가지치기(pruning)를 진행하여 과적합(overfitting)을 방지한다(Choi, Park, et al., 2017). 본 연구에서는 R studio의 “rpart” 라이브러리를 사용하여 의사결정나무를 이용한 호우피해 예측함수를 개발하였다.

2.2.2 랜덤포레스트

랜덤포레스트(Random forest)는 말 그대로 (의사결정)나무들이 많이 있는 모형으로, 의사결정나무 모형을 다수 만들어 예측력을 높이는 방법이다(Yoo, 2015). 의사결정나무는 하나의 데이터 집합(data set)에서 한 번의 훈련용 데이터(training data)를 생성하고 학습하여 하나의 의사결정나무 모형으로 예측하지만, 랜덤포레스트는 하나의 데이터 집합에서 여러 개의 훈련용 데이터를 만들어 여러 번의 학습을 통해 여러 개의 의사결정나무를 생성하여 이들을 결합한 결과를 통해 예측력을 향상시키는 특징이 있다(Choi, 2017). 또한 랜덤포레스트는 종속변수가 연속형인 경우 독립변수의 개수가 m개이면, 각 분할에서 랜덤으로 m/3개의 변수를 선택하여 나무(tree)를 구성한다(Breiman, 2001). 본 연구에서는 R studio의 “randomForest” 라이브러리를 사용하여 랜덤포레스트를 이용한 호우피해 예측함수를 개발하였다.

2.2.3 서포트 벡터 머신

서포트 벡터 머신(Support Vector Machine, SVM)은 Vapnik (1995)에 의해 제시된 이후 문서분류, 고객분류 등 여러 가지 문제에서 분류 결과가 매우 정확하고, 다양한 형태의 자료에 적용이 가능한 방법으로 알려져 있다(Choi et al., 2013). SVM은 선형으로 서로 다른 클래스의 벡터들을 그 사이의 거리에 대해 최대의 마진(margin)으로 분류할 수 있는 서포트 벡터들로 이루어진 초평면을 찾는 방법이다(Lee, Chung, et al., 2016). 선형으로 분류 불가능한 데이터에 대해서는 커널함수(kernel function)를 이용하여 평면에서 고차원 공간에 사상시킨 뒤 분류하는데, 대표적인 커널함수의 종류로는 Polynomial, Sigmoid, Radial Basis Function(RBF) 등이 있다(J.S. Park, 2006; Kim et al., 2012). SVM은 주로 분류(classification) 문제의 예측에 사용되며, SVM에 ϵ -무감도 손실함수(ϵ -insensitive loss function)를 도입하여 회귀(regression) 분석에 사용할 수 있도록 확장한 방법을 Support Vector Regression (SVR)이라 한다(Kim et al., 2012). 즉, SVM은 분류 문제에서 데이터를 ‘+1’ 클래스와 ‘-1’ 클래스로 구분하는 데 사용되지만, 임의의 실수값을 예측할 수 있도록 SVM을 일반화한 방법이 SVR이다(C.K. Park, 2006). 따라서, SVR을 이용하여 호우로 인한 피해액을 예측하였고,

커널함수로 비교적 성능이 뛰어나다고 알려진 RBF를 적용하였다(Tay and Cao, 2001; C.K. Park, 2006; Park and Kim, 2007). 본 연구에서는 R studio의 “e1071” 라이브러리를 사용하여 서포트 벡터 머신을 이용한 호우피해 예측함수를 개발하였다.

2.3 예측력 평가 방법

본 연구에서 제시한 호우피해 예측함수들의 예측력을 평가하는 과정은 다음과 같다.

- ① 1994~2011년 데이터를 학습용 데이터(training set)로, 2012~2016년 데이터를 평가용 데이터(test set)로 구분함
- ② 학습용 데이터(전체 자료의 약 70%)로 호우피해 예측 함수를 개발 함
- ③ 개발된 호우피해 예측함수를 평가용 데이터(전체 자료의 약 30%)에 적용하여 계산한 예측값과 평가용 데이터의 실제값을 비교하여 예측력을 평가함

본 연구에서는 예측력 평가의 지표로 root mean squared error (RMSE)를 표준화한 normalized-root mean squared error (NRMSE)를 사용하였다. NRMSE는 분자인 RMSE를 분모인 실제값의 범위(최대값 - 최소값)로 나눈 값으로, 0에 가까울수록 오차의 정도가 작음을 의미한다. NRMSE를 식으로 나타내면 Eq. (2)와 같다.

$$NRMSE(\%) = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\text{Max}(y_i) - \text{Min}(y_i)} \times 100 \quad (2)$$

여기서, y_i 는 실제값, $\hat{y}_i$ 는 예측값을 나타내고, $\text{Max}(y_i)$ 와 $\text{Min}(y_i)$ 는 실제값의 최대값과 최소값을 나타낸다.

3. 대상지역 선정 및 자료 구축

3.1 대상지역 선정

행정안전부에서 발간한 재해연보를 통해 1994년부터

2016년까지의 호우피해 발생현황을 조사하였고, 이를 기상청의 예보지역을 기준으로 구분하여 Fig. 2와 같이 지역별 호우피해 발생 횟수 및 피해액을 표현하였다. Fig. 2에서 나타난 바와 같이 수도권(서울, 인천, 경기도) 지역이 호우로 인한 피해가 가장 빈번하게 발생하였고, 강원도 영서 지역에 이어 두 번째로 많은 피해를 입은 것으로 나타났다. 따라서 호우피해 발생 횟수가 가장 많고, 피해액도 높은 수도권 지역을 본 연구의 대상지역으로 선정하였다.

