Journal of Environmental Science International
28(8); 701~707; August 2019

pISSN: 1225-4517 eISSN: 2287-3503
https://doi.org/10.5322/JESI.2019.28.8.701

TECHNICAL NOTE

기상 및 토양정보가 고랭지배추 단수예측에 미치는 영향

권태용 · 김래용[1] · 윤상후[1]*

대구대학교 일반대학원 통계학과, [1]대구대학교 수리빅데이터학부

The Effect of Highland Weather and Soil Information on the Prediction of Chinese Cabbage Weight

Taeyong Kwon, Rae Yong Kim[1], Sanghoo Yoon[1]*

Department of Statistics, Daegu University, Gyeongsan 38453, Korea

[1]Division of Mathematics and big data science, Daegu University, Gyeongsan 38453, Korea

Abstract

Highland farming is agriculture that takes place 400 m above sea level and typically involves both low temperatures and long sunshine hours. Most highland Chinese cabbages are harvested in the Gangwon province. The Ubiquitous Sensor Network (USN) has been deployed to observe Chinese cabbages growth because of the lack of installed weather stations in the highlands. Five representative Chinese cabbage cultivation spots were selected for USN and meteorological data collection between 2015 and 2017. The purpose of this study is to develop a weight prediction model for Chinese cabbages using the meteorological and growth data that were collected one week prior. Both a regression and random forest model were considered for this study, with the regression assumptions being satisfied. The Root Mean Square Error (RMSE) was used to evaluate the predictive performance of the models. The variables influencing the weight of cabbage were the number of cabbage leaves, wind speed, precipitation and soil electrical conductivity in the regression model. In the random forest model, cabbage width, the number of cabbage leaves, soil temperature, precipitation, temperature, soil moisture at a depth of 30 cm, cabbage leaf width, soil electrical conductivity, humidity, and cabbage leaf length were screened. The RMSE of the random forest model was 265.478, a value that was relatively lower than that of the regression model (404.493); this is because the random forest model could explain nonlinearity.

Key words : Chinese cabbage, Random forest, Regression model, Ubiquitous sensor network

1. 서 론

농산업은 기상에 많은 영향을 받으므로 농작물을 재배하는 농업인의 위험을 줄이기 위해선 정확한 기상정보가 제공되어야 한다. 한반도의 기상정보는 기상재해로부터 국민의 생명과 재산을 보호하고 공공의 복리 증진을 목적으로 기상청에서 수집하고 있다. 기상정보 수집 목적이 한반도의 전반적인 대기 현상 파악이므로 농산업에 해당하는 작물재배지 기상정보 생산에는 한계점이 있다. 예를 들어 고랭지배추는 해발고도가 높은 강원도에서 재배

Received 16 May, 2019; **Revised** 1 July, 2019;
Accepted 21 July, 2019
***Corresponding author**: Sanghoo Yoon, Division of Mathematics and big data science, Daegu University, Gyeongsan 38453, Korea
Phone : +82-53-850-6421
E-mail : statstar@daegu.ac.kr

되고 있으나 강원도에 설치된 기상정보의 최대해발고도는 대관령이 772 m로 800 m 이상 해발고도에서 재배되는 고랭지배추의 농업기상정보 생산에는 어려움이 있다. 이에 해발고도가 950 m 이상인 고랭지배추 재배지의 기상 및 토양정보를 유비쿼터스 네트워크 장비(Ubiquitous Sensor Network, USN)로 수집하고 있다(Cho et al., 2018).

고랭지 농업은 해발고도 400 m 이상의 고원이나 산지 등에서 이루어지는 농업이다. 고랭지 지역은 기온이 낮고 일조시간이 길어 감자, 옥수수, 배추 등이 주로 재배된다. 이 중 배추는 우리나라 4대 채소 중 하나로 김치의 소비량이 많은 한국 경제의 가격안정을 위해 생산량 관리가 필요한 작물이다. 강원도의 고랭지 밭에서 생산되는 배추는 전국 생산량의 93%를 차지하고 있는데 이는 강원도의 심한 일교차가 배추의 맛과 식감에 긍정적인 영향을 미치기 때문이다.

