Check for updates

Journal of KIIT. Vol. 22, No. 9, pp. 11-22, Sep. 30, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.9.11

고랭지배추의 생육 시기별 주산지 기상정보를 활용한 출하시기 도매가격 예측모형

허다솜*, 정승권**

Wholesale Price Prediction Model of Highland Cabbage at the Time of Shipment using Weather Information from the Main Production Area by Growth Season

Dasom Hur*, Seung Kwon Jung**

본 원고는 행정안전부 재난안전 공동연구 기술개발사업의 지원을 받아 수행된 연구
(2022-MOIS63-001(RS-2022-ND641011))의 내용을 기반으로 작성되었습니다

요 약

가뭄은 비정상적으로 적은 강우량이 장기간 지속되면서 물이 부족하게 됨에 따라 물 부족으로 토양의 수분율을 감소시켜 배추재배에 직간접적인 영향을 미칠 수 있다. 본 연구에서는 고랭지배추의 생육 시기를 파종기, 정식기, 수확기로 나누어 도매가격에 미치는 영향을 분석한 뒤 SVM(Support Vector Machine), Random Forest, Gradient Boosting 분류모델을 기반으로 고랭지배추 전년 동일 대비 도매가격 변동률을 예측하여 최적의 모델과 변수를 제시하였다. 고랭지배추 도매가격 예측 결과, 정확도가 가장 높았던 모델은 강릉과 대관령 모두 Random Forest였다. 생육 시기 중 고랭지배추에 가장 큰 영향을 주는 시기는 생육기이며 변수는 SPI(Standardized Precipitation Index)1, SPI2인 것으로 분석되었다. 테스트 성능은 세 모델 중 Random Forest가 가장 우수한 성능을 보였다. 또한 두 지역 모두 지역별 중요도가 높은 입력변수 10개 정도의 사용으로도 과적합을 방지하며 예측을 수행할 수 있음을 확인하였다.

Abstract

Drought can directly impact highland cabbage yields by reducing soil moisture due to abnormally prolonged low rainfall. Hence, this study examines the significance of climate factors on cabbage wholesale price yearly changes and presents prediction models based on SVM, Random Forest, and Gradient Boosting classification. While the training performance of the highland cabbage wholesale price yearly price change prediction model showed high accuracy for Gradient Boosting, the best test performance among the three models was observed in Random Forest. Additionally, it was confirmed that using around 10 highly significant input variables in both regions could prevent overfitting and enable effective predictions.

Keywords

classification model, drought impact, growth stages, highland cabbage, wholesale price prediction

* (재)국제도시물정보과학연구원 연구원
- ORCID: https://orcid.org/0009-0003-1290-8841
** (재)국제도시물정보과학연구원 연구위원(교신저자)
- ORCID: https://orcid.org/0000-0003-3961-0570

· Received: May 27, 2024, Revised: Sep. 12, 2024, Accepted: Sep. 15, 2024
· Corresponding Author: Seung Kwon Jung
International Center for Urban Water Hydroinformatics Research & Innovation
Tel.: +82-32-852-5731, Email: skjung6779@gmail.com

Ⅰ. 서 론

무, 고추, 마늘과 함께 우리나라 4대 채소로 꼽히는 배추는 김치의 주재료로 2022년 배추 생산량은 221만 1천 톤으로 추정된다. 작형 별로 봄배추, 여름 배추, 가을배추, 겨울 배추로 구분되며 여름에는 고랭지배추가 생산되어 도매시장에 유통된다. 강원도는 한국 고랭지배추 생산량의 85.6%를 차지하며 고랭지배추 재배면적도 90.1%로 매우 집중되어 한국 고랭지배추 시세에 미치는 영향이 가장 큰 지역이다[1].

가뭄은 비정상적으로 적은 강우량이 장기간 지속되면서 물이 부족하게 됨에 따라 물 부족으로 토양의 수분율을 감소시켜 토양의 영양분 섭취 저하 및 성장 지연, 조기에 씨앗을 생산하여 상품 가치 하락, 불규칙한 물 공급으로 잎 가장자리가 갈색으로 변해 죽는 팁번(Tip burn) 발생, 해충 및 질병에 더욱 취약해짐 등 배추재배에 직간접적인 영향을 미칠 수 있다. 생산량은 도매가격 형성에 영향을 줄 수 있으며 가뭄 상황에 따른 배추 가격의 민감성은 농업 경제와 소비자 그리고 국가의 농업 및 식품 정책에 큰 영향을 미칠 수 있다.

따라서 본 연구는 생육시기별 주산지 기상정보를 활용한 고랭지배추 도매가격 예측을 목적으로 한다. 본 연구결과는 주산지 기상정보가 농산물 생산에 미치는 영향을 분석하는데 활용될 수 있을 것이며, 가뭄 등 기상재해 발생 시 기상정보를 활용하여 농산물의 생산량과 도매가격의 변동률을 분석하여 대응대책 수립에 기여할 수 있을 것이다.

본 논문의 구성을 다음과 같다. 2장은 관련연구로 통계적 방법을 이용한 농산물 가격 및 생산량 예측연구와 머신러닝 기법을 이용한 농산물 가격 및 생산량 예측연구의 연구 동향을 살펴본다. 또한 기존 관련연구가 갖는 한계점 및 본 연구의 목적과 차별성에 대해 기술한다. 3장에서는 본 논문에서 제안하는 고랭지배추 예측모델에 대해 자세하게 기술한다. 4장에서는 배추 생육 시기를 고려한 feature selection 결과와 도매가격 예측 결과를 통해 SVM, Random Forest, Gradient Boosting 분류모델의 성능평가에 대해 논한다. 마지막으로 5장에서는 결론에 대해 기술하고 본 연구가 갖는 한계점 및 향후 연구에 대하여 기술한다.

본 연구는 고랭지배추의 전년 동일 대비 도매가격 변동률과 파종기, 정식기, 생육기의 기상 및 가뭄지수 자료로 구성된 변수 간 영향을 분석하였다. 분석을 통해 관계가 높게 나타난 변수들을 중심으로 구축한 SVM(Support Vector Machine), Random Forest, Gradient Boosting 분류모델을 기반으로 고랭지배추 전년 동일 대비 도매가격 변동률을 예측하여 최적의 모델과 변수를 제시하였다.