3.2 종속변수 구성

수도권 지역의 호우로 인한 피해를 예측하는 함수를 개발하기 위해 종속변수(dependent variable)로 재해연보의 1994년부터 2016년까지의 시군구별/이벤트별 호우피해액 자료를 활용하였다. 재해연보에는 태풍으로 인한 피해도 집계되어 있기 때문에 태풍으로 인한 호우피해도 고려할 수 있지만, 태풍을 동반한 호우피해와 집중호우로 인한 호우피해는 다른 특징을 가지고 있는 것으로 판단하였기 때문에 본 연구에서는 집중호우로 인한 호우피해만을 대상으로 하였다. 재해연보에 집계된 피해액의 경우 동일한 금액이라도 과거의 화폐가치와 현재의 화폐가치가 다르기 때문에 과거의 금액을 현재의 가치로 환산해야 한다(예를 들어 1994년도의 100만원의 가치와 2016년의 100만원 가치는 차이가 있음). 재해연보에서는 과거의 화폐가치를 현재의 화폐가치로 환산하기 위해 생산자 물가지수를 이용하여 2016년 기준으로 금액 환산지수를 산정하고, 이를 과거의 금액에 곱하여 현재(2016년)의 가치로 환산한다. 본 연구에서도 생산자 물가지수를 이용하여 과거의 화폐가치를 현재의 화폐가치로 환산하였다. 종속변수인 호우피해액을 그대로 사용하면 자료의 변동성이 크기 때문에 적절한 호우피해 예측함수를 개발하기 어렵다. 따라서 본 연구에서는 호우피해액에 자연로그를 취하여 대수변환을 통해 변동성을 줄여주고 함수를 개발한 후에, 자연로그단위로 개발된 함수식에서 나온 결과를 역로그하여 실제 피해액을 도출하였다.

Fig. 2. Incidence of Heavy Rain Damage and Total Heavy Rain Damage

3.3 독립변수 구성

호우피해 예측함수의 독립변수(independent variable)로 ‘기상자료개방포털’에서 제공하는 기상청 종관기상관측장비(Automated Synoptic Observing System, ASOS)의 일별 기상관측자료를 사용하였다. ‘기상자료개방포털’에서는 총 57가지의 일별 기상관측자료를 제공하는데 이 중에서 호우 피해와 관련 없는 일 최심신적설, 일최심적설량 등의 요소들과 결측이 많아서 사용이 불가능한 요소들을 제외하고, Table 1과 같이 총 27가지의 일별 기상관측자료를 독립변수로 사용하였다.

3.2절에서 구성한 독립변수인 호우피해액 자료는 시군구 별로 구성되어 있고, 독립변수인 기상관측자료는 기상관측소의 지점별로 구성되어 있기 때문에 독립변수를 종속변수의 공간적 범위로 변환시킬 필요가 있다. 따라서 수문분야에서 주로 사용되는 티센다각형법을 적용하여 지점별 기상관측자료를 시군구별 자료로 변환하였다(Yoon, 2007; Lee, 2017; Kim, Choi, Kim, et al., 2017; Choi et al., 2018).

기상관측자료를 사용하는 시점에 대해 크게 두 가지 경우(당일의 기상관측자료 사용, 과거의 기상관측자료 사용)를 생각해 볼 수 있다. 먼저 당일의 기상관측자료와 당일에 발생하는 호우피해가 연관성이 높기 때문에, 예측하고자 하는 당일의 기상관측자료를 독립변수로 사용하여 함수를 개발한다면 호우피해를 예측하는 성능은 뛰어날 것이다. 가장 이상적인 방법이지만, 시간적 한계(8/27 23시에 8/28의 피해를 사전에 예측하기 위해서는 8/28 00시~23시의 기상자료가 8/27 23시에 필요하지만 관측이 불가능함)가 있기 때문에 당일의 기상관측자료를 사용하는 것은 불가능하다고 판단된다. 기상관측자료 대신 기상예측자료를 사용하는 방안을 고려할 수 있지만, 아직까지 소수의 기상요소만을 예측하고 있고 자료의 불확실성이 높기 때문에 피해를 예측하는데

어려움이 있다고 판단하였다. 특히 최근 발생하는 집중호우의 경우 좁은 지역에 짧은 지속시간에 내리는 경우가 많기 때문에 이를 예측하는 것은 쉽지 않은 것으로 분석되고 있다(Song et al., 2016). 특히 실제 호우피해의 경우 피해 발생 시점 전후의 고해상도(시 혹은 10분 단위)의 강수자료가 호우피해액과 상관성이 높다고 볼 수 있지만, 현재까지의 과학기술로는 고해상도의 신뢰성 있는 예측자료를 획득하는 것이 어렵기 때문에 과거의 일별 기상관측자료를 이용하는 방법을 고려하였다. 과거 기상관측자료는 비교적 신뢰도가 높으며 활용 가능한 기상관측자료의 종류가 풍부하기 때문에, 과거의 기상관측자료를 사용하여 호우피해 예측함수를 개발하는 것이 타당하다고 판단하였다. 따라서 본 연구에서는 당일의 기상관측자료 대신 과거 1~7일별 기상관측자료를 사용하여 당일의 호우피해를 예측하는 함수를 개발하였다. 피해 발생 전의 일주일치 기상자료를 수집하여 Fig. 3과 같이 각각 1일전의 기상자료(1), 1일전과 2일전의 기상자료(1-2), 1일전, 2일전, 3일전의 기상자료(1-3) ··· 6일전과 7일전의 기상자료(6-7), 7일전 기상자료(7)로 총 28가지의 기상관측자료 형태를 이용하였다.

