Daily Rainfall Prediction using Two-Stage Modeling with Boosting Classification on Statistical Downscaling

Agung Satrio Wicaksono[1], Hari Wijayanto[2], Agus Mohamad Soleh[3]
{agung_satriowicaksono@apps.ipb.ac.id[1], hari@apps.ipb.ac.id[2],agusms@apps.ipb.ac.id[3]}

Department of Statistics, IPB University, Bogor, 42176, Indonesia[1]
Department of Statistics, IPB University, Bogor, 42176, Indonesia[2]
Department of Statistics, IPB University, Bogor, 42176, Indonesia[3]

Abstract. Statistical Downscaling (SD) techniques can be used to predict local rainfall data by using the General Circulation Model (GCM) output data as large-scale global data. Previous research concluded that SD techniques in two-stage modeling with classification using monthly rainfall data can reduce errors in one-stage modeling with Partial Least Square Regression (PLSR). In this study, SD techniques in two-stage modeling with classification are used to predict daily rainfall data. First, the robustness of Boosting method in classification was used to determine the occurrence of rainfall in a day. Second, the PLSR method was used to predict amount of rainfall in rainy days predicted by Boosting method. The capability of the model is tested in four stations all located in West Java Province. Results obtained from 5-fold Cross Validation with 2 repeats clearly show that the RMSEP value will be decrease if the classification accuracy value increase.

Keywords: Boosting, PLSR, RMSEP, statistical downscaling.

1 Introduction

Rainfall has an important role in the hydrological cycle[1]. Research about the rainfall predictions can be used in various fields, including agricultural sectors because it determines the availability of water for plants. Indonesia is located on the equator and has a tropical climate, and agriculture has a major role in the economy of its population. Therefore, rainfall predictions can be used to support economic success in Indonesia.

Statistical Downscaling (SD) is a technique to obtain a model that would be able to analyze the relationship between large-scale global data and small-scale local data [2]. SD techniques can be used to predict local rainfall data by using the General Circulation Model (GCM) output data as large-scale global data [3]. GCM output data consists of several grids with resolutions measured in units of time and in certain domain [4]. In SD modeling, there is a multicollinearity problem because many predictors used are correlated with each other. Several SD techniques have been developed to overcome this problem, including by using Partial Least Square Regression (PLSR) [5].

Two-step modeling of statistical downscaling by integrating the stages of classification and prediction has been developed [6]. In predicting monthly rainfall data, several classification methods such as Classification Tree, Logistic Regression and Ordinal Logistic Regression produce a smaller Root Mean Square of Error Prediction (RMSEP) when

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290524

compared to one-stage modeling with PLSR. It is interesting to examine the relationship between changes in the accuracy value at the classification stage to changes in the RMSEP value at the prediction stage. It is expected that an increase in the classification accuracy value can make the RMSEP value decrease, so the prediction results will be better.

Boosting is one of the ensemble tree methods which is a development of the classification tree method. The Ensemble tree combines the results of estimation values from several trees to produce an estimated value [7]. In Boosting, trees are built iteratively using regression trees and it converts weak learners into a strong learner[8]. The main difference is that it takes the predictive error of the previous tree and use the residual as the dependent variable and then creates the tree and again determine the residual. The final outcome is the weighted value of each tree and classifier, and the weight is dependent on the accuracy.

The objective of this study is applied SD techniques in two-stage modeling with classification to predict daily rainfall data. First, the robustness of boosting method in classification is used to determine the occurrence of rainfall in a day. Second, the PLSR is used to predict rainfall in each class. By increasing the accuracy of the classification, it is expected to have a good effect on the rainfall prediction, so the final results will be better.

2 Materials

Daily rainfall data series for the time period 2011-2018 recorded at four stations in West Java Province, namely Bandung, Bogor, Citeko and Jatiwangi. This data used to show the capability of the two-step modeling for predict the daily rainfall in each station. The covariate for each model were obtain from the National Center of Environmental Prediction (NCEP) reanalysis data set and called Climate Forecast System (CFS) data. CFS data is a model that describes the global interaction between land, sea and air [9]. The parameter used from CFS data is temperature, with 6x6 domain (36 predictors) and the 0.5 degrees distances between grids. Daily rainfall data is taken from the National Climatology and Geophysics Department's website (BMKG) at http://dataonline.bmkg.go.id/. CFS data is taken from Research Data Archive (RDA) Computational & Information System Lab's website at http://rda.ucar.edu.

3Methods

The steps used to develop the two-step modeling are described below in steps:

1. Combining the rainfall data at each station(BMKG data) with 36 temperature variables from CFS data that correspond to their location.
2. Classification modeling with Boosting to predict occurrence of rainfall in a day.

   Adaptive Boosting (AdaBoost) (Freund et al. 1996) is the most popular Boosting algorithm to overcome the problem of binary classification cases by changing weak learners to strong learners in each iteration. In general, the AdaBoost algorithm consists of the following stages:

   a. Initialize $w_b(i)=\frac{1}{n}$, i=1,2,…,*n*, with *n* is the number of observation.

   b. For each iteration *b*=1,2,…,*B*:

i. Define $C_b(x_i)=\{1,2,\ldots,k\}$ with using $w_b(i)$ where k is the number of category

ii. Calculate the error value (e_b):

$$e_b=\sum_{i=1}^{n} w_b(i)\, I\big(C_b(x_i)\neq y_i\big)\#(1)$$

iii. Calculate α_b:

$$\alpha_b=\frac{1}{2}\ln\left(\frac{1-e_b}{e_b}\right)\#(2)$$

iv. Update the weight:

$$w_{b+1}(i)= w_b(i)\exp\left(\alpha_b I\big(C_b(x_i)\neq y_i\big)\right)\#(3)$$

v. Normalize the weight.

c. *Output* is the final splitter:

$$C_f(x_i)= \arg\max_{j\in Y}\sum_{b=1}^{B}\alpha_b I(C_b(x_i)=j).\#(4)$$

3. PLSR model is used to predict amount of rainfall in a day in rainy days predicted by Boosting method.

   PLSR is a technique that reduces the predictor to a smaller component than the uncorrelated component and performs the least squares regression on this component, not on the original data. PLSR is very useful when predictors are highly collinear. PLSR does not assume that the predictors are fixed, unlike multiple regression. This means that predictors can be measured by errors and making PLSR more robust for measuring uncertainty.

4. Calculate the RMSEP value obtained from the difference between the actual value and the predict value obtained.

$$\mathrm{RMSEP}=\sqrt{\frac{\sum_{i=1}^{n}\big(Y_{(-i)}-\hat{Y}_{(-i)}\big)^2}{n}}\#(5)$$

where,
i : amount of testing data
$Y_{(-i)}$: observed value
$\hat{Y}_{(-i)}$: predicted value
n : number of observations
$\bar{Y}$: mean of observed value

The capability of the Boosting method will be compared with two-stage SD modeling through the Classification Tree method and also Logistic Regression. PLSR is also used in both methods to predict the amount of daily rainfall. 5-Fold Cross Validation and 2 repetition were used to test the robustness of the models.

4Results and Discussion

After combining the rainfall data at each station with 36 temperature variables from CFS data, rainfall data are classified into two groups with categorical scales: rain (0) and not rain (1). Before predicting the occurrence of rain, the proportion of the class is shown to see the characteristics of the data.

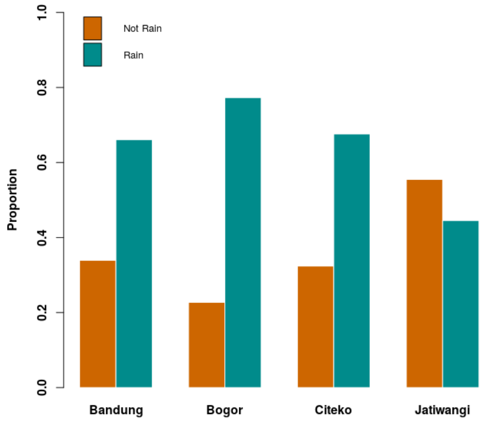

Fig. 1. Rain Proportion in Each Area

Figure 1 shows the proportion for each group in each observation. The proportion between the rain group and the non-rain group in Bandung, Bogor and Citeko is less balanced. It means that the data have a class imbalanced problem. The problem can be solved by using Synthetic Minority Oversampling Technique (SMOTE). The SMOTE method makes some synthetic data for minority groups, so the number of observations in minority groups becomes more numerous, and also more balanced with the majority class [10]. **Figure 2** shows the proportion in each class after handling the class imbalanced problem with the SMOTE method.