배추의 가격관리를 위해선 배추의 단수예측모형이 필요하다. Kim et al.(2015)은 일 최고온도와 일 최저온도를 이용한 생육도일을 통해 구중을 예측하는 회귀모형을 제시하였다. Kim and Yun(2015)은 비선형 회귀모형을 이용하여 온도구간에 따른 배추의 생육을 정량화하였다. 배추의 생육은 생육기와 결구기로 나누었으며 토양수분이나 일사량 등은 평년수준으로 간주하여 모형을 설계하였다. Ahn et al.(2014)은 고랭지배추의 수량을 온도, 습도, 강수량 등의 기상요인을 활용한 로지스틱 회귀모형으로 구축하였다. 배추의 단수예측모형은 주로 회귀모형을 기반으로 개발되었다. 회귀모형은 자료의 복잡한 비선형성을 반영하지 못하므로 서포트 벡터 머신, 랜덤 포레스트, 신경망 모형 등의 기계학습을 통해 이를 해결하기 위한 연구가 수행되고 있다.

기계학습은 자료의 양이 증가할수록 자료의 복잡한 비선형성이 훈련되어 종속변수의 예측 정확도가 높아진다. 기계학습 중 랜덤 포레스트는 분류와 회귀 분석 등에 사용되는 앙상블 학습 방법으로 다수의 결정 트리로부터 비선형의 평균 예측치를 생산할 수 있다.Oh(2013)는 환자의 재원일수에 영향을 미치는 환자의 주요특성을 랜덤 포레스트로 도출하였고, Min et al.(2017)은 대전시 공공 자전거 수요 예측모형을 랜덤 포레스트로 개발하였으며, Kim(2017)은 태풍 발생 여부를 랜덤 포레스트로 연구하였다.

본 연구에서는 고랭지배추 주산지에 설치된 USN 장비에서 수집되는 기상정보와 생육정보를 이용하여 고랭지배추 단수예측모형을 개발하고자 한다. 단수예측모형으로 회귀모형과 랜덤 포레스트가 고려되었다. 모형의 예측성능을 평가하기 위해 예측값과 관측값의 평균제곱오차(Root Mean Squared Error, RMSE)를 계산하였다.

2. 연구 방법

선행연구에서 배추의 단수예측을 위한 구중은 주로 회귀모형으로 개발되었다(Ahn et al., 2014; Kim et al., 2015; Kim and Yun, 2015). 본 연구에서는 선행연구에서 사용된 회귀모형 외에 랜덤 포레스트를 이용하여 배추의 단수를 예측하고자 한다. 단수예측의 실효성을 위해 1주일 전에 수집된 기상자료와 생육자료를 모형에 사용하여 배추재배 시기별 배추의 구중을 예측하였다.

연구에 사용된 배추의 생육정보에 대한 측정방법 및 용어는 다음과 같다. 연구에 사용된 생육정보는 엽장(cabbage leaf length), 엽수(the number of cabbage leaves), 엽폭(cabbage leaf width), 주폭(cabbage width) 그리고 구중(cabbage weight)이다. 구중은 배추의 무게(kg), 엽장은 배추 겉잎의 세로 길이(cm), 엽수는 1 cm 미만인 잎을 제외한 배춧잎의 개수(매)이다. 엽폭은 배추의 겉잎에서 가장 넓은 부분의 가로길이(cm)이고, 주폭은 배추의 전체 가로 폭의 길이(cm)이다.

2.1. 유비쿼터스 네트워크 장비 (USN)

USN은 센서 네트워크를 이용하여 유비쿼터스 환경을 구축하기 위한 기술이다. 다양한 센서에서 정보를 수집할 수 있으므로 USN 기술을 통해 기상관측 및 농작 환경 감시 등을 위한 통합관측환경을 구축할 수 있다. 농촌경제연구원에서는 10개의 센서로 고랭지배추 재배지의 기상정보 및 토양정보를 수집하여 실시간으로 모니터링하는 시스템을 2015년 구축하였다.

