Comparing Decision Tree, Random Forest and Boosting in Identifying Weather Index for Rice Yield Prediction

Mohammad Masjkur[1], Ken Seng Tan[2]
{masjkur@apps.ipb.ac.id[1], kstan@uwaterloo.ca[2]}

Department of Statistics, Bogor Agricultural University, Bogor, Indonesia[1],Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada[2]

Abstract. Modeling relationship of weather index and yield losses is a basis for developing weather-based index crop insurance. The data mining approach may overcome some limitations of traditional regression approaches to identify a weather index for predicting crop yield. The purpose of study is to evaluate performance Decision Tree, Random Forest and Boosting in identifying most important weather index for rice crop yield prediction. The study using district level of rice yield data of 8 locations within the annually period of 1991 – 2014 in Java region. The corresponding weather data consist of 48 weather variables including timescale Standardized Precipitation Index (SPI), Growing Degree Days (GDD), and Vapor Pressure Deficit (VPD) for growing season, respectively. Results show that Boosted Regression Tree is the best model compared to Regression Tree and Random Forest for rice yield prediction. The most important weather index is Growing Degree Days on growing season I (GDD I) and Growing Degree Days on growing season III (GDD III).The threshold values of GDD I > 2100^0C and GDD III > 2150^0C would trigger rice yield losses.

Keywords: boosting, decision tree, random forest, rice insurance, weather index.

1 Introduction

In recent years, rice farmers in Indonesia frequently facing crop failure due to extreme climate. The crop failure almost happen annually and sharply increase in extreme climate events. The crop failure due to drought more severe because it happens in wider area and with longer time, while flood events is more local and with shorter time. The crop failure risk would decrease rice production and farmer income and threaten national food security [1].

Crop insurance is an alternative solution as a risk management tool. It is one of finance scheme related to risk transfer in farming system. Crop insurance is an instrument to sustain crop production and protect farmer against yield losses. By crop insurance, production process could be maintained following the best management practices of crop farming [2].

One type of crop insurance is index-based insurance. The insurance will be paid if an events (drought, flood) happen in a region based on weather index recorded. The index is used as a tool for identifying regional yield losses rather than individual losses of farmer. A weather index crop insurance has been developed, because highly correlation between climate event with yield losses [3], [4], [5].

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290475

[6] using quadratic regression model developed a Weather Index Insurance (WII) for rice production based on a rainfall index. [7] using Panel Geographically Weighted Regression model constructed a WII district specific indices for rice production based on monthly Palmer Drought Severity Index (PDSI) and determine the threshold value for the insurance plan. [8] using linear regression proposed weather index insurance based on Standardized Precipitation Index (SPI) for cereal crops in West Africa. On the other hand, [9] reported stepwise regression model of crop yield with climate change variables including Growing Degree Days (GDD) and Vapor Pressure Deficit (VPD). However, many problems arise when using a linear regression model. The assumptions are relationships between response and predictor is linear, error terms is independent and have constant variance, no outliers and the predictors not correlated. The lack of overcoming these problems could reduce significantly the reliability of model [10].

Machine-learning algorithm is a non-parametric approach may overcome some limitations of traditional regression approaches for predicting crop yield. [11] found that Random Forest model outperform the Multiple Linear Regression in predicting crop yield to environmental variables, including climate and soil characteristics. Moreover, [12] observed Boosted Regression Tree (BRT) and Random Forest (RF) yielded similar predicted Soil Organic Carbon spatial patterns.

The purpose of this paper is to evaluate performance of Decision Tree, Random Forest and Boosting in identifying most important Weather Index for rice crops yield prediction.

2 Data

This study using district level of rice yield data obtained from the Centre of Statistical Bureau (BPS) and the Centre of Agricultural Data and Information System (Pusdatin). The related-weather data including daily minimum and maximum temperature and rainfall data obtained from National Weather Service of Indonesia (BMKG) and Agro-climate and Hydrology Research Institute (Balitklimat). The data consist of 8 locations, i.e. Majalengka, Tegal, Cilacap, Semarang, Banyuwangi, Sidoarjo, Gresik, and Sumenep district in Java region within the period of 1991 – 2014.

2.1 Standardized Precipitation Index (SPI)

The drought indices SPI is based on the probability occurrence of precipitation cumulated over any time scale. The long-term precipitation data for any location is fitted to a probability distribution, usually a gamma distribution which is then transformed into a standardized normal distribution [13], [14]. The drought categories of SPI values are summarized in Table 1.

