Yield Prediction of Paddy Rice with Machine Learning

Yuichiro Maeda[1], Taichi Goyodani[2], Shunsaku Nishiuchi[3] and Eisuke Kita[1,4]

[1]Graduate School of Informatics
[2]Graduate School of Information Science
[3]Graduate School of Bioagricultural Sciences
Nagoya University
Nagoya, Japan
[4]Graduate School of System Informatics
Kobe University
Kobe, Japan

Abstract— *In this paper, we describe the yield prediction of paddy rice so as to be useful for planning the rice cultivation schedule. The use of XGBoost defines the prediction model of the rice yield from the weather information, the cultivation data and the location information of the paddy rice field. The weather information includes the daily maximum temperature, the daily minimum temperature and sunshine time, which are integrated to define the explanatory variables. Cultivation data includes 13 variables such as year, seedling type, transplant method and so on. The best accuracy is estimated as 74.4% when the weather information is integrated at two intervals such as planting date to heading date and heading date to ripening date. The discussion on the variable importance of explanatory variables for the prediction accuracy revealed that the weather information was very effective in yield prediction.*

Keywords: Machine learning, Agri-informatics, Rice cultivation, Artificial intelligence

1. Introduction

Japanese agriculture production tends to decline mainly after the total output reached 11.7 trillion yen in 1984[1], [2]. The decrease and the rapid aging of agricultural workers are also very difficult problems. According to the reference [3], the agricultural working population will decrease to about 2 million in 2020, which is 35% to 1990. In the near future, aging of agricultural workers will progress still more, but new farmers will be fewer. According to the National Fresh Farming Consultation Center nationwide[4], the most important difficulty for new farmers is "low income" of the agriculture business. In fact, only one quarter of new farmers earn income enough to cover livelihoods. In addition, "Immature of technology" is the second most important problem. In addition to these problems, some researchers have pointed out that the yield decreases with the abnormally high temperature due to global warming in recent years[5], [6], [7]. For solving these problems, the introduction of data science and the information science to the agriculture has been studied widely.

The aim of this study is to discuss the yield prediction of the paddy rice in Japan by using the data mining technique. The yield data of paddy rice is taken as the objective variable. The weather information, the cultivation data and the location information are as the explanatory variables. The weather information includes the daily maximum temperature, the daily minimum temperature and sunshine time. Weather information uses the integrated value from rice planting date to heading date, rice heading date to ripening date and so on. Cultivation data includes 14 variables such as year, seedling type, transplant method and so on. The use of XGboost defines the prediction model of the objective variable with respect to the explanatory variables[12], [13], [14].

The remaining part of this paper is organized as follows. In section 2, describes the proposed algorithm and explains verification experiment. In section 3, describes experimental results, In section 4, describes the summary.

2. Prediction Algorithm

2.1 Rice Cultivation Process

The rice stage from germination to harvesting is divided into nursery stage, tillering stage, panicle development stage, and ripening stage.

The nursery season means the stage from germination of rice fields to rice planting. Normally in Japan, seedlings are raised in 20 to 30 days using nursery unit or plastic houses instead of paddy fields.

In the tilling stage, the stems are divided into several parts. Tilling affects yield.

In the panicle development stage, rice makes a leaf which is the last leaf, and makes panicle which is the source of panicle. The formation of panicles begins about one month before heading. This period is called the panicle development period.

After heading, rice will reach its ripening stage.

2.2 Gradient Boosting

XGBoost is a kind of gradient boosting. In gradient tree boosting, updating weights of leaves of tree ensemble models, derivation an optimal model that can minimize the evaluation formula. As an outline, predicted values of the tree ensemble model are calculated as follows.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{n} K = f_k(x_i), f_k \in \mathcal{F} \qquad (1)$$

Note that

$$\mathcal{F} = f(x) = w_{q(\mathbf{x})}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T) \qquad (2)$$

is regression tree space. x_i is input, $\hat{y}_i$ is output, q tree structure, and T is number of leaves in tree. each f_k matches the independent tree structure q and the weight w. w_i represents the score of the ith leaf. This predicted value can be evaluated by

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \qquad (3)$$

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel ^2 \qquad (4)$$

where l is the loss function to find the difference between the predicted value $\hat{y}_i$ and the target value y_i

Ω representing the complexity of the model is a regularization term and has the function of smoothing the weight to avoid over learning.