2.2. 회귀모형

회귀모형은 종속변수와 독립변수 간에 존재하는 선형관련성을 분석하기 위한 모형이다. 본 연구에서는 작물의 생육에 영향을 미치는 기상요인(X)을 독립변수로 사용하였고 고랭지 배추의 수확량을 의미하는 구중을 종속

Table 1. The pearson correlation coefficients between the weight of cabbage and growth variables

	Cabbage leaf length	The number of cabbage leaves	Cabbage leaf width	Cabbage width
Cabbage weight	0.333	0.555	0.381	0.458

변수(Y)로 사용하였다. 종속변수와 여러 개의 독립변수 사이의 직선관계식은 식(1)과 같다.

$$Y=\beta_0+\beta_1 X_1+\cdots+\beta_i X_i, \quad (1)$$

여기서 β_0는 절편을 나타내며, $\beta_1, \cdots, \beta_i$는 기울기를 나타낸다. 회귀모형의 적합 정도는 결정계수(R^2)가 사용되며, 식(2)와 같이 계산된다.

$$R^2=1-\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}_i\right)^2}, \quad (2)$$

여기서 n은 관측 자료의 수이며, k는 독립변수의 수이다. 결정계수는 항상 $0 \le R^2 \le 1$의 값을 가진다.

적합한 회귀모형을 위해선 3가지 가정을 만족해야 한다. 첫 번째로 회귀모형을 구성하는 오차의 정규성이다. 본 연구의 오차의 정규성은 Shapiro-Wilk 검정으로 확인하였다. 두 번째로 회귀모형을 구성하고 있는 오차들은 독립성을 만족해야 한다. 일반적으로 Durbin-Watson 검정 통계량을 통해 오차들의 자기상관성을 확인한다. 마지막으로 오차의 등분산성이다. 오차들의 등분산성 가정을 확인하기 위해 spread-level plot을 통해 오차의 분산이 변하는지 확인하였다.

2.3. 랜덤 포레스트

Breiman(2001)이 제안한 랜덤 포레스트(random forest)는 여러 개의 의사결정나무들로 구성된 앙상블 모형이다. 앙상블이란 여러 개의 예측모형을 만든 후 이 예측모형들을 하나로 결합하여 최종 하나의 예측 결과를 생성하는 기법이다. 대표적인 앙상블 기법으로 배깅(bagging)과 부스팅(boosting)이 있다. 배깅은 bootstrap aggregating의 약자로 주어진 자료로부터 부스트랩 표본을 여러 번 생성하여 각 부스트랩 표본의 예측 결과를 결합하여 최종 예측모형을 만드는 방식이다. 부스트랩 표본의 의사결정나무의 상관성이 낮을수록 예측오차가 작아지므로 일반적으로 의사결정나무보다 더 좋은 성능을 보인다. 본 연구에서는 포트란으로 구현된 Breiman의 배깅코드를 R로 개발한 'randomForest' 패키지를 이용하였다(Liaw et al., 2002).

2.4. 연구 자료

본 연구에서 사용된 자료는 고랭지 배추의 생육을 관리하기 위해 설치된 USN 장비 5개소에서 수집된 기상 및 토양정보다. USN 장비는 강릉시의 안반덕 3곳(해발고도 1,057 m, 1,052 m, 1,006 m), 태백시의 귀네미 1곳(해발고도 953 m)과 매봉산 1곳(해발고도 1,156 m)에 설치되어 2015년부터 자료를 수집하고 있다(Fig. 1).