We generated Standardized Precipitation Index (SPI) for time period 1, 3, 4, 6, 12 months and for growing season I (January-April), II (May-August), and III (September-December) by summing the observed precipitation data. For instance,calculation of 4-month timescale for April (named SPI-4-4) based on summation precipitation for January, February, March, and April. Computation of SPI using software of US National Drought Mitigation Center (US-NDMC) available at https://drought.unl.edu/droughtmonitoring/SPI/SPIProgram.aspx[15], [16].

Table 1. The drought categories of SPI values.

SPI value	Drought Category
≥2.0	extremely wet
1.5 to 1.99	very wet
1.0 to 1.49	moderately wet
-.99 to .99	near normal
-1.0 to -1.49	moderately dry
-1.5 to -1.99	severely dry
≤-2.0	extremely dry

2.2Growing Degree Days (ºC)

Growing degree days (GDD) measures accumulated exposure to heat over the growing season and provides a usefull approach for estimating the growth and development of plants during the growing season[9], [17].The Growing Degree Days (GDD) for each growing season measured as follows:

$$GDD = \sum_{i=1}^{N} \left[\frac{T_{max} + T_{min}}{2} - T_b \right]$$

where T_{max} = daily maximum temperature, T_{min} = daily minimum temperature, T_b = Base temperature = 10ºC.

2.3Vapor Pressure Deficit (ºC)

Vapor pressure deficit (VPD) calculate the difference between how much water the air can hold when it is saturated and how much water it currently holds. VPD is exponentially related to temperature. When VPD is very high, water requirements are also high which is critical for photosynthesis [9]. The Vapor Pressure Deficit (VPD) for growing season measured as follows:

$$VPD = \left(e^{\frac{17.269\ T_h}{237.3+T_h}} - e^{\frac{17.269\ T_l}{237.3+T_l}} \right)$$

whereT_h = daily maximum temperature, T_l = daily minimum temperature.

2.4Rice Yield (qha^{-1})

Rice yield changes in time depend on several factors, such as weather factors, new management practices and technologies [15]. To capture the effect of weather factors, a linear time trend was removed from yield data and the proportional yield deviation (YD) calculated as follows:

$$YD = 100 \times (log(Yield) - log(Trend))$$

[9].

3Methodology

We analyze relationship of rice yield and weather variables using regression tree, random forest and boosting models. The predictor variables consisting of 48 weather variables

including SPI for time period 1, 3, 4, 6, 12 months, GDD and VPD for growing season I, II, and III, respectively. We use tree, RandomForest and gbm R package [18].

The algorithm for building a regression tree is as follows [10],

1. A large regression tree was fitted to the training data set.
2. A 10-fold cross-validation was used to select the best number terminal nodes.
3. Applying pruning to the large regression tree.

The random forest algorithm of regression tree is as follows,

1. Bootstrapped samples with size=500 was used for building several decision trees.
2. Choosing random subset 16 predictors for split candidates from all 48 predictors.
3. Predictor variables are evaluated by total decreasing by splitting variable of node impurity, averaged across all trees [10].

Boosting for regression tree working as follows,

1. Fitting a tree $\hat{t}^b(x)$with 4splits to the training data for $b = 1, 2, \ldots, 5000$ and residual

$$r_i = y_i$$

2. Adding a shrinkage parameter λ =0.001 to the new tree to get $\hat{t}(x)$and residual update,

$$\hat{t}(x) \leftarrow \hat{t}(x) + \lambda \hat{t}^b(x) \quad (1)$$

$$r_i \leftarrow r_i - \lambda \hat{t}^b(x) \quad (2)$$

3. The boosted regression tree is

$$\hat{t}(x) = \sum_{b=1}^{B} \lambda \hat{t}^b(x) \quad (3)$$

[10], [19].

We selected randomly 75% of dataset for training ('training dataset') and the other 25% as the 'testing dataset'. The best model performances selected based on Root Mean Square Error (RMSE),

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N}$$

where, y_i the observations of testing dataset and $\hat{y}_i$ the predictions. Smaller RMSE indicate a better model.

