JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 7, pp. 57-65, July 2023
https://doi.org/10.9708/jksci.2023.28.07.057

Research on Selecting Influential Climatic Factors and Optimal Timing Exploration for a Rice Production Forecast Model Using Weather Data

Jin-Kyeong Seo*, Da-Jeong Choi*, Juryon Paik*

*Student, Dept. of Digital Information & Statistics, Pyeongtaek University, Pyeongtaek, Korea
*Student, Dept. of Digital Information & Statistics, Pyeongtaek University, Pyeongtaek, Korea
*Professor, Dept. of Digital Information & Statistics, Pyeongtaek University, Pyeongtaek, Korea

[Abstract]

Various studies to enhance the accuracy of rice production forecasting are focused on improving the accuracy of the models. In contrast, there is a relative lack of research regarding the data itself, which the prediction models are applied to. When applying the same dependent variable and prediction model to two different sets of rice production data composed of distinct features, discrepancies in results can occur. It is challenging to determine which dataset yields superior results under such circumstances. To address this issue, by identifying potential influential features within the data before applying the prediction model and centering the modeling around these, it is possible to achieve stable prediction results regardless of the composition of the data. In this study, we propose a method to adjust the composition of the data's features in order to select optimal base variables, aiding in achieving stable and consistent predictions for rice production. This method makes use of the Korea Meteorological Administration's ASOS data. The findings of this study are expected to make a substantial contribution towards enhancing the utility of performance evaluations in future research endeavors.

▸ **Key words**: Rice Yield, Multiple Linear Regression, Minimum Feature Variables, Climatic Factors, Optimal Timing Exploration

[요 약]

쌀 생산량 예측의 정확성을 높이기 위한 대다수의 연구는 모델의 정확도 증진에 초점이 맞춰져 있다. 이에 비해, 예측 모델을 적용할 대상 데이터 자체에 관한 연구는 상대적으로 미흡하다. 쌀 생산량 데이터에 동일한 종속변수와 예측 모델을 사용하여 다른 특성들로 구성된 두 부류의 데이터에 적용하면, 결과의 차이가 발생하는데 이때 어느 데이터 셋이 더 우수한지 판단하기는 어려운 일이다. 이러한 문제를 해결하기 위해, 예측 모델 적용 전에 데이터 내에서 예측 결과에 큰 영향을 미칠 가능성이 있는 특성들을 선별하고, 이를 중심으로 모델링을 수행하면, 데이터의 구성이 다르더라도 안정적인 예측 결과를 얻을 수 있을 것이다. 본 연구에서는 기상청의 종관기상관측(ASOS) 데이터를 활용하여, 쌀 생산량의 안정적이고 일관된 예측을 위해 데이터 구성 특성들의 조정을 통해 최적의 기반 변수를 선별하는 방법에 대해 제안한다. 본 연구의 결과는 향후 다른 연구에서 성능평가의 유용성을 높이는 데 기여할 것으로 기대한다.

▸ **주제어**: 쌀생산량, 다중선형회귀, 최소특성변수, 날씨요인, 최적시기탐색

• First Author: Jin-Kyeong Seo, Corresponding Author: Juryon Paik
*Jin-Kyeong Seo (sjk2700@naver.com), Dept. of Digital Information & Statistics, Pyeongtaek University
*Da-Jeong Choi (dajung2020@ptu.ac.kr), Dept. of Digital Information & Statistics, Pyeongtaek University
*Juryon Paik (jrpaik@ptu.ac.kr), Dept. of Digital Information & Statistics, Pyeongtaek University
• Received: 2023. 07. 06, Revised: 2023. 07. 24, Accepted: 2023. 07. 25.
• This paper is an extension of that won the Best Paper Award at the 2022 Korea Society of Computer and Information Summer Conference.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 불안정한 쌀 수급불균형 문제는 국내 농가의 생산성과 수익성에 부정적인 영향을 미치고 있으며, 이는 궁극적으로 농업 경제 전반에 큰 위협으로 작용한다. 정부는 쌀 수급 안정을 위해, 생산량 감소로 가격이 급등하면 공공비축미를 시장에 공급하며, 반대로 생산량이 과잉하여 가격이 급락하면 초과분을 매입하여 시장을 격리한다. 이러한 유연 대책은 정확한 예측에 기반을 두어야 한다. 올해의 쌀 생산량과 수요량을 예측하여, 부족 또는 과잉 상황을 판단한 후, 이에 따른 매입량 또는 공급량을 사전에 결정하고 불안정한 상황에 대비하는 것이다. 그러나 현재의 예측 생산량은 실제 생산량과 차이가 있다. 통계청에서 발표한 최근 5년간의 보도자료[1]를 보면, 2018년 7천 톤, 2019년 3만 5천 톤, 2020년 12만 4천 톤, 2021년 5만 5천 톤, 2022년 4만 톤의 오차가 발생했다. 매년 일관된 신뢰 수준의 예측이 이루어지지 않는 현 상황은 불안정한 시장을 만드는 주요 요인이 된다.

