A REGIONAL SCALE MAPPING OF BAMBOO GROVE DISTRIBUTION USING MACHINE LEARNING WITH MULTI-TEMPORAL LANDSAT-8 OLI IMAGERY

Hiroto Shimazaki [1]

[1] National Institute of Technology (NIT), Kisarazu College, Kisarazu City, Chiba Prefecture, 292-0041 Japan

Email: shimazaki@c.kisarazu.ac.jp

KEY WORDS: Random Forest, C5.0, eXtreme Gradient Boosting

ABSTRACT: This study created a map of bamboo grove distribution over a regional scale, aiming to support the forest management planning in Japan. First, the classification performance of Machine Learning methods was evaluated using multi-temporal Landsat-8 OLI imagery acquired in the period from 2013/04/26 to 2019/12/23. Representative six Machine Learning methods such as Artificial Neural Network (ANN), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Random Forest (RF), C5.0 and eXtreme Gradient Boosting (XGB) were employed for predictive modeling of land use / land cover (LULC) classes including a bamboo grove class. 500 points of ground reference data were used to calibrate the model parameters and to validate their classification performance. The classification performance of each model was validated based on the 10 times repeated 10-fold cross validation method with Cohen's kappa. The result showed that the highest performance was achieved by RF (0.818), closely followed by XGB (0.814), C5.0 (0.803), SVM (0.799), ANN (0.756), and finally kNN (0.739). Variable importance metrics suggested that the NIR and SWIR portions observed during spring and autumn seasons were the key to distinguish bamboo groves from other landcover classes. Based on the result, the RF predictive model was used to create a regional scale map of bamboo grove distribution. The map created is expected to be used as a source of information for understanding the status of bamboo grove distribution.

1. INTRODUCTION

Although bamboo is important to the culture and tradition of Japan, rapid increase of unmanaged bamboo grove has brought us with a variety of regional scale problems. Specifically, such problems include increase of breeding and foraging habitat potentially preferred by vermin; increase of risk of sediment-related disasters; decrease in biodiversity due to decrease in forest illuminance caused by dense canopy and simplification of forest structure; inducement of illegal dumping of garbage; and deterioration of traditional rural landscapes. To address these problems relating to the rapid expansion of bamboo grove, a systematic forest management planning is expected to be formulated based on an understanding of the distribution of bamboo groves over a regional scale.

Satellite remote sensing is expected to have potential to identify the extent and distribution of bamboo grove over a regional scale. A previous study suggested that optical remote sensing data of near-infrared (NIR) and short-wavelength infrared (SWIR) could have potential of discriminating bamboo grove from other landcover types (Koizumi et al., 2003). The other study suggested that the Land Use / Land Cover (LULC) classification using multi-temporal optical remote sensing data is of promising approach to achieve a more accurate result and to avoid the adverse effects of cloud contamination (Hashimoto et al., 2014). However, most of the conventional LULC maps derived from satellite remote sensing have lacked the LULC class corresponding to bamboo grove in their classification schemes.

A wide variety of LULC classification methods have been developed by combining a wide range of data and classification algorithms. In recent years, a vast amount of data from Earth observation satellites such as Landsat-8 and Sentinel-2 has been available free of charge (Belward and Skøien, 2014; Harris and Baumann, 2015). In addition, there has been a remarkable development of Machine Learning algorithms that show excellent performance in classifying large multidimensional data sets, and those algorithms are increasingly being applied to the LULC classification task (Lu and Weng, 2007; Otukei and Blaschke, 2010; Maxwell et al., 2018; Abdi, 2019). A previous study on LULC classification that considered bamboo grove as a LULC class has compared classification performance of several Machine Learning methods (Mochizuki and Murakami, 2016). However, the details of data pre-processing and the parameter settings in classification algorithms were not reported well, and more detailed studies are needed to find an optimal LULC classification algorithm for mapping bamboo grove distribution.

