International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Optimizing Weather Variables for Rice Yield Prediction Using Linear Regression

Juryon Paik
Department of Data Information and Statistics, Pyeongtaek University, Pyeongtaek-si, Gyeonggi-do, South Korea-17869
Email address: jrpaik@ptu.ac.kr

Abstract*—This study proposes a multivariate linear regression model to forecast rice yield using optimal weather features and periods. Utilizing weather data from ASOS (Automated Synoptic Observing System) and rice yield data from KOSIS, the study identified key features affecting rice production through stepwise selection and addressed multicollinearity issues. The results demonstrated that six features, including field area, maximum wind speed, and total sunshine duration, significantly influenced rice yield predictions. The optimal prediction period was determined to be March through August, with August being the most impactful month. These findings provide valuable insights for agricultural planning and resource allocation, highlighting the need for advanced forecasting models.*

***Keywords*—** *Feature selection: multiple linear regression: rice yield: weather condition*

I. INTRODUCTION

Rice is a staple food crop consumed by billions of people worldwide, making its production critical for food security and economic stability. Accurately forecasting rice yield is essential for optimizing agricultural practices, ensuring adequate resource allocation, and mitigating risks associated with climate variability. However, predicting rice yield is challenging due to the complex interactions between weather conditions, field characteristics, and other environmental factors.

Traditional rice yield forecasting models often rely on a broad set of variables without sufficient optimization, which can lead to overfitting or reduced accuracy. Recent advancements in regression analysis have demonstrated the potential for feature selection techniques to enhance prediction performance. Specifically, stepwise selection methods have been effective in identifying the most significant variables while eliminating irrelevant ones, thus improving model efficiency and interpretability.

This study focuses on utilizing multiple linear regression analysis to develop a reliable rice yield prediction model. By leveraging weather data and field area information, the study aims to identify the optimal set of features that significantly impact rice production. Furthermore, this research explores the temporal influence of weather data, identifying the specific periods and months that have the greatest impact on rice yield accuracy.

The findings of this study provide valuable insights into the optimization of weather variables and temporal factors, contributing to the development of more accurate and practical forecasting models for rice yield.

II. RELATED WORKS

Several international studies have explored the use of regression models and feature selection methods for predicting agricultural yields, emphasizing the impact of environmental factors such as weather and soil conditions.

Tao et al. [1] developed a regression-based approach to predict rice yields in East Asia, incorporating multiple weather factors such as solar radiation, humidity, and temperature. Their research highlighted the importance of identifying optimal time windows for weather data to improve prediction accuracy. This aligns with the findings of the present study, which identified March to August as the most critical period for rice yield prediction.

Lobell et al. [2] demonstrated the utility of statistical models in understanding the relationship between climate variables and crop yields. Their study employed linear regression to evaluate the impact of temperature and precipitation on wheat and maize production across multiple regions. The results emphasized the importance of season-specific weather variables in enhancing yield prediction accuracy, which is consistent with this study's focus on optimizing the weather period for rice production forecasting.

Reddy and Puttanna [3] applied stepwise regression to identify significant meteorological factors influencing soybean yield in India. Their research demonstrated that stepwise selection effectively reduced the complexity of prediction models while maintaining high accuracy. This study adopts a similar approach, reaffirming the value of stepwise selection in agricultural modeling.

A study by Ahmed et al. [4] addressed the issue of multicollinearity in regression models for wheat yield prediction in Pakistan. By employing variance inflation factor (VIF)-based elimination techniques, their study improved model reliability and interpretability. This approach mirrors the methods used in the present study to address multicollinearity among weather variables.

Chakraborty et al. [5] compared linear regression with machine learning approaches, including random forests and support vector machines, for predicting rice yields in South Asia. While linear regression provided interpretable results, machine learning models achieved higher prediction accuracy. Their findings suggest that integrating advanced modeling techniques with the selected features and weather periods identified in this study could further enhance rice yield forecasting. These studies collectively underscore the critical role of weather data and feature selection in agricultural yield

Juryon Paik, "Optimizing Weather Variables for Rice Yield Prediction Using Linear Regression," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 7, pp. 43-46, 2024.

prediction. They highlight the importance of refining models to address issues such as multicollinearity and leveraging advanced techniques to achieve higher prediction accuracy, providing valuable insights for future research directions.