Ⅱ. 관련 연구

2.1 통계적 방법을 이용한 농산물 가격 및 생산량 예측연구

통계적 방법을 이용한 농산물 가격 및 생산량 예측 연구의 추세를 살펴보면 다음과 같다. 강원도를 사례로 기후가 고랭지배추 생산에 미치는 영향을 분석하였으며[1] 작물의 주산지별 농경지 단위의 상세화 된 기상자료와 작물 수량 자료를 이용하여 배추, 무, 고추, 마늘, 양파를 대상으로 단수 예측이 가능한 다중회귀모형을 구축하였다[2]. 또한 가뭄 발생 시점의 단위 면적당 배추 생산량을 변화에 저항하는 능력으로 정의하였으며 가뭄 리질리언스를 분석하기 위하여 충청북도/충청남도 배추재배 지역의 시군별 단위 면적당 배추 생산량, SPEI (Standardized Precipitation Evapotranspiration Index) 지수, 통계 연보 연도별 가뭄 발생 수 자료를 이용하여 SPEI 지수 기반 가뭄 발생 시나리오 구성 후 단위 면적당 배추 생산량을 비교하였다[3].

2.2 머신러닝 기법을 이용한 농산물 가격 및 생산량 예측연구

머신러닝 기법을 이용한 농산물 가격 및 생산량 등의 예측 연구 추세를 살펴보면 다음과 같다.

랜덤 포레스트 알고리즘을 적용하여 작물 생산지 기상 환경에 따른 농산물의 가격이 어떻게 변화하는지 예측하는 모델을 구축하였으며[4], 인도 정부의 농민 복지부에서 발표한 2010년 ~ 2022년 데이터를 사용하여 14개 작물을 대상으로 농산물 재배비용을 예측하였다.

농산물 재배 비용은 인건비와 종자, 비료, 살충제, 관개 비용을 포함하며 머신러닝 회귀모델 11종(Linear, Random Forest, SVR, Gradient Boosting, Ridge, LassoCV, K-neighbours, XGB, LGB, SGD, Decision Tree)을 사용하여 예측하였다. 예측 결과, 농산물 재배 비용은 Random Forest와 Gradient Boosting, K-neighbours, XGB, Decision Tree 모델을 통해 정확도 높게 예측할 수 있는 것으로 분석되었다[5]. 중국 진샹 지역의 마늘 가격 데이터를 대상으로 로지스틱 회귀, SVM, XGBoost와 같은 분류 알고리즘을 사용하여 마늘 가격 추세를 두 종류로 분류하고 예측한 결과, 정확도 값은 각각 62.6%, 71.4% 및 72.9%였으며, 분류가 증가할수록 정확도는 감소하였지만, 그 중 XGBoost 알고리즘은 3단계, 4단계 및 5단계 예측에서 로지스틱 회귀 및 SVM 알고리즘보다 성능이 우수한 것으로 분석되었다[6].

데이터 셋 구축 및 데이터의 전처리가 성능에 미치는 영향을 분석하기 위한 연구로는 모델 학습 전 XGBoost 알고리즘을 사용하여 Feature Selection 수행 후 호주의 벼 수확량 예측에 적용한 결과, 기존 대비 정확도 높게 예측할 수 있는 것으로 분석되었으며[7], 인도에서 가장 소비가 많은 원예 상품인 토마토, 감자, 양파를 대상으로 전통적인 통계적(ARIMA, ETS) 방법과 머신 러닝(MLP, SVM, LSTM)을 이용한 방법 및 두 방식을 혼합한 하이브리드 방법을 사용하여 농산물의 가격 예측을 시도하였다. 예측 결과, 모든 농산물 가격 시계열 데이터에서 어떤 방법도 최상의 예측을 제공하지 못하는 것으로 나타났으며 이는 다양한 농산물 가격을 예측하기 위해 각각의 농산물에 맞는 다양한 방법이 고려해야 한다는 것을 시사한다[8].

2.3 시사점

기존의 연구는 다양한 분석모형을 적용하여 도매가격을 예측하거나 기상요인을 고려하여 생산성을 추정하는 연구를 진행하였으며 시계열 예측에 관한 연구의 경우 데이터의 변동성과 주기성을 효과적으로 포착하고 설명할 수 있음을 보여주었으나 농작물의 생육 시기별 영향을 미치는 요인분석이 부족하다는 한계점이 있다.

따라서 본 연구에서는 농산물의 생육 시기별 가격변동 영향 분석 적용 가능성을 확인하고자 하였다. 농산물의 생육 시기를 파종기, 정식기, 생육기로 나누어 feature selection을 이용하여 고랭지배추 전년 동일 대비 도매가격 변동률에 미치는 영향을 분석하였다.

분석 결과를 반영하여 구축한 SVM, Random Forest, Gradient Boosting 분류모델을 기반으로 고랭지배추 전년 동일 대비 도매가격 변동률을 예측하여 최적의 모델과 변수를 제시하였다.

III. 연구 설계

3.1 고랭지배추 생육 시기 정의

고랭지배추의 성장 과정은 표 1과 같이 크게 네 단계, 즉 파종기, 정식기, 생육기, 출하기로 나눌 수 있다. 파종기는 농작물을 심는 초기 단계로 이 시기에는 올바른 파종 깊이, 적절한 수분, 그리고 토양 온도가 중요하다. 정식기는 씨앗이 먼저 어린 모종 상태로 자란 후, 다시 농장이나 밭에 옮겨 심는 과정을 말한다. 이 과정에서 중요한 것은 모종의 건강, 적절한 간격으로 식재, 그리고 이식 충격을 최소화하는 것이다. 생육기는 밭에 심어진 작물이 성장하는 시기로 이 기간의 영양 관리, 물 관리, 그리고 병해충 방제가 중요하다. 출하기는 수확 준비가 되는 시기로 수확이 늦어지면 작물의 질이 떨어질 수 있고, 너무 일찍 수확하면 미숙한 상태에서 수확될 위험이 있어 적절한 수확 시기를 결정하는 것이 중요하다.