4. 호우피해 예측함수 개발 및 예측력 평가

4.1 호우피해 예측함수 개발

본 연구에서는 호우로 인한 피해를 사전에 예측하기 위해 기상관측자료를 이용하여 해당 지역의 호우피해액을 예측하는 함수를 호우피해 예측함수라고 정의하였고, 기존의 선행연구들에서 주로 사용하였던 선형회귀모형과 본 연구에서 제시한 3가지의 머신러닝 기법을 이용하여 총 4가지 호우피해 예측함수를 개발하였다.

Table 1. List of Independent Variables

Category	Variables (unit)	Category	Variables (unit)
Precipitation	Precipitation duration (hr)	Wind	Maximum instantaneous wind speed (m/s)
	10 minutes maximum precipitation (mm)		Maximum wind speed (m/s)
	1 hour maximum precipitation (mm)		Average wind speed (m/s)
	Daily precipitation (mm)		Wind match (100m)
Humidity	Average dew point temperature (°C)	Atmospheric pressure	Average local pressure (hPa)
	Minimum relative humidity (%)		Maximum sea level pressure (hPa)
	Average relative humidity (%)		Minimum sea level pressure (hPa)
	Average vapor pressure (hPa)		Average sea level pressure (hPa)
Temperatures	Average temperature (°C)	Insolation and insolation duration	Possible duration of sunshine (hr)
	Minimum temperature (°C)		Duration of sunshine (hr)
	Maximum temperature (°C)	Cloud	Average total cloud amount (1/10)
Evaporation	Large evaporation (mm)		Average middle and lower layers cloud amount (1/10)
	Small evaporation (mm)	Fog	Fog duration time (hr)
	9-9 Precipitation (mm)		

Fig. 3. Type of Independent Variables

4.1.1 선형회귀모형

선형회귀모형은 독립변수들이 가지고 있는 높은 상관성으로 인해 다중공선성 문제가 발생하는데, 이를 해결하기 위해 주성분 분석을 이용하여 소수의 주성분으로 축소시켰다. 일반적으로 주성분의 개수를 결정하기 위해 누적분산을 이용하는데, 본 연구에서는 누적분산이 90% 이상일 때의 주성분만을 독립변수로 사용하였다. 또한 예측력 높은 호우 피해 예측함수를 개발하기 위해 변수선택법으로 종속변수와 유의한 연관성을 갖는 독립변수만을 선택하는 단계선택 방법을 이용하였다.

4.1.2 의사결정나무

의사결정나무는 평균제곱오차를 최소화하는 분기점을 찾아 데이터를 이분화하는 과정을 반복적으로 실행하며 나무를 생성한다. 데이터를 이분화하는 과정이 반복될수록 과적합 될 가능성이 높아지기 때문에, 교차검증을 이용하여 오차를 최소화하는 나무의 크기를 결정해야 한다. 본 연구에서는 10겹 교차검증(10-fold cross validation)을 통해 나무의 크기를 표현하는 비용 복잡도 인자(complexity parameter, CP)별로 평균 오차를 계산하였다. 가장 낮은 평균 오차가 나오는 CP값을 찾아서 가지치기를 진행하였고, 이를 통해 최적의 의사결정나무 모형을 결정하였다.

4.1.3 랜덤포레스트

랜덤포레스트는 종속변수가 연속형인 경우 독립변수의 개수가 m 개이면, 각 분할에서 랜덤으로 $m/3$ 개의 변수를 선택하여 나무(tree)를 구성하고, 일반적으로 500개의 각기 다른 나무를 생성하여 결과를 도출한다. 그러나 나무의 개수가 너무 많아지면 과적합 될 위험이 높기 때문에, 평균제곱오차를 최소화하는 크기의 나무 개수를 찾아서 최적의 랜덤포레스트 모형을 결정하였다.

4.1.4 서포트 벡터 머신

서포트 벡터 머신의 경우 회귀(regression) 문제인 임의의 호우 피해액을 예측할 수 있도록 Support Vector Regression (SVR)로 변환하여 사용하였고, 커널함수는 회귀문제에 비교적 정확도가 높다고 알려진 Radial Basis Function (RBF)를 사용하였다(실제로 본 연구에서도 Polynomial, Sigmoid를 모두 이용하여 예측력을 평가하였지만, RBF에 비해 성능이 떨어지기 때문에 RBF에 대한 결과만 제시하였음). 서포트 벡터 머신은 회귀함수의 일반화를 결정짓는 모수인 $cost$ 와 ϵ -무감도 손실함수의 ϵ 값을 조정하여 최적의 모수를 선택하고 이를 반영한 함수를 개발한다(Kim et al., 2012). 본 연구에서는 기존 연구들을 참고하여 10겹 교차검증을 통해 $cost$ 의 범위를 $2^{(-7)}$, ϵ 의 범위를 0~1까지 0.1 간격으로 조정하면서 최적의 모수를 선택하였다(Tay and Cao, 2001; Park and Kim, 2007; Kim and Ahn, 2010; Lesmeister, 2017).

4.2 예측력 평가

4가지 방법론(선형회귀모형, 의사결정나무, 랜덤포레스트, 서포트 벡터 머신)과 28가지의 기상관측자료 형태를 이용하여 총 112종의 호우피해 예측함수를 개발하였다. 2012년부터 2016년까지의 실제 피해액과 모형별로 예측된 피해액을 비교하여 예측력을 평가한 결과는 Table 2와 같다.