4Results and Discussion

4.1 Decision Tree

A large regression tree show that the important variables in the tree are SPI-6-10, GDD-I, GDD-III, SPI-1-4, SPI-6-11, SPI-1-8, SPI-3-4, VPD-I and SPI-1-11 (**Figure 1**). The 10-fold cross-validation present that the CV error takes its minimum value for a four-node tree (**Figure 2**). The pruning tree with four terminal nodes is displayed in **Figure 3**.

Based on the best subtrees, SPI-6-10 is the most important factor determining rice yield, followed by GDD-I and GDD-III (**Figure 3**). For region with SPI-6-10 < -1.58, the predicted mean rice yield is 131.40 qha^{-1}. The region with SPI-6-10 ≥ -1.58 is further subdivided by GDD-I and then by GDD-III. The tree stratification of weather variables is as follows: $R_1 = \{X | SPI6.10 < -1.58\}$, $R_2 = \{X | SPI6.10 \geq -1.58, GDD1 \geq 2145.75\}$, $R_3 = \{X | SPI6.10 \geq -1.58, GDD1 < 2145.75, GDD3 < 2180.95\}$, $R_4 = \{X | SPI6.10 \geq -1.58, GDD1 < 2145.75, GDD3 \geq 2180.95\}$. The related mean rice yield is 131.40, 59.27, 109.50, and 71.16 qha^{-1}, respectively. The RMSE of regression tree is 36.53.

Fig. 1. The large regression tree of weather variables and rice yield.

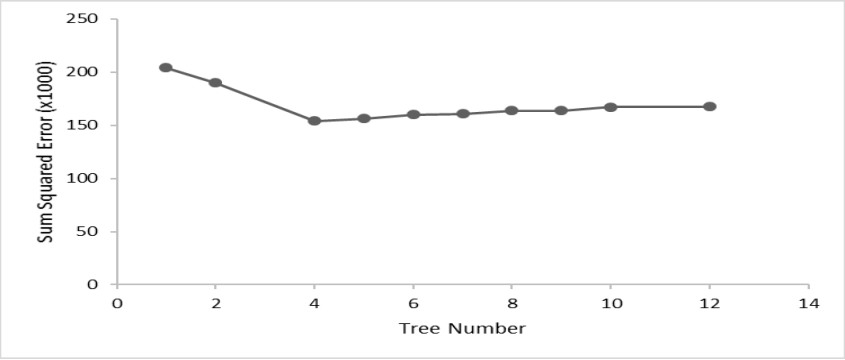


Fig. 2. Cross validation error rate of several tree number.

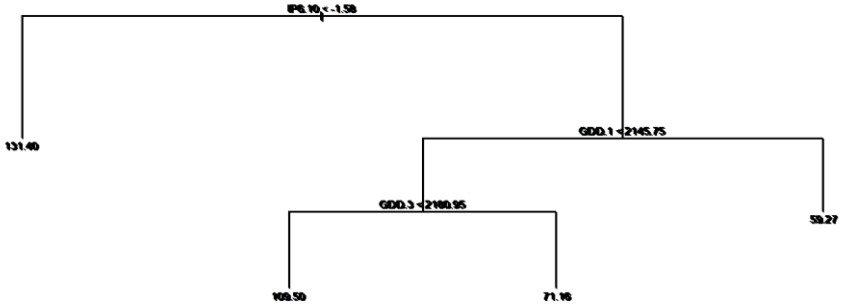


Fig. 3. Best subset regression tree of weather variables and rice yield.

4.2Random Forest

The variable importance plot of random forest indicate that GDD-I is the most important weather variables correlated with the rice yield followed by SPI-6-10 and SPI-1-8 (**Figure 4**). The RMSE of random forest is 27.43.