표 1. 고랭지배추의 생육 시기 구분표
Table 1. Highland cabbage growth stages

Sowing	Transplanting	Growing	Harvesting
mid March ~ late April	early April ~ mid May	late May ~ mid Jun	mid Jun ~ late October

3.2 데이터

분석지역은 강원도 고랭지 농사 지역으로 선정하

였으며 본 연구에서는 고랭지배추의 주요 재배지역으로 강릉, 대관령 지역을 선정하여 재배면적과 생산량을 고려하였다. 배추 도매가격 데이터는 농넷에서 제공하는 가락시장 경락 가격(원/단위 : 10kg, 품질 : 상품) 자료를 활용하였다. 기상 데이터는 기상청에서 제공하는 2003년~2022년의 ASOS(종관기상관측) 자료를 활용하였다. 분석지역을 중심으로 ASOS(강릉, 대관령)에서 제공하는 기온, 강수 등의 일 자료를 생육 시기별로 재생산하였다. 기온에 따른 고랭지배추의 생육 민감도를 비교하기 위하여 일평균기온, 일최고기온, 일최저기온 자료를 이용하였다. 가뭄지수 데이터는 기상청에서 제공하는 2003~2022년의 기상관측소별 표준강수지수 자료를 활용하였다. 고랭지배추의 출하 시기는 7월 초부터 10월 말이다. 2003년 7월부터 2022년 10월까지의 기상요인, 가뭄지수 데이터를 정식기, 생육기, 결구기로 나누어 이동평균법을 이용하여 자료를 구축하였다. 전년 동일 대비 도매가격 변동률은 데이터를 4구간으로 분류하여 전년 동일 대비 큰 하락, 하락, 상승, 큰 상승으로 구분하였다. 2003년부터 2018년까지 구축한 자료를 학습시켜 2019년, 2020년, 2021년, 2022년의 7월부터 10월까지의 고랭지배추 출하시기 전년 동일 대비 도매가격 변동률을 예측하였다.

3.3 데이터 구축 방법

본 연구의 데이터 구축 절차는 그림 1과 같다. 농넷 가락시장 경락 가격, 강릉 및 대관령 지역 기상청 종관기상관측 자료, SPI 가뭄지수 자료를 수집하였다. 본 연구에서 고려한 기상요인은 평균기온, 최저기온, 최고기온, 일강수량, 평균 이슬점온도, 최소 상대습도, 평균 상대습도, 합계 일사, 평균 지면온도, 최저 초상온도이며, 가뭄지수는 SPI1, SPI2, SPI3, SPI4를 대상으로 수집하였다. 그 후 수집자료 전처리를 수행하였으며 기상요인, 가뭄지수 데이터를 파종기, 정식기, 생육기로 나누어 이동평균법을 이용한 자료를 구축하였다. 농산물 도매가격 자료를 이용하여 전년 동일 대비 도매가격 변동률 산출 및 변동률의 4구간을 산출하였으며 생육시기별 기상요인을 구분하여 데이터를 구축하였다.

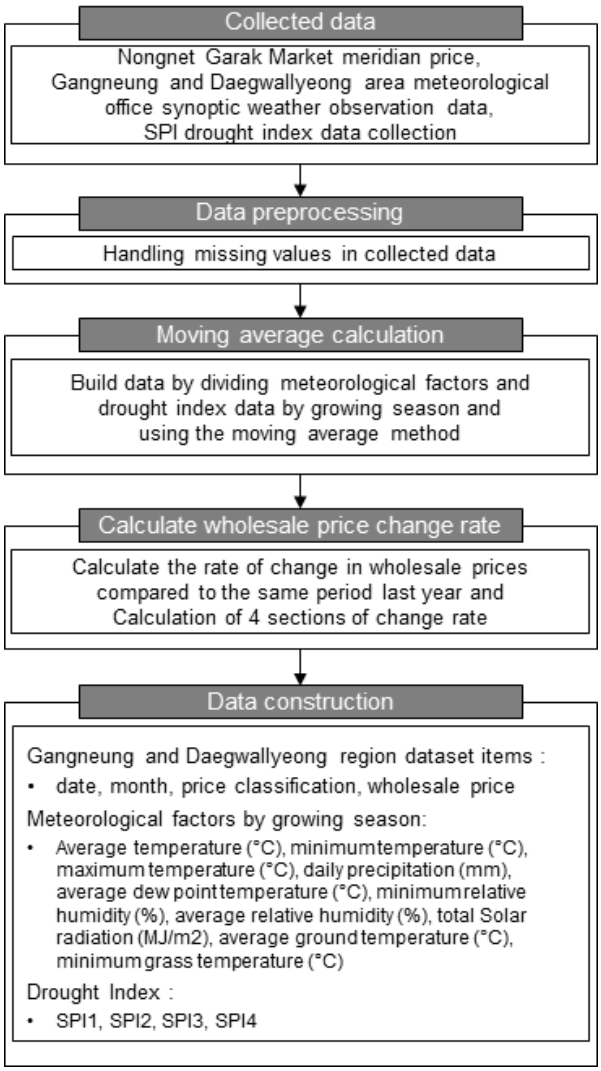

그림 1. 데이터 구축 절차
Fig. 1. Data construction procedure

3.4 데이터 셋 구축 결과

본 연구는 강원도 강릉 및 대관령을 분석지역으로 선정하였으며 2003~2022년의 자료를 수집하였다. 데이터의 전처리는 파이썬 언어의 Pandas 및 Numpy 라이브러리를 이용하였다.

수집 및 구축 결과 강릉 1,424 rows × 44 columns, 대관령 1,503 rows × 44 columns 형태의 데이터가 구성되었다.

3.5 입력변수

전년 동일 대비 도매가격 변동률과 고랭지배추의 파종기, 정식기, 생육기의 기후 영향을 분석하기 위

하여 고랭지배추의 도매가격과 기상요인, 가뭄지수 간 시기별 상관관계를 가뭄 판단지표로 정의하여 생산량에 영향을 미치는 최적의 시기를 분석하였다. 도매가격의 경우 전년 동일 대비 도매가격 변동률을 구한 뒤 사분 범위를 활용하여 4구간으로 분류하였다.