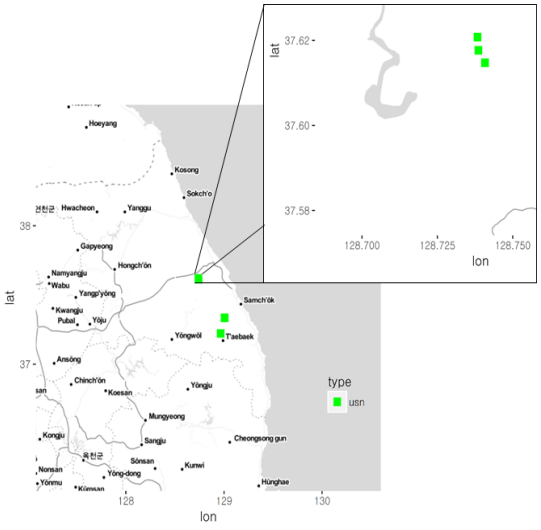

Fig. 1. The location of ubiquitous sensor network.

고랭지배추 생육에 영향을 미치는 요인을 센서로 수집하기 위한 USN은 재배지의 가장자리에 펜스와 함께 설치되어 있다(Fig. 2). 설치장소는 고도가 높아 벼락의 위험성이 있으며, 낮은 기온과 높은 습도로 전자 장비에

Table 2. The pearson correlation coefficient between the weight of cabbage and meteorological variables

	1d ago	2d ago	3d ago	1d-2d average	1d-3d average	7d average
Soil moisture 30 cm	0.410	0.340	0.313	0.415	0.409	**0.513**
Soil moisture 60 cm	0.295	0.164	0.137	0.262	0.261	**0.406**
Soil moisture 90 cm	0.317	0.181	0.230	0.245	0.277	**0.404**
Soil temperature	-0.506	**-0.542**	-0.497	-0.493	-0.504	-0.475
Insolagtion	-0.033	**0.225**	-0.156	0.100	-0.022	0.102
Humidity	-0.153	-0.104	0.074	-0.233	-0.184	**-0.255**
Temperature	-0.349	**-0.556**	-0.436	-0.416	-0.437	-0.308
Wind speed	0.032	0.139	**0.155**	0.119	0.094	-0.085
Precipitation	-0.422	-0.374	-0.130	-0.454	-0.449	**-0.512**
Soil electrical conductivity	-0.152	**-0.221**	-0.193	-0.163	-0.175	-0.116

수분으로 인한 이상이 발생할 가능성이 존재한다. Cho et al.(2018)은 USN으로 수집된 고랭지배추 재배지의 품질관리 알고리즘을 개발하였다. 본 연구는 품질관리 알고리즘으로 정상자료와 비정상자료로 구분한 후, 분석의 편의성을 위해 1시간 자료를 1일 단위로 변환하였다. 강우량은 누적값이 사용되었고, 강우량을 제외한 기상요인과 토양요인은 평균값이 사용되었다.

Fig. 2. The ubiquitous sensor networks established for chinese cabbage.

측정된 기상자료는 일사량, 습도, 대기온도, 풍속, 강우량이고 토양자료는 토양수분 30 cm, 토양수분 60 cm, 토양수분 90 cm, 토양온도, 토양전도율이다. 연구기간은 총 3년으로 2015년부터 2017년까지 수집된 자료를 사용하였다. 강원도 고랭지 배추는 일반적으로 9월부터 수확하므로 수확 시기에 따른 배추의 생육 예측을 위해 7월에서 10월까지 USN 장비에서 수집된 자료를 이용하였다.

3. 결과 및 고찰

고랭지배추 생육에 영향을 미치는 기상 및 토양정보는 당일보다는 이전 시점의 정보가 영향을 미친다. 본 연구에서는 1일 전부터 7일 전까지 기상 및 토양변수와 생육변수 간 상관성을 분석하여 상관성이 높게 나온 1일, 2일, 3일, 그리고 7일 기상 및 토양정보를 변수로 생성하여 분석을 진행하였다. 일주일 전 생육변수가 일주일 후 단수를 의미하는 구중에 미치는 영향을 살펴보기 위해 상관분석을 실시한 결과는 Table 1이다. 엽수(r=.555), 주폭(r=.458), 엽폭(r=.381), 엽장(r=.333) 순으로 구중과 상관성을 보이고 있다. 일주일 동안 관측된 기상 및 토양변수가 고랭지배추의 구중에 미치는 영향을 상관계수로 표현하면 Table 2이다. 구중과 상관성이 높게 나온 변수를 정리하면 7일 평균 토양수분 30 cm, 7일 평균 토양수분 60 cm, 7일 평균 토양수분 90 cm, 2일 전 토양온도, 2일 전 일사량, 7일 평균 습도, 2일 전 온도, 3일 전 풍속, 7일 평균 강수량, 2일 전 토양전도율이다. 본 연구에서는 구중에 가장 높은 상관관계가 나타나는 변수를 선택하여 회귀모형을 세웠다.