This study aims to create a regional scale map of bamboo grove distribution, through evaluation of the performance of Machine Learning methods for mapping bamboo grove using multi-temporal Landsat-8 OLI imagery acquired in the period from 2013/04/26 to 2019/12/23. Classification performance of representative six Machine Learning methods such as Artificial Neural Network (ANN), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Random Forest (RF), C5.0 and eXtreme Gradient Boosting (XGB) will be compared to find optimal predictive model for mapping a bamboo grove distribution. A regional scale map of bamboo grove distribution derived from an optimal predictive model is expected to be used as a source of information for understanding the status of bamboo grove distribution.

2. MATERIAL AND METHODS

2.1 Study Area and Remote Sensing Data

The study area (Figure 1), which was in the middle of the main island of Japan and was covered by the extent of an mosaiced Landsat-8 OLI scene, was selected as a typical region having problems relating to rapid expansion of bamboo grove. Multi-temporal Landsat-8 OLI imagery of the study area was obtained using USGS Earth Explorer for the period from 2013/04/26 to 2019/12/23. Digital Numbers (DN) recorded in two different types of imagery products, Level-1 precision and terrain corrected product (L1TP) and Level-1 systematic terrain corrected product (L1GT), were used as spectral and temporal features for LULC classification.

2.2 Land Use / Land Cover Classes and Ground Reference Data

Initially, the total of 12 LULC classes were set to be classified based on the 11 classes considered in the High-Resolution Land Use and Land Cover Map of Japan (version 16.09), adding a new LULC class corresponding to bamboo grove. However, the area covered by deciduous coniferous forest and snow and ice was very small or could be found only temporarily in the study area. Therefore, these two LULC classes were considered negligible, and the remaining 10 LULC classes were selected as LULC classes to be identified (Figure 2).

Representative 50 location points within the area dominantly covered by each of the 10 LULC classes were selected as ground reference data (Figure 2), and a total of 500 location points of ground reference data were stored in KML format.

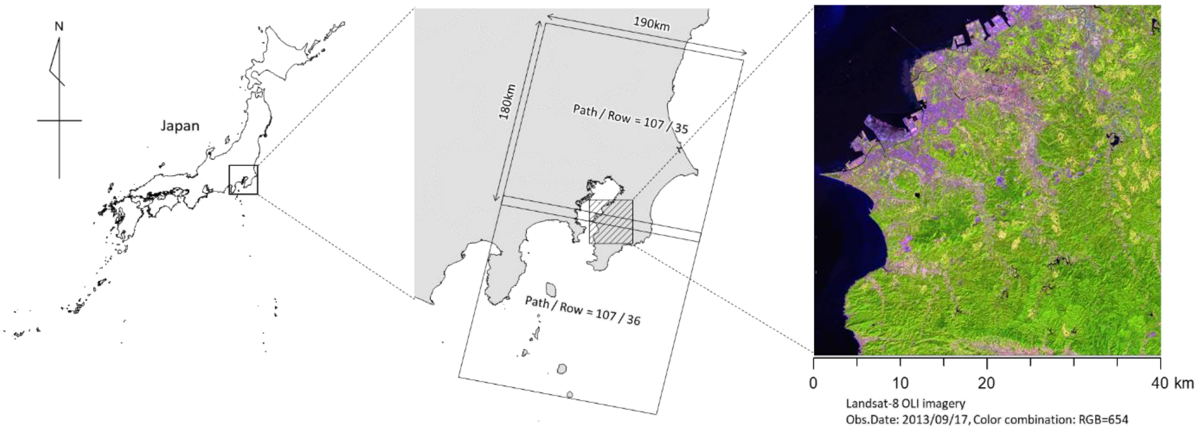
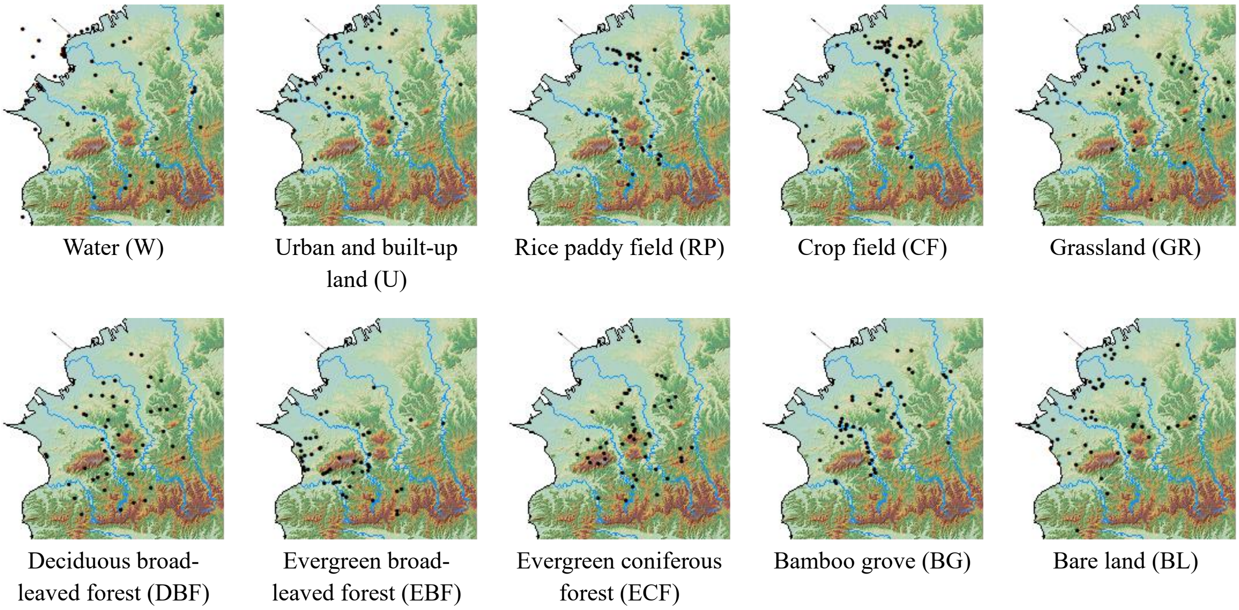


