Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 40, No. 5, 449-456, 2022
https://doi.org/10.7848/ksgpc.2022.40.5.449

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

GIS를 이용한 토양정보 기반의 배추 생산량 예측 수정모델 개발

Development of a modified model for predicting cabbage yield based on soil properties using GIS

최연오[1] · 이재현[2] · 심재후[3] · 이승우[4]
Choi, Yeon Oh · Lee, Jaehyeon · Sim, Jae Hoo · Lee, Seung Woo

Abstract

This study proposes a deep learning algorithm to predict crop yield using GIS (Geographic Information System) to extract soil properties from Soilgrids and soil suitability class maps. The proposed model modified the structure of a published CNN-RNN (Convolutional Neural Network-Recurrent Neural Network) based crop yield prediction model suitable for the domestic crop environment. The existing model has two characteristics. The first is that it replaces the original yield with the average yield of the year, and the second is that it trains the data of the predicted year. The new model uses the original field value to ensure accuracy, and the network structure has been improved so that it can train only with data prior to the year to be predicted. The proposed model predicted the yield per unit area of autumn cabbage for kimchi by region based on weather, soil, soil suitability classes, and yield data from 1980 to 2020. As a result of computing and predicting data for each of the four years from 2018 to 2021, the error amount for the test data set was about 10%, enabling accurate yield prediction, especially in regions with a large proportion of total yield. In addition, both the proposed model and the existing model show that the error gradually decreases as the number of years of training data increases, resulting in improved general-purpose performance as the number of training data increases.

Keywords : Deep Learning, Crop Yield Prediction, CNN, RNN, LSTM, Hyperparameter Optimization

초 록

본 연구는 GIS를 통해 토양정보를 수집하고 가공하여 농산물 생산량을 예측하는 모델을 제안한다. 농산물 생산량 예측 딥러닝 알고리즘은 공개된 CNN-RNN 농산물 생산량 예측 모델 구조를 변경하여 국내 농산물 자료 환경에 적합하도록 새롭게 구축하였다. 기존모델은 두 가지 특징을 가지고 있는데 첫 번째는 농산물의 생산량을 해당 필지 값이 아닌 당해 평균값으로 대체한다는 것이고 두 번째는 예측하는 연도의 데이터까지 학습한다는 것이다. 새로운 모델은 해당 필지의 값을 그대로 사용하여 데이터의 정확성을 확보하고 예측하고자 하는 연도 이전의 데이터만 가지고 학습할 수 있도록 네트워크 구조를 개선하였다. 제안한 CNN-RNN 모델은 1980년부터 2020년까지의 기상정보, 토양정보, 토양적성도, 생산량 데이터를 학습하여 김장용 가을배추의 지역별 단위면적당 생산량을 예측한다. 2018년부터 2021년까지 4개 연도별 자료에 대하여 계산하고 생산량을 예측한 결과, 테스트 데이터셋에 대한 오차백분율이 약 10% 내외로 실제값과 비교하여 정확도 높은 생산량 예측이 가능했고, 특히 전체 생산량 비중이 큰 지역에서의 생산량은 비교적 근접하게 예측하는 것으로 분석되었다. 또한 제안모델과 기존모델은 모두 학습자료 연도 수가 증가할수록 점점 오차가 작아지므로 학습데이터가 많아질수록 범용 성능은 향상되는 결과를 나타낸다.

핵심어 : 딥러닝, 농산물 생산량 예측, CNN, RNN, LSTM, 하이퍼파라미터 최적화

Received 2022. 10. 13, Revised 2022. 10. 17, Accepted 2022. 10. 20
1) Research Engineer, Lodics Co.,LTD (E-mail: hnchoi@lodics.com)
2) Research Engineer, Lodics Co.,LTD (E-mail: jhyenlee@lodics.com)
3) Research Engineer, Lodics Co.,LTD (E-mail: simjh@lodics.com)
4) Corresponding Author, Member, CEO, Lodics Co.,LTD (E-mail: james.lee@lodics.com)

1. 서 론

농산물 가격 안정화를 위해서는 농산물의 생산량을 예측해야 할 필요가 있다. 농작물 생산량 예측은 기상자료 기반의 선형회귀 모델을 사용하는 것으로부터 공간자기상관을 포함한 비선형모델 등 다양한 회귀모형 적용이 이루어졌다(Choi, 2016). 이후 기계학습과 빅데이터 기술을 도입한 연구들이 진행되어 Lee *et al*. (2017)는 미국 중서부를 대상으로 최적의 농산물 예측 시스템을 가려내기 위하여 MARS (Multivariate Adaptive Regression Splines), SVM (Support Vector Machine), RF (Random Forest), ERT (Extremely Randomized Trees), ANN (Artificial Neural Network), DNN (Deep Neural Network) 총 6가지의 인공지능을 비교 분석하여 DNN이 가장 우수한 성능을 보였으며 수확하기 한 달 전에 옥수수와 콩의 생산량을 예측하는 것이 가능하다는 것을 확인했다(Kim *et al*., 2019). Kim and Kim(2021)은 농산물 생산성과 연관 있는 63개 요인을 선정하고, 예측 모델 연구를 위해 머신러닝 알고리즘 중 Ridge Regression, Random Forest, XGBoost 3가지의 알고리즘에 대하여 MAE와 RMSE 값으로 평가한 결과, XGBoost가 가장 성능이 좋은 것으로 나타났다(Kim and Kim, 2021).