4.2.1 선형회귀모형 결과

Table 2에 제시한 선형회귀모형의 결과를 살펴보면 피해 발생 4일전부터 6일전까지의 기상관측자료를 사용하여 개발한 함수가 가장 좋은 예측력을 나타냈다. 피해 발생 4일전부터 6일전까지의 기상관측자료를 기반으로 선형회귀모형을 이용하여 개발한 함수의 경우 총 81개(1일당 27개×3일)의 독립변수가 존재하는데 이렇게 많은 독립변수를 모두 사용하여 함수를 개발할 경우 일반적으로 다중공선성에 문제가

Table 2. Evaluation of Prediction Performance by Model

(Unit: NRMSE (%))

Time	Linear regression model	Decision tree	Random forest	Support Vector Machine
1 day ago	8.439%	8.726%	8.556%	8.499%
1-2 days ago	8.502%	8.829%	8.496%	8.249%
1-3 days ago	8.432%	17.891%	8.366%	8.509%
1-4 days ago	8.446%	22.123%	8.426%	8.461%
1-5 days ago	8.426%	9.726%	8.487%	8.548%
1-6 days ago	8.400%	10.171%	8.478%	8.616%
1-7 days ago	8.641%	8.252%	8.513%	8.663%
2 days ago	8.524%	8.592%	8.272%	7.469%
2-3 days ago	8.517%	9.126%	8.292%	8.773%
2-4 days ago	8.445%	8.770%	8.417%	8.636%
2-5 days ago	8.389%	9.016%	8.490%	8.660%
2-6 days ago	8.461%	9.972%	8.486%	8.731%
2-7 days ago	8.537%	8.252%	8.497%	8.708%
3 days ago	8.389%	17.902%	8.266%	8.396%
3-4 days ago	8.328%	8.702%	8.400%	8.659%
3-5 days ago	8.315%	9.164%	8.480%	8.637%
3-6 days ago	8.352%	9.160%	8.494%	8.454%
3-7 days ago	8.520%	10.431%	8.501%	8.774%
4 days ago	8.296%	8.514%	8.532%	8.649%
4-5 days ago	8.339%	8.962%	8.546%	8.556%
4-6 days ago	8.265%	11.644%	8.554%	8.613%
4-7 days ago	8.559%	10.491%	8.535%	8.632%
5 days ago	8.468%	14.110%	8.611%	8.615%
5-6 days ago	8.362%	9.396%	8.582%	8.683%
5-7 days ago	8.607%	14.244%	8.641%	8.624%
6 days ago	8.423%	8.860%	8.556%	8.000%
6-7 days ago	8.593%	7.984%	8.619%	8.667%
7 days ago	8.545%	8.556%	8.552%	8.664%

발생한다. 이를 해결하기 위해 주성분 분석을 이용하여 누적분산이 90% 이상을 차지하는 총 14개의 주성분을 추출하였고, 이를 단계선택법을 이용하여 총 10개의 주성분을 선택하여 함수를 결정하였다. 다중공선성 문제를 판별할 때 분산팽창요인(Variance Inflation Factors, VIF)이 10 이상인 경우 다중공선성이 높다고 판별하는데, 서로 독립인 주성분만을 이용하여 함수를 개발하였기 때문에 VIF값은 모두 1인 것으로 나타나 다중공선성 문제는 없는 것으로 판별되었다.

4.2.2 의사결정나무 결과

의사결정나무의 경우 NRMSE가 7.98%~22.12%로 결과의 변동성이 큰 편인 것으로 나타났고, 피해 발생 6일전부터 7일전의 기상관측자료를 이용하여 개발한 함수가 가장 좋은 예측력을 나타냈다. 1일전부터 7일전과 2일전부터

7일전까지의 기상관측자료를 이용하여 개발한 함수에서도 좋은 예측력을 나타냈고, 1일전부터 4일전까지의 기상관측자료를 이용하여 개발한 함수에서 가장 낮은 예측력을 나타냈다. 가장 좋은 예측력을 나타낸 피해 발생 6일전부터 7일전까지의 기상관측자료를 사용하여 함수를 개발하는 과정을 살펴보면, Table 3과 Fig. 4에서 제시한 바와 같이 10겹 교차검증을 통해 나무의 크기를 표현하는 비용 복잡도 인자(CP)별로 평균 오차를 계산하였다. 가장 낮은 오차(error)가 나오는 CP값을 찾아서 나무의 크기를 결정하고, 가지치기(pruning)를 실시하여 최적의 함수를 개발하였다. Fig. 5의 (a)는 가지치기 전의 의사결정나무, (b)는 가지치기 후의 최종 의사결정나무를 나타낸다. 가지치기 후의 최종 의사결정나무를 살펴보면 7일전의 평균기온과 강수 계속 시간, 6일전의 풍정합이 호우피해를 예측하는데 가장 중요한 요소로 판별되었다.

Table 3. Error by Size of Tree (Decision tree)

Size of tree	CP	Error	Standard deviation
1	0.086307	1.002517	0.046799
2	0.081955	0.957296	0.046638
3	0.039258	0.844123	0.040948
4	0.037971	0.835601	0.041834
5	0.031135	0.800892	0.041437
6	0.024597	0.78714	0.04056
7	0.022127	0.782182	0.040339
8	0.021651	0.796914	0.041044
9	0.021466	0.786018	0.040552
10	0.013071	0.774175	0.041045
11	0.012809	0.784677	0.042327
12	0.011822	0.785269	0.041702
13	0.011419	0.777074	0.041582
14	0.010905	0.773166	0.041287
15	0.010245	0.772005	0.041586
16	0.010075	0.776087	0.041289
17	0.01	0.775189	0.041247

Fig. 4. Error by Size of Tree (Decision tree)

(a) Decision tree before pruning

(b) Decision tree after pruning

Fig. 5. Decision Tree by Pruning (Before and after)