Figure 1. Study area and an example of Landsat-8 OLI imagery



Water (W)

Urban and built-up land (U)

Rice paddy field (RP)

Crop field (CF)

Grassland (GR)

Deciduous broad-leaved forest (DBF)

Evergreen broad-leaved forest (EBF)

Evergreen coniferous forest (ECF)

Bamboo grove (BG)

Bare land (BL)

Figure 2. Ground reference data points (black dots) for the 10 LULC classes over the study area.

2.3 Machine Learning Methods

Representative six Machine Learning methods such as Artificial Neural Network (ANN), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Random Forest (RF), C5.0 and eXtreme Gradient Boosting (XGB) were employed for predictive modeling of the 10 LULC classes. DNs extracted by 500 location points of ground reference data were used as the predictor variables, and the LULC class labels were used as observed outcomes. Calibration and validation of the predictive models were executed using R caret package (Kuhn and Johnson, 2013), which provides us with a set of functions that attempt to streamline the process for creating predictive models. The classification performance of each model was validated based on the 10 times repeated 10-fold cross validation method with Cohen's kappa.

Classification performance of the predictive models could be changed depending not only on the types of predictor variables to be used, but also on the pre-processing of the predictor variables (Kuhn and Johnson, 2013). Although there have been many types of pre-processing methods, this study focused on the following six standard methods:

(1) Impute predictor variables (IMP)

The predictor variables (DNs) affected by cloud contamination were considered as missing values and were supplemented by the median DNs of other locations in the same scene.

(2) Remove non-informative predictor variables (RMV)

Non-informative predictor variables with near-zero variance were removed from a set of predictor variables.

(3) Remove correlated predictor variables (RMV)

Highly correlation predictor variables with correlation coefficient over 0.75 were removed from a set of predictor variables.

(4) Transform predictor variables using principal component analysis (PCA)

Principal component analysis (PCA) was executed to transform the predictor variables to a smaller sub-space where the new variