Journal of Wetlands Research
Vol. 25, No. 3, August 2023, pp. 167-175

ISSN 1229-6031 (Print) / ISSN 2384-0056 (Online)
DOI https://doi.org/10.17663/JWR.2023.25.3.167

머신러닝 기반의 강우추정 방법 개발

한희찬* · 김창주* · 김동현**†

*조선대학교 토목공학과
**인하대학교 수자원시스템연구소

Development of Machine Learning Based Precipitation Imputation Method

Heechan Han* · Changju Kim* · Donghyun Kim**†

*Department of Civil Engineering, Chosun University,
**Department of Water Resources System, Inha University

(Received : 30 June 2023, Revised : 14 July 2023, Accepted : 14 July 2023)

요 약

강우 데이터는 습지관리, 수문모의, 수자원 관리와 같은 다양한 분야에서 활용되는 필수 입력자료 중 하나이다. 강우 데이터를 활용하여 효율적인 수자원관리를 위해서는 기본적으로 데이터의 결측률을 최소화 시킴으로써 최대한 많은 데이터를 확보하는 것이 필수적이다. 또한 미계측 지역에 대한 강우 데이터를 확보한다면 보다 효율적인 수문모의가 가능하다. 그러나 결측 강우 데이터는 주로 통계학적 기법에 의해 추정되어 왔다. 본 연구의 목적은 데이터 간의 상관관계를 기반으로 새로운 데이터를 예측할 수 있는 머신러닝 알고리즘을 활용하여 결측 강우 데이터를 복원할 수 있는 새로운 방법을 제안하고자 한다. 또한, 기존의 통계적 방법들과 비교하여 머신러닝 기법의 결측 강우 데이터 복원을 위한 활용가치를 평가하고자 한다. 평가를 위해 대표적인 머신러닝 알고리즘인 Artificial Neural Network (ANN)과 Random Forest (RF)을 적용하였다. 강우의 발생 유무를 분류하는 성능은 RF 알고리즘이 ANN 알고리즘보다 강우 발생유무의 분류 정확도가 높은 것으로 나타났다. 분류 모형의 평가 지표인 F1-score나 Accuracy값이 RF는 0.80, 0.77인 반면에, ANN은 0.76, 0.71로 계산되었다. 또한 강우량을 추정하는 성능 역시 RF가 ANN 알고리즘보다 보다 높은 정확도를 보였다. RF과 ANN 알고리즘의 RMSE은 2.8mm/day과 2.9mm/day이고, R^2값은 0.73, 0.68으로 계산되었다.

핵심용어 : 강우추정, 머신러닝 알고리즘, Artificial Neural Network, Random Forest

Abstract

Precipitation data is one of the essential input datasets used in various fields such as wetland management, hydrological simulation, and water resource management. In order to efficiently manage water resources using precipitation data, it is essential to secure as much data as possible by minimizing the missing rate of data. In addition, more efficient hydrological simulation is possible if precipitation data for ungauged areas are secured. However, missing precipitation data have been estimated mainly by statistical equations. The purpose of this study is to propose a new method to restore missing precipitation data using machine learning algorithms that can predict new data based on correlations between data. Moreover, compared to existing statistical methods, the applicability of machine learning techniques for restoring missing precipitation data is evaluated. Representative machine learning algorithms, Artificial Neural Network (ANN) and Random Forest (RF), were applied. For the performance of classifying the occurrence of precipitation, the RF algorithm has higher accuracy in classifying the occurrence of precipitation than the ANN algorithm. The F1-score and Accuracy values, which are evaluation indicators of the classification model, were calculated as 0.80 and 0.77, while the ANN was calculated as 0.76 and 0.71. In addition, the performance of estimating precipitation also showed higher accuracy in RF than in ANN algorithm. The RMSE of the RF and ANN algorithms was 2.8 mm/day and 2.9 mm/day, and the values were calculated as 0.68 and 0.73.

†To whom correspondence should be addressed.
Department of Water Resources System, Inha University
E-mail : yesdktpdi@naver.com

• **Heechan Han** Department of Civil Engineering, Chosun University/Assistant professor(heechan@chosun.ac.kr)
• **Changju Kim** Department of Civil Engineering, Chosun University/bachelor' Course(changaround@naver.com)
• **Donghyun Kim** Department of Water Resources System, Inha University/Post-doc(yesdktpdi@naver.com)

Key words : Precipitation estimation, Machine learning algorithms, Artificial Neural Network, Random Forest

1. 서 론

강우 데이터는 수문기상, 환경, 농업, 자연재해, 그리고 수자원 시스템 분야에서 가장 필수적인 기본 요소 중 하나이다 (New et al., 2001; Sadat-Noori et al., 2020; 오승철 외 5인, 2022). 또한 강우 데이터는 수문학적 분석에서 활용되는 필수 입력자료 중 하나로 관측 데이터의 품질에 따라 수문모형을 이용한 모의 결과물의 정확도가 결정된다고 할 수 있다. 습지 관리 측면에서도 강우 데이터는 매우 중요한 역할을 한다. 인공 및 자연습지 관리를 위한 수문 혹은 수질 모의를 하는 경우 강우데이터는 모델링 과정의 기초 자료로 활용되기 때문에 강우 데이터의 높은 정확성이 요구되는 바이다 (박기수 외 2인, 2013). 따라서, 강우 관측소별로 강우 데이터의 품질을 어떻게 관리하느냐에 따라 수문 모형을 활용한 수자원, 습지 관리의 효율성이 결정될 수 있다 (Larson et al., 1974; 권태용 외 3인, 2022).