배추 생산량 예측에 대해서는 Lee and Moon(2015)이 선형 방정식을 사용하여 농산물의 생산 단수, 면적, 기후 요소를 예측하였고, 생육도일에 따른 배추의 생장 모델과 데이터를 직접 만들고 회귀분석 이용하여 생산 수량을 예측하는 식을 도출하였다(Lee and Moon, 2015). Kim and Kim(2015)은 배추 생산성에 영향을 미치는 기상요인을 정량화하고 재배부적합지역을 선별하여 최대생산량을 얻는 정식시기 도출 연구를 수행하였다. 또한 Lee *et al*. (2017)은 봄과 가을배추의 정식시기에 따라 실시간으로 측정되는 생육지표 값과 재배기간에의 기상요소를 기반으로 배추의 생산량을 예측하는 연구를 수행한 바 있다.

최근 들어서는 DNN을 사용한 농산물 예측 시스템 연구가 활발히 이루어지고 있다. Khaki and Wang(2019)은 DNN 농산물 예측 모델을 설계하여 2017년 농산물 생산량을 예측하였는데, 이는 비교군이었던 Lasso (The Least Absolute Shrinkage and Selection Operator), SNN (Shallow Neural Networks), RT (Regression Tree)보다 월등히 좋은 성능을 보여주었다(Khaki *et al*., 2019). 그다음 해에는 CNN과 RNN을 이용한 딥러닝 모델을 설계하여 2018년 미국 오하이오주의 콘벨트 옥수수 생산량을 예측했고, 당시 다른 딥러닝 모델보다 우수한 예측력을 보여주었다(Khaki *et al*., 2020).

국내 농산물 예측 알고리즘은 지금까지 특정 작물에 대한 변수를 일일이 계산한 회귀모형이 주축이 되어 왔다. 본 연구는 기존 Khaki *et al*. (2020)의 모델을 우리나라 환경에 맞게 수정 및 개선하였다. 학습 데이터는 GIS 정보를 이용해 세계 어디에서나 얻을 수 있는 토양정보와 기상정보 및 국내 작물에 특화된 지역별 토양적성도를 사용한다. 토양정보와 기상정보는 특정 작물에 국한되지 않으며 토양적성도는 과수류부터 인삼류까지 국내에서 재배되는 총 64가지 작물에 대한 데이터를 제공한다. 본 연구는 데이터의 수나 농산물의 종류에 상관없이 적용 및 확장할 수 있는 국내 작물에 특화된 새로운 농산물 생산량 예측 Framework를 제안한다.

2. 연구 데이터

2.1 사용 데이터

농산물 예측을 위한 딥러닝 학습 데이터 구축은 한 지역에 대해 장시간 축적된 데이터가 필요하다. 그런데, 장기간 자료 축적되었다고 해도 현대적인 농업의 기준과 시스템이 마련된 1980년대 이후부터의 데이터만 수집 가능하므로 한 지역에 대한 데이터는 줄어들 수밖에 없다. 이 문제를 해결하기 위해 한 지역이 아닌 다양한 지역의 데이터를 모아 하나의 데이터셋을 구성한다.

다양한 지역을 하나의 데이터셋으로 사용하면 수집할 수 있는 모든 데이터의 연도와 지역이 일치하지 않는 문제가 발생한다. 데이터가 누락되어 있거나 지역이 새로 생기고 없어지는 등의 경우가 이에 해당한다. 본 연구에서는 학습 데이터에 지역과 연도가 동일하게 존재하는 데이터만 추출하여 사용하며 기상정보, Soilgrids, 토양적성도, 생산량을 학습 데이터로 사용한다.

기상정보는 기상청에서 API (Application Programming Interface)로 제공하는 누적 강수량(mm), 일조량(hr), 지면 온도(℃), 최고기온(℃), 최저기온(℃), 증기압(hPa) 총 6가지를 사용한다. 일별 기상정보는 주별 평균값을 사용하는데, 주별 평균값임에도 불구하고 여전히 값이 누락된 데이터는 해당 기상인자의 평균값으로 대체한다. 읍면동 별로 수집한 데이터는 통계청에서 제공하는 생산량의 지역적 단위와 일치시키기 위해 시군구별 평균을 계산한다.