P_t는 현재 가격, $P_{t-365days}$는 1년 전 같은 일자의 가격을 의미하며 전년 동일 대비 도매가격 변동률 R_t의 산출 식 (1)은 아래와 같다.

$$R_t = \frac{P_t}{P_{t-365days}} - 1 \tag{1}$$

전년 동일 대비 도매가격 변동률의 4구간(C_t = 0, 1, 2, 3)은 세 개의 분위(R_{Q1} = -0.402, R_{Q2} = -0.034, R_{Q3} = 0.429)를 사용하여 분류한다. 분류 식 (2)는 아래와 같다.

$$\begin{aligned} C_t &= 0 \text{ if } R_t < R_{Q1} \\ C_t &= 1 \text{ if } R_{Q1} \le R_t < R_{Q2} \\ C_t &= 2 \text{ if } R_{Q2} \le R_t < R_{Q3} \\ C_t &= 3 \text{ if } R_{Q3} \le R_t \end{aligned} \tag{2}$$

입력변수는 아래의 표 2와 같다. 기상요인, 가뭄지수 자료의 경우 고랭지배추가 출하되는 7월~10월을 기점으로 파종기(15주 전~17주 전), 정식기(7주 전~14주 전), 생육기(3주 전~6주 전)로 나누어 재생산하였다. 파종기의 경우 도매시장 판매일로부터 105일~119일 전, 정식기의 경우 43일~104일 전, 생육기의 경우 21일~42일 전으로 설정한 뒤 이동평균법을 이용하여 해당 생육 기간별 가뭄지수 데이터와 기상 데이터를 구축하였다.

3.6 분석 방법

농산물의 생육 시기를 파종기, 정식기, 생육기로 나누어 feature selection을 이용하여 고랭지배추 전년 동일 대비 도매가격 변동률에 미치는 영향을 분석하였다. 분석 후 feature selection을 통해 도출된 결과를 반영하여 구축한 SVM, Random Forest, Gradient Boosting 분류모델을 기반으로 전년 동일 대비 큰 하락, 하락, 상승, 큰 상승을 예측하였다. Random Forest와 Gradient Boosting의 경우 max_depth와 n_estimators를 각기 다르게 구분하여 최적의 파라미터 값을 도출하고자 하였다.

max_depth는 결정 트리의 최대 깊이를 결정하는 변수이다. 즉, 트리가 얼마나 깊게 성장할 수 있는지를 정하는 값이다. 이 파라미터는 모델이 학습 데이터에 대해 얼마나 복잡하게 학습할 수 있는지를 제한한다. 깊이가 깊어질수록 모델은 더 많은 분할을 하여 데이터의 특징을 더 세밀하게 학습할 수 있지만, 너무 깊으면 학습 데이터에 과적합 될 위험이 있다. max_depth 값이 너무 낮으면 모델이 데이터의 중요한 특성을 포착하지 못할 수 있고, 너무 높으면 과적합의 위험이 커진다.

n_estimators는 Random Forest, Gradient Boosting과 같은 앙상블 기반 모델에서 사용되며, 모델이 구축할 결정 트리의 개수를 의미한다. 더 많은 수의 트리는 일반적으로 더 나은 성능을 제공하지만, 계산 비용이 증가하고 모델을 훈련하는 시간도 길어진다. 또한 특정 지점 이후에는 성능 향상이 미미해질 수 있다. 따라서 n_estimators는 과적합을 줄이고 모델의 안정성을 높이는 데 도움이 될 수 있지만, 너무 많은 트리는 훈련 시간과 메모리 사용량을 증가시킬 수 있으므로 적절한 값을 설정하는 것이 중요하다. 본 연구에서는 이와 같은 특성을 반영하여 표 3과 같이 Random Forest는 12가지의 모델로 구성하였으며 Gradient Boosting은 10가지의 모델로 구성하여 각각의 정확도를 비교 분석하였다.

분석에 사용한 세 모델 중 SVM은 분류 및 회귀 작업에 사용되는 지도 학습 알고리즘이다. 분류모델은 데이터가 비추어진 공간에서 경계로 표현되는데 SVM 알고리즘은 그중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다.

Random Forest는 분류 및 회귀 작업에 널리 사용되는 앙상블 학습 방법의 일종으로 여러 의사결정 트리를 결합하여 전반적인 예측 성능을 향상하고 과적합 위험을 줄이는 알고리즘이다.

Gradient Boosting은 분류 및 회귀 작업 모두에 사용되는 지도 학습 알고리즘이다. 일반적으로 의사 결정 트리와 같은 여러 약한 학습자를 순차적으로 교육하여 강력한 예측 모델을 구축하는 알고리즘이다.

표 2. 입력변수와 및 설명
Table 2. Input variables and description

Category	Variable code	Description
Price	price_change_rate	(Current Price - Price from the Same Date Last Year) / Price from the Same Date Last Year
SPI	Sowing_SPI1	SPI1 from 15 weeks prior to the current date
	Sowing_SPI2	SPI2 from 15 weeks prior to the current date
	Sowing_SPI3	SPI3 from 15 weeks prior to the current date
	Sowing_SPI4	SPI4 from 15 weeks prior to the current date
	Planting_SPI1	SPI1 from 7 weeks prior to the current date
	Planting_SPI2	SPI2 from 7 weeks prior to the current date
	Planting_SPI3	SPI3 from 7 weeks prior to the current date
	Planting_SPI4	SPI4 from 7 weeks prior to the current date
	Growing_SPI1	SPI1 from 3 weeks prior to the current date
	Growing_SPI2	SPI2 from 3 weeks prior to the current date
	Growing_SPI3	SPI3 from 3 weeks prior to the current date
	Growing_SPI4	SPI4 from 3 weeks prior to the current date
Climate	Sowing_avg_temp	14-day moving average of average temperature (°C), 15 weeks prior to the current date
	Sowing_min_temp	14-day moving average of lowest temperature (°C), 15 weeks prior to the current date
	Sowing_max_temp	14-day moving average of highest temperature (°C), 15 weeks prior to the current date
	Sowing_d_prcp	14-day moving average of daily precipitation (mm), 15 weeks prior to the current date
	Sowing_avg_dew_point_temp	14-day moving average of dew point temperature (°C), 15 weeks prior to the current date
	Sowing_min_relative_humidity	14-day moving average of minimum relative humidity (%), 15 weeks prior to the current date
	Sowing_avg_relative_humidity	14-day moving average of average relative humidity (%), 15 weeks prior to the current date
	Sowing_Total_solar_radiation	14-day moving average of total solar radiation (MJ/m2), 15 weeks prior to the current date
	Sowing_avg_ground_temp	14-day moving average of average ground temperature (°C), 15 weeks prior to the current date
	Sowing_min_grass_temp	14-day moving average of lowest grass temperature (°C), 15 weeks prior to the current date
	Planting_avg_temp	62-day moving average of average temperature (°C), 7 weeks prior to the current date
	Planting_min_temp	62-day moving average of lowest temperature (°C), 7 weeks prior to the current date
	Planting_max_temp	62-day moving average of highest temperature (°C), 7 weeks prior to the current date
	Planting_d_prcp	62-day moving average of daily precipitation (mm), 7 weeks prior to the current date
	Planting_avg_dew_point_temp	62-day moving average of dew point temperature (°C), 7 weeks prior to the current date
	Planting_min_relative_humidity	62-day moving average of minimum relative humidity (%), 7 weeks prior to the current date
	Planting_avg_relative_humidity	62-day moving average of average relative humidity (%), 7 weeks prior to the current date
	Planting_Total_solar_radiation	62-day moving average of total solar radiation (MJ/m2), 7 weeks prior to the current date
	Planting_avg_ground_temp	62-day moving average of average ground temperature (°C), 7 weeks prior to the current date
	Planting_min_grass_temp	62-day moving average of lowest grass temperature (°C), 7 weeks prior to the current date
	Growing_avg_temp	21-day moving average of average temperature (°C), 3 weeks prior to the current date
	Growing_min_temp	21-day moving average of lowest temperature (°C), 3 weeks prior to the current date
	Growing_max_temp	21-day moving average of highest temperature (°C), 3 weeks prior to the current date
	Growing_d_prcp	21-day moving average of daily precipitation (mm), 3 weeks prior to the current date
	Growing_avg_dew_point_temp	21-day moving average of dew point temperature (°C), 3 weeks prior to the current date
	Growing_min_relative_humidity	21-day moving average of minimum relative humidity (%), 3 weeks prior to the current date
	Growing_avg_relative_humidity	21-day moving average of average relative humidity (%), 3 weeks prior to the current date
	Growing_Total_solar_radiation	21-day moving average of total solar radiation (MJ/m2), 3 weeks prior to the current date
	Growing_avg_ground_temp	21-day moving average of average ground temperature (°C), 3 weeks prior to the current date
	Growing_min_grass_temp	21-day moving average of lowest grass temperature (°C), 3 weeks prior to the current date