Table 3. The result of the regression model for the weight of cabbage

	Estimate	S. E.	Stand. Est.	t	Pr(>|t|)	VIF
(Intercept)	7315.663	3037.321		2.409	0.025***	
The number of cabbage leaves	**29.061**	**8.693**	**0.667**	**3.343**	**0.003*** **	**3.514**
Cabbage leaf width	-83.998	79.480	-0.209	-1.057	0.303	3.440
Soil moisture 30 cm	-5.666	7.706	-0.180	-0.735	0.470	5.272
Soil moisture 60 cm	6.206	5.449	0.204	1.139	0.268	2.837
Soil temperature	-101.175	96.736	-0.219	-1.046	0.308	3.861
Insolagtion	-152.566	1289.612	-0.014	-0.118	0.907	1.303
Humidity	-23.335	19.674	-0.182	-1.186	0.249	2.079
Temperature	-45.464	57.449	-0.168	-0.791	0.438	3.988
Wind speed	**413.782**	**187.170**	**0.297**	**2.211**	**0.038***	**1.591**
Precipitation	**-2805.035**	**1121.137**	**-0.450**	**-2.502**	**0.021***	**2.854**
Soil electrical conductivity	**-81.493**	**33.210**	**-0.333**	**-2.454**	**0.023***	**1.625**
F=6.106, p-value<0.001, R^2=0.762, adj R^2=0.637						

*p<0.05, **p<0.01, ***p<0.001

일주일 전 생육변수와 일주일간 기상 및 토양변수로 구중을 예측하는 회귀모형을 세우면 Table 3이다. 회귀모형은 통계적으로 유의미하였고(F=6.106, p<0.001), 결정계수(R^2)는 0.762로 생육변수와 기상변수로 구중의 76.2%가 설명되고 있다. 회귀모형의 오차는 정규성(W=0.977, p=0.681), 독립성(D-W=2.093, p=0.291), 그리고 등분산성(χ^2=0.052, p=0.819)을 모두 만족하였다. 추가로 회귀모형의 다중공선성을 파악하기 위해 분산팽창지수(Variance Inflation Factor, VIF)를 확인하였다. 엽장, 주폭, 토양수분 90 cm는 다중공선성이 존재하여 제거하였다. 이를 제외한 모든 값이 10보다 작아 독립변수 간 상관성이 존재하지 않았다. 통계적으로 유의미한 독립변수는 일주일 전 엽수(t=3.343, p=0.003), 3일 전 풍속(t=2.211, p=0.038), 7일 평균 강수량(t=-2.502, p=0.021), 2일 전 토양전도율(t=-2.454, p=0.023)이다.

본 연구에서는 회귀모형 외 랜덤 포레스트로 배추 단수예측모형의 비선형성을 살펴보았다. 랜덤 포레스트의 최적화를 위해 의사결정나무의 수(ntree)와 고려할 변수의 개수(mrty)를 조율(tune)하였다. 조율된 의사결정나무의 수는 200이고 의사결정나무의 변수의 수는 8이다. 추가로 구중에 영향을 미치는 중요 변수를 찾기 위해 랜덤 포레스트를 통해 일주일 전 생육변수와 일주일 동안 기상 및 토양변수의 평균제곱편차(Mean Square Error, MSE) 차이를 변수중요도를 통해 살펴보았다(Fig. 3). 변수의 중요도는 주폭, 엽수, 2일 전 토양온도 순으로 높았고, 7일 평균 강수량, 2일 전 온도, 7일 평균 토양수분 30 cm, 엽폭, 2일 전 토양전도율, 7일 평균 습도, 엽장 순서로 중요했다.