Soilgrids란 World Soil Information(ISRIC)에서 제공하는 전 세계 토양 속성 예측 시스템이다(Poggio *et al*., 2021). Soilgrids는 250미터의 공간 해상도에서 6개의 표준 깊이 간격에 따라 물의 pH, 토양 유기 탄소 함량, 벌크 밀도, 거친 파편 함량, 모래 함량, 미사 함량, 점토 함량, 양이온 교환 용

량(CEC), 총 질소 밀도, 토양 유기 탄소 밀도, 유기 탄소 스톡, 총 11가지 토양 속성값을 제공한다. Soilgrids는 Fig. 1과 같이 지도로 표출할 수 있는 TIFF 형식으로 제공된다. GIS를 활용하여 속성값을 환원하는 과정은 다음과 같다. 우선 Soilgrids로부터 필요한 항목별로 VRT (Virtual XML)와 OVR (Overview VRT)파일을 다운로드한다. 다운받은 VRT 파일과 OVR의 메타데이터를 이용하여 우리가 원하는 지역의 TIFF 파일 URL 경로를 추출한다. 미리 확보한 국내 읍면동 좌표를 기준으로 해당 좌표별 TIFF 파일을 가져온다. 마지막으로 읍면동 좌표의 경계와 중심점 값을 평균하여 최종 속성값을 구한다. 이러한 방법으로 6가지 깊이 별 11가지 속성정보를 추출하여 학습 파라미터로 사용하였다.

국립농업과학원에서 제공하는 작물별 토양적성도는 작물별 생육에 유리한 토양 환경 분석을 통해 특정 지역을 Fig. 2와 같이 최적지, 적지, 가능지, 저위생산지, 기타로 구분한다 (National Honam Agricultural Experiment Station, 2003). 토양적성도를 학습 데이터로 사용한 이유는 토양적성도 값이 기존 농산물 생산량 관련 연구에서 중요하게 여겨지는 경사, 토질, 배수등급, 침식등급, 토성 등 작물 생장의 핵심 요인을 분석하여 나온 값이기 때문이다. 또한 단순한 지표 값들의 나열이 아니라 농업전문가들의 판단기준을 더해 산출된 값으로 그 신뢰성이 높다. 토양적성도는 공공데이터에서 Open API를 통해 제공한다. 읍면동별 최적지, 가능지, 저위생산지, 기타의 면적을 시군구 단위로 합친 뒤 각 해당 면적을 백분율로 환산하여 사용하였다.

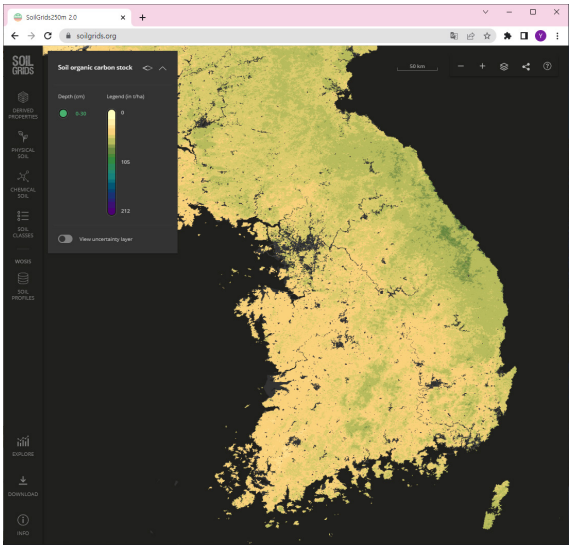

Fig. 1. Soil organic carbon stock map from Soilgrids.org

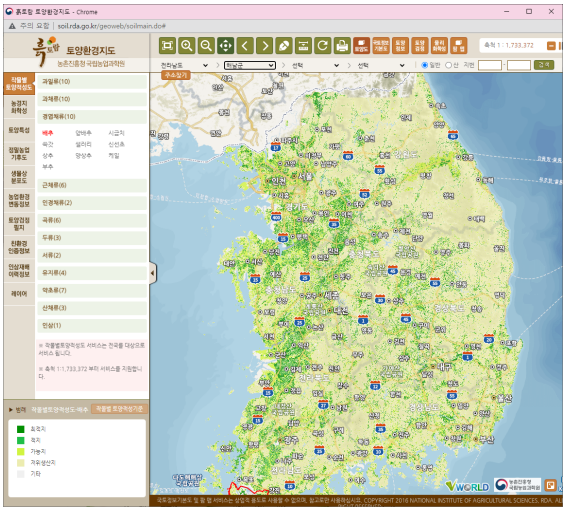

Fig. 2. Soil suitability class from Korean Soil Information System

배추의 시군구별 생산량은 통계청에서 제공하는 API를 통해 수집한다. 10a당 생산량을 사용하고 단위는 kg이다. 이 중에 2015년에 새로이 편입한 세종특별시와 누락데이터가 포함된 제주특별시는 학습에서 처음부터 제외했다. 생산량은 CNN으로 특징을 추출한 기상정보, Soilgrids, 토양적성도 값과 함께 RNN layer에 들어가 시계열적 생산량 예측에 사용되었다.

2.2 배추 데이터의 특성

본 연구는 농산물 중에서도 특히 가격변동이 심한 배추를 대상으로 한다. 배추 데이터의 특성을 파악하기 위해 탐색적 데이터 분석과 통계적 데이터 분석을 했다. 대상이 된 데이터는 통계청에서 얻을 수 있는 1980년부터 2021년까지의 노지가을배추 생산량이다. 배추데이터 및 학습을 위한 데이터 선택에 앞서 배추 데이터가 가진 특성을 살펴볼 필요가 있다. 우리나라 노지가을배추 생산량 그래프는 Fig. 3과 같다. 비교를 통해 데이터의 특성을 더 명확하게 분석하기 위해 기존모델의 옥수수 데이터와 그 차이를 비교하였다.