표 3. 예측모델 세부사항 구분
Table 3. Prediction model details

Model	Model code	Parameter
SVM(Support Vector Machine)	svm	-
Random forest classifier	rf1	n_estimators=15, max_depth=3
	rf2	n_estimators=15, max_depth=4
	rf3	n_estimators=15, max_depth=5
	rf4	n_estimators=20, max_depth=3
Random forest classifier	rf5	n_estimators=20, max_depth=4
	rf6	n_estimators=20, max_depth=5
	rf7	n_estimators=25, max_depth=3
	rf8	n_estimators=25, max_depth=4
	rf9	n_estimators=25, max_depth=5
	rf10	n_estimators=30, max_depth=3
	rf11	n_estimators=30, max_depth=4
	rf12	n_estimators=30, max_depth=5
Gradient boosting classifier	gb1	n_estimators=4, max_depth=2
	gb2	n_estimators=4, max_depth=3
	gb3	n_estimators=4, max_depth=4
	gb4	n_estimators=5, max_depth=2
	gb5	n_estimators=5, max_depth=3
	gb6	n_estimators=5, max_depth=4
	gb7	n_estimators=8, max_depth=2
	gb8	n_estimators=8, max_depth=3
	gb9	n_estimators=8, max_depth=4
	gb10	n_estimators=8, max_depth=5

3.7 성능 측정 기준

기계학습에서는 분류모델의 성능을 정확하게 평가하는 것이 중요하다. 분류모델은 범주형 클래스 레이블을 예측하는 데 사용되며 성능은 일반적으로 특정 측정항목을 사용하여 평가된다.

이 중 Accuracy는 가장 직관적이고 일반적으로 사용되는 측정항목이다. 본 연구에서는 Accuracy를 전체 예측에 대한 일치된 예측의 비율로 정의하였으며 계산 식 (3)은 아래와 같다.

$$Accuracy = \frac{Nmber\,of\,Correct\,\mathrm{P}redictions}{Total\,Nmber\,of\,\mathrm{P}redictions} \quad (3)$$

본 연구에서는 Python 라이브러리 중 하나인 sklearn의 score함수를 이용하여 모델의 분류 정확도를 측정하였다.

Ⅳ. 연구 결과

4.1 배추 생육 시기를 고려한 feature selection 결과

고랭지배추 전년 동일 대비 도매가격 변동률과 생육 시기를 고려한 파종기, 정식기, 생육기의 기상 및 가뭄지수 자료로 구성된 변수 간 feature selection을 진행하였다. 강릉의 변수 간 중요도는 표 4와 같다.

표 4. 생육 시기를 고려한 강릉 지역 입력변수 중요도 구분표
Table 4. Importance table of feature selection considering the growth stages in Gangneung

Index	Variable	Importance
1	Growing_SPI1	59.67
2	Growing_SPI2	49.16
3	Growing_min_grass_temp	48.50
4	Planting_SPI1	46.61
5	Growing_d_prcp	38.36
6	Growing_SPI4	37.56
7	Growing_SPI3	35.28
8	Planting_SPI2	15.86
9	Planting_d_prcp	15.72
10	Planting_SPI3	14.72
11	Planting_min_grass_temp	12.34
12	Growing_min_temp	11.92
13	Sowing_SPI3	11.82
14	Sowing_SPI4	7.24
15	Growing_avg_temp	7.14
16	Growing_avg_dew_point_temp	6.32
17	Planting_SPI4	5.92
18	Sowing_SPI2	5.40
19	Growing_min_relative_humidity	5.29
20	Growing_Total_solar_radiation	4.41
21	Sowing_min_relative_humidity	4.31
22	Growing_max_temp	4.12
23	Sowing_avg_relative_humidity	2.92
24	Growing_avg_ground_temp	2.09
25	Planting_min_temp	1.87
26	Planting_avg_relative_humidity	1.68
27	Sowing_SPI1	1.54
28	Planting_Total_solar_radiation	1.46
29	Planting_max_temp	1.43
30	Planting_avg_temp	1.35
31	Sowing_d_prcp	1.07
32	Sowing_max_temp	0.91
33	Sowing_min_grass_temp	0.87
34	Planting_avg_ground_temp	0.44
35	Sowing_Total_solar_radiation	0.31
36	Sowing_avg_temp	0.30
37	Planting_min_relative_humidity	0.28
38	Sowing_avg_dew_point_temp	0.15
39	Growing_avg_relative_humidity	0.09
40	Sowing_avg_ground_temp	0.09
41	Sowing_min_temp	0.06
42	Planting_avg_dew_point_temp	0.01

대부분 생육기와 정식기의 변수들이 영향을 미치는 것으로 확인되었다. 가장 중요도가 높은 변수는 생육기의 SPI1이며 SPI2, 최저 초상온도, 일강수량, SPI4, SPI3, 최저기온, 평균기온의 중요도가 높았으며, 이는 생육기의 기상 요소 및 표준강수지수의 변화가 고랭지배추의 성장과 그에 따른 도매가격에 영향을 미칠 수 있음을 나타낸다. 정식기의 경우 SPI1의 중요도가 가장 높았으며 SPI2, 강수량, SPI3이 중요한 변수로 확인되었다.