강우의 시공간적 변동성은 수 많은 인자들과 직간접적으로 연계되어 있기 때문에 미계측 강우자료에 대해 직접 관측이 아닌 수치 모형을 이용하여 강우의 발생과 강우량을 산정하는 것은 매우 복잡한 과제 중 하나이다 (Bellido-Jiménez et al., 2021). 현재 국내에서 운용되고 있는 강우 관측소의 경우에도 미계측 된 강우 데이터가 존재함으로써 강우 데이터의 활용에 제한이 생기는 경우가 있다 (Teegavarapu et al., 2008; 장상민 외 4인 2018). 따라서, 이러한 미계측 데이터의 추정 및 보완은 보다 효과적인 수재해 방지, 수자원 관리를 위한 필수 과제 중 하나이다. 일반적으로, 미계측 강우를 산정하기 위해서 Kriging, Thiessen, 등우선법, 그리고 역거리관측법 등 다양한 수문학적 방법들이 적용되고 있다. 이러한 방법들은 산악효과나 강우 관측소의 분포 상태 등을 고려하지 못하기 때문에 측정하는 지역에 따라 강우 추정 오차가 커질 수 있다는 한계가 있다 (황석환 외 2인, 2022).

이러한 한계를 보완하기 위해 격자형 강우 데이터를 개발하는 사례가 증가하고 있다. 일반적으로 강우레이더나 인공위성 기술을 활용하여 강우 데이터를 생성하고 있다 (김성준 외 2인, 1999; 김병식 외 3인, 2007; 강나래 외 4인, 2013; 김경탁 외 1인, 2013). 국내의 경우 현재 기상청 및 환경부에서 약 20대 이상의 강우 레이더를 운용중이고, Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS; Funk et al., 2015), Climate Prediction Center morphing method (CMORPH; Joyce et al., 2004), Tropical Rainfall Measuring Mission (TRMM; Huffman et al., 2007), Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN; Sorooshian et al., 2000)와 같은 고해상도의 인공위성 기반의 강우 데이터가 다양한 연구분야에서 활용되고 있다. 하지만, 격자형 강우 데이터의 경우 지상 강우가 커버하지 못하는 복합 산악 지형과 같은 지역의 강우 데이터를 수월하게 활용할 수 있다는 장점이 있는 반면에 지상 강우에 비해 정확도가 다소 떨어진다는 단점이 있다 (Kim and Han, 2021).

최근에는 데이터 관측 시스템과 빅데이터 기술의 발전과 활용 가능한 데이터의 양이 증가함에 따라 머신러닝을 활용한 사례가 증가하고 있다 (Rosenblatt, 1958). 머신러닝은 데이터 사이의 관계를 기반으로 분류, 회귀, 그리고 예측 문제에 주로 사용되는 기법 중 하나이다. 대표적인 머신러닝 기법으로는 인공신경망 (Aritificial Neural Network; ANN)이 있는데, ANN은 인간의 뇌구조를 바탕으로 개발된 알고리즘이다. 또한 Random Forest (RF), Support Vector Machine, K-means Clustering technique와 같은 알고리즘이 널리 활용되고 있다.

머신러닝을 이용해서 강우의 결측값을 한 연구는 다음과 같다. Polishchuk et al. (2021)은 RF 알고리즘을 활용하여 Australia 지역의 강수 유무를 사전에 예측할 수 있는 모형을 개발하였다. 해당 모형은 강우 발생에 영향이 있는 기상 인자들을 선별하여 모형의 입력자료로 활용하였다. 그 결과 예측 결과는 약 85%의 정확도를 보여주었다. Sahoo & Ghose (2022)는 K-nearest neighbor, self-organizing maps, RF, feed-forward neural network 머신러닝 알고리즘을 활용하여 India 지역의 결측 강우를 추정하는 연구를 수행하였다. 4가지 알고리즘 중 feed- forward neural network 알고리즘의 정확도가 가장 우수한 것을 확인하였다. 이러한 연구들은 강우 예측 분야에서 머신러닝 알고리즘의 활용 가능성을 보여주었다.

따라서, 본 연구에서는 최근 주목받고 있는 AI 기술 중 하나인 머신러닝 기술 중 ANN과 RF 알고리즘을 활용하여 광주광역시의 주요 강우 관측소에서 발생하는 미계측 강우의 추정 방법을 제안하고자 한다. 본 연구는 강우의 발생 유무뿐만 아니라 강우량을 함께 추정할 수 있는 모형을 제안하였다.

2. 방법론

2.1 머신러닝 알고리즘

본 연구에서는 광주광역시에 위치한 주요 강우 관측 지점들을 대상으로 미계측 된 일강우 데이터를 추정하고자 한다. 여기서, 데이터 추정 기술이란 미계측 강우의 양과 더불어 강우의 발생 유무도 함께 추정할 수 있는 기술을 의미한다. 이를 위해 인공신경망(ANN), 랜덤포레스트(RF) 머신러닝 알고리즘을 적용하였다. 이 두 가지 알고리즘을 이용한 데이터의 분류 및 예측 성능은 이전의 많은 연구에서 확인이 되었기 때문에 본 연구의 방법론으로 채택하였다. 각 알고리즘의 입력자료로는 강우와 연관성이 높은 것으로 알려진 기온, 습도, 그리고 관측소의 고도 등 타기상 인자 및 지형 특성을 활용하여 보다 효과적인 강우 추정 기술을 개발하고자 한다.

또한, 기존의 통계적 기반의 방법들과 함께 비교를 통해 본 연구에서 개발된 머신러닝 기반의 강우 추정 기술의 성능을 평가하고자 한다.