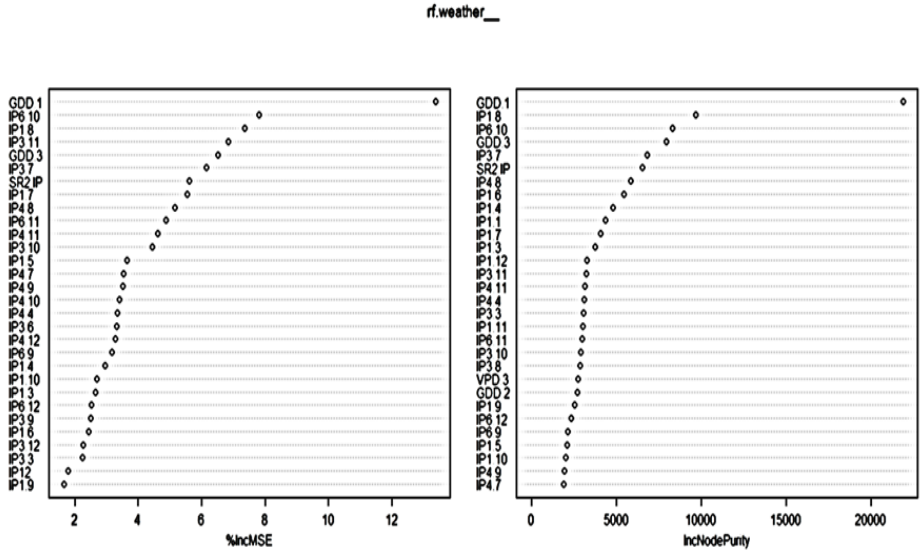


Fig. 4. The important variables plot of random forest model

4.3Boosted Regression Tree

Figure 5 show the relative importance of variables of boosted regression tree. The GDD-I is the most important factor determining rice yield followed by GDD-III and SPI-1-4. The RMSE of boosted regession tree is 24.64. The partial dependence plot of GDD-I and GDD-III show that thresholdvalues of GDD-I > 2100^0C and GDD-III > 2150^0Cwould trigger rice yield losses (**Figure 6**).

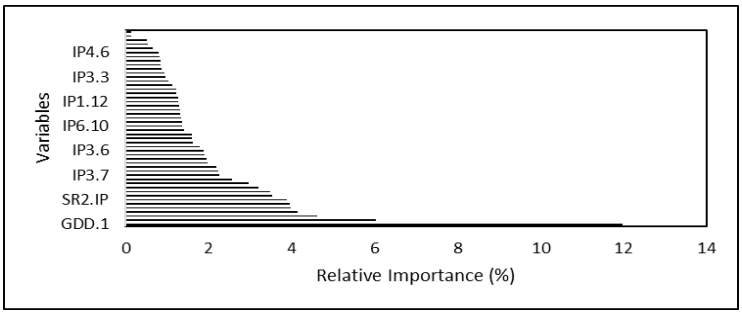
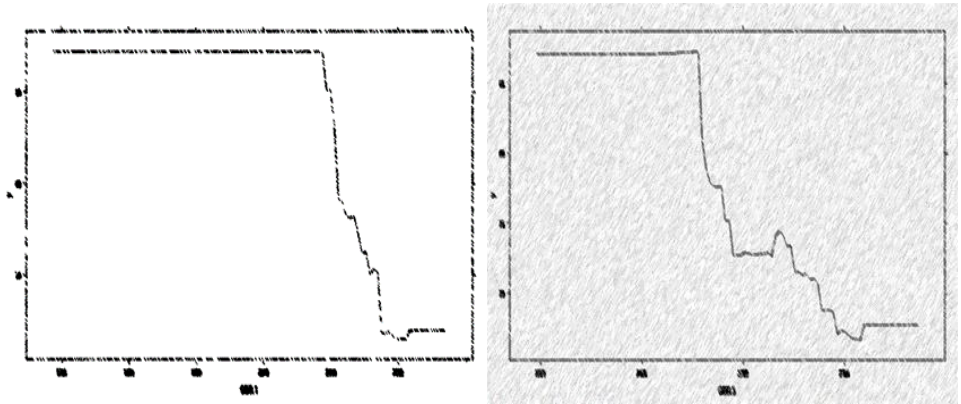


Fig. 5. Relative importance of each variables of boosted regression tree



(a) **(b)**

Fig. 6. The partial dependence plot of **(a)** GDD-I and **(b)** GDD-III and rice yield of Boosted regression tree.

4.4Comparing Decision Tree, Random Forest and Boosted Regression Tree

We find that based on RMSE, Boosted Regression Tree (BRT) is the best model compared to Random Forest (RF) and Decision Tree (DT) in identifying weather index for rice yield prediction. The complete results show that Boosted Regression Tree (BRT) < Random Forest (RF) < Decision Tree (DT) (Tabel 2).