2.3 Explanatory and Objective Variables

We describe an algorithm for the prediction of yield. The objective variable is the yield of rice. In rice cultivation, weather information is one of factors that affect yield. The references indicate that high temperature leads to yield loss and that the yield becomes smaller when the amount of solar radiation is low under high temperature conditions[8]. The other reference pointed out that paddy rice is affected by daylight hours and temperature[15]. Also, the possibility that crop data such as planting density is involved in yield. From the above, In this experiment, we use two explanatory variables weather information and cultivation data for making prediction model of yield. In addition, location information for each test site was also introduced in this experiment.

The explanatory variables of the prediction model include

A weather information,
B cultivation data, and
C location information.

Weather information includes

1) daily maximum temperature
2) daily minimum temperature and
3) sunshine time.

Table 1: The details of explanatory variables used for prediction model

Test	Integration interval	Number of explanatory variables.
Test 1	Planting date to ripening date.	3
Test 2	Planting date to heading date, and heading date to ripening date.	6
Test 3	Planting date to heading date, and heading date to ripening date, which are split in two interval.	12

They are defined as the integrated values from rice planting date to ripening date.

Cultivation data includes the following 13 variables.

1) Year of rice cultivation
2) Seedling type or sowing style
3) Transplant method or direct seedling field condition
4) Fertilization level
5) Basal fertilizer
6) The number of additional supplementary fertilizer
7) The quantity of additional supplementary fertilizer
8) The height of the rice plant
9) The length of the rice ear
10) Planting density
11) The number of days from sowing date to planting date
12) The number of days from planting date to heading date
13) The number of days from heading date to ripening date

Location information includes the latitude and longitude of each test site.

3. Results and Discussions

3.1 Test Settings

Weather information includes daily maximum temperature, daily minimum temperature and sunshine time. They are integrated in some intervals. In order to discuss the effect of the integral interval for the prediction accuracy, three intervals are compared. They are shown in Table 1.

In Test 1, the weather variables are integrated from planting date to ripening date.

In Test 2, the integral intervals are from planting date to heading date and from heading date to ripening date, which are represented as “max_tmp_1” and “max_tmp_2”, respectively.

In Test 3, the integral intervals of Test 2 are divided into equal two intervals, which are as “max_tmp_1”, “max_tmp_2”, “max_tmp_3” and “max_tmp_4”.

3.2 Accuracy Evaluation

It is required that the prediction accuracy of yield is within 10% error. Therefore, in this experiment, the prediction with

Table 2: XGBoostt's parameter in yield prediction

$Parameter$	Test 1	Test 2	Test 3
max_depth	8	9	9
min_child_weight	6	10	10
$learning\_rate$	0.1	0.1	0.15
$subsample$	1	1	1
$colsample\_bytree$	0.5	0.5	1

Table 3: Effect of integral intervals for accuracy in 2011-2015

category	Test 1	Test 2	Test 3
Average	7.35%	7.25%	7.01%
Median	6.16%	6.41%	5.94%
Max	26.53%	29.83%	27.13%
N	614	614	614
N_over 10%	150	135	144
N_over 20%	17	22	14
Acc	72.8%	74.4%	74.3%

the prediction accuracy within the error of 10% is regarded as the correct answer, and the accuracy is calculated based on how well it was predicted from the whole data. The prediction accuracy is defined by the following equation.

$$Accuracy = \frac{T}{T+F}\ (\%) \tag{5}$$

where the variables T and F represent the number of data that could be predicted with an error of 10% and the number of data predicted with an error of 10% or more, respectively.

3.3 Yield Prediction Result

XGBoost parameters for the model of yield prediction are listed in Table 2 [12], [13].

The prediction results of three tests are compared in Table3. The label "Average", "Median" and "Max" denote the average value, the median error and the maximum error of the absolute error between the actual value and the predicted value from 2011 to 2015, respectively. The label "N" is the number of predicted data. The label "N_over 10%" and "N_over 20%" mean the over 10% error and the over 20% error between the actual value and the predicted value, respectively. Finally, the label "Acc" is the ratio of the number of data that achieved the required accuracy.

Table 4: Prediction accuracy of yield in 2011-2015 (Test 1)

year	2011	2012	2013	2014
Average	7.24%	6.73%	7.42%	7.34%
Median	6.12%	5.97%	6.19%	6.16%
Max	21.67%	17.75%	26.54%	26.51%
N	164	146	135	169
N_over 10%	32	41	31	46
N_over 20%	2	0	7	8
Acc	79.27%	71.92%	71.86%	68.05 %

Table 5: Prediction accuracy of yield in 2011-2015 (Test 2)

year	2011	2012	2013	2014
Average	6.93%	6.86%	7.25%	7.29%
Median	6.25%	6.25%	6.41%	6.46%
Max	24.41%	19.29%	29.83%	23.801%
N	164	146	135	169
N_over 10%	35	31	26	43
N_over 20%	6	0	10	6
Acc	75.0%	78.77%	73.33%	71.01 %