예측의 정확도를 향상하려면 예측에 결정적인 영향을 주는 특성들을 식별하고, 이를 중심으로 효율적인 예측 모델을 개발해야 한다. 현재 쌀 생산량 예측에 관한 다양한 연구들이 진행되고 있지만, 이들은 대부분 모델의 정확도를 최적화하는 방향에 초점을 맞추고 있다 [2][3][4][5]. 이에 비해, 예측 모델이 적용될 데이터 자체에 관한 연구는 상대적으로 부족한 상황이다. 만약 예측 모델을 적용하기 전 중요한 결과를 도출할 가능성이 큰 특성들을 미리 식별하고, 이들을 기반으로 예측 모델을 구성한다면, 데이터의 어느 부분을 사용하더라도 안정적인 결과를 얻을 수 있을 것이다. 이는 데이터의 다양성을 고려하면서도 효과적인 예측을 가능케 하는 새로운 방안을 제시한다.

본 연구는 쌀 생산량의 안정적이고 균일한 예측을 수행하기 위해, 데이터를 구성하고 있는 특성 변수들에 대한 조정을 실시하여 예측 모델에 적용할 수 있는 최소한의 특성 변수의 목록을 구성하고자 한다. 이를 위해 다중 선형 회귀분석을 이용한 쌀 생산량 예측 모델에 영향을 미치는 유의미한 특성들을 정의하고, 해당 특성들을 사용하여 모델 정확도가 가장 높은 날씨 시기까지 탐색한다. 논문의 구성은 다음과 같다. 2장에서는 국내 관련 연구 동향과 본 연구에서 사용할 변수 선택법에 대해 알아본다. 3장에서는 예측 모델의 오차를 최소화하면서 가장 적은 개수의 특성 집합을 선택하는 과정에 관해서 기술한다. 4장에서는 선별된 소수의 특성 집합을 활용하여 쌀 생산량 예측에서 높은 정확도를 보장하는 최적의 날씨 시기를 탐색한다. 마지막으로, 향후 연구 방향을 제시한다.

II. Preliminaries

1. Related works

농산물의 생산량 및 가격 예측에 관한 연구들은 수십 년에 걸쳐 다양한 작물들에 대해 진행됐다. 이러한 연구 중에서도, 미곡 생산량 예측에 관한 연구는 작물 생육 모델을 사용한 초기 연구로부터 출발하여, 전통적인 통계 기법을 활용한 연구를 거쳐 현재는 머신러닝과 딥러닝 모델에 기반한 최신 연구들로 발전하였다. 각 연구에서 예측에 사용한 기법이나 모델에 따라 활용한 입력 데이터의 종류는 다르지만, 모든 연구가 공통으로 활용한 데이터는 바로 기상 데이터였다.

1999년 작물생육모델을 이용해 국내 쌀 생산량 예측연구를 수행한 조경숙과 윤진일[2]은 입력 기상 데이터로 1988년부터 1997년까지의 일별 최고기온, 최저기온, 강수량, 수평면일사량을 선정해 벼 생육모형 CERES-rice에 적용하였다. 위성 데이터에 기반한 다중회귀모형을 활용한 홍석영 외 10인[3]의 연구는 MODIS의 NVID 위성 데이터에 2002년부터 2010년까지의 평균온도, 평균 누적 강수량, 평균 누적 일조시간을 추가하였을 뿐만 아니라 지형적 특성 반영을 위해 Cressman 기법을 적용한 일별 강수량, 일별 일사량까지 사용하여 총 5종류의 기상 데이터를 입력 데이터로 추가하였다.

2016년 마종원 외 3인은 위성 데이터 기반 딥러닝 모델을 적용한 연구[4]를 진행했다. 해당 연구에서 활용한 기상 데이터는 2000년부터 2013년까지 4월부터 9월 사이의 온도, 강수량, 일사량 데이터를 최대, 최저, 평균으로 세분화한 9개의 특성 변수를 산출하였다. 해당 특성 변수들에 MODIS 위성 데이터를 더하여 2개의 은닉층으로 구성된 인공신경망 모델과 회선 신경망 모델인 LeNet을 사용하여 쌀 생산량을 예측 후 비교 및 분석을 수행하였다.

2019년 허지나 외 3인은 논문[5]에서 머신러닝 모델에 기상 변수를 이용해 국내 도별 쌀 생산량 예측을 수행하였다. 1986년부터 2018년까지 6월부터 10월 사이의 월평균 최고온도, 월평균 최저온도, 월평균 온도, 강수량, 일조시간을 기상 데이터로 선정 후 도별 상관관계가 높은 3개의 기상 특성을 재선정하여 다중 회귀모형을 구축하였다. 11겹 교차검증으로 도별 쌀 생산량 예측을 수행하였다. 해당 연구는 지역 간 서로 다른 기상 상황을 고려해 적합한 독립변수를 적용하고 도별 생산량을 예측한다는 특징을 갖

는다. 그러나 단순 상관계수를 고려한 변수 선택이라는 점에서, 예측 모델에 따른 변수 선택이라는 본 연구와 차이점을 갖는다.

변수 선택의 유의미성을 보이기 위해 이도영 외 4인은 대파 및 양파가격 예측 모델의 변수 선정 관련 연구[6]를 수행하였다. Lasso 회귀분석을 적용하였으며 특성 변수로는 날씨, 수입/수출량, 재배면적을 사용하였다. 선정된 변수들의 유의미성을 판단하기 위해 기본변수만을 이용한 연구자들의 선행연구 결과와 기본변수에 Lasso를 적용한 모델의 성능을 비교하였다. Lasso를 통한 변수 선정이 유효하다는 결론을 제시했으며, 품목별로 적합한 특성 변수가 상이했다는 결론 또한 도출하였다.