Tabel 2. RMSE of Regression tree, Random Forest and Boosted regression tree

Model	RMSE
Decision Tree	36.53
Random Forest	27.43
Boosted Regression Tree	24.64

5Conclusion

Based on the above findings, it was concluded that Boosted Regression Tree is the best model compared to Regression Tree and Random Forest in identifying weather index for rice yield prediction. The most important weather index is Growing Degree Days on growing season I (GDD-I) and Growing Degree Days on growing season III (GDD-III). The threshold values of GDD-I > 2100^0C and GDD-III > 2150^0C would trigger rice yield losses.

References

[1]Kartiwa B, Surmaini E, Sosiawan H, Rejekiningrum P.: Dampak Perubahan Iklim terhadap Keragaan Sumberdaya Air. Badan Penelitian dan Pengembangan Pertanian, Kementerian Pertanian, Jakarta (2013)

[2]Pasaribu SM.: Penerapan Asuransi Pertanian Di Indonesia. Badan Penelitian dan Pengembangan Pertanian, Kementerian Pertanian, Jakarta(2014)

[3]Choudhury A, Jones J, Okine A, Choudhury R.: Drought-Triggered Index Insurance Using Cluster Analysis of Rainfall Affected by Climate Change. Journal of Insurance Issues, 39(2): 169-186(2016)

[4]Shi H , Jiang Z.: The efficiency of composite weather index insurance in hedging rice yield risk: evidence from China. Agr. Econ. 47:319–328(2016)

[5]Stoppa A, Manuamorn, Pomme O.: Weather index insurance for maize production in Eastern Indonesia : a feasibility study. Washington, D.C. : World Bank Group(2017)

[6]Estiningtyas W, Boer R, Las I, Buono A, Rakhman A.: Deliniasi Risiko Iklim dan Evaluasi Model Hubungan Curah Hujan dan Produksi Padi dalam Mendukung Pengembangan Asuransi Berbasis Indeks Iklim (Climate Index Insurance) Pada Sistem Usahatani Berbasis Padi. Jurnal Ilmu Pertanian Indonesia, 16(3): 198-208(2011)

[7]Kusuma A, Noy I, Jackson B.: A Viable and Cost-Effective Weather Index Insurance for Rice in Indonesia. CESifo Working Paper No. 6530. Munich Society for the Promotion of Economic Research - CESifo GmbH(2017)

[8]Okpara JN, Afiesimama EA, Anuforom AC, Owino A, Ogunjobi KO.: The applicability of Standardized Precipitation Index: drought characterization for early warning system and weather index insurance in West Africa. Nat Hazards 89:555–583 (2017)

[9]Roberts MJ, Schlenker W, Eyer J.: Agronomic Weather Measures in Econometric Models of Crop Yield with Implications for Climate Change. Amer. J. Agr. Econ. 95(2): 236–243; doi: 10.1093/ajae/aas047(2012)

[10] James G, Witten D, Hastie T, Tibshirani R.: An Introduction to Statistical Learning: with Applications in R. Springer, New York (2017)

[11]Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE.: Random Forests for Global and Regional Crop Yield Predictions. PLoS ONE 11(6): e0156571. doi:10.1371/journal.pone.0156571(2016)

[12]Yang RM, Zhang GL, Liu F, Lu YY, Yang F, Yang F, Yang M, Zhao YG, Li DC.: Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators 60: 870–878(2016)

[13]Guttman, N. B.: Accepting the standardized precipitation index: a calculation algorithm. Journal of the American Water Resources Association, 35(2), 311-322. doi:10.1111/j.1752-1688.1999.tb03592.x (1999)

[14]WMO.: Standardized Precipitation Index User Guide. Geneva: World Meteorological Organization(2012)

[15]Kattelus M, Salmivaara A, Mellin I, Varisa O, Kummu M.: An evaluation of the Standardized Precipitation Index for assessing inter-annual rice yield variability in the Ganges–Brahmaputra–Meghna region. Int. J. Climatol. 36: 2210–2222(2016)

[16] Praveen KV, Patel SR, Choudhary JL, Bhelawe S.: Heat Unit Requirement of Different Rice Varieties Under Chhattisgarh Plain Zones of India. J Earth Sci Clim Change 5: 165. doi:10.4172/2157-7617.1000165 (2013)

[17] R Core Team.: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (2018)

[18] Hastie T, Tibshirani R, Friedman J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York(2017)

Acknowledgments. The first author acknowledges the READI Mentorship Program of University of Waterloo, Canada.