2.1.1 인공신경망 (Artificial Neural Networks; ANN)

인공신경망(ANN)은 머신러닝 기술 중 가장 널리 사용되고 있는 기술 중 하나로 인간의 뇌의 구조를 바탕으로 개발된 알고리즘이다 (Rosenblatt, 1958). 인공신경망은 크게 세 가지 층 (입력층, 은닉층, 그리고 출력층)으로 구성되어 있다. 입력층과 출력층은 각 1개의 층을 포함하고 있지만, 은닉층의 경우는 주어진 데이터의 특성과 사용자에 따라 2개 이상의 층으로 구성할 수 있다. 여기서 2개 이상 다수의 은닉층으로 이루어져 있는 다층 퍼셉트론(multilayer perceptron)이라고 한다. 은닉층과 각 층이 포함하고 있는 노드의 개수가 지나치게 증가함에 따라 계산과정이 복잡하고, 과적합 문제가 발생할 수도 있기 때문에 최적의 매개변수 값을 찾는 것이 인공신경망을 효과적으로 활용하기 위한 중요한 과정이라고 할 수 있다. 인공신경망을 수학적 공식으로 나타내면 Eq.(1)과 같다.

$$\begin{aligned} f_1 &= f(b_1 + w_{11}X_1 + w_{21}X_2 + \cdots + w_{n1}X_n) \\ f_2 &= f(b_2 + w_{12}X_1 + w_{22}X_2 + \cdots + w_{n2}X_n) \\ &\vdots \\ f_m &= f(b_m + w_{1m}X_1 + w_{2m}X_2 + \cdots + w_{nm}X_n) \\ Y &= f(B_{out} + w_{1,out}X_1 + w_{2,out}X_2 + \cdots + w_{n,out}X_n) \end{aligned} \quad (1)$$

여기서 X는 입력 변수, f는 활성화 함수, w는 각 층 사이의 가중치를 나타낸다. b는 입력층 및 은닉층에서 그리고 B는 출력층에서 발생하는 바이어스(Bias)를 의미한다. 마지막으로 Y는 인공신경망에서 최종적으로 출력되는 결과값을 의미한다. Fig. 1은 인공신경망 알고리즘의 개념도를 나타내고 있다.

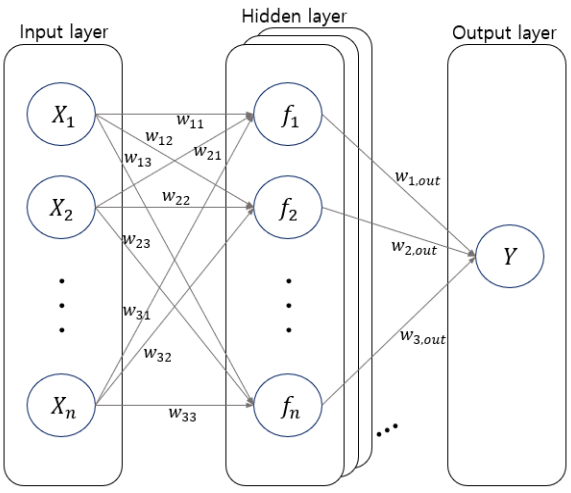


Fig. 1. Conceptual diagram of ANN algorithm

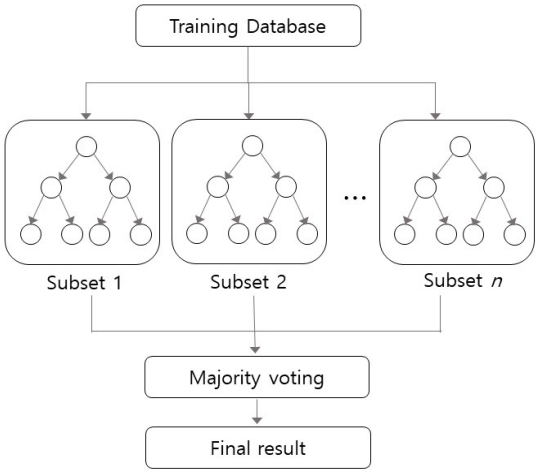


Fig. 2. Conceptual diagram of RF algorithm

2.1.2 랜덤포레스트 (Random Forest; RF)

랜덤포레스트 기법(RF)은 대표적인 머신러닝 알고리즘 중 하나로 주로 의사결정을 위한 분류나 회귀를 수행하는데 사용된다. 랜덤포레스트 알고리즘은 다수의 의사결정 나무가 결합된 형태로 각 의사결정 나무를 통해 제시된 출력결과를 바탕으로 최종 출력값을 결정하는 구조로 이루어져 있다 (Breiman, 2001). 또한, 방대한 양의 데이터를 다루는데 빠른 처리 속도와 높은 정확도를 제시한다는 장점이 있다 (Kim et al., 2019; Tyralis et al., 2019).

랜덤포레스트 알고리즘은 학습과정에서 다수의 의사결정 나무를 기반으로 다수의 원칙 (majority voting)을 바탕으로 최종 출력값을 제시하며, 주요 매개변수는 max_depth, max_features, bootstrap의 적용 여부, 그리고 n_estimator 등으로 볼 수 있다. 다른 머신러닝 알고리즘과 마찬가지로 랜덤포레스트 알고리즘 역시 과접합 문제를 피하기 위해 각 매개변수들의 최적값을 찾는 것이 중요하다. 이를 위해 K-Fold CV, RandomizedSearchCV 등 다양한 cross validation 방법들이 활용 될 수 있다. 랜덤포레스트 알고리즘의 개념도를 Fig. 2와 같이 나타내었다.

2.2 강우 추정 모형 개발

본 연구에서 제안하는 강우 추정 과정은 총 4가지 단계를 포함하고 있고 각 단계에 대한 설명은 다음과 같다 (Fig 3).

(1) 6개 강우 관측소(광주, 무등산, 광산, 조선대, 과기원, 풍암)에서 관측된 12년(2010 – 2021)동안의 시강우 데이터를 수집 및 전처리를 수행한다. 이와 더불어 대상지점인 광주 관측소 및 나머지 지점들의 강우 데이터 간의 상관관계를 분석한다.