쌀을 포함한 대다수 농산품목의 생산량 예측 연구에 있어서 기상 데이터 선정은 선행연구에 기반하거나 연구자의 판단으로 이루어진다. 기술한 관련 연구들 역시 사용한 기상 데이터의 큰 틀은 같으며, 최저/최고/평균 등 세부적인 사항에 있어서만 차이가 존재한다. 따라서, 기상 데이터는 쌀 생산량 예측연구에서 필수적으로 사용되는 특성이라고 볼 수 있다. 본 연구에서는 논문 [5]의 한계점인 단순 상관관계에 기반한 변수 선택 과정을 개선해 변수 집합을 선정하고, 연구 [6]에서 남긴 시사점을 토대로 쌀 생산량 예측에 적합한 최소한의 특성 변수 구성과 예측 정확도 향상을 위한 최적 날씨 시기 탐색도 수행한다. 이를 위해 본 연구의 선행논문 [7][8]을 확장하여 수행하였다.

2. Variables Selection

변수 선택법은 주어진 모델에서 유의미한 변수를 선택하거나 불필요한 변수를 소거해 모델에서 사용할 변수를 최적화하기 위해 사용한다[9]. 일반적으로, 전진 선택법(forward selection), 후진 소거법(backward elimination), 그리고 단계적 선택법(stepwise selection)으로 나뉜다. 본 장에서는 참고문헌을 기반으로 각 변수 선택법을 정리하고 해당 논문의 연구 결과를 반영해 본 연구에서 적용할 변수 선택법을 선정한다.

전진 선택법은 예측 모델에 적용할 독립변수의 개수를 0부터 시작(독립변수를 전혀 사용하지 않는 모델)해서 유의미한 독립변수를 순차적으로 하나씩 추가하는 방법이다. N-1개의 독립변수로 구성된 이전 변수 집합에 비해 N번째 새로운 변수가 추가된 집합의 모델적합도 증가 폭이 기준치 이상을 상회하지 못할 때 변수 추가는 중지하고 변수 선택이 종료된다. N-1개로 구성된 독립변수 집합이 최종 선택된다. 전진 선택법은 이해하기 쉽고 빠른 변수 선택이 가능하다는 장점이 존재하지만, 한 번 선택된 변수는 다시 제거될 수 없는 단점을 갖는다.

후진 소거법은 전진 선택법과 반대되는 변수 선택법이다. 모든 독립변수를 적용한 모델에서 시작하여 유의미하지 않은 독립변수를 하나씩 제거하며 모델적합도를 비교해나간다. N개로 구성된 독립변수 집합에서 어떤 독립변수 1개를 제거한 N-1개로 구성된 변수 집합을 적용한 모델적합도가 일정 기준치 이하로 크게 감소할 때 변수 제거가 종료되고 최종 N개로 구성된 독립변수가 남는다. 후진 소거법은 모든 변수를 고려하기 때문에 유의미한 변수가 제거될 가능성이 작아 상대적으로 안전한 방법이라는 장점이 존재하지만, 학습 초기 오랜 계산 시간이 소요된다는 점과 전진 선택법과 마찬가지로 한 번 제거된 변수는 다시 추가되지 않는다는 단점을 갖는다.

마지막으로, 단계적 선택법은 전진 선택법과 후진 소거법의 단점을 보완해 결합한 방법으로 한 번 선택되거나 소거된 변수를 다시 고려하여 양방향으로 변수를 선택하는 방법이다. 상수항만 포함된 모형에서 시작하여 전진 선택법과 후진 소거법을 번갈아 시행하며 새롭게 선택된 변수가 유의미하지 않을 때까지 진행한다[10]. 단계적 선택법은 한 번 선택되거나 소거된 변수가 고정되지 않기 때문에 결과가 비교적 안정적이며, 이전의 전진 선택법과 후진 소거법보다 상대적으로 많은 경우의 수를 탐색함으로써 보다 최적의 결과를 얻을 수 있다. 그러나 더 많은 경우의 수를 계산하는 만큼 더 긴 시간이 소요되어 학습이 느리다는 단점이 있다.

논문 [9]에 따르면 세 가지 변수 선택법에 대해 오류 개선율을 기준으로 내림차순 정렬했을 때, 단계적 선택법−후진 소거법−전진 선택법 순으로 결과를 나타낸다. 이는 변수 감소 비율을 기준으로 내림차순 했을 때의 결과와 같다. 다만 두 가지 기준 모두에서 차이에 대한 충분한 통계적 유의성은 확보되지 않는다. 단순히 계산 시간만을 따진 효율성 관점에서는 전진 선택법−후진 소거법−단계적 선택법 순으로 빠르게 결과를 얻을 수 있다. 반면에, 예측 정확도 향상 관점에서는 단계적 선택법−후진 소거법−전진 선택법 순으로 결과가 확인되었다. 본 연구에서는 시간은 오래 소요되지만, 오류 개선율과 예측 정확도 측면에서 가장 높은 성능을 보이는 단계적 선택법을 채택하여 쌀 생산량 예측 모델에 적용한다.