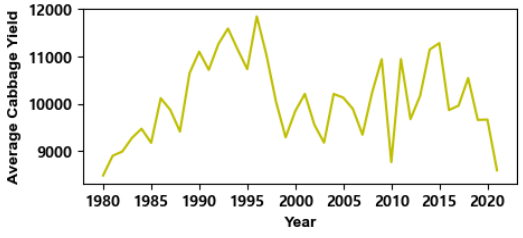

Fig. 3. Average yield of field grown autumn cabbage, South Korea from 1980 to 2021(Unit : kg per 10 acre)

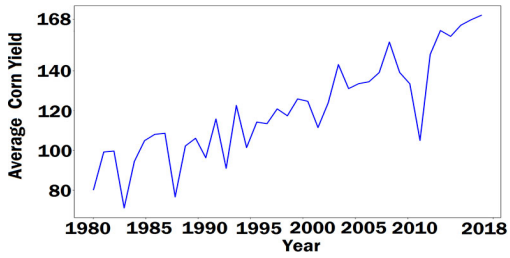

Fig. 4. Average corn yield in Corn Belt, USA from 1980 to 2018 (Unit : bushels per acre)

첫 번째로 기존모델에서 사용한 옥수수 데이터는 콘벨트의 옥수수 수확량을 예측 목표로 하고 있다. Fig. 4 콘벨트의 경우 토지의 대부분이 옥수수밭이어서 옥수수라는 작물이 대표성을 갖지만, 한국에서는 배추를 많이 생산하는 지역일지라도 배추가 지역의 대표성을 띠지는 않는다. 즉, 작은 면적의 농장에 비해 해당 필지에 대한 토양정보나 기상정보와의 밀접성이 떨어진다. 따라서 농지의 면적이 작고 지엽적인 한국의 경우 농산물 생산량 예측을 위해서는 더욱더 필지와 밀접성이 높은 세밀한 정보가 필요하다. 두 번째로 노지가을배추 10a당 평균 생산량이 계속 증가하는 추세가 아니다. 1997년에 IMF의 여파로 배추생산량이 급격히 줄었고 98년과 99년에도 하락 추세를 이어갔다. 2000년 이후에 조금씩 회복되는 추세를 보이지만 2003년에는 중국산 김치 수입의 증가로, 2010년에는 태풍 및 폭염으로 배추생산량이 급감했다. 가장 최근인 2021년에는 그 수치가 1980년도나 2010년과 비슷한 최저 수준으로 떨어졌다. 세 번째로 그래프가 들쭉날쭉하다. 이것은 전년도에 배추 생산량이 많으면 이듬해에는 배추를 생산하지 않는 농민들의 심리와 사회적 요인이 반영되었기 때문이다.

다음으로 통계적 데이터 분석을 했다. 지역별 연도별로 노지가을배추 생산량을 비교한 결과 1980년에는 경기도, 충청남도, 경상북도가 각각 18.9%, 16.3%, 15.1%로 가장 큰 비율을 차지했다. 이후에 경기도와 경상북도의 비율은 점점 작아지고 전라남도와 충청남도의 생산 비율이 점점 커지기 시작하여 90년도에는 충청남도가 가장 큰 비율을 차지했다. 2000년대부터는 전라남도가 치고 올라오기 시작하여 2010년에는 전라남도가 22.7%로 주산지가 되었다. 그 추세는 2021년까지 이어져 현재 전라남도는 가을노지배추 생산량의 27.9%를 차지할 정도로 커졌다.

서울특별시, 부산광역시, 대구광역시, 인천광역시, 광주광역시, 대전광역시, 울산광역시, 세종특별자치시, 제주특별시가 전체 배추생산량에 미치는 영향은 미미하다. 덧붙여 세종특별시는 2015년에 배추생산량 통계에 새로 편입되었고 그 생산량이 10,000kg을 넘기지 못할 정도로 영세하다. 따라서 4년 연속의 데이터를 학습하고 다음 연도의 생산량을 예측하는 본 모델에 적용하기에는 학습 데이터가 없다는 점과 생산량 결괏값이 너무 작아 오히려 모델에 혼란을 가중시킬 가능성 때문에 학습 데이터에서 배제했다. 마찬가지로 제주도 역시 2017년과 2020년도의 데이터가 0으로 아예 누락되어 이상치로 분류하고 전체 학습 데이터에서 배제했다.