(2) 두 가지 머신러닝 알고리즘 (ANN, RF)의 training과 test를 위해 전체 강우 데이터의 70%, 30%를 활용하고, 각 알고리즘의 매개변수 최적화를 수행한다. 여기서 데이터

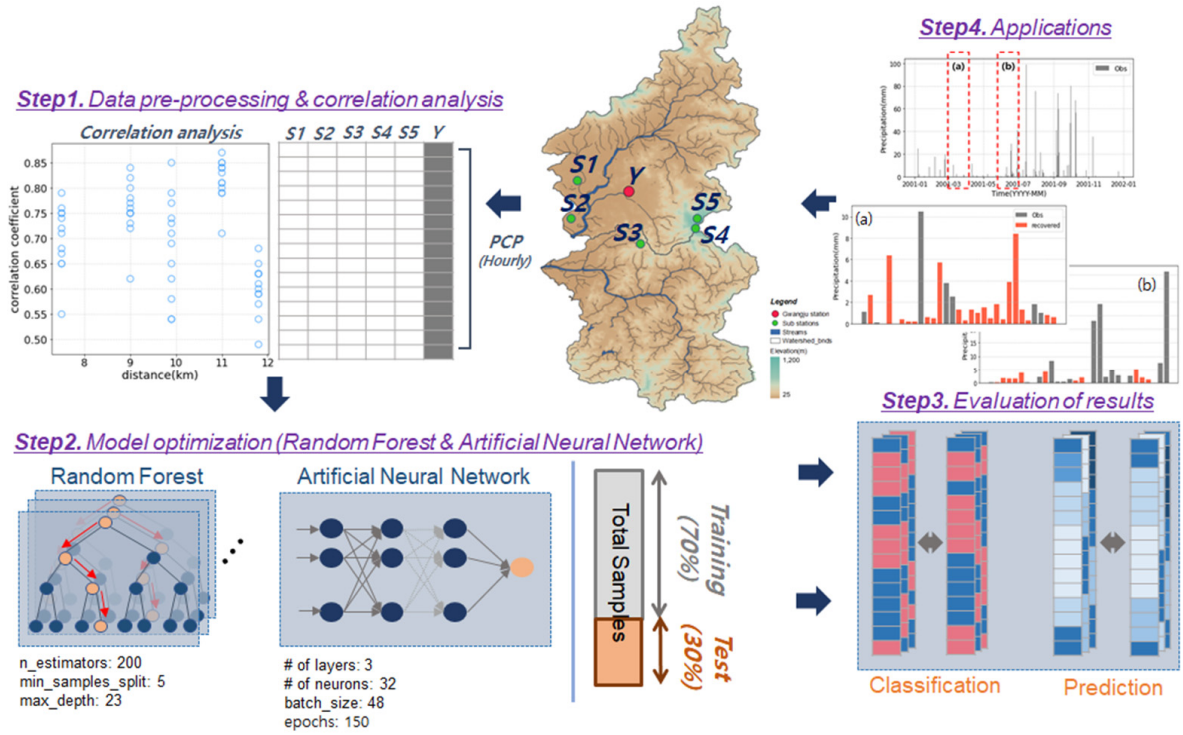


Fig 3. Flowchart of this study

training과 test는 강우 미발생 및 발생 데이터의 균형을 맞추기 위해 전처리 과정을 수행하였다.

(3) 최적화된 두 모형을 이용하여 강우의 유무를 판단할 수 있는 분류 모형과 강우의 양을 추정할 수 있는 예측 모형을 개발하고, 통계학적 평가 지표들을 활용하여 모형으로부터 추정된 강우의 특성을 평가한다.

(4) 마지막으로, 성능이 검증된 최종 알고리즘을 통해 추정된 강우량을 실제 미계측된 시점에 적용한다.

2.3 모형의 평가 지표

강우의 결측치를 보완하기 위해 인공신경망과 랜덤포레스트 알고리즘의 성능을 평가하기 위해 본 연구는 2가지 성능 분석 지표(Eqs. (2)-(3))를 이용하였다.

결정계수 (coefficient of determination; R^2)는 모형의 적합성을 나타내는 척도로, 종속변수의 분산 중에서 독립변수로 설명되는 비율을 의미한다.

$$R^2 = \left[\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}} \right]^2 \quad (2)$$

Root Mean Squared Error (RMSE)는 예측값과 관측값의 차이를 다룰 때 사용되는 지표로 정밀도를 표현하는데 주로 쓰인다. 평균제곱근오차로서 예측값과 관측값의 차에 대한 제곱평균제곱근으로 나타낸다.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2} \quad (3)$$

여기서 x_i와 y_i는 강우의 모의값 및 관측값을 의미하고, $\bar{x}$와 $\bar{y}$는 모의된 강우량과 관측 강우량의 평균값을 의미한다. N은 평가에 사용된 강우 데이터의 개수를 나타낸다.

본 연구에서는 딥러닝 알고리즘의 강우 추정 성능을 평가하기 위해 강우의 양 뿐만 아니라 강우의 발생을 추정할 수 있는지에 대해서도 함께 고려하였다. Table 1은 오차행렬(confusion matrix)를 나타내고 있다. 오차행렬은 관측소에서의 강우 관측 여부와 알고리즘의 모의 결과간의 비교를 위한 행렬이다. 즉, 관측소에서 강우가 실제로 관측되었을 때 알고리즘 역시 강우의 발생을 올바르게 추정하였는지 여부를 판단하는 기준을 제시한다. 오차행렬로부터 계산된 4가지 지표 (TP, FP, FN, TN)을 이용하여 정확도(Accuracy; Eq. (4)), 정밀도 (Precision; Eq. (5)), 재현율 (Recall; Eq.(6)), 그리고 F1-Score(Eq. (7))을 구할 수 있다.