III. Features Extraction

본 장에서는 날씨 데이터를 비롯하여 필요한 원시 데이터를 수집하고 정제하여 추출하는 전반적인 과정을 기술한다. 완성된 데이터 셋에 2장에서 설명한 단계적 선택법을 적용하여 유의미한 변수를 선정한다. 최종적으로, 남은 변수에 대해 다중공선성 검사를 진행 후 변수 선택을 종료한다.

1. Collecting and Cleaning Data

1.1 Weather data

사용한 날씨 데이터는 기상청의 종관기상관측 (ASOS, Automated Synopic Observing System) 데이터[11]이다. 농업과 밀접한 기상 현상에 대한 농업기상관측 (AAOS, Automated Agricultural Observing System) 데이터를 제공하지만, 해당 데이터는 전국 농촌 지역을 대상으로 한 것이 아니라 11개 제한된 지점 정보만을 포함한다. 반면 ASOS 데이터는 총 103개 지점에 대한 정보를 제공하며 전 지역에 대한 전국 데이터이다. 본 연구에서는 ASOS 데이터를 사용하였으며, 충분한 데이터 확보를 위해 2012년부터 2021년까지 10년의 데이터를 일별로 수집하였다. 일별 데이터를 사용하지만, 벼 재배와 상관없는 기간은 불필요하다. 농촌진흥청에 따르면 벼 모내기는 5월에 시작하고 벼 수확은 10월에 마친다. 지역별 편차가 존재할 수 있지만 평균적으로 5월부터 10월까지의 날씨 데

Table 1. Variable Explanation Related to Weather or Rice

Variable name	Explanation	Variable name	Explanation
place	observation area	avg_sea_press	average sea level pressure
date	observation date	possi_sunshine	the duration of possible sunshine
low_temp	lowest temperature	total_du_sunshine	total duration of sunshine
high_temp	highest temperature	oh_max_solar_radi_time	1 hour maximum solar radiation time
high_temp_time	highest temperature time	oh_max_am_sunshine	1 hour maximum amount of solar radiation
precip_duration	precipitation duration	total_solar_radi	total solar radiation
tm_max_precip	10 minute maximum precipitation	day_max_dep_snow	day maximum depth of snowfall
tm_max_precip_time	10 minute maximum precipitation time	day_max_dep_snow_time	day maximum depth of snowfall time
oh_max_precip	1 hour maximum precipitation	day_max_dep_old_snow	day maximum depth of old snowfall
oh_max_precip_time	1 hour maximum precipitation time	day_max_dep_old_snow_time	day maximum depth of old snowfall time
day_precip	day precipitation	total_th_new_snow	Total 3 hours of new snow
max_ins_wind_speed	maximum instantaneous wind speed	avg_am_cloud	average amount of clouds
max_ins_wind_speed_direc	maximum instantaneous wind speed direction	avg_mid_low_am_cloud	average amount of clouds in the middle layer and lower layer
max_ins_wind_speed_time	maximum instantaneous wind speed time	avg_ground_temp	average ground temperature
max_wind_speed	max wind speed	low_grass_temp	lowest temperature above the grass
max_wind_speed_direc	max wind speed direction	avg_fivec_soil_temp	average 5cm soil temperature
max_wind_speed_time	time of maximum wind speed	avg_tenc_soil_temp	average 10cm soil temperature
avg_wind_speed	average wind speed	avg_twec_soil_temp	average 20cm soil temperature
air_distan_sum	the sum of the distances the air has flowed	avg_thirc_soil_temp	average 30cm soil temperature
avg_dew_point_temp	average dew point temperature	avg_fifc_soil_temp	average 0.5m soil temperature
min_rel_humid	minimum relative humidity	avg_onem_soil_temp	average 1.0m soil temperature
min_rel_humid_time	minimum relative humidity time	avg_onefivem_soil_temp	average 1.5m soil temperature
avg_rel_humid	average relative humidity	avg_threem_soil_temp	average 3.0m soil temperature
avg_vapor_press	average vapor pressure	avg_fivem_soil_temp	average 5.0m soil temperature
avg_loc_press	average local pressure	total_mass_evapor	total mass evaporation
high_sea_press	highest sea level pressure	total_small_evapor	total small evaporation
high_sea_press_time	highest sea level pressure time	ntn_precip	nine to nine precipitation
low_sea_press	lowest sea level pressure	fog_duration	fog duration
low_sea_press_time	lowest sea level pressure time	area	field area
yield	rice yield		

이터만을 추출하여 사용한다.

1.2 Rice data

쌀 생산량 데이터는 국가통계포털(KOSIS)이 제공하는 미곡 생산량(백미, 90.4%) 데이터[12]를 이용한다. 사용되는 미곡 생산량 데이터는 날씨 데이터와 동일 기간인 2012년부터 2021년까지의 10년 동안의 데이터이다. 전체 쌀 생산량은 연도별로 취합한 값이 아닌, 지역별로 계산된 생산량 통계를 이용 후, 해당 지역 날씨 데이터와 결합하였다. 이는 지역별로 서로 다른 날씨 데이터에 맞추어 좀 더 정확한 생산량을 제공하기 위함이다. 또한, 생산된 논의 면적변수도 지역의 쌀 생산량에 맞춰 함께 결합하였다. 이는 동일 단위에 대한 쌀 생산량을 제공하기 위해서이다.