대관령의 변수 간 중요도는 표 5와 같다. 대부분 생육기와 파종기의 변수들이 영향을 미치는 것으로 확인되었다. 그림 2와 같이 가장 중요도가 높은 변수는 생육기의 SPI1이었으며 다른 변수들에 비해 중요도가 두 배 이상 높은 것으로 나타났다. 생육기의 변수들이 중요한 변수로 확인된 것은 강릉과 같지만 파종기의 변수들이 영향을 미치는 차이점이 나타났다. 가장 중요도가 높은 변수는 생육기의 SPI1이며 SPI2, SPI3, 강수량, 최저기온, 평균기온, 최고기온의 순서로 중요도가 높았으며, 파종기의 경우 SPI3과 SPI4, SPI2가 중요한 변수로 확인되었다.

표 5. 생육 시기를 고려한 대관령 지역 입력변수 중요도 구분표

Table 5. Importance table of feature selection considering the growth stages in Degwallyeong

Index	Variable	Importance
1	Growing_SPI1	56.51
2	Sowing_SPI3	26.82
3	Sowing_SPI4	23.66
4	Growing_SPI2	19.31
5	Growing_SPI3	16.02
6	Sowing_SPI2	13.53
7	Growing_d_prcp	12.87
8	Growing_min_temp	11.29
9	Growing_avg_temp	10.69
10	Growing_max_temp	10.54
11	Growing_avg_dew_point_temp	9.36
12	Growing_SPI4	8.56
13	Growing_min_grass_temp	7.04
14	Planting_Total_solar_radiation	5.16
15	Planting_SPI1	5.12
16	Planting_max_temp	3.41
17	Planting_SPI4	3.14
18	Planting_avg_temp	3.02
19	Planting_min_temp	2.50
20	Growing_avg_ground_temp	2.45
21	Planting_avg_ground_temp	2.32
22	Sowing_d_prcp	2.10
23	Planting_d_prcp	1.88
24	Planting_min_grass_temp	1.88
25	Planting_avg_relative_humidity	1.63
26	Planting_SPI2	1.51
27	Planting_min_relative_humidity	1.36
28	Sowing_avg_relative_humidity	0.90
29	Planting_avg_dew_point_temp	0.53
30	Sowing_SPI1	0.47
31	Sowing_max_temp	0.44
32	Sowing_min_grass_temp	0.25
33	Sowing_min_relative_humidity	0.24
34	Planting_SPI3	0.19
35	Sowing_avg_dew_point_temp	0.12
36	Sowing_min_temp	0.10
37	Sowing_Total_solar_radiation	0.09
38	Growing_min_relative_humidity	0.08
39	Growing_Total_solar_radiation	0.07
40	Sowing_avg_ground_temp	0.02
41	Sowing_avg_temp	0.02
42	Growing_avg_relative_humidity	0.01

4.2 고랭지배추 도매가격 예측 결과

Feature selection 결과를 반영한 고랭지배추 도매가격 변동률을 예측한 결과는 표 6, 표 7, 표 8, 표 9와 같다. 중요도를 고려하여 입력변수가 5개, 10개, 15개, 20개, 25개, 30개 일 경우를 구분하여 각각의 예측 정확도를 측정하였다. 사용한 모델은 SVM, Random Forest, Gradient Boosting 분류모델이다. Random Forest와 Gradient Boosting의 경우 max_depth와 n_estimators를 각기 다르게 구분하여 최적의 파라미터 값을 도출하고자 하였다.

강릉의 경우 트레이닝 결과 정확도가 가장 높았던 모델은 Gradient Boosting Classifier (n_estimators=8, max_depth=5)였으며 정확도는 0.94였다. 입력변수별 정확도 또한 해당 모델이 가장 높았으며, 각 모델별 정확도는 입력변수가 30개 일 때 가장 높게 나타났다.

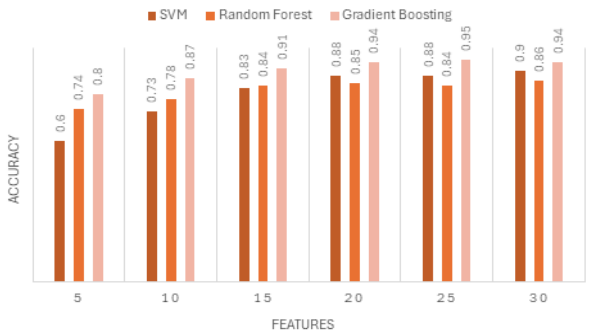

그림 2. 강릉 지역 고랭지배추 도매가격 예측모형 훈련 정확도 그래프

Fig. 2. Graph of train accuracy in Gangneung

표 6. 강릉 지역 고랭지배추 도매가격 예측모형 훈련 정확도 구분표
Table 6. Train accuracy in Gangneung

Features	SVM	Random forest	Gradient boosting
5	0.6	0.74	0.8
10	0.73	0.78	0.87
15	0.83	0.84	0.91
20	0.88	0.85	0.94
25	0.88	0.84	0.95
30	0.9	0.86	0.94

테스트 결과 정확도가 가장 높았던 모델은 Random Forest Classifier (n_estimators=15, max_depth=5)였으며 정확도는 0.46이었다. 입력변수가 5개와 10개일 경우 Gradient Boosting, 15개, 20개, 30개일 경우 Random Forest의 정확도가 가장 높게 나타났다. 특정 모델에 집중되었던 트레이닝 결과와는 달리 입력변수 개수에 따라 정확도가 높은 모델이 다양한 양상을 보이는 것으로 분석되었으며 모델별 정확도 또한 트레이닝 결과와 달리 입력변수가 10개 일 때 대부분 높게 나타났다.