III. FEATURE EXTRACTION

This section details the process of collecting and preprocessing the necessary data, followed by the application of the stepwise selection method to identify significant variables. Finally, multicollinearity tests are conducted on the remaining variables to complete the feature selection process.

A. Data Collction and Cleaning

The weather data used in this study were obtained from the Automated Synoptic Observing System (ASOS) provided by the Korea Meteorological Administration (KMA) [6]. While KMA also offers Agricultural Automated Weather Observation System (AAOS) data, which are more directly relevant to agriculture, the AAOS dataset is limited to 11 observation points and thus insufficient for nationwide analysis. Conversely, the ASOS dataset includes information from 103 observation points, covering a wide range of locations across South Korea, including major cities such as Seoul, Busan, Daegu, Incheon, Gwangju, and Daejeon, as well as regions in Gyeonggi-do, Gangwon-do, Chungcheong-do, Jeolla-do, Gyeongsang-do, and Jeju.

For this study, the ASOS dataset was selected due to its broader coverage and greater data availability. Daily weather data spanning ten years, from 2012 to 2021, were collected to ensure a robust dataset. Since rice cultivation is limited to a specific growing season, non-relevant periods were excluded. Based on guidelines from the Rural Development Administration, rice planting begins in May, and harvesting concludes in October. Consequently, weather data from May to October were extracted and used for analysis.

The rice yield data were sourced from the Korea Statistical Information Service (KOSIS) and correspond to milled rice with a milling ratio of 90.4% [7]. The milling ratio refers to the proportion of brown rice converted into white rice after the bran layer is removed. Prior to 2011, the government applied a milling ratio of 92.9%, but to enhance the accuracy of rice statistics, the standard was revised to 90.4% to reflect market trends more accurately.

The dataset spans the same ten-year period (2012–2021) as the weather data, ensuring consistency between the datasets. Rather than using aggregated annual production values, rice yield statistics were calculated regionally and then merged with corresponding regional weather data. This approach accounts for regional weather variations and their impact on rice production, improving the precision of the analysis. Additionally, field area variables associated with each region's rice production were integrated into the dataset to ensure a consistent basis for evaluating production levels.

The compiled weather and rice yield datasets consist of a total of 61 variables. Prior to applying the stepwise selection method, preliminary preprocessing was conducted. Among the compiled variables, the dependent variable(yield) and two independent variables, observation location(place) and observation date(date), were excluded from the analysis. This left a total of 58 variables eligible for stepwise selection. Of these, six precipitation-related variables that lacked daily data - precip_duration(precipitation duration), tm_max_precip(10-minute maximum precipitation), tm_max_precip_time(time of 10-minute maximum precipitation), oh_max_precip(1-hour maximum precipitation), oh_max_precip_time(time of 1-hour maximum precipitation), and day_precip (daily precipitation)—were removed.

After the removal of these variables, 52 variables remained. Subsequently, 12 additional variables with more than 50% missing values were excluded. Finally, the missing values in the remaining 40 variables were imputed with zeros, completing the preprocessing step.

B. Stepwise Selection

The stepwise selection method was conducted at a significance level of 5%, following the steps outlined below:

1. Each variable was individually fitted to a linear regression model.
2. The model with the lowest p-value was selected.
3. If the p-value of the selected model was less than 5%, the corresponding variable was included in the model.
4. If the p-value was greater than or equal to 5%, the variable selection process was terminated.
5. Variables already included in the model were re-evaluated by refitting the linear regression model.
6. The variable with the highest p-value among the selected variables was identified.
7. If this p-value was greater than or equal to 5%, the variable was removed.
8. If the p-value was less than 5%, the variable selection process was completed.

The entire process, encompassing steps 1 through 8, was defined as one iteration (step). When the stepwise selection method was applied to the 40 variables in this study, a total of 30 iterations were performed. Starting with 40 variables, 11 were removed, resulting in a final selection of 29 variables. TABLE I provides a list of the variables selected through the stepwise selection process.

TABLE I. Variables after processing stepwise selection.