4.2.3 랜덤포레스트 결과

랜덤포레스트의 결과를 살펴보면 피해 발생 3일전의 기상관측자료인 27종만을 이용하여 호우피해를 예측한 함수가 예측력이 가장 높은 것으로 나타났고, 2일전의 기상관측자료만을 이용하여 개발한 함수에서도 좋은 예측력을 나타냈다. 가장 좋은 예측력을 나타낸 피해 발생 3일전의 기상관측자료만을 사용하여 함수를 개발하는 과정을 살펴보면, 독립변수의 개수가 27개이기 때문에 랜덤으로 9개(27/3)의 변수를 선택하여 500개의 각기 다른 나무를 생성하여 랜덤포레스트를 이용한 함수를 개발하였다. 함수의 최적화를 위해 Fig. 6에서 제시한 바와 같이 오차를 최소화하는 크기의 나무 개수를 확인하였고, 나무의 개수가 310개일 때 오차가 최소화되기 때문에 랜덤으로 9개의 변수를 선택하여 310개의 각기 다른 나무를 생성하여 최적의 랜덤포레스트를 이용한 호우피해 예측함수를 도출하였다. 변수별 중요도를 나타낸 Fig. 7을 살펴보면 3일전의 최고기온, 최대 순간 풍속, 최대 풍속 등이 호우피해를 예측하는데 가장 중요한 변수인 것으로 분석되었다.

4.2.4 서포트 벡터 머신 결과

서포트 벡터 머신의 경우 피해 발생 2일전의 기상관측자료인 27종만을 이용하여 호우피해를 예측한 함수가 예측력이 가장 높은 것으로 나타났고, 그 다음으로 6일전의 기상관측자료만을 이용하여 개발한 함수에서도 좋은 예측력을 나타냈다. 가장 좋은 예측력을 나타낸 피해 발생 2일전의 기상관측자료만을 사용하여 함수를 개발하는 과정을 살펴보면, 오차를 최소로 하는 최적의 모수조건은 cost가 4, ϵ 값은 0.3인 것으로 나타났다.

4.3 결과 요약

- (1) 본 연구에서는 28가지의 기상관측자료 형태와 4가지 방법론을 결합하여 총 112종의 호우피해 예측함수를 개발하였다.

Fig. 6. Error by Size of Tree (Random forest)

- (2) 예측력을 평가한 결과를 살펴보면, 피해 발생 2일전의 기상관측자료만을 이용하여 서포트 벡터 머신으로 개발된 호우피해 예측함수가 가장 높은 예측력을 가지고 있는 것으로 평가되었다.
- (3) 서포트 벡터 머신의 경우 가장 높은 예측력을 보였지만, 어떤 독립변수가 피해를 예측하는데 중요한 역할을 했는지 알아볼 수 없는 모형의 한계점이 있었다.
- (4) 피해 발생 2일전의 기상관측자료를 이용하여 서포트 벡터 머신으로 개발된 함수, 피해 발생 6일전과 7일전의 기상관측자료를 이용하여 의사결정나무 모형으로 개발된 함수, 6일전의 기상관측자료를 이용하여 서포트 벡터 머신으로 개발된 함수 순으로 높은 예측력을 가지는 것으로 나타났다.
- (5) 예측력 평가 결과 상위에 있는 함수들은 모두 머신러닝 기법으로 개발된 함수로 기존의 선형회귀모형에 비해 호우피해를 잘 예측하는 것으로 나타났다.
- (6) 기상관측자료의 형태별로 예측력을 평가해보면 피해 발생 2일전의 기상관측자료만을 이용하거나 6일 전의 기상관측자료만을 이용한 함수가 평균적으로 높은 예측력을 나타냈다.
- (7) 본 연구에서 제시한 모든 모형의 code는 Table 4에 제시하였고, 아래의 링크에서도 다운 받을 수 있다.
<https://github.com/karesma0/Heavy-rain-damage-prediction-function-1.git>

5. 결 론

본 연구에서는 수도권(서울, 인천, 경기도) 지역의 호우피해를 사전에 예측하기 위해 선형회귀모형과 머신러닝 기법(의사결정나무, 랜덤포레스트, 서포트 벡터 머신)을 이용하여 호우피해 예측함수를 개발하였다. 재해연보의 호우피해액을 종속변수로 사용하였고, 기상관측자료를 각각 1일전,

Fig. 7. Variable Importance Plot (Random forest)

1~2일전 . . . 7일전 형태로 구축하여 독립변수로 사용하였다. 4가지 방법론(선형회귀모형, 의사결정나무, 랜덤포레스트, 서포트 벡터 머신)과 28가지의 기상관측자료 형태를 이용하여 총 112종의 호우피해 예측함수를 개발하였고, 각각의 함수를 평가용 데이터에 적용하여 예측력을 평가하였다. 예측력 평가 결과 2일전의 기상관측자료만을 이용하여 서포트 벡터 머신으로 개발한 함수가 가장 좋은 예측력을 가진 것으로 나타나 수도권 지역의 최종 호우피해 예측함수로 선택하였다. 이는 기존의 선행연구에서 주로 사용하였던 선형회귀모형에 비해 예측 성능이 향상된 것으로 나타났다. 또한 예측력 평가에서 상위권에 있는 함수들은 대부분 머신러닝 기법을 이용한 함수들인데 이는 기존의 선형회귀모형에서는 고려할 수 없었던 호우피해의 비선형성을 머신러닝 기법을 통해 일정 부분 고려하였기 때문인 것으로 판단된다.