Table 6: Prediction accuracy of yield in 2011-2015 (Test 3)

year	2011	2012	2013	2014
Average	7.01%	7.03%	6.72%	6.96%
Median	5.92%	5.94%	5.54%	5.92%
Max	27.13%	20.26%	27.13%	23.44%
N	164	146	135	169
N_over 10%	30	27	37	50
N_over 20%	2	1	7	4
Acc	80.49%	80.82%	67.4%	68.05 %

The prediction accuracy of the yield should be less than 10% of the relative error. Overall accuracy is the highest for the prediction model of Test 2. Test 1 and Test 3 are better than Test 2 for average, median and max. Although the prediction model of Test 2 is the best in this experiment, we can see that the number predicted with a prediction error of 20% or more is Test 2 the most.

Tables 4, 5 and 6 show the prediction results of Test 1, Test 2, and Test 3, respectively. The prediction accuracy in 2011 is the highest in test 1, and it is less than 70% only in 2014. In Test 2, the prediction accuracy of 2012 was the highest, and the prediction accuracy was over 70% throughout. In Test 3, prediction accuracy in 2011 is the highest, and the prediction accuracy for 2013 and 2014 is less than 70%.

Finally, it is concluded that the best model is Test 2. Focusing on Median, however, we can see that the model of Test 3 is excellent throughout all of 2011 - 2014. And focusing on Max, we can see that the accuracy of the Test 2 model is not as good as the Test 1 and Test 3. Although the overall accuracy is highest for the prediction model of Test 2,

Table 7: The importance variable (Test 1)

Rank	Variable	Importance
1	max_tmp	0.12
2	sunshine_time	0.10
3	min_tmp	0.097
4	The height of rice plant	0.088
5	year	0.087
6	latitude	0.073
7	The number of days from planting date to heading date	0.068
8	The length of the rice ear	0.067
9	longitude	0.051
10	The number of days from heading date to ripening date	0.043

when we look at the prediction error ratio and the maximum error, we found an interesting result that the prediction model of Test 1, Test 2 is superior.

The prediction accuracy in 2013 and 2014 is lower than in 2011 and 2012. The report[16] says that in 2013, it became a low temperature trend nationwide from the middle of April to the beginning of May, but it was covered with warm air at other times, and the temperature fluctuation was great. In 2014, sunshine time between March to May were the longest since 1946 and in late March and late May, warm air flowed from the south, the temperature significantly exceeded the average year. It may think that the change in weather information different from the annual and the division method of weather information in this experiment may not be able to capture the change appropriately. We can think that the facts that prediction accuracy of prediction model of test 2 was the best in this experiment, the low accuracy of 2013 and 2014. Although N_over 10% are 77 of Test 1 and 77 of Test 3, 69 of test 2. As a result, it is thought that the overall prediction accuracy was improved.

From the above analysis, the prediction accuracy was improved by dividing the integration period of temperature and sunshine time and adding it to the explanatory variable. In other words, by dividing the integration period, it is thought that prediction became possible by considering fine displacement of weather information of each data.

As a result, the division method of the integration period of Test 2 left the best accuracy, but in addition to this, it is necessary to try out division methods of various integration patterns and investigate the most effective integration method for yield prediction.

3.4 Variable importance of explanatory variable

Tables 7, 8 and 9 show the importance variable ranking of top 10 on each predicted model.

Table 8: The importance variable (Test 2)

Rank	Variable	Importance
1	sunshine_time_1	0.093
2	min_tmp_1	0.088
3	min_tmp_2	0.084
4	sunshine_time_2	0.082
5	max_tmp_1	0.077
6	max_tmp_2	0.075
7	The height of rice plant	0.073
8	year	0.061
9	The length of the rice ear	0.054
10	latitude	0.053

Table 9: The importance variable (Test 3)

Rank	Variable	Importance
1	The height of the rice plant	0.086
2	sunshine_time_1	0.063
3	sunshine_time_2	0.0604
4	sunshine_time_4	0.058
5	sunshine_time_3	0.057
6	min_tmp_1	0.0504
7	max_tmp_4	0.050
8	min_tmp_4	0.049
9	min_tmp_2	0.048
9	min_tmp_3	0.048

We can understand that the importance of weather information is high in all prediction models. Among them, when we look at Test 2 and Test 3 which divides the integration interval of weather information, sunshine_time_1 indicating the sunshine time from rice planting date to heading day in Test 2 is ranked first, and in test 3, it can be seen that the sunshine time in the first half of the period from the rice planting date to the heading date strongly influences the yield. From this, it can be seen that integration of sunshine time from rice planting day to heading day is important in yield prediction.