3. 제안 모델

Khaki *et al.* (2020)가 발표한 모델에서는 1980년부터 2018년도까지 약 5만 건의 지역별 생산량 데이터를 사용하였다. 그러나 데이터셋을 구성할 때 y값에 해당 지역의 생산량을 사용한 것이 아니라 같은 년도의 모든 지역에 대한 평균 생산량으로 바꾸었다. 이는 지역별 데이터를 사용하여 지역 전체 값을 예측하기 위한 알고리즘이다. 또한 특정 연도의 생산량을 예측하기 위해서는 예측하려는 연도의 데이터까지 학습시켜야 한다는 문제가 있다. 예측하려는 연도의 데이터는 아직 존재하지 않는 데이터로 그 역시 예측값이나 전년도 값으로 대체해야 하기 때문에 데이터가 부정확해진다. 본 연구가 제안하는 수정 모델은 앞에서 언급한 두 가지 문제점을 개선하기 위해 평균 생산량 대신 y값을 그대로 사용하고 예측하기 전 t년도의 데이터를 가지고 이듬해의 생산량을 예측한다.

3.1 배추 데이터의 상관 분석

학습 파라미터로 사용하려는 데이터가 얼마나 생산량과 밀접하게 관련이 있는지 파악하기 위하여 단위 면적당 생산량과 기상정보, Soilgrids, 토양적성도 변수 간의 상관 분석(Pearson Correlation Coefficient)을 했다. 참조 연구에서는 상관성 분석 대신 역전파 할 때 가중치를 업데이트한 파라미터에 1을 더하는 방식으로 어떤 인자가 생산량 예측 결괏값에 가장 큰 영향을 주었는지를 파악하였다. 그러나 본 연구는 특정 학습 파라미터를 사용하는 것 자체가 얼마나 유효한 결과를 도출할 수 있는지를 검증하기 위해 학습 이전에 데이터 자체에 대한 상관 분석을 했다. 귀무가설은 "생산량과 해당 변수는 관계가 없다"고 유의성을 판단하기 위하여 p-value를 구했다. P-value는 통계적 유의성을 판단하기 위해 사용하는 기준으로 사회과학 분야에서는 0.05를 그 값으로 한다. P 값이 0.05 미만이면 귀무가설을 기각하고 변수 간 유의한 차이가 있다는 결론 내릴 수 있다. 본 분석에서는 $p < .001$ 이면서 상관계수가 ± 0.3 이상인 변수에 대하여 단위면적당 생산량과 유의미한 상관관계가 있다고 판단한다. 유의미한 상관관계가

있는 변수들의 목록은 Table 1과 같다. 단위 면적당 생산량과 양의 상관관계를 보이는 변수는 해당 배추밭의 '생산력'을 측정하는 척도일 것으로 판단한다. 60-100cm 깊이에서의 물의 pH가 높을수록 생산력이 좋다. 배추의 토양적성도 가능지 면적이 클수록, 토양의 종합적 요인이 배추 생장에 유리할수록 생산력이 좋다. 60-100cm 깊이에서 토양이 모래-흙의 중간 토질일수록 생산력이 좋다. 일반적으로 음의 상관관계가 있는 변수들은 Table 2와 같다. Table 2는 다음과 같이 해석할 수 있다. 약한 음의 상관관계를 보이는 변수는 pH 7에서 양이온 교환용량, 유기 탄소 함량, 질소, 토양 유기 산소, 유기 탄소 밀도 총 5가지가 있다. 각 함유량에 반비례하여 생산량이 증가한다.

Table 1. Cor index and p-Value of positively correlated yield and training data

Sort	Variable	Cor
Soilgrids	phh2o_mean_60-100cm	0.384
soil suitability	possible area	0.379
Soilgrids	phh2o_mean_100-200cm	0.330
Soilgrids	silt_mean_60-100cm	0.313
weather	3rd week of August vapor pressure	0.308
weather	4th week of December vapor pressure	0.307
weather	2nd week of January vapor pressure	0.307
weather	1st week of February vapor pressure	0.302
weather	1st week of September vapor pressure	0.300
weather	4th week of July vapor pressure	0.300

Table 2. Cor index and p-value of negatively correlated yield and training data

Sort	Variable	Cor
Soilgrids	cec_mean_0-5cm	-0.409
Soilgrids	ocs_mean_0-5cm	-0.397
Soilgrids	ocs_mean_5-15cm	-0.397
Soilgrids	ocs_mean_15-30cm	-0.397
Soilgrids	ocs_mean_30-60cm	-0.397
Soilgrids	ocs_mean_60-100cm	-0.397
Soilgrids	ocs_mean_100-200cm	-0.397
Soilgrids	cec_mean_5-15cm	-0.393
Soilgrids	nitrogen_mean_5-15cm	-0.346
Soilgrids	cec_mean_15-30cm	-0.330
Soilgrids	soc_mean_15-30cm	-0.324
Soilgrids	cec_mean_30-60cm	-0.313
Soilgrids	cec_mean_60-100cm	-0.311
Soilgrids	ocd_mean_0-5cm	-0.304
Soilgrids	soc_mean_5-15cm	-0.302