Table 1. Confusion matrix for evaluation of classification results

Detected	Observation	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score = 2 \times \frac{\mathrm{Precision}*Recall}{\mathrm{Precision}+Recall} \quad (7)$$

Accuracy는 전체 데이터 중 올바르게 분류된 건의 비율을 나타내고, Accuracy 값이 높을수록 해당 알고리즘의 정확도가 높다는 것을 의미한다. 하지만, 데이터의 특성에 따라 다소 신뢰성이 떨어지는 결과를 제시할 수 있다는 한계가 있다. 이러한 한계점으로 인해 Precision과 Recall, F1-Score와 같은 다른 지표들도 함께 고려하는 것이 일반적이다. 특히, F1-Score는 데이터의 왜곡성, 편향성 등에 영향이 적고, Precision, Recall을 모두 고려하여 분류 모델을 평가하는 매우 적합한 지표이다. F1-Score는 0 – 1 사이의 값을 가지며 1에 가까울수록 알고리즘의 성능이 좋다고 판단할 수 있다.

3. 연구 대상 지역 및 자료

3.1 연구 대상 지역

본 연구에서는 대한민국 광주광역시를 대상으로 결측 강우 데이터에 대한 보완 연구를 수행하였다. 대상지역의 경우 약 501km^2의 면적을 포함하고 있으며, 총 7개의 기상관측소가 운용중이다. 대상 지역의 대표 기상관측소인 광주기상관측소의 경우 관측이 시작된 1939년 이후 연 평균 강우량은 약 1,314mm이며, 7-8월에 가장 많은 강우량이 관측되고 있다. 연구 지역의 고도는 25 – 1,200m로 이루어져 있고, 연구지역 내에 6개의 기상관측소가 위치하고 있다.

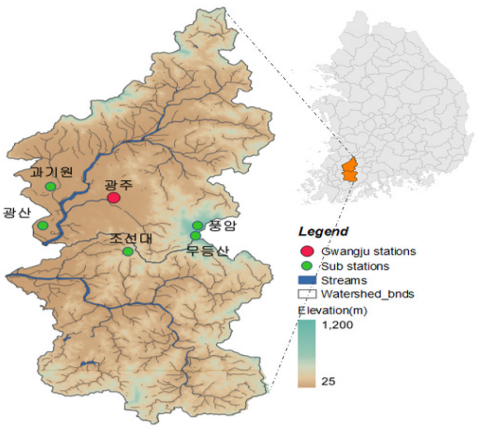


Fig. 4. Study area of this study. Red dot is target station and green dots are sub stations

3.2 대상지점의 강우 자료 수집

본 연구에서는 기상청 기상자료개방포털 홈페이지(https://data.kma.go.kr/)를 통해 제공하는 시간별 강수량 자료를 사용하였다. 기상청에서 현재 운용중인 1곳(광주)의 종관기상관측소(ASOS)와 5곳(무등산, 광산, 과기원, 조선대, 풍암)의 방재기상관측소(AWS)에서 관측된 기상자료를 사용하였다(Fig 4). 연구지역에 설치된 총 7개의 관측소 중에서 광주남구 관측소의 경우 2018년도부터 관측이 시작된 지점으로 본 연구에 사용하기에 자료가 부족하여 사용하지 않고 충분한 자료가 확보되어 있는 6개 지점의 기상자료만을 본 연구에 사용하였다.

각 관측소는 30m(광산)부터 912m(무등산)까지 다양한 고도에 위치하고 있기 때문에 평야부터 산악지형까지 각기 다른 고도별 기상자료를 활용할 수 있다는 장점이 있다. 본 연구에서는 2010년부터 2021년까지 총 11년 동안의 1시간 단위 강수량 데이터를 활용하였다.

Fig 5는 광주 관측소와 그 외 5개 관측소에서 관측된 강우 데이터의 상관관계를 보여주고 있다. 연 강우량의 경우 상관지수가 약 0.5 – 0.9 사이로 나타났고, 월 강우량의 경우 상관지수가 0.1 – 0.9 사이로 대체로 넓은 범위를 보였다. 월 강우량의 경우 5개 관측소 모두 건기보다는 우기 동안 강우별 상관지수가 높은 것을 확인할 수 있었다.

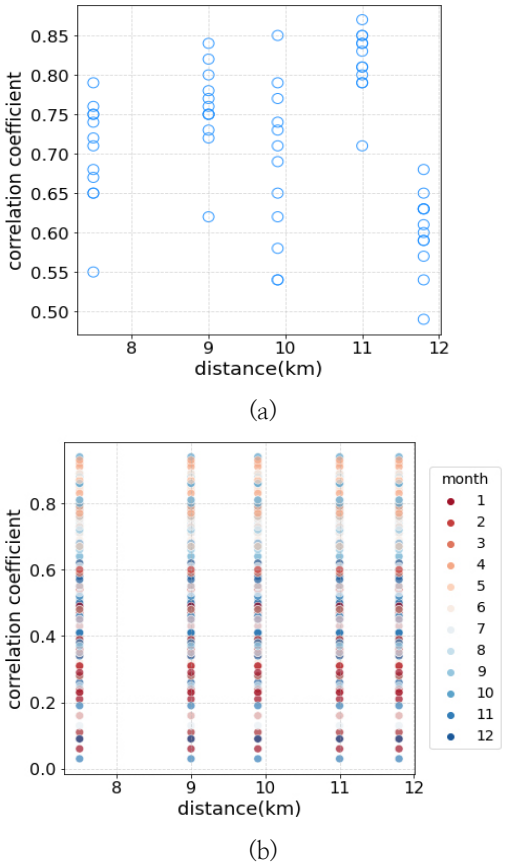


Fig. 5. Scatter plots illustrating the correlation coefficient of (a)annual and (b)monthly precipitation between the Gwangju station and five stations