기상관측자료의 형태별로 예측력을 평가한 결과 1일 전부터 7일 전까지의 기상관측자료를 모두 사용하는 형태와 같이 사용할 수 있는 기상관측자료를 많이 이용할수록 예측력이 높아질 것으로 예상했지만, 결과에선 피해 발생 전 특정 짧은 기간의 기상관측자료만을 이용한 함수가 좋은 결과를 보이는 것으로 나타났다. 특히 가장 성능이 우수한 것으로 평가된 함수들은 피해 발생 2일전의 기상관측자료 기반 서포트 벡터 머신, 6일전과 7일전의 기상관측자료 기반 의사결정나무, 6일전의 기상관측자료 기반 서포트 벡터 머신 등으로 대부분 1일치의 기상관측자료나 2일치의 기상관측자료만을 기반으로 개발한 함수들이 비교적 좋은 예측력 성능을 보였다. 이는 일자별 기상현상간의 상관성이 높기 때문에 비교적 긴 기간의 기상관측자료를 사용하는 경우 독립변수들 간의 상관성이 높아지면서 오히려 모형의 예측력 향상에 악영향을 미치기 때문인 것으로 판단된다. 따라서 자연현상 및 피해의 비선형을 고려할 수 있는 머신러닝 기법을 이용하여 피해 발생 전의 특정 짧은 기간에 발생하는 전조 증상을 파악할 수 있는 연구가 추가적으로 필요할 것으로 판단된다.

Table 4. R Code of this Study

<pre># You can download my code that I developed through this study at the address below. https://github.com/karesma0/Heavy-rain-damage-prediction-function-1.git Train_D <- read.csv("1_training set.csv") Test_D <- read.csv("2_test set.csv") library(rpart) library(partykit) library(randomForest) library(caret) library(e1071) NRMSE <- function(yi, yhat_i){ (sqrt(mean((yi - yhat_i)^2)))/(max(yi)-min(yi)) } for(j in 1:7) { for(i in j:7) { Train_1 <- Train_D[, c(1, (2+27*(j-1)):(1+27*i))] ## Linear Regression model(with PCA) Train_sub <- Train_1[,-c(1)] preProc <- preProcess(Train_sub, method = c("center", "scale", "zv", "nzv", "pca"), thresh = 0.9) PC_train <- predict(preProc, newdata = Train_sub) Damage <- Train_1[,1] Train_2 <- cbind.data.frame(Damage, PC_train) LR_train <- lm(Damage~., Train_2) LR_model <- step(LR_train, direction = "both") ## Decision Tree model set.seed(180926) tree <- rpart(Damage ~., data=Train_1) min_cp <- tree\$cpstable[which.min(tree\$cpstable['xerror']),] Tree_model <- prune(tree, cp = min_cp[1]) ## Random Forest model set.seed(180926) RF <- randomForest(Damage ~., data = Train_1) RF_model <- randomForest(Damage ~., data = Train_1, ntree = which.min(RF\$mse))</pre>	<pre>## Support Vector Machine model svm.radial <- tune(svm, Damage ~., data = Train_1, kernel = "radial", ranges = list(cost = 2^(0:7), epsilon = seq(0,1,0.1))) SVM_model <- svm.radial\$best.model ## Predict heavy rainfall damage Val_1 <- Test_D[, c(1, (2+27*(j-1)):(1+27*i))] PC_val <- predict(preProc, newdata = Val_1) sim_LR <- predict(LR_model, newdata = PC_val) sim_Tree <- predict(Tree_model, newdata = Val_1) sim_RF <- predict(RF_model, newdata = Val_1) sim_SVM <- predict(SVM_model, newdata = Val_1) ## Reverse Log Transform real_sim_LR <- exp(sim_LR) real_sim_Tree <- exp(sim_Tree) real_sim_RF <- exp(sim_RF) real_sim_SVM <- exp(sim_SVM) ## Evaluate data NRMSE_LR <- NRMSE(exp(Val_1\$Damage), real_sim_LR) NRMSE_Tree <- NRMSE(exp(Val_1\$Damage), real_sim_Tree) NRMSE_RF <- NRMSE(exp(Val_1\$Damage), real_sim_RF) NRMSE_SVM <- NRMSE(exp(Val_1\$Damage), real_sim_SVM) ## Print result Result <- cbind(NRMSE_LR, NRMSE_Tree, NRMSE_RF, NRMSE_SVM) Name <- c("1-1", "1-2", "1-3", "1-4", "1-5", "1-6", "1-7", "2-1", "2-2", "2-3", "2-4", "2-5", "2-6", "2-7", "3-1", "3-2", "3-3", "3-4", "3-5", "3-6", "3-7", "4-1", "4-2", "4-3", "4-4", "4-5", "4-6", "4-7", "5-1", "5-2", "5-3", "5-4", "5-5", "5-6", "5-7", "6-1", "6-2", "6-3", "6-4", "6-5", "6-6", "6-7", "7-1", "7-2", "7-3", "7-4", "7-5", "7-6", "7-7") hoho <- paste0('output/') hoho[i+7*(j-1)] <- paste0('output/',Name[i+7*(j-1)],'.csv') write.csv(Result, hoho[i+7*(j-1)]) } }</pre>

본 연구에서는 크게 3가지 한계점이 있는데, 우선 본 연구는 과거의 호우피해 자료와 기상관측자료를 이용하여 통계적인 방법으로만 피해를 예측하였다. 그러나 최근에는 기후변동과 기후변화로 인해 기존에 발생하지 않았던 극한 기상현상과 심각한 피해가 빈번하게 발생하고 있는데, 변화되는 기상현상을 반영할 수 있는 물리적인 모형을 고려하지 못한 한계점이 있다. 두 번째로 본 연구에서는 호우피해를 특별한 구분 없이 모두 고려하였는데, 장마전선으로 인한 피해, 국지성 집중호우로 인한 피해 등 호우피해 발생 원인별로