3.2 수정 모델 구조

본 연구는 Khaki and Wang(2019)이 만든 농산물 생산량 예측 Framework의 구조를 변경한 모델을 제안한다. CNN으로 특징을 추출하고 RNN으로 시계열적 특성을 반영하는 모델의 전체적 구조는 그대로 채택하되 그 외 데이터, 네트워크, 데이터 학습 구조 및 하이퍼파라미터 값은 모두 새롭게 구성했다(Fig. 5). Base 모델의 입력 데이터는 기상, 토양, P(기간별 정식 비율), 당해 평균 생산량 Y_{mean}인 반면 제안한 모델의 입력 데이터는 기상, 토양, AD(토양적성도), 지역별 고유 생산량 값 Y를 그대로 사용했다. Base 모델은 예측하고자 하는 년도 k의 4년 전 데이터부터 예측 연도 k의 데이터까지 입력 데이터로 사용했다. 특히 k년도의 입력 데이터와 Y값은 아직 존재하지 않으므로 기상, 토양, P는 예측값으로 넣고 Y값은 전년도 값으로 대체하였다. 반면 새로 제안하는 모델은 예측하려는 년도를 k로 두고 k-4부터 k-1까지의 실제 데이터만을 입력 데이터로 사용하였다.

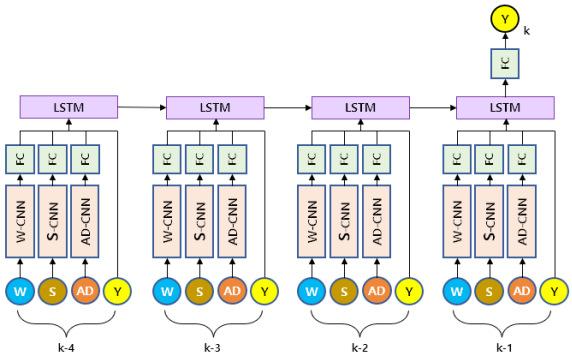

Fig. 5. Structure of Proposed Model

W, S, Y, AD는 각각 기상정보, Soilgrids, 생산량, 토양적성도를 의미한다. W는 날씨 정보를 의미하고 일 년을 주(week)로 표현하여 52개의 데이터로 구성되어 있다. 각 데이터는 일별로 수집한 뒤 주별 평균을 내어 사용했다. 이런 방식으로 총 6개의 날씨 요소를 각각 W-CNN에 넣어 특징을 추출한 뒤 다시 하나로 합쳐 Dense Layer에 넣고 60개의 값을 뽑아낸다. Soilgrids 데이터도 마찬가지 방법으로 12개의 값을 뽑아낸다. AD는 토양적성도 비율을 의미하고 특징을 추출하는 AD-CNN을 거쳐 4개의 값으로 도출된다. Yield는 수확량을 의미하는데 이 값은 따로 Conv Layer를 거치지 않고 그대로 사용했다. 이렇게 축약된 데이터는 LSTM Layer에 들어가게 된다. LSTM의 이전 셀의 정보가 잊히지 않고 그대로 전달되는 특징을 이용해 4년 치 데이터를 학습하고 마지막 5번째 연도의 생산량을 예측한다.

Table 3. Structure of W-CNN to extract features of weather information (FS, NF, S, and Padding are kernel size, number of filters, stride, and padding, respectively)

W-CNN				
Input size	52x1			
Layer name	**FS**	**NF**	**S**	**Padding**
Conv1	3	16	1	valid
Conv2	3	16	1	valid
Average pooling 2	2	-	1	valid
Conv3	3	32	1	valid
Conv4	3	32	1	valid
Average pooling 2	2	-	1	valid
Conv5	3	64	1	valid
Conv6	3	64	1	valid
Average pooling 2	2	-	1	valid
Conv7	3			
Fully-connected	64			
Fully-connected	32			
Fully-connected	16			
Output size	16x1			

Table 4. Structure of S-CNN to extract characteristics of Soilgrids

S-CNN				
Input size	6x1			
Layer name	**FS**	**NF**	**S**	**Padding**
Conv1	3	16	1	valid
Conv2	3	32	1	valid
Average pooling 2	2	-	2	valid
Output size	16x1			

Table 5. Structure of AD-CNN to extract characteristics of soil suitability classes

AD-CNN				
Input size	5x1			
Layer name	**FS**	**NF**	**S**	**Padding**
Conv1	3	16	1	valid
Conv2	3	16	1	valid
Fully-connected	4			
Output size	4x1			

4. 실험 및 분석

전체 학습할 수 있는 데이터는 572개이고 이 중에서 5년 연속의 데이터를 가지고 있는 지역을 대상으로 무작위한 연도 5개를 뽑아 test 데이터셋으로 사용하였다. 하나의 test 데이터셋은 5년 연속 데이터로 구성된다. 5개의 데이터셋에는 총 25개의 데이터가 사용된다. 더 많은 데이터를 사용할 수도 있지만 본 모델의 구조상 5년도 연속으로 존재하는 데이터만 학습 가능하다는 점에서 25개 데이터 공백은 단순히 25개만 비는 것이 아니라 test 연도의 4년 앞, 뒤 및 중간 데이터 모두를 사용할 수 없게 만들기 때문에 학습용 데이터 보존을 위해 test 데이터를 줄였다. 대신 2018년부터 2021년도를 예측 데이터로 두어 test 데이터의 부족함을 보완하고 실제 예측 성능 향상을 꾀했다. 2018년을 예측하는 모델의 경우 2017년도의 데이터까지, 2019년 모델은 2018년도의 데이터까지 학습하는 방법으로 예측 연도 전년도까지의 데이터만 학습하도록 하였다.