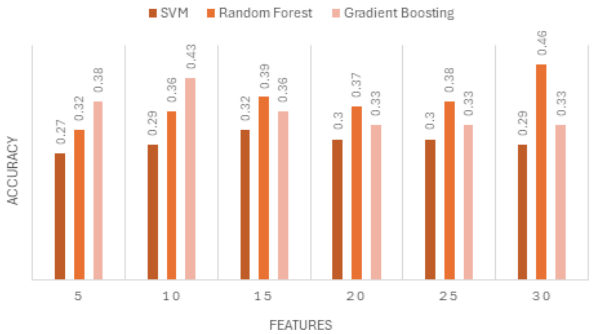

그림 3. 강릉 지역 고랭지배추 도매가격 예측모형 예측 정확도 그래프
Fig. 3. Graph of test accuracy in Gangneung

표 7. 강릉 지역 고랭지배추 도매가격 예측모형 예측 정확도 구분표
Table 7. Test accuracy in Gangneung

Features	SVM	Random forest	Gradient boosting
5	0.27	0.32	0.38
10	0.29	0.36	0.43
15	0.32	0.39	0.36
20	0.3	0.37	0.33
25	0.3	0.38	0.33
30	0.29	0.46	0.33

대관령의 경우 트레이닝 결과 정확도가 가장 높았던 모델은 Gradient Boosting Classifier (n_estimators=8, max_depth=5)였으며 정확도는 0.933이었다. 입력변수별 정확도 또한 해당 모델과 Gradient Boosting Classifier(n_estimators=8, max_depth=4)이 가장 높았으며, 모델별 정확도는 입력변수가 10개와 30개 일 때 가장 높게 나타났다.

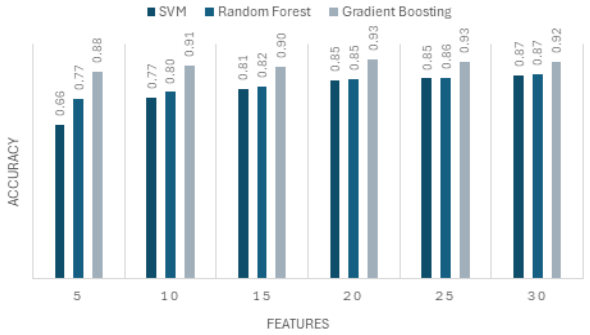

그림 4. 대관령 지역 고랭지배추 도매가격 예측모형 훈련 정확도 그래프
Fig. 4. Graph of train accuracy in Daegwallyeong

표 8. 대관령 지역 고랭지배추 도매가격 예측모형 훈련 정확도 구분표
Table 8. Train accuracy in Daegwallyeong

Features	SVM	Random forest	Gradient boosting
5	0.66	0.77	0.88
10	0.77	0.80	0.91
15	0.81	0.82	0.90
20	0.85	0.85	0.93
25	0.85	0.86	0.93
30	0.87	0.87	0.92

테스트 결과 정확도가 가장 높았던 모델은 Random Forest Classifier(n_estimators=15, max_depth=4)였으며 정확도는 0.38이었다. 입력변수가 5개, 15개, 20개일 경우 Random Forest의 정확도가 높았으며, 입력변수가 25개, 30개일 경우 Gradient Boosting의 정확도가 가장 높게 나타났다. 특정 모델에 집중되었던 트레이닝 결과와는 달리 입력변수 개수에 따라 정확도가 높은 모델이 다양한 양상을 보이는 것으로 분석되었으며 모델별 정확도는 트레이닝 결과와 같은 입력변수가 30개 일 때 대부분 높게 나타났다.

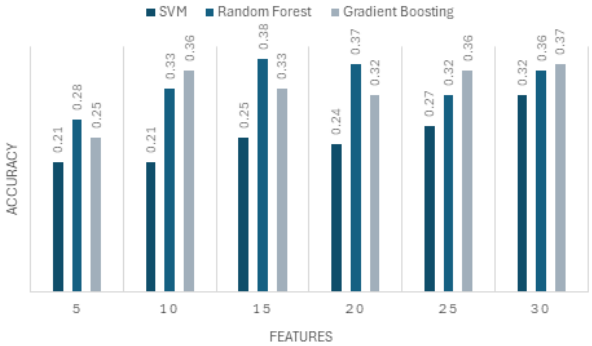

그림 5. 대관령 지역 고랭지배추 도매가격 예측모형 예측 정확도 그래프

Fig. 5. Graph of test accuracy in Daegwallyeong

표 9. 대관령 지역 고랭지배추 도매가격 예측모형 예측 정확도 구분표

Table 9. Test accuracy in Daegwallyeong

Features	SVM	Random forest	Gradient boosting
5	0.21	0.28	0.25
10	0.21	0.33	0.36
15	0.25	0.38	0.33
20	0.24	0.37	0.32
25	0.27	0.32	0.36
30	0.32	0.36	0.37

사분위로 구분한 변동률 구간을 랜덤하게 예측한다면 정확도는 0.25가 되겠지만 본 연구의 트레이닝 결과 평균 정확도는 강릉 0.712 대관령 0.75, 테스트 결과 평균 정확도는 강릉 0.32, 대관령 0.27로 랜덤한 예측보다 높은 것으로 나타났다.

Random Forest와 Gradient Boosting 모델의 경우 max_depth가 증가할수록 복잡도가 높아져 Train의 정확도가 증가했지만, 그에 비하여 Test 정확도는 크게 증가하지 않은 것을 확인할 수 있었다. 일반적으로 변수를 늘리면 훈련 데이터에 대한 모델의 정확도가 향상되지만, 지역별 중요도가 높은 입력변수 10개 정도의 사용으로도 과적합을 방지하며 예측을 수행할 수 있음을 확인하였다.

Ⅴ. 결론 및 향후 과제

고랭지배추 전년 동일 대비 도매가격 변동률과 생육 시기를 고려한 파종기, 정식기, 생육기의 기상 및 가뭄지수 자료로 구성된 변수 간 feature selection을 진행한 결과, 강릉의 경우 대부분 생육기와 정식기의 변수들이 영향을 미치는 것으로 확인되었다. 가장 중요도가 높은 변수는 생육기의 SPI1로 분석되었다. 대관령의 경우 대부분 생육기와 파종기의 변수들이 영향을 미치는 것으로 확인되었으며 가장 중요도가 높은 변수는 생육기의 SPI1로 분석되었다.