Variable	Description
avg_am_cloud	average amount of clouds
area	field area
max_wind_speed	max wind speed
max_ins_wind_speed	maximum instantaneous wind speed
possi_sunshine	the duration of possible sunshine
low_temp	lowest temperature
high_temp	highest temperature
total_du_sunshine	total duration of sunshine
low_temp_time	lowest temperature time
avg_ground_temp	average ground temperature
avg_rel_humid	average relative humidity
low_sea_press	lowest sea level pressure
avg_loc_press	average local pressure
avg_sea_press	average sea level pressure
low_grass_temp	lowest temperature above the grass
high_temp_time	highest temperature time
max_ins_wind_speed_di	maximum instantaneous wind speed

Juryon Paik, "Optimizing Weather Variables for Rice Yield Prediction Using Linear Regression," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 7, pp. 43-46, 2024.

rec	direction
avg_mid_low_am_cloud	average amount of clouds in the middle layer and lower layer
day_precip	day precipitation
min_rel_humid	minimum relative humidity
tm_max_precip_time	10 minute maximum precipitation time
avg_dew_point_temp	average dew point temperature
avg_temp	average temperature
precip_duration	precipitation duration
tm_max_precip	10 minute maximum precipitation
high_sea_press_time	highest sea level pressure time
low_sea_press_time	lowest sea level pressure time
max_ins_wind_speed_time	maximum instantaneous wind speed time
high_sea_press	highest sea level pressure

C. Multicollinearity Elimination

The stepwise selection process resulted in a total of 29 variables being selected; however, this number of variables still posed a risk of multicollinearity. Therefore, it was necessary to address multicollinearity before proceeding with regression analysis. To achieve this, normalization was performed for variables with different units, and the coefficients(coef) of each variable were examined. The coefficient values represent the magnitude of each variable's influence on the dependent variable, where larger absolute values indicate greater impact.

Six variables with relatively low absolute coefficient values (below 3000) were identified: high_sea_press_time (highest sea level pressure time), low_sea_press_time(lowest sea level pressure time), max_ins_wind_speed_time (maximum instantaneous wind speed time), tm_max_precip_time (time of 10-minute maximum precipitation), max_ins_wind_speed_direc (direction of maximum instantaneous wind speed), and low_temp_time (lowest temperature time). To confirm whether these variables had minimal actual impact on the dependent variable (yield), correlation coefficients were analyzed. The results revealed exceptionally low correlations, leading to the conclusion that these variables had negligible influence on the dependent variable. As a result, these six variables were excluded, and the remaining 23 variables were subjected to further multicollinearity analysis and resolution.

In general, multicollinearity is considered present when the variance inflation factor (VIF) exceeds 10. To ensure that multicollinearity among the variables is below this threshold, the following steps were performed iteratively to finalize the variables for the prediction model. The procedure follows the multicollinearity elimination method described in Reference [8]:

1. Calculate the VIF for all variables.
2. Remove the variable with the highest VIF.
3. Refit the linear regression model with the remaining variables and check the statistical significance of all variables.
4. Remove variables that are not statistically significant at a 5% significance level.
5. If all variables have VIF values below 10, stop the process.

This process was repeated for the 23 variables, resulting in a total of 12 iterations. During this procedure, 17 variables with high multicollinearity were removed, leaving six variables in the final model. [Table 2] presents the final selected variables along with their corresponding VIF values.

TABLE II. Variables after eliminating multicollinearity.

Variable	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

IV. Best Weather Time

This section explores the optimal weather period for achieving the highest prediction accuracy in rice yield forecasting using the variables presented in TABLE II. To identify this period, weather data from May to October over a 10-year span (2012–2021) were divided into equal time intervals, and regression models were fitted to past data. The root mean square error (RMSE) was calculated to compare the accuracy of the models, providing insights into which time periods preceding rice yield measurements have the greatest impact.

The data were split into training (80%) and testing (20%) subsets, and multiple linear regression analysis was performed. RMSE, as the evaluation metric, was used to assess model performance, with lower values indicating higher prediction accuracy. By adjusting the retrospective periods, the optimal time frame with the most significant influence on rice yield was identified.