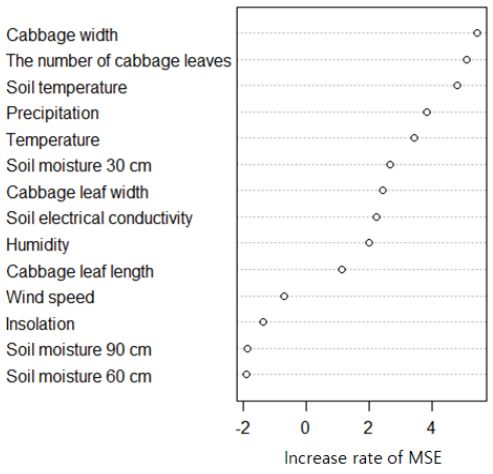

Fig. 3. The importance of variables in the random forest.

생육변수와 기상 및 토양변수로 고랭지배추의 구중을 회귀모형과 랜덤 포레스트로 세웠을 때 예측성능을 비교

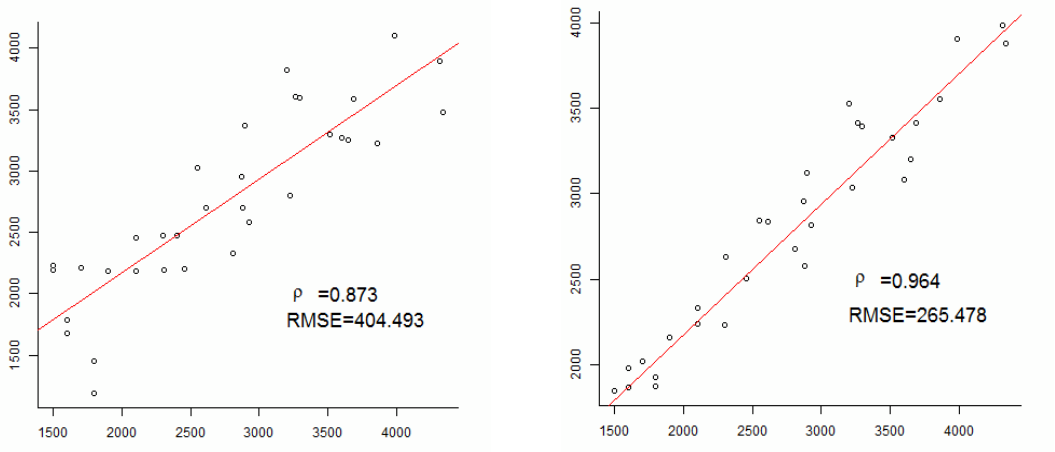

Fig. 4. Comparison of a linear regression model and a random forest.

Table 4. The comparison of the importance of variables

Variable importance	Linear regression model	Random forest
1	**The number of cabbage leaves**	Cabbage width
2	**Precipitation**	The number of cabbage leaves
3	**Soil electrical conductivity**	Soil temperature
4	**Wind speed**	Precipitation
5	Soil temperature	Temperature
6	Cabbage leaf width	Soil moisture 30 cm
7	Soil moisture 60 cm	Cabbage leaf width
8	Humidity	Soil electrical conductivity
9	Soil moisture 30 cm	Humidity
10	Temperature	Cabbage leaf length
11	Insolation	Wind speed
12		Insolation
13		Soil moisture 90 cm
14		Soil moisture 60 cm
RMSE	404.493	265.478

하기 위해 구중의 관측값과 예측값의 산점도와 RMSE를 그리면 Fig. 4이다. 랜덤 포레스트의 RMSE는 265.5이고 회귀모형의 RMSE는 404.5로 비선형성을 반영할 수 있는 랜덤 포레스트가 회귀모형보다 예측성능이 상대적으로 우수하였다.