구분하여 예측함수를 개발한다면 예측력 향상에 도움이 될 것으로 판단된다. 마지막으로 과거의 기상관측자료가 아닌 좀 더 상관성이 높은 당일의 기상자료를 사용하지 못한 것이 연구의 한계점이다. 본 연구에서 사용한 기상관측자료 대신 기상예측자료를 사용하는 방안을 고려할 수 있지만, 아직까지 소수의 기상요소만을 예측하고 있고 자료의 불확실성이 높기 때문에 피해를 예측하는데 어려움이 있다고 판단하였다. 특히 최근에 발생하는 국지성 집중호우의 경우 예측하는 것이 상당히 어렵기 때문에, 불확실성을 내포

하고 있는 기상예측자료를 이용하여 함수를 개발하는 방법 보다는 불확실성을 제거한 기상관측자료를 이용하는 방안을 제시하였다. 그러나 레이더 강우자료를 이용하거나 향후에 기상예측자료의 신뢰도가 향상 된다면 이를 반영한 연구를 진행 할 수 있을 것이다.

본 연구의 결과는 선행연구에서 고려하지 못한 다양한 기상관측자료를 사용하였고, 기존의 선형회귀모형에 비해 머신러닝 기법이 호우피해 예측함수의 예측력을 극대화하는데 효과적임을 제시하였다. 또한 피해 발생 전의 전조 증상으로 볼 수 있는 과거의 기상관측자료만을 활용하여 호우피해를 예측할 수 있는 방안을 제시하였다. 본 연구결과를 활용하여 대규모 호우피해를 사전에 예측하고 대비 차원의 재난관리를 실시한다면, 피해를 저감하는데 도움이 될 수 있을 것으로 기대한다.

감사의 글

본 연구는 정부(행정안전부)의 재원으로 재난안전기술개발 사업단의 지원을 받아 수행된 연구임[MOIS-재난-2015-05].

References

- Asim, K.M., Martínez-Álvarez, F., Basit, A., and Iqbal, T. (2017). Earthquake magnitude prediction in Hindukush region using machine learning techniques. *Natural Hazards*, Vol. 85, No. 1, pp. 471-486.
- Bae, J.S. (2014). *A study on priority of determinants of career decision level and career preparation behavior in high school students based on decision tree analysis*. Ph.D. dissertation, Soonchunhyang University.
- Bae, M.S. (2018). Cause of flood damage in central Chungbuk Province in 2017. *Chungbuk Issues & Trends*, Vol. 31, pp. 8-13.
- Breiman, L. (2001). Random Forests. *Machine Learning*, Vol. 45, No. 1, pp. 5-32.
- Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). *Classification and regression trees*. CRC press.
- Choi, C.H., Kim, J.H., Kim, J.S., Kim, D.H., Bae, Y.H., and Kim, H.S. (2018). Development of heavy rain damage prediction model using machine learning based on big data. *Advances in Meteorology*, Vol. 2018, Article ID 5024930.
- Choi, C.H., Kim, J.S., Kim, J.H., Kim, H.Y., Lee, W.J., and Kim, H.S. (2017). Development of heavy rain damage prediction function using statistical methodology. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 3, pp. 331-338.
- Choi, C.H., Kim, J.S., Lee, M.J., Kim, J.H., Lee, W.J., and Kim, H.S. (2017). Development of heavy rain damage prediction function using principal component analysis and logistic regression model. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 6, pp. 159-166.
- Choi, C.H., Park, K.H., Park, H.K., Lee, M.J., Kim, J.S., and Kim, H. S. (2017). Development of heavy rain damage prediction function for public facility using machine learning. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 6, pp. 443-450.
- Choi, C.K. (2017). *Evaluation of flood impact variables and development of flood damage function: Case study for residential buildings and contents*. Ph.D. dissertation, Inha University.
- Choi, H.S., Park, H.W., and Park, C.Y. (2013). Support vector machines for big data analysis. *Journal of the Korean Data & Information Science Society*, Vol. 24, No. 5, pp. 989-998.
- Davis, S.A., and Skaggs, L.L. (1992). *Catalog of residential depth-damage functions used by the army corps of engineers in flood damage estimation*. Army Engineer Institute For Water Resources, Alexandria, VA.
- Han, D.G., Choi, C.H., Kim, D.H., Jung, J.W., Kim, J.W., and Kim, S.J. (2016). Determination of flood reduction alternatives for responding to climate change in Gyeongang watershed. *Journal of Wetlands Research*, Vol. 18, No. 2, pp. 154-165.
- Kalra, A., Miller, W.P., Lamb, K.W., Ahmad, S., and Piechota, T. (2013). Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins. *Hydrological Process*, Vol. 27, No. 11, pp. 1543-1559.
- Kang, S.M., Park, M.J., Kim, S.H., and Kim, S.J. (2007). A study on the mitigation of inundation damage using flood inundation analysis model FLUMEN: For the part of Jinwicheon reach. *Journal of the Korean Society of Civil Engineers*, Vol. 27, No. 6B, pp. 583-590.
- Kim, B.J., Kim, G.H., and Kim, H.S. (2016). Statistical analysis of hazen-williams C and influencing factors in multi-regional water supply system. *Journal of Korea Water Resource Association*, Vol. 49, No. 5, pp. 399-410.
- Kim, D.H., Hong, S.J., Kim, J.W., Han, D.G., Hong, I.P., and Kim, H.S. (2015). Water quality analysis of Hongcheon river basin under climate change. *Journal*