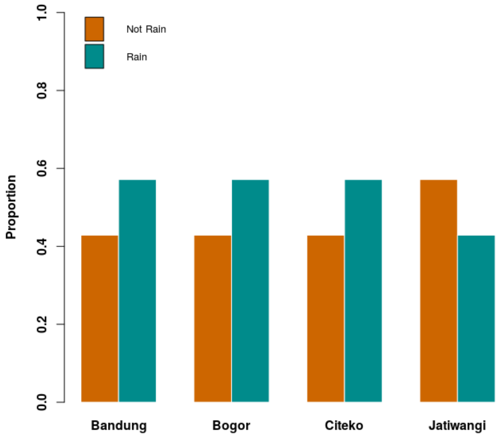

Fig. 2. Rain Proportion After SMOTE in Each Area

After handling the class imbalanced problem, proportion of the minority class becomes more balanced with the majority class. The next step is classification modeling to predict occurrence of rainfall in a day. Boosting results will be compared with Classification Tree and Logistic Regression methods. These results are shown in Table 1.

Table 1. Classification Results

Rainfall Station	Statistics	Classification Method		
		Classification Tree	Logistic Regression	Boosting
Bandung	Accuracy	0.745	0.747	0.771
	Sensitivity	0.848	0.782	0.854
	Specificity	0.545	0.680	0.611
Bogor	Accuracy	0.742	0.672	0.736
	Sensitivity	0.856	0.718	0.838
	Specificity	0.354	0.513	0.390
Citeko	Accuracy	0.728	0.700	0.747
	Sensitivity	0.829	0.735	0.845
	Specificity	0.516	0.627	0.543
Jatiwangi	Accuracy	0.745	0.743	0.764
	Sensitivity	0.750	0.629	0.781
	Specificity	0.741	0.834	0.750

Table 1 shows that the accuracy of Boosting method is superior in almost all observation areas except the Bogor region. However, the sensitivity and specificity of the Boosting method is not always directly proportional to the results of accuracy. Furthermore, with increasing classification accuracy, it is expected to have a good effect on the RMSEP value.

Table 2 shows that after increasing the accuracy at the classification stage with the Boosting method, the RMSEP value decreases. Only in the Bogor area where the RMSEP value is not the best. This happens because the accuracy of Boosting in the Bogor area is not the best too. Smaller value RMSEP indicates that the prediction results from the modeling are getting better. This shows that the two-stage SD modeling with classification will produce a smaller RMSEP value if the classification accuracy obtained is higher.

Table 2.RMSEP Value

Rainfall Station	Classification Method		
	Classification Tree	Logistic Regression	Boosting
Bandung	13.911	13.924	13.721
Bogor	20.524	21.099	20.711
Citeko	17.103	17.487	16.750
Jatiwangi	17.070	17.084	16.810

5Conclusions

Performance of the two-stage SD modeling with Boosting method in classification and PLSR method in predict amount of rainfall in a day has been assessed in this paper. Results obtained from 5-fold Cross Validation with 2 repeats clearly show that the RMSEP value will be decrease if the classification accuracy value increase. In predicting the occurrence of rainfall in Bandung, Citeko and Jatiwangi area, Boosting has a better accuracy than Classification Tree and Logistic Regression.

References

[1] M. Coenders-Gerrits, “The role of interception in the hydrological cycle,” 2010.

[2] E. Zorita and H. Von Storch, “The Analog Method as a Simple Statistical Downscaling Technique: Comparison With More Complicated Methods,” *J. Clim.*, vol. 12, pp. 2474–2489, 1999.

[3] A. V. M. Ines and J. W. Hansen, “Bias correction of daily GCM rainfall for crop simulation studies,” 2006.

[4] J. Noilhan and P. and Lacarrere, “GCM grid-scale evaporation from mesoscale modeling,” *J. Clim.*, vol. 8, no. 2, pp. 206–223, 1995.

[5] A. H. Wigena, “Regresi kuadrat terkecil parsial multi respon untuk statistical downscaling,” *Forum Stat. dan Komputasi*, vol. 16, no. 2, pp. 12–15, 2011.

[6] S. H. Pour, S. Shahid, and E.-S. Chung, “A hybrid model for statistical downscaling of daily rainfall,” *Procedia Eng.*, vol. 154, pp. 1424–1430, 2016.

[7] Y. Freund, R. E. Schapire, and M. Hill, “Experiments with a new boosting algorithm,” *icml*, vol. 96, pp. 148–156, 1996.

[8] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

[9] A. M. Soleh, A. H. Wigena, A. Djuraidah, and A. Saefuddin, “Statistical downscaling to predict monthly rainfall using linear regression with L_1 regularization (LASSO),” *Appl. Math. Sci.*, vol. 9, no. 108, pp. 5361–5369, 2015.

[10] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.