The analysis began with six months of weather data (from May to October) and gradually shifted back in six-month intervals to assess RMSE values. Adjustments were limited to January to ensure the models only incorporated data from the current year. Moreover, data adjustments were handled to ensure weather periods were compared within the same geographic locations. TABLE III presents the RMSE values calculated for weather data in six-month intervals between January and October over the 10-year period. Using the baseline period of May to October, the trained model yielded an RMSE of 36,105.249. Among the five models with adjusted periods, the model utilizing weather data from March to August achieved the lowest RMSE of 35,403.141. This result indicates that shifting the baseline period back by two months, to incorporate data from March to August, provides the most accurate predictions. Consequently, the March-to-August period is identified as the season with the greatest influence on rice production.

TABLE III. RMSE values with every 6 months.

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

Once the optimal period was identified, it became necessary to determine which specific month within that period had the greatest influence on rice yield prediction accuracy. Monthly RMSE values for the rice yield prediction

Juryon Paik, "Optimizing Weather Variables for Rice Yield Prediction Using Linear Regression," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 7, pp. 43-46, 2024.

model were calculated using data from March to August, as shown in TABLE IV. The results indicated that August had the lowest RMSE, suggesting that weather conditions in August have the most significant impact on rice production within a year.

TABLE IV. Monthly RMSE values from March to August.

Month	RMSE
March	36053.771
April	36700.675
May	36145.712
June	36253.564
July	36289.125
August	35525.532

V. CONCLUSIONS

This study identified the minimum set of features for a rice yield prediction model using multiple linear regression analysis and determined the weather period and specific month with the greatest impact on prediction accuracy. Starting with 58 feature variables, 18 variables with a high proportion of missing values were removed. Through stepwise selection, an additional 11 variables were eliminated. Among the remaining 29 variables, six with negligible influence on the dependent variable(yield) were excluded, followed by the removal of 17 variables with high multicollinearity. Ultimately, six key features were selected and incorporated into the model.

Using the selected six features, the study explored the weather period that had the most significant impact on rice yield prediction accuracy. The results revealed that the optimal weather period was from March to August. Furthermore, among the months within this period, August was found to have the lowest RMSE, indicating that it had the most substantial influence on rice production.

This research successfully identified six key weather-related features that positively affect yield prediction and determined the optimal weather period for achieving high predictive accuracy. However, the multiple linear regression approach employed in this study is a relatively basic method with limitations in addressing certain regulatory issues. Future research should focus on integrating the identified key features and optimal weather periods into advanced regression techniques to develop models with even higher prediction accuracy and conduct comparative analyses to identify the most effective methods.

REFERENCES

[1] F. Tao, Z. Zhang, and S. Zhang. “Response of rice yields to climate variability in East Asia,” *Climatic Change*, 112(3-4), pp. 601-614, 2012.
[2] D. B. Lobell, W. Schlenker, and J. Costa-Roberts. “Climate trends and global crop production since 1980,” *Science*, 333(6042), pp. 616-620, 2013.
[3] K. S. Reddy, and K. Puttanna. “Use of stepwise regression analysis to model crop yield prediction,” *Indian Journal of Agricultural Sciences*, 86(9), pp. 1125-1129, 2016.
[4] M. Ahmed, S. Akhtar, and M. Rehman. “Eliminating multicollinearity for reliable yield prediction using regression models,” *Pakistan Journal of Agricultural Sciences*, 56(2), pp. 105-113, 2019.
[5] S. Chakraborty, P. Paul, and N. Adhikary. “Comparison of statistical and machine learning approaches for rice yield prediction in South Asia,” *Journal of Agronomy and Crop Science*, 206(3), pp. 478-490, 2020.
[6] Korea Meteorological Administration. Automated Synoptic Observing System. Available from: https://data.kma.go.kr
[7] Korea Statistical Information Service, Agricultural Production Survey:Rice Production(white rice. 90.4%), 2022. Available from: https://kosis.kr
[8] S. Jung. “On the Multicollinearity in a Linear Regression Analysis,” Gwangju Health University, Thesis Collection, vol.22, pp.293-302, 1997.

Juryon Paik, “Optimizing Weather Variables for Rice Yield Prediction Using Linear Regression,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 7, pp. 43-46, 2024.