- of *Wetlands Research*, Vol. 17, No. 4, pp. 348-358.
- Kim, J.H., Kim, H.J., Lee, S.O., and Cho, Y.S. (2007). Numerical simulation of flood inundation with quadtree grid. *J. Korean Soc. Hazard Mitig.*, Vol. 7, No. 2, pp. 45-52.
- Kim, J.S., Choi, C.H., Kim, D.H., Lee, M.J., and Kim H.S. (2017). Development of heavy rain damage prediction function using artificial neural network and multiple regression model. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 6, pp. 73-80.
- Kim, J.S., Choi, C.H., Lee, J.S., and Kim, H.S. (2017). Damage prediction using heavy rain risk assessment: (2) Development of heavy rain damage prediction function. *J. Korean Soc. Hazard Mitig.*, Vol. 17, No. 2, pp. 371-379.
- Kim, S.J., Ryoo, E.C., Jung, M.K., Kim, J.K., and Ahn, H.C. (2012). Application of support vector regression for improving the performance of the emotion prediction model. *Journal of Intelligence and Information Systems*, Vol. 18, No. 3, pp. 185-202.
- Kim, S.W., and Ahn, H.C. (2010). Development of an intelligent trading system using support vector machines and genetic algorithms. *Journal of Intelligence and Information Systems*, Vol. 16, No. 1, pp. 71-92.
- Kim, Y.M., Hwang, M.N., Kim, T.H., Jeong, C.H., and Jung, D.H. (2015). Big data mining for natural disaster analysis. *Journal of the Korean Data & Information Science Society*, Vol. 26, No. 5, pp. 1105-1115.
- Kusiak, A., Wei, X., Verma, A.P., and Roz, E. (2012). Modeling and prediction of rainfall using radar reflectivity data: A data-mining approach. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 4, pp. 2337-2342.
- Lee, B.J. (2017). Analysis on inundation characteristics for flood impact forecasting in Gangnam drainage basin. *Atmosphere*, Vol. 27, No. 2, pp. 189-197.
- Lee, H.H., Chung, S.H., and Choi, E.J. (2016). A case study on machine learning applications and performance improvement in learning algorithm. *Journal of Digital Convergence*, Vol. 14, No. 2, pp. 245-258.
- Lee, H.S., and Lim, J.H. (2005). *SPSS 12.0 manual*. Bobmunsa.
- Lee, J.S., Eo, G., Choi, C.H., Jung, J.W., and Kim, H.S. (2016). Development of rainfall-flood damage estimation function using nonlinear regression equation. *Journal of the Korea Society of Disaster Information*, Vol. 12, No. 1, pp. 74-88.
- Lesmeister, C. (2017). *Mastering machine learning with R* (2nd ed.). Packt Publishing Ltd.
- Munich Re. (2002). *Winter storms in Europe: Analysis of 1990 losses and future loss potentials*.
- Murane, R.J., and Elsner, J.B. (2012). Maximum wind speeds and US hurricane losses. *Geophysical Research Letters*, Vol. 39, No. 16, L16707.
- Park, C.K. (2006). Estimating software development cost using support vector regression. *Journal of Korean Management Science Review*, Vol. 23, No. 2, pp. 75-91.
- Park, J.S. (2006). *An empirical comparison between support vector regression and neural networks*. Master's thesis, Dongguk University.
- Park, U.Y., and Kim, G.H. (2007). A study on predicting construction cost of apartment housing projects based on support vector regression at the early project stage. *Journal of the Architectural Institute of Korea, Structure & Construction*, Vol. 23, No. 4, pp. 165-172.
- Prahl, B.F., Rybski, D., Kropp, J.P., Burghoff, O., and Held, H. (2012). Applying stochastic small-scale damage functions to German winter storms. *Geophysical Research Letters*, Vol. 39, No. 6, Article No. L06806.
- Radhika, Y., and Shashi, M. (2009). Atmospheric temperature prediction using support vector machines. *International Journal of Computer Theory and Engineering*, Vol. 1, No. 1, pp. 55-58.
- Rasouli, K., Hsieh, W.W., and Cannon, A.J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, Vol. 414-415, pp. 284-293.
- Sakr, G.E., Elhajj, I.H., Mitri, G., and Wejinya, U.C. (2010). Artificial intelligence for forest fire prediction. *2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pp. 1311-1316.
- Sharma, N., Sharma, P., Irwin, D., and Shenoy, P. (2011). Predicting solar generation from weather forecasts using machine learning. *2011 IEEE International Conference on Smart Grid Communications*, pp. 528-533.
- Song, J.H., Kim, H.S., Hong, I.P., and Kim, S.U. (2006). Parameter calibration of storage function model and flood forecasting: (1) Calibration methods and evaluation of simulated flood hydrograph. *Journal of the Korean Society of Civil Engineers*, Vol. 26, No. 1B, pp. 27-38.

Song, Y.S., Lim, C.H., Joo, J.G., and Park, M.J. (2016). A study on heavy rain forecast evaluation and improvement method. *J. Korean Soc. Hazard Mitig.*, Vol. 16, No. 2, pp. 113-121.

Tay, F.E., and Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, Vol. 29, No. 4, pp. 309-317.

Vapnik, V.N. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

Yoo, J.E. (2015). Random forests, an alternative data mining technique to decision tree. *Journal of Educational*

Evaluation, Vol. 28, No. 2, pp. 427-448.

Yoon, Y.N. (2007). *Hydrology*. Chungmungak.

Zhai, A.R., and Jiang, J.H. (2014). Dependence of US hurricane economic loss on maximum wind speed and storm size. *Environmental Research Letters*, Vol. 9, No. 6, Article No. 064019.

Received	October 19, 2018
Revised	October 24, 2018
Accepted	November 1, 2018