Table 1은 기상 데이터와 미곡 생산량 데이터를 종합한 변수들과 각 변수에 대한 설명을 보인다. 해당 변수들에 단계적 선택법을 적용하기 전, 최소한의 전처리를 수행한다. 종속변수인 생산량(yield)과 독립변수인 관측지점(place) 그리고 관측날짜(date) 변수는 예측에 있어 필요한 변수이므로 단계적 선택법에서 제외한다. 58개의 남은 변수 중에서 날씨와 연관된 변수들 중 매일의 데이터를 얻을 수 없는 비와 관련된 6개의 변수– precip_duration (우기), tm_max_precip (최대강수량 10분), tm_max_precip_time (최대강수시간 10분), oh_max_precip (최대강수량 1시간), oh_max_precip_time (최대1시간 강수시간), day_precip (일 강수량) –은 제외한다. 제거 후 남은 52개의 변수에서 결측값이 50% 이상 존재하는 12개의 변수를 다시 제거하고 최종으로 남은 40개 변수의 결측값들은 0으로 치환하여 전처리를 마친다.

2. Stepwise Selection

40개 변수를 대상으로 다음 단계에 따라 단계적 선택법을 수행하였다. 유의수준은 5%로 한다. ① 모든 변수를 개별적으로 하나씩 사용하여 선형회귀 모델에 적합한다. ② 가장 작은 유의확률을 가진 모델을 선택한다. ③ 단계 ②에서 산출된 모델의 유의확률이 5% 미만이라면 해당 변수를 선택한다. ③′ 만약 단계 ②에서 선택된 모델의 유의확률이 5% 이상이라면 변수 선택을 종료한다. ④ 선택된 변수 안에서 순환하며 선형회귀 모델에 적합시킨다. ⑤ 가장 큰 유의확률 값을 가진 모델을 선택한다. ⑥ 모델의 유의확률이 5% 이상이라면 해당 변수를 제거한다. 만약 ⑥′ 모델의 유의확률이 5% 미만이라면 변수 선택을 종료한다. ① ~ ⑥ 단계를 모두 수행한 한 과정을 1 step이라 할 때, 40개 변수 전체에 대해서 단계적 선택법을 적용하여 총 30회의 step이 수행되었다. 각 step을 수행하는 동안 변수 11개가 제거되고 최종 29개의 변수가 선택되었다. Table 2는 단계적 선택법을 통해 선정된 변수들이다.

Table 2. Variables after Processing Stepwise Selection

Variable name	Explanation	Variable name	Explanation
avg_am_cloud	average amount of clouds	high_temp_time	highest temperature time
area	field area	max_ins_wind_speed_direc	maximum instantaneous wind speed direction
max_wind_speed	max wind speed	avg_mid_low_am_cloud	average amount of clouds in the middle layer and lower layer
max_ins_wind_speed	maximum instantaneous wind speed	day_precip	day precipitation
possi_sunshine	the duration of possible sunshine	min_rel_humid	minimum relative humidity
low_temp	lowest temperature	tm_max_precip_time	10 minute maximum precipitation time
high_temp	highest temperature	avg_dew_point_temp	average dew point temperature
total_du_sunshine	total duration of sunshine	avg_temp	average temperature
low_temp_time	lowest temperature time	precip_duration	precipitation duration
avg_ground_temp	average ground temperature	tm_max_precip	10 minute maximum precipitation
avg_rel_humid	average relative humidity	high_sea_press_time	highest sea level pressure time
low_sea_press	lowest sea level pressure	low_sea_press_time	lowest sea level pressure time
avg_loc_press	average local pressure	max_ins_wind_speed_time	maximum instantaneous wind speed time
avg_sea_press	average sea level pressure	high_sea_press	highest sea level pressure
low_grass_temp	lowest temperature above the grass		

3. Multicollinearity Elimination

단계적 선택법 후 총 29개의 변수가 선정되었지만, 여전히 많은 변수로 데이터가 구성되기 때문에 다중공선성 발생 가능성이 있다. 이를 해결하기 위해 각각 다른 단위를 가지는 변수들에 대해서 정규화 수행 후 각 변수의 계수 coef를 파악한다. 계수 coef는 해당 변수가 종속변수에 미치는 영향의 크기로 계수의 절댓값이 클수록 영향력이 큰 변수이다. 변수들의 coef 값을 계산하여 가장 큰 차이를 보이는 6개의 변수를 선정하였다. 이들 선택된 변수들은 모두 계수의 절대값이 3000 미만인 것으로 확인되었다 해당 변수들은 high_sea_press_time(최고해면 기압시각), low_sea_press_time(최저해면기압시각), max_ins_wind_speed_time(최대순간풍속시각), tm_max_precip_time(10분최다 강수량시각), max_ins_wind_speed_direc(최대 순간풍속풍향), low_temp_time(최저 기온시각)이다. 종속변수 생산량(yeild)에 대한 해당 변수들의 영향력을 검증하기 위해, 상관계수를 분석하였으며, Table 3에 제시된 값처럼 매우 낮은 상관관계를 보였다. 가장 우측에 제시된 값이 종속변수 생산량과의 상관계수 값이다. 해당 변수들을 제거하고 변수 23개를 대상으로 다중공선성을 해결한다.