새로 제안한 모델의 성능을 분석하기 위해 기존모델의 핵심 아이디어인 y값을 평균한 값으로 대체한 데이터와 y값을 그대로 사용한 데이터 두 가지를 학습시키고 평균 오차를 비교하였다(Table 6). 가중치 초기화 method로는 HeNormal을 사용했다. Optimizer로는 Adam을 사용했고 초기 학습률은 2.46E-5로 설정하였다. epoch는 3830으로 주었다. Loss는 MAE(Mean Absolute Error)를 metric으로는 Pearson COR (Correlation Coefficient)를 사용했다. 각 W-CNN과 S-CNN 및 AD-CNN에서는 ReLU를 활성화 함수로 사용하였고 해당 CNN에서 추출한 특징을 모아 Fully-connected layer에 넣을 때는 ELU를 사용했다.

제안모델과 기존모델의 결과 모두 연도가 증가할수록 점점 오차가 작아지는 것으로 보아 학습하는 데이터가 많아질수록 범용 성능은 향상된다는 것을 알 수 있다. 제안모델과 기존모델의 test 결과를 비교하면 2018년도를 제외하고는 제안한 모델의 오차가 더 작다. 이를 통해 y값을 그대로 사용하는 제안모델이 평균값으로 치환한 기존모델에 비해 범용 성능이 더 우수하다는 것을 알 수 있다. Prediction 결과는 2019년과 2021년도에 대해 기존모델의 오차가 더 작다. 이것은 기존모델이 당해의 평균값을 예측하는 데 특화되었기 때문에 예측값의 분산이 작아서 실제값에 더 가까운 값을 예측하게 된 것이다.

하이퍼파라미터를 달리하여 여러 모델을 학습시켰을 때 test 데이터의 최저 오차는 392.62kg, 평균 오차 백분율은 10% 내외로 대부분 근사하게 예측할 수 있었으나, 2021년 예측 모델만은 다른 어떤 모델보다도 유독 오차가 컸다. 그 이

Table 6. Comparison of Test and Prediction Error(Unit: kg)

Year	Ground Truth	Base model		Proposed model	
		Test	Prediction	Test	Prediction
2018	10547.00	1573.42(14.92%)	1136.74(10.78%)	1597.72(15.15%)	**752.24(7.13%)**
2019	9663.69	1724.32(17.84%)	**986.39(10.21%)**	1560.64(16.15%)	1091.64(11.30%)
2020	9670.37	1385.18(14.32%)	1126.98(11.65%)	**848.4(8.77%)**	1000.62(10.35%)
2021	8598.28	1317.50(15.32%)	1631.56(18.98%)	**698.12(8.12%)**	2222.61(25.85%)

유는 첫 번째로 학습 데이터 부족으로 과적합 되었기 때문이고 두 번째로 2021년의 가을노지배추 생산량이 기존의 패턴을 벗어나 큰 폭으로 하락하였기 때문이다. 통계청에 따르면 2020년 7~9월 대비 배추 가격이 하락한 영향으로 2021년 재배면적이 줄었고, 2021년 배추가 형성되는 시기인 9~10월에 고온과 병해가 발생해 10a당 생산량이 감소했다고 한다. 또한 기존의 패턴은 대부분 한 연도의 생산량이 좋으면 이듬해에는 생산량이 떨어지지만, 2021년은 2019, 2020년에 이어 3년 연속 생산량이 하락하고 있다.

5. 결 론

Table 6은 2018년부터 2021년까지 실제 생산량과 기존모델, 제안모델의 결괏값을 비교하고 오차를 백분율로 나타낸다. 생산량의 경우 2018년을 기점으로 2019년부터 2021년까지 생산량이 지속해서 줄어들고 있는 것을 확인할 수 있다. Test 결과는 학습 전에 완전히 분리한 test 데이터셋으로 검증한 결과로 모델의 범용성을 확인하기 위해 사용되었다. Prediction 결과 역시 학습데이터와 완전히 분리된 또 다른 검증 데이터셋으로써 실제로 예측을 얼마나 잘하는지를 확인하고자 사용되었다. 결과를 전반적으로 분석하면 생산량 예측의 평균 오차는 대략 10%(1,000kg) 내외이다. 배추 한 포기를 약 3kg로 가정한다면 1,000kg은 약 330포기 정도라고 볼 수 있다. 2021년의 예측 결과와 같이 기존의 패턴에서 벗어나는 경우에도 그 오차는 660포기 정도로 가을노지배추 정식 전에 작황을 예측하는 용도로써 무리가 없을 것으로 판단한다. 또한 기존 패턴을 벗어난 2021년과 같은 데이터를 누적하여 기존 모델을 지속해서 전이 학습시키기 때문에 데이터가 쌓여갈 수록 새로운 패턴에 대한 예측력은 점점 좋아질 것으로 기대된다.