4. 머신러닝 알고리즘을 이용한 강우 추정

4.1 ANN 및 RF 알고리즘 매개변수 추정

본 연구에서 사용된 2개의 머신러닝 알고리즘인 ANN과 RF를 강우 추정에 적용하기 위해 각 알고리즘의 대표 매개변수를 결정하였다. 먼저 ANN 알고리즘의 경우 batch size와 epoch 개수를 결정하기 위해 batch size는 8 – 96 사이, 그리고 epoch 개수는 10 – 1000 사이의 값을 무작위로 선정하여 최적의 매개변수 조합을 찾는 과정을 수행하였다. 또한 RF 알고리즘의 경우 RandomizedSearchCV 방법을 사용하여 RF 알고리즘의 대표 매개변수인 n_estimators, mex_depth, max_features, 그리고 min_samples_split의 최적값을 찾는 과정을 수행하였다. RandomizedSearchCV란 각 매개변수들 값의 구간을 정의한 뒤, 범위 내의 매개변수 임의값들을 알고리즘에 적용하여 최적의 조합을 찾아내는 방법이다. 가능한 매개변수의 모든 조합을 적용하는 Gridsearch와 달리 RandomizedSearchCV 방법은 시간소비를 최소화 시켜 효율성을 높일 수 있는 방법이다. 위의 과정을 통해 추정 된 두 알고리즘의 매개변수는 Table 2와 같다.

Table 2. Parameters of ANN and RF algorithms

ANN	
number of layers	3
number of neurons	32
batch size	48
epochs	150
RF	
n_estimators	200
min_samples_split	5
max_depth	23
max_features	‘auto’

4.2 강우 발생 추정 성능 비교

본 연구에서는 먼저 ANN, RF 알고리즘이 강우의 발생을 추정하기 위한 활용성을 확인하였다. 이를 위해 2010년 이후 광주 관측소에서 강우의 발생이 관측된 6,811시간에 대해 머신러닝 알고리즘의 강우 유무 추정 성능을 평가하였다. Fig 6는 confusion matrix를 도식화한 그림이다. 그림에서 볼 수 있듯이, RF 알고리즘의 FP, TP, FN, TN은 774, 1862, 174, 1274로 나타났고, ANN 알고리즘은 1061, 1895, 141, 987로 나타났다.

Confusion matrix 결과값을 바탕으로 Accuracy, Recall, Precision, F1 score를 계산한 결과는 Table 3과 같다.

Table 3에 나타났듯이, RF 알고리즘이 ANN 알고리즘보다 강우의 발생유무 분류 정확도가 다소 높은 것을 알 수 있었다. Recall을 제외한 모든 지표에서 높은 성능이 나타났다.

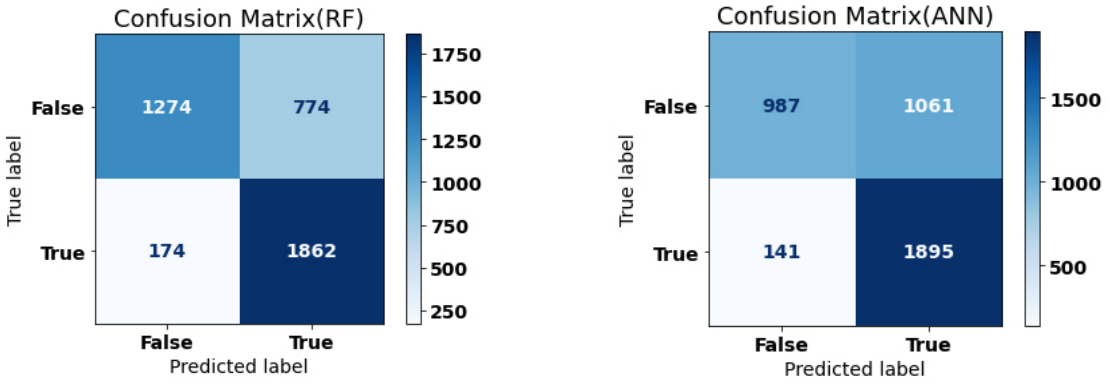
Table 3. Confusion matrix of ANN and RF algorithms.

ANN	
Accuracy	0.71
Recall	0.93
Precision	0.64
F1-score	0.76
RF	
Accuracy	0.77
Recall	0.91
Precision	0.71
F1-score	0.80

Recall은 ANN이 0.93, RF가 0.91로 유사한 값으로 계산되었다.