일반적으로, 다중공선성이 존재함을 의미하는 기준은 다중공선성 평가 지표인 VIF(Variance Inflation Factor) 값이 5 또는 10보다 클 때이다. VIF 값이 1에 가까울수록 해당 변수가 다른 변수와의 상관관계가 낮아 다중공선성의 문제가 적음을 의미하고, VIF 값이 1보다 크게 증가하면 다중공선성 문제가 있다는 신호로 받아들일 수 있다. 본 논문에서는 10을 초과하면 다중공선성이 존재한다고 판단하여 제거한다. 본 연구에서는 참고문헌 [13]에서 제시한 다중공선성 발생 변수 제거법을 따른다. ① 모든 변수의 다중공선성을 계산한다. ② 다중공선성이 가장 큰 변수를 제거한다. ③ 남은 변수를 선형회귀 모델에 적합 후 전체 변수의 유의확률을 확인한다. ④ 유의수준 5%에서 유의하지 않은 변수가 있다면 제거한다. ⑤ 만약 모든 변수의 다중공선성이 10보다 작다면 변수 제거를 종료한다. 전체 23개의 변수를 대상으로 다중공선성이 10보다 작을 때까지 ① ~ ⑤ 단계를 반복한 결과 총 12번의 루틴이 수행되었다. 다중공선성을 갖는 17개의 변수가 제거되고 최종적으로 6개의 변수가 남았다. Table 4는 최종적으로 선택된 변수들과 해당 변수들의 다중공선성 지표값이다.

Table 4. Variables after Multicollinearity Elimination

Variable Name	VIF
area	3.252657
max_wind_speed	6.106088
avg_am_cloud	2.431192
total_du_sunshine	2.811281
min_rel_humid	6.290695
precip_duration	1.318146

다중공선성 제거 후 최종적으로 선택된 특성 변수 6개 사이의 상관계수 값을 계산하였다. Fig 2 테이블은 각 변수 사이의 상관계수를 나타내며 Fig 3은 상관계수 간의 관계를 색상이 진한 부분은 강한 상관관계를, 연한 부분은 약한 상관관계를 나타내는 히트맵이다.

Table 3. Six Variables with Low Correlation Values

	high_sea_press_time	low_sea_press_time	max_ins_wind_speed_time	10 minute maximum precipitation time	max_ins_wind_speed_direc	lowest temperature time	yield
high_sea_press_time	1.000000	-0.626487	-0.116506	-0.141265	0.076104	0.149037	0.008283
low_sea_press_time	-0.626487	1.000000	0.211036	0.140985	-0.072747	-0.161255	-0.000081
max_ins_wind_speed_time	-0.116506	0.211036	1.000000	0.090655	0.026434	-0.096400	0.003972
10 minute maximum precipitation time	-0.141265	0.140985	0.090655	1.000000	-0.082819	0.150080	-0.012582
max_ins_wind_speed_direc	0.076104	-0.072747	0.026434	-0.082819	1.000000	0.033620	0.064215
lowest temperature time	0.149037	-0.161255	-0.096400	0.150080	0.033620	1.000000	0.000122
yield	0.008283	-0.000081	0.003972	-0.012582	0.064215	0.000122	1.000000

	area	max_wind_speed	avg_am_cloud	total_du_sunshine	min_rel_humid	precip_duration
area	1.000000	0.039676	-0.109385	-0.020204	0.049784	-0.073591
max_wind_speed	0.039676	1.000000	0.079189	0.031454	-0.016286	0.158421
avg_am_cloud	-0.109385	0.079189	1.000000	-0.381658	0.374870	0.388109
total_du_sunshine	-0.020204	0.031454	-0.381658	1.000000	-0.606422	-0.269320
min_rel_humid	0.049784	-0.016286	0.374870	-0.606422	1.000000	0.243515
precip_duration	-0.073591	0.158421	0.388109	-0.269320	0.243515	1.000000

Fig. 2. Correlation Values of the Selected Six Variables

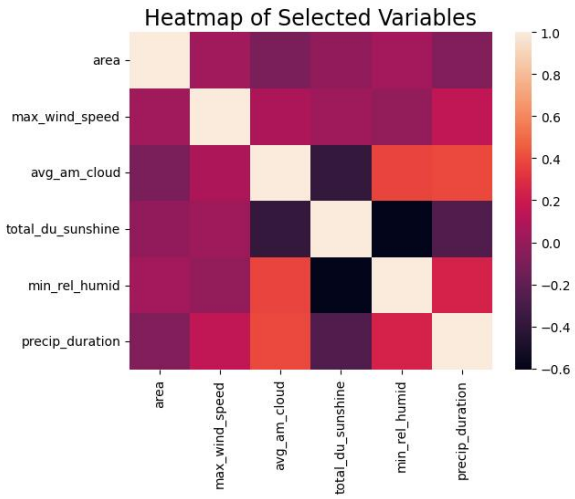

Fig. 3. Heatmap of Six Variables

IV. Exploration of Optimal Weather Periods

본 장에서는 3장의 변수 선택법 및 다중공선성 제거 과정 후 최종 선택된 Table 4의 변수들로 쌀 생산량 예측을 수행한다. 예측 모델을 수행한 환경은 구글 Calab이며 분석방식은 다중선형 회귀분석을 채택한다. 다중선형 회귀분석은 다수의 독립변수가 종속변수에 미치는 영향의 정도를 개별적으로 파악하는 것이 가능할 뿐만 아니라, 예측값이 도출된 이유를 설명할 수 있으므로, 쌀 생산량에 유의미한 영향을 미치는 최소 개수의 특성 변수를 찾고자 하는 본 연구에 부합한다.