In the temperature information, the importance of the explanatory variable of the integration of the maximum temperature was high in Test 1, while the importance of the explanatory variable of the accumulation of the minimum temperature was high in Test 2 and Test 3. Also, in test 3, the point that the highest variable importance is the height of the rice plant is different from other prediction models.

In the cultivation data, variable importance such as year, the height, the length is high. In the location information, latitude is ranked in as the top of the variable importance. It seems that climate and solar radiation amount change greatly due to the difference in latitude.

4. Conclusion

In this research, we presented the yield prediction model of paddy rice. We use XGBoost for defining the model between the yield and the explanatory variables. As explanatory variables, we used three informations such as weather information, cultivation data and location information.

We integrated the weather information in some integral intervals of the interval from rice planting date to ripening date. In addition, we use three kinds of explanatory variables depending on the difference in accumulation interval, and evaluated the influence on yield prediction.

In order to discuss the effect of the integral intervals for the accuracy, three integral intervals were compared. The results showed that the best prediction accuracy was observed at two integral intervals; one is from planting date to heading date and the other is from heading date to ripening date. Its prediction accuracy was 74.4%.

We discussed the variable importance of explanatory variables for the prediction accuracy. The results showed that the weather information was very effective in yield prediction.

Acknowledgment

A part of this work was supported by the NARO Bio-oriented Technology Research Advancement Institution (R&D matching funds on the field for Knowledge Integration and Innovation). Besides, the data on japonica paddy rice are provided from the Aichi agriculture research center, Aichi, Japan.

References

[1] Ministry of Agriculture. “Annual agricultural production output and production agricultural income." (in japanese) 2017/12/24 accesced http://www.e-stat.go.jp/SG1/stat/List.do(2008)

[2] National Agricultural Cooperative Association. “Think about Japanese ingredients 2013. Current situation of agriculture in Japan" in japanese)2017/12/24 accesed https://www.zennoh.or.jp/japan_food/02.html

[3] Matsuhisa, Tsutomu. “Future statistics of the farmers’ population and agricultural labor force in Japan." Agricultural Integrated Research 46.2(1992): 89-112

[4] National Fresh Farming Consultation Center nationwide. “Survey on the storage condition of new farmers in 2016" (in Japanese) 2017/12/24 accesed https://www.nca.or.jp/Be-farmer/statistics/pdf/OChagC5X8b3V3NsIcbsm201704071333.pdf

[5] Wakamatsu, Ken-ichi. “Effects of high air temperature during the ripening period on the grain quality of rice in warm regions of Japan." Bulletin of the Kagoshima Prefectural Institute for Agricultural Development. Agricultural Research (Japan) (2010).

[6] Hasegawa, Toshihiro, et al. “Recent warming trends and rice growth and yield in Japan." MARCO Symposium on Crop Production under Heat Stress: Monitoring, Impact Assessment and Adaptation. National Institute for Agro-Environmental Studies, Tsukuba, Japan. 2009.

[7] Okada, Masashi, et al. “A climatological analysis on the recent declining trend of rice quality in Japan." J. Agric. Meteorol 65.4 (2009): 327-337.

[8] Shimono, Hiroyuki. “Modelling, Information and Environment-Impact of Global Warming on Yield Fluctuation in Rice in the Northern Part of Japan." Japanese Journal of Crop Science 77.4 (2008): 489.

[9] Yokozawa, Masayuki, Iizumi, Toshichika and Okada, Masaki. “Large scale projection of climate change impacts on variability in rice yield in japan" (2009)

[10] Horie, T., et al. “The rice crop simulation model SIMRIW and its testing." Modeling the impact of climate change on rice production in Asia (1995): 51-66.

[11] Wakiyama, Yasuyuki, Kimio Inoue, and Kou Nakazono. “A simple model for yield prediciton of rice based on vegetation index derived from satellite and AMeDAS data during ripening period." Journal of Agricultural Meteorology (Japan) (2003).

[12] Scalable and Flexible Gradient Boosting 2017/12/24 accessed https://xgboost.readthedocs.io/en/latest//

[13] Jain, Aarshay. “Complete guide to parameter tuning in xgboost." (2016).

[14] Chen, Tianqi, and Carlos Guestrin. “Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.

[15] Nagano, Atsushi J., et al. “Deciphering and prediction of transcriptome dynamics under fluctuating field conditions." Cell 151.6 (2012): 1358-1369.

[16] JMA/ Climate system monitoring annual report (in Japanese) 2017/12/24 accesed https://www.data.jma.go.jp/gmd/cpd/diag/nenpo/