다만 현재 모델은 학습 데이터에 대해서는 완성도가 높지만, validation에 대한 cor는 학습에 한계가 있는 것으로 보아 아직 모델의 네트워크나 하이퍼파라미터를 개선할 여지가 남아있다. 과적합을 피하기 위해서 가중치에 L1, L2 규제 및 dropout을 추가하고 네트워크의 크기를 적절하게 줄이는 방법을 취할 수 있다. 또한 데이터 증강을 통해 새로운 데이터를 생성하는 것도 성능 향상에 좋은 방법이 될 수 있다. 그리고 무엇보다 배추의 생산력에 영향을 미치는 자연재해와 심리 사회적 요인을 학습 데이터에 반영하고 배추가 생산되는 읍면동 값에 밀접한 토양정보를 사용한다면 더욱 좋은 성능을 보일 것이다.

현재 연구는 앞으로 두 가지 다른 방향으로 발전시킬 수 있다. 첫째로 모든 지역을 평균한 생산량과 상관관계가 높은 지역만 학습하여 모든 지역에 대한 전체 생산량을 예측하도록 하는 것이다. 이 방법은 지역별로 각기 다른 패턴을 배제하고 비슷한 패턴만 학습시킴으로써 예측의 정확도를 높일 수 있다. 두 번째는 필지별 데이터만 현재 모델에 전이 학습 시켜 정확도가 높은 필지별 생산력을 예측하는 것이다. 현재 통계청에서는 작물별로 1980년도부터 누적된 약 680개의 데이터를 제공하는데 이는 시군구 17개에 지나지 않아 데이터의 양이 적을 수밖에 없다. 이에 읍면동 단위 혹은 필지에 해당하는 자료를 얻을 수 있다면 데이터의 양을 늘려 모델의 예측 정확도를 높이는 방법이 될 수 있다. 이를 위해서는 앞으로 필지별 혹은 읍면동 별로 누적된 생산량 데이터 확보가 우선되어야 할 것이다. 또는 해외에서 데이터를 얻어 모델의 정확도를 개선해 나갈 것이다.

감사의 글

본 연구는 중소벤처기업부의 기술혁신개발사업의 일환으로 수행하였음. [S2982642]

References

Choi, S.C. (2016), *Crop Yields Estimation Using Spatial Panel Regression Model,* Master's thesis, Chonnam National University, Gwangju, Korea, 33p. (in Korean with English abstract)

Kim, N., Ha, K.J., Park, N.W., Cho, J., Hong, S., and Lee, Y.W. (2019), A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern united states, 2006-2015, *ISPRS International Journal of Geo-Information*, Vol. 8, 240. https://doi.org/10.3390/ijgi8050240

Kim, J.H. and Kim, K.D. (2015), An outlook on chinese cabbage production by cultivation type under the RCP8.5 projected climate, *Proceedings of The Korean Society of Agricultural and Forest Meteorology Conference-2015,* 25 August, Jeonju, Korea, pp. 183-186. (in Korean with English abstract)

Kim, S.W. and Kim, Y.h. (2021), A study on the application of machine learning algorithm to predict crop production, *Journal of the Korea Academia-Industrial cooperation Society,* Vol. 22, No. 7, pp. 403-408. (in Korean with English abstract) https://doi.org/10.5762/KAIS.2021.22.7.403

Khaki, S., and Wang, L. (2019), Crop yield prediction using deep neural networks, *Frontiers in Plant Science*, Vol. 10, article 621. https://doi.org/10.3389/fpls.2019.00621

Khaki, S., Wang, L. and Archontoulis, S. V, (2020), A CNN-RNN framework for crop yield prediction, *Frontiers in Plant Science,* Vol. 10, article 1750. https://doi.org/10.3389/fpls.2019.01750

Poggio L., de Sousa L.M., Batjes H.N., Heuvelink B.M.G., Kempen B., Ribeiro E., and Rossiter D., SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, *SOIL*, Vol. 7, pp. 217-240, 2021 https://doi.org/10.5194/soil-7-217-2021

Lee, J.H., Lee, H.J., Kim, S.K., Lee, S.G., Lee, H.S., and Choi, C.S. (2017), Development of growth models as affected by cultivation season and transplanting date and estimation of prediction yield in kimch cabbage, *Journal of BioEnvironment Control*, Vol. 26, No. 4, pp. 235-241. (in Korean with English abstract) https://doi.org/10.12791/KSBEC.2017.26.4.235

Lee, J.G. and Moon, A. (2015), Yield forecasting method for smart farming, *Proceedings of the Korean Institute of Information and Commucation Sciences Conference-2015.* 26 October, Busan, Korea, pp. 619-622. (in Korean with English abstract)

National Honam Agricultural Experiment Station (2003), Soil Survey Theory and Practical Skills, National Honam Agricultural Experiment Station, National Institute of Agricultural Sciences, Jeollabuk-do Iksan (in Korean)