더 정확한 예측값을 위해 쌀 생산에 있어 가장 중요한 특징인 날씨 시기를 탐색하여 예측값을 산출한다. 쌀 생산 시기 데이터와 같은 2012년부터 2021년까지 10년 기간 중 모내기를 시작하는 5월부터 추수를 시작하는 10월 데이터를 기준으로 한다. 동일한 시간 간격으로 데이터를 나누고 이를 과거로 회귀시켜 쌀 생산량 회귀모델의 RMSE를 비교한다. 이러한 접근법은 쌀 수확 며칠 전의 데이터가 생산량에 어떠한 정도로 큰 영향을 끼치는지 파악하기 위한 목적을 갖는다. 모델을 학습하고 평가하기 위해 훈련 데이터(80%)와 테스트데이터(20%)로 분리하였으며, 평가 지표로는 평균제곱근오차 (RMSE; Root Mean Square Error)를 사용하였다.

30일을 한 달로 하여 5월에서 10월까지 6개월 동안의 날씨 데이터를 시작으로 과거 6개월씩 조정하며 RMSE 값을 계산하였다. 당해의 데이터만을 적용하기 위해 과거로의 조정은 1월까지만 진행하였으며 쌀 생산지점 데이터를 기준으로 동일 지점 내에서만 시기가 조정되도록 처리하였다. Table 5는 10년 동안의 10월부터 1월까지의 데이터를 6개월 간격으로 예측한 RMSE 값이다. 5월에서 10월까지의 기본 벼농사 시기를 기준으로 하여 훈련된 모델의 RMSE 값은 36105.249이다. 6개월 간격으로 시기를 조정한 총 5개의 모델 중 가장 작은 RMSE 값을 갖는 모델은 3월부터 8월까지의 날씨 데이터를 이용한 모델로 35403.141이다. 따라서 기본 벼농사 시기에서 2개월 과거 데이터인 3월에서 8월 데이터를 적용한 모델의 쌀 생산량 예측 정확도가 가장 높으며, 이는 쌀 생산에 가장 많은 영향을 미치는 계절적인 시기가 3월부터 8월이라는 것을 의미한다.

Table 5. RMSE Values with Every 6 Months Backwards

Duration	RMSE
October ~ May	36105.249
September ~ April	35792.653
August ~ March	35403.141
July ~ February	35839.121
June ~ January	35833.844

Fig 4는 3월부터 8월까지의 데이터를 활용하여 구축한 모델의 검증 결과를 보여주며, 이는 98.06%의 높은 정확도를 달성함으로써 본 연구의 모델이 신뢰성을 확보하였음을 입증한다.

```
lin_reg.score(X_train, y_train)
lin_reg.score(X_test, y_test)

0.9805946943763436
```

Fig. 4. Accuracy of Selected model

최적의 날씨 시기를 파악했다면, 쌀 생산량 예측에 있어 가장 높은 정확도를 보이는 날씨 데이터가 어느 월에 해당하는지 판단할 필요가 있다. 3월부터 8월까지의 데이터를 월 단위로 적용하여 쌀 생산량 예측 모델의 RMSE 값을 계산하였다 그 결과, Table 6에서 보이는 값이 도출되었다. 따라서, 1년 중 8월의 날씨가 쌀 생산량에 가장 큰 영향을 미치는 것으로 결론지을 수 있다.

Table 6. Monthly RMSE Values

Month	RMSE
March	36053.771
April	36700.675
May	36145.712
June	36253.564
July	36289.126
August	35525.532

V. Conclusions

본 연구는 다중선형 회귀분석을 활용하여 쌀 생산량 예측 모델에 적용될 최소한의개수의 특성 변수들을 선정하였으며, 이를 통해 쌀 생산량 예측 정확도에 가장 큰 영향을 미치는 날씨 시기와 그에 해당하는 특정 월을 도출하였다. 총 58개의 특성 변수 중에서, 결측값 비중이 높은 18개의 변수를 제외하였고, 이후 단계적 선택법을 통해 11개의 변수를 추가로 제거하였다. 남은 29개의 변수 중에서, '생산량'에 거의 영향을 미치지 않는 6개의 변수와 높은 다중공선성을 보이는 17개의 변수를 각각 제외함으로써, 최종적으로 6개의 특성 변수만을 모델에 적용하였다.

선택된 6개의 변수를 사용하여 쌀 생산량 예측의 정확도를 높이는데 가장 유리한 날씨의 시기를 탐색한 결과, 3월부터 8월까지의 데이터가 가장 적합하다는 결론을 내렸다. 또한, 이 기간 중에서도 8월 데이터가 가장 좋은 예측 결과를 보였다. 이어지는 연구에서는 8월의 기상 특성 중 어떤 요소들이 쌀 생산량에 가장 큰 영향을 미치는지에 대해 분석하고자 한다.