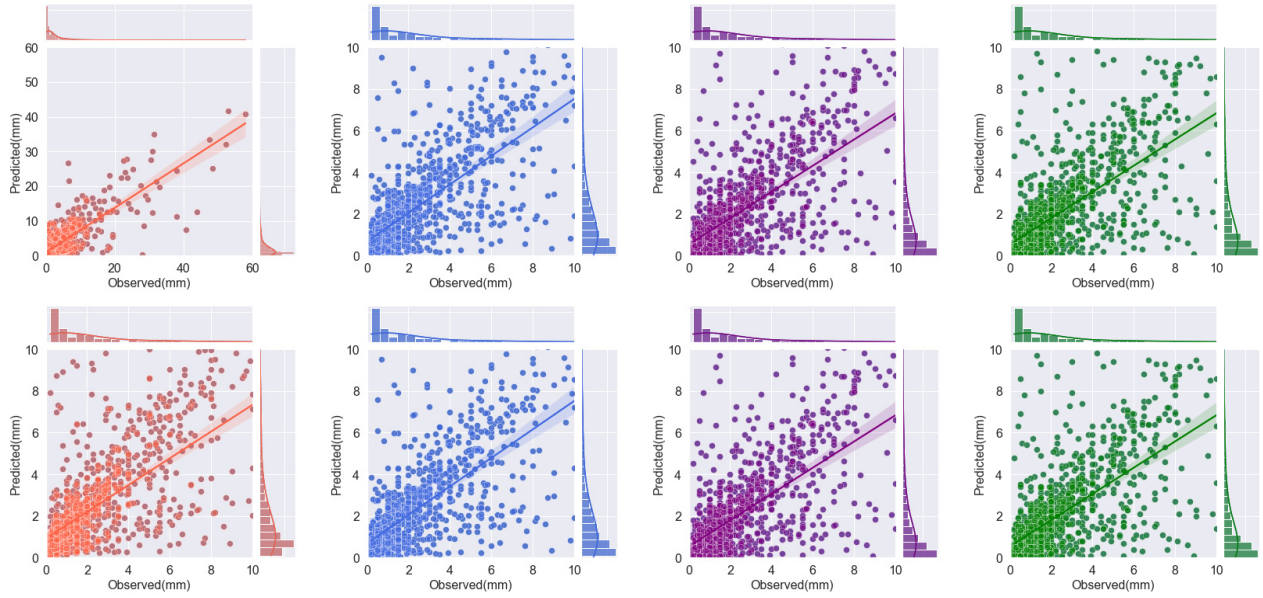
4.3 강우량 추정 성능 비교

본 연구에서 구축한 ANN, RF 알고리즘을 이용하여 강우 추정 성능을 확인하기 위해 기존의 수문학적 방법인 Inverse Distance Method (IDW)와 산술평균법(arithmetic mean; AVG)을 이용하여 산정한 강우 추정 결과와 함께 비교하였다. IDW는 미지의 관측소와 인접한 관측소 사이의 거리의 특성을 가중치로 고려하여 강우값을 추정하는 방법이고, AVG 방법은 인접한 관측소에서 관측된 강우의 평균값을 미지의 관측소에서의 강우값으로 정의하는 방법이다.



(a) (b)

Fig. 6. Classification results of two algorithms (a) RF and (b) ANN



(a) RF (b) ANN (c) IDW (d) AVG

Fig 7. Scatter plots of references and estimated precipitation from (a)RF, (b)ANN, (c)IDW, and (d)AVG. (Upper): precipitation ranged from 0 to 60mm, (Bottom): precipitation ranged from 0 to 10mm.

Fig. 7은 본 연구에서 4가지 방법을 통해 추정한 강우와 실제 관측된 강우 데이터의 scatter plots을 나타내고 있다. 강우량의 전체 범위가 0 – 60mm임을 감안하여 0–10mm 범위와 전체 범위에 대해 구분하여 평가하였다. 전체 강우(0–60mm)의 경우 R^2의 범위가 0.69 – 0.73으로나타났으며, 0–10mm범위에서는 0.51 – 0.55로 나타났다. 특히, RF의 R^2는 0.73인 반면에 IDW와 AVG는 0.70으로 계산되었다. 또한, 4가지 방법에 의해 추정된 강우량의 RMSE 값은 RF가 2.8mm/day이고, ANN가 2.9mm/day인 반면에 IDW와 AVG는 3.1mm/day로 계산되었다. 지표들의 값의 차이가 크지는 않지만, RF의 성능이 기존 방법보다는 다소 우수함을 확인할 수 있었다.

5. 결 론

본 연구에서는 데이터 기반 기술 중 하나인 머신러닝 알고리즘을 활용하여 강우를 추정하고 더 나아가 결측 데이터를 보완할 수 있는 모형에 대한 연구를 수행하였다. Target 지점과 그 주변 강우 데이터 간의 상관관계를 바탕으로 RF와 ANN 알고리즘을 활용하여 비교적 간단하면서 정확도가 높은 알고리즘을 개발하였다.

결측 강우를 산정하기 위해 사용되고 있는 기존의 수문학적 방법 IDW 및 AVG 방법과 비교를 통해 머신러닝 알고리즘의 강우 추정을 위한 활용성을 평가하였다. 이를 위해 강우의 발생 유무뿐만 아니라 강우량을 함께 추정하는 모형을 개발하였다. 강우의 발생 유무를 추정하는 결과는 정확도가 약 0.76 – 0.8 정도로 나타났고, ANN보다는 RF 알고리즘이 다소 높은 정확도를 보였다.

강우량을 추정하는 결과는 기존의 방법의 정확도보다 높거나 유사한 것으로 나타났다. 추정하고자 하는 강우의 양이 작을수록 강우량의 추정 성능은 다소 낮은 것으로 평가되었다. 하지만 기존 방법에 비해 추정 성능이 뒤처지지 않음을 확인하였고, 머신러닝이 결측 강우 복원을 위한 대체 방법으로 활용될 수 있음을 확인할 수 있었다. 본 연구의 결과를 바탕으로 추가적인 데이터와 알고리즘 적용을 통해 강우의 복원 및 예측 정확도를 향상 시킬 수 있는 개선 방법에 대한 추가적인 연구가 요구되는 바이다.

감사의 글

이 논문은 2022학년도 조선대학교 학술연구비의 지원을 받아 연구되었음.