입력변수가 5개, 10개, 15개, 20개, 25개, 30개 일 경우를 구분하여 feature selection 결과를 반영한 고랭지배추 도매가격 변동률을 예측한 결과, 강릉의 경우 입력변수가 30개일 때 Random Forest Classifier(n_estimators=15, max_depth=5) 모델의 정확도가 0.46으로 가장 높았지만, 입력변수가 10개일 때 Gradient Boosting (n_estimators=8, max_depth=4) 모델의 정확도가 0.43으로 큰 차이는 보이지 않았다. 마찬가지로 대관령의 경우 입력변수가 20개일 때 Random Forest Classifier(n_estimators=15, max_depth=4) 모델의 정확도가 0.38로 가장 높았지만, 입력변수가 10개일 때 Gradient Boosting (n_estimators=4, max_depth=4) 모델의 정확도가 0.36으로 큰 차이는 보이지 않았다. 이는 고랭지배추 도매가격 변동률을 예측하기 위하여 중요도가 높은 입력변수 10개 정도의 사용으로 과적합을 방지하며 예측을 수행할 수 있다는 것을 의미한다.

이와 같은 분석결과는 농산물 도매가격 형성에 결정적인 영향을 미치는 생육시기별 기상요인을 도출하고, 작물의 가격을 정확하게 예측하게 되면 농산물 도매가격의 급등 및 급락이 예상될 경우 그에 적합한 대응정책을 마련하는데 도움이 될 수 있다. 또한 농산물의 품목을 확대하여 분석한다면 각 주산지의 지자체에서 농산물을 판매할 최적의 시기 선정 및 관리로 도매가격의 폭등 및 폭락을 대처하고 가격의 안정화에 도움을 줄 수 있어 대응방안 마련 시 인근 지자체와 협력적 정책의 기초자료로 활용할 수 있다.

연구의 한계점으로, 전년도 출하기 가격 하락과 연작피해에 따른 휴경, 양배추 및 약용작물 등 작목 전환으로 인한 배추 재배(의향)면적 변화, 인건비 상승 등으로 인한 작목 전환을 반영하지 못하였다. 이를 고려한다면 도매가격 변동율 예측 결과가 달라질 수 있다.

본 연구는 강원도 고랭지 농업지역인 강릉과 대관령 지역을 대상으로 2003년~2022년의 데이터를 사용하였으나, 향후 주 생산 지역과 기간을 확대하여 분석이 필요할 것으로 사료된다.

본 연구에서는 고랭지배추의 생육기간을 파종기, 정식기, 생육기로 나누어 기상 및 표준강수지수가 도매가격에 미치는 영향과 주요 요인을 분석하였다. 결과를 반영하여 (기계학습 회귀모델) 모델을 기반으로 도매가격을 예측하였다 본 연구의 한계로 전년도 출하기의 가격 하락, 경작지 휴경, 다른 작물로의 전환 등으로 인한 배추 재배면적 변화를 반영할 경우 도매가격 변동 예측 결과가 달라질 수 있다. 기후변화 및 가뭄 상황에 따른 배추 가격의 민감도는 농업 경제, 소비자, 국가의 농식품 정책에 큰 영향을 미칠 수 있으며, 본 연구를 통해 농산물의 생육 기간별 가격변화 영향 분석의 적용 가능성을 확인하였다.

References

[1] S. H. Lee and I. H. Heo, "Impact of Climate on Yield of Highland Chinese Cabbage in Gangwon Province", Journal of the Korean Geographical Society, Vol. 53, No. 3, pp. 265-282, Jun. 2018.

[2] C. H. Lim, G. S. Kim, E. J. Lee, S. Heo, T. Kim, Y. S. Kim, and W. K. Lee, "Development on crop yield forecasting model for major vegetable crops using meteorological information of main production area", Journal of Climate Change Research, Vol. 7, No. 2, pp. 193-203, Jun. 2016. https://doi.org/10.15531/KSCCR.2016.7.2.193.

[3] S. Jung, H. Kang, and S. Maeng, "Analysis for Drought Resilience of Monoculture on Climate Change", Crisisonomy, Vol. 11, No. 1, pp. 63-81, Apr. 2015.

[4] H. O. Choe, H. Yoe, M. H. Lee, and J. W. Park, "A Study on the Price Prediction of Agricultural Products in the Wholesale Market According to the Environment of the Production Area", Journal of Knowledge Information Technology and Systems, Vol. 17, No. 6, pp. 1285-1295, Dec. 2022. https://doi.org/10.34163/jkits.2022.17.6.020.

[5] P. Bari and L. Ragha, "Machine learning-based extrapolation of crop cultivation cost", Inteligencia Artificial, Vol. 27, No. 74, pp. 80-101, May 2024. https://doi.org/10.4114/intartif.vol27iss74pp80-101.

[6] F. Sun, X. Meng, H. Zhang, Y. Wang, and P. Liu, "Prediction of Weekly Price Trend of Garlic Based on Classification Algorithm and Combined Features", Horticulturae, Vol. 10, No. 4, pp. 347, Mar. 2024. https://doi.org/10.3390/horticulturae10040347.

[7] A. Clarke, D. Yates, C. Blanchard, M. Z. Islam, R. Ford, S. Rehman, and R. Walsh, "The effect of dataset construction and data pre-processing on the eXtreme Gradient Boosting algorithm applied to head rice yield prediction in Australia", Computers and Electronics in Agriculture, Vol. 219, pp. 108716, Apr. 2024. https://doi.org/10.1016/j.compag.2024.108716.

[8] S. K. Purohit, S. Panigrahi, P. K. Sethy, and S. K. Behera, "Time series forecasting of price of agricultural products using hybrid methods", Applied Artificial Intelligence, Vol. 35, No. 15, pp. 1388-1406, Sep. 2021. https://doi.org/10.1080/08839514.2021.1981659.

저자소개

허 다 솜 (Dasom Hur)

2014년 2월 : 강릉원주대학교 도시계획부동산학과(경제학사)
2021년 8월 : 서울대학교 환경계획학과 석사수료 (도시계획학)
2022년 2월 ~ 현재 : (재)국제도시물정보과학 연구원
관심분야 : 기후변화, 환경계획, 도시환경

정 승 권 (Seung Kwon Jung)

1998년 2월 : 충북대학교
토목공학과(공학사)
2000년 2월 : 충북대학교
토목공학과(공학석사)
2021년 2월 : 강원대학교
방재전문대학원(공학박사)
2017년 9월 ~ 현재 :
(재)국제도시물정보과학연구원 연구위원
관심분야 : 기후변화, 재난관리, 도시환경