구중과 일주일 전 생육변수 및 일주일간 기상과 토양변수의 회귀모형과 랜덤 포레스트를 비교하면 Table 4이다. 회귀모형의 토양전도율은 통계적으로 유의미했으나, 랜덤 포레스트에서는 중요도가 낮은 변수로 나타났다. 엽수와 7일 평균 강수량은 두 모형에서 중요 변수로 확인된다. 단수예측모형의 RMSE를 비교한 결과 비선형성이 고려된 랜덤 포레스트가 배추의 단수예측모형에

적절하다고 판단된다.

Kim et al.(2015)은 온도기반의 생육도일과 배추의 생육변수(엽폭, 초고, 구고, 구폭)로 구중을 예측하는 다중회귀모형을 제안하였다. 본 연구는 온도 외에도 USN 장비에서 관측되는 다양한 기상정보와 토양정보를 이용하여 구중 예측모형을 세웠다. 분석 결과 종속변수와 독립변수 간 비선형 관계를 설명할 수 있는 랜덤 포레스트가 구중 예측에 적절한 모형으로 판단된다. 추가로 변수중요도를 통해 향후 구중 예측에 필요한 기상 및 토양변수를 제시하였다.

4. 결 론

본 논문에서는 고랭지배추 주산지 5개 지점에 설치된 USN 장비에서 수집된 기상 및 토양정보와 배추 생육정보를 이용하여 고랭지배추의 단수예측모형을 세웠다. 단수예측모형으로 회귀모형과 랜덤 포레스트가 고려되었으며 RMSE를 통해 모형의 성능을 비교하였다. 구중에 영향을 미치는 중요변수는 회귀모형에서 엽수, 풍속, 강수량, 토양전도율이고 랜덤 포레스트에서 주폭, 엽수, 토양온도, 강수량, 온도, 토양수분 30 cm, 엽폭, 토양전도율, 습도, 엽장 순이다. 엽수와 강수량은 회귀모형과 랜덤 포레스트에서 모두 높은 영향력을 미치는 변수이므로 배추의 단수예측에 중요한 요인으로 판단된다. RMSE의 경우 회귀모형은 404.493이고 랜덤 포레스트는 265.478로, 랜덤 포레스트가 고랭지배추 단수예측모형에 적절하였다.

REFERENCES

Ahn, J. H., Kim, K. D., Lee, J. T., 2014, Growth modeling of chinese cabbage in an alpine area, Kor. J. of Agric. and Forest Meteorology, 16, 309-315.

Brieman, L., 2001, Random forests, Machine Learning, 45, 5-32.

Cho, C., Hwang, G., Yoon, S., 2018, Development ubiquitous sensor network quality control algorithm for highland cabbage, Kor. J. of Agric. and Forest Meteorology, 20, 337-347.

Jin, J. H., Oh, M. A., 2013, Data analysis of hospitalization of patients with automobile insurance and health insurance: a report on the patient survery, J. of the Kor. Data Anal. Soc., 15, 2457-2471.

Kim, H. S., 2017, Typhoon qccurrence prediction using random forest technique, Kor. Meteorological Soc., 162-163.

Kim, J. H., Yun, J. I., 2015, A Thermal time - based phenology estimation in Kimchi cabbage, Kor. J. of Agric. and Forest Meteorology, 17, 333-339.

Kim, K. D., Suh, J. T., Lee, J. N., Yoo, D. L., Kwon, M., Hong, S. C., 2015, Evaluation of factors related to productivity and yield estimation based on growth characteristics and growing degree days in highland Kimchi cabbage, Kor. Soc. for Hort. Sci., 33, 911-922.

Liaw, A., Wiener, M., 2002, Classification and regression by randomForest, R news, 2, 18-22.

Min, J. W., Mun, H. S., Lee, Y. S., 2017, Demand forecast for public bicycles (“Tashu”) in Daejeon using random forest, Kor. Infor. Sci. Soc., 969-971.

• 권태용, 대구대학교 일반대학원 통계학과 박사과정
hero6504@naver.com
• 김래용, 대구대학교 수리빅데이터학부 조교수
raeyongkim@daegu.ac.kr
• 윤상후, 대구대학교 수리빅데이터학부 조교수
statstar@daegu.ac.kr