References

Bellido-Jiménez, J. A., Gualda, J. E., & García-Marín, A. P. (2021). Assessing machine learning models for gap filling daily rainfall series in a semiarid region of Spain. Atmos. Vol. 12, pp. 1158. https://doi.org/10.3390/atmos12091158

Breiman, L. (2001). Random forests. Mach. Learn. Vol. 45, pp. 5–32.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... and Michaelsen, J. (2015). The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. Sci. Data,

Vol. 2, pp. 1-21. https://doi.org/10.1038/sdata.2015.66

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., ... and Stocker, E. F. (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. J. Hydrometeorol., Vol. 8, pp. 38-55. https://doi.org/10.1175/JHM560.1

Hwang, S, Kang, N, & Yoon, J. (2022). Error Generation Characteristics of the Areal Rainfall Estimation Interpolation Method Using Rainfall Radar Data. J. Korean Soc. Hazard Mitigation, Vol. 22, pp. 273-283. https://doi.org/10.9798/KOSHAM.2022.22.6.273 [Korean Literature]

Jang, S., Yoon, S., Lee, S., Lee, T., & Park, K. (2018). Evaluation of drought monitoring using satellite precipitation for un-gaged basins. J. Korean Soc. Agric. Eng. Vol. 60, pp. 55-63. [Korean Literature]

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P. (2004) CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrometeorol., Vol. 5, pp. 487-503. https://doi.org/10.1175/1525-7541(2004)005%3C0487:CAMTPG%3E2.0.CO;2

Kang, N. R., Noh, H. S., Lee, J. S., Lim, S. H., & Kim, H. S. (2013). Runoff simulation of an urban drainage system using radar rainfall data. J. Wetlands Res. Vol. 15, pp. 413-422. [Korean Literature]

Kim, J., Han, H., Johnson, L. E., Lim, S., & Cifelli, R. (2019). Hybrid machine learning framework for hydrological assessment. J. Hydrol. 577, 123913. https://doi.org/10.1016/j.jhydrol.2019.123913

Kim, B. S., Hong, J. B., Kim, H. S., & Yoon, S. Y. (2007). Combining radar and rain gauge rainfall estimates for flood forecasting using conditional merging method. In World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat (pp. 1-16).

Kim, S. J., Shin, S. C., & Suh, A. S. (1999). Satellite rainfall monitoring: recent progress and its potential applicability. Korean Journal of Agricultural and Forest Meteorology, Vol. 1, pp. 142-150. [Korean Literature]

Kim, J., & Han, H. (2021). Evaluation of the CMORPH high-resolution precipitation product for hydrological applications over South Korea. Atmos. Res. 258, 105650. https://doi.org/10.1016/j.atmosres.2021.105650

Kim, K. T, & Kim, J. H. (2013). Introduction of rainfall observation data and utilization cases using artificial satellites. Water and the Future: J. Korean Soc. Water Res. Vol. 46, pp. 66-75. [Korean Literature]

Kwon, T., Yoon, S., Shin, H., & Yoon, S. (2022). The selection of radar merging by watershed for quantitative precipitation estimation: At the Han river basin. Korean Data Inf. Sci. Soc. Vol. 33, pp. 1021-1030. https://doi.org/10.7465/jkdi.2022.33.6.1021 [Korean Literature]

Larson, L. W., & Peck, E. L. (1974). Accuracy of precipitation measurements for hydrologic modeling. Water Resour. Res. Vol. 10, pp. 857-863. https://doi.org/10.1029/WR010i004p00857

New, M., Todd, M., Hulme, M., & Jones, P. (2001). Precipitation measurements and trends in the twentieth century. Int. J. Climatol.: J. R. Meteorolog. Soc. Vol. 21, pp. 1889-1922. https://doi.org/10.1002/joc.680

Mosaffa, H., Sadeghi, M., Mallakpour, I., Jahromi, M. N., & Pourghasemi, H. R. (2022). Application of machine learning algorithms in hydrology. In Comput. earth and Environ. Sci. pp. 585-591. https://doi.org/10.1016/B978-0-323-89861-4.00027-0

Oh, S., Kim, W., Kang, M., Yoon, H., Yang, J., & Choi, M. (2022). An analysis of land displacements in terms of hydrologic aspect: satellite-based precipitation and groundwater levels. J. Korea Water Res. Assoc. Vol. 55, pp. 1031-1039. 10.3741/JKWRA.2022.55.12.1031 [Korean Literature]

Park, K., Niu, S., & Kim, Y. (2013). Reduction efficiency of the stormwater wetland from animal feeding-lot. J. Wetlands Res. Vol. 15, pp. 79-90. [Korean Literature]

Polishchuk, B., Berko, A., Chyrun, L., Bublyk, M., & Schuchmann, V. (2021). The rain prediction in Australia based Big Data analysis and machine learning technology. In 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), Vol. 1, pp. 97-100.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, Vol. 65, pp. 386.

Sadat-Noori, M., Glamore, W., & Khojasteh, D. (2020). Groundwater level prediction using genetic programming: the importance of precipitation data and weather station location on model accuracy. Environ. Earth Sci. Vol. 79, pp. 1-10. https://doi.org/10.1007/s12665-019-8776-0

Sahoo, A., & Ghose, D. K. (2022). Imputation of missing precipitation data using KNN, SOM, RF, and FNN. Soft Computing, Vol. 26, pp. 5919-5936. https://doi.org/10.1007/s00500-022-07029-4

Shen, C., Chen, X., & Laloy, E. (2021). Broadening the use of machine learning in hydrology. Frontiers in Water, 3, 681023. https://doi.org/10.3389/frwa.2021.681023

Sorooshian, S., Hsu, K. L., Gao, X., Gupta, H. V., Imam,

B., and Braithwaite, D. (2000) Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. Bull. Am. Meteorol. Soc., Vol. 81, pp. 2035-2046. https://doi.org/10.1175/1520-0477(2000)081%3C2035:EOPSSE%3E2.3.CO;2

Teegavarapu, R. S., & Pathak, C. (2008). Infilling of rain gage records using radar (NEXRAD) data: Influence of spatial and temporal variability of rainfall processes. In World Environmental and Water Resources Congress 2008: Ahupua'A pp. 1-9. https://doi.org/10.1061/40976(316)406

Tyralis, H., Papacharalmpous, G., and Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, Vol. 11, pp. 910. https://doi.org/10.3390/w11050910