이러한 연구 과정을 통해, 다양한 날씨 조건 중에서 생산량 예측 모델에 가장 긍정적인 영향을 미치는 상위 6개의 특성을 도출하였으며, 생산량 예측의 정확도를 향상시킬 수 있는 날씨의 시기를 파악할 수 있었다. 특히, 최소한의 특성을 분석함으로써 객관적인 비교를 가능하게 한 점이 본 연구의 중요성을 대표한다. 그러나, 본 연구에서 사용한 다중선형 회귀분석은 기본적인 회귀기법으로서, 규제와 관련된 한계가 있다. 이를 보완한 회귀기법[14]이 제안되어 있으며, 앞으로의 연구에서는 이러한 기법들을 통해 최소 특성 변수들과 최적의 날씨 시기 데이터를 적용, 가장 높은 예측 정확도를 보이는 모델을 찾는 비교분석 연구를 수행하고자 한다.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2021R1F1A1064073)

REFERENCES

[1] Statistics Korea, "Rice Yield Survey Result Press Release & Rice Expected Yield Survey Result Press Release," 2018~2022, Available from: https://kostat.go.kr

[2] K. S. Cho, J. I. Yun, "Regional Crop Evaluation and Yield Forecast of Paddy Rice Based on Daily Weather Observation," Korean Journal of Agricultural and Forest Meteorology, Vol.1, No.1, 1999, pp.12-19.

[3] S. Y. Hong, J. N. Hur, J. B. Ahn, J. M. Lee, B. K. Min, C. K. Lee, Y. H. Kim, K. D. Lee, S. H. Kim, G. Y. Kim, K. M. Sim, "Estimation Rice Yield Using MODIS NDVI and Meteorological Data in Korea," Korean Journal of Remote Sensing, Vol.28, No.5, 2012, pp.509-520. DOI:https://doi.org/10.7780/kjrs.2012.28.5.4

[4] J. W. Ma, C. H. Nguyen, K. D. Lee, J. Heo, "Convolutional Neural Networks for Rice Yield Estimation Using MODIS and Weather Data: A Case Study for South Korea," Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, Vol.34, No.5, 2016, pp.525-534. DOI: https://doi.org/10.7848/ksgpc.2016.34.5.525

[5] J. N. Hur, K. M. Shim, Y. S. Kim, K. K. Kang, "Estimation of Rice Yield by Province in South Korea Based on Meteorological Variables," Journal of the Korean Earth Science Society, Vol.40,

No 6, 2019, pp.599-605. DOI: https://doi.org/10.5467/JKESS.2019.40.6.599

[6] D. Y. Lee, Y. W. Yang, J. H. Lee, J. H. Park, M. G. Kang, "A Study on Selection of Variables in Forecasting Model of Agricultural Products Price Using Lasso Regression," Journal of D-Culture Archives, Vol.4, No.2, 2021, pp.123-136. DOI: 10.23089/jdca.2021.4.2.009

[7] J. K. Seo, D. J. Choi, K. Ko, J. Paik, "Analyzing Significant Variables from a Linear Regression-Based Prediction Model for Rice Prices," Proceedings of the 66th Korea Society of Computer Information Summer Conference, Vol.30, No.2, 2022, pp.39-42. UCI :G100:I100-KOI(KISTI1.1013/PCD.CFKO202232249589462

[8] D. J. Choi, J. K. Seo, K. Ko, J. Paik, "Prediction of Rice Prices and Search for a Period of Weather Affecting the Prices Based on a Linear Regression Model," Proceedings of the 66th Korea Society of Computer Information Summer Conference, Vol.30, No.2, 2022, pp.37-38. UCI:G100:I100-KOI(KISTI1.1013/PCD.CFKO202232249595463)

[9] N. H. Rye, H. S. Kim, P. S. Kang, "Evaluating Variable Selection Techniques for Multivariate Linear Regression," The Journal of Korean Institute Of Industrial Engineers, Vol.42, No.5, 2016, pp.314-326. DOI: https://doi.org/10.7232/JKIIE.2016.42.5.314

[10] H. S. Yong, H. J. Lee, S. I. Lee, "A Study to Explore Factors Related to Sleep Duration among Students Based on 2009 Young Penal Data from Korea Employment Information Service," Journal of Korean Data Analysis Society, Vol.15, No.2, 2013, pp.785-798. UCI: G704-000930.2013.15.2.030

[11] Korea Meteorological Administration. Automated Synoptic Observing System. Available from: https://data.kma.go.kr

[12] Korea Statistical Information Service, Agricultural Production Survey:Rice Production(white rice. 90.4%), 2022, Available from: https://kosis.kr

[13] S. Jung, "On the Multicollinearity in a Linear Regression Analysis," Gwangju Health University, Thesis Collection, Vol.22, 1997, pp.293-302

[14] H. S. Nalwa, Y. T. Jeong, Y. J. Choi, "Statistical Research Methods in Physical Education: Focusing on Regression Analysis," The Journal of Humanities and Social science, Vol.13, No.3, 2022, pp.1495-1510. DOI: http://dx.doi.org/10.22143/HSS21.13.3.105

Author

Jin-Kyeong Seo received the B.S. degree in Department of Digital Information and Statistics at Pyeongtaek University, Korea in 2022.

Da-Jeong Choi is an undergraduate student in Department of Digital Information and Statistics at Pyeongtaek University, Korea.

Juryon Paik received the B.E. degree in Information Engineering from Sungkyunkwan University, Korea in 1997. She worked at the Samsung SDS company about a year. She achieved her M.E. and Ph.D. degrees in Computer Engineering from the sane university in 2005 and 2008, respectively. Dr. Paik joined the faculty of the Department of Digital Information and Statistics at Pyeongtaek University, Pyeongtaek-si, Korea, in 2016. She is currently an assistant professor. She is currently interested in recommeder systems, big data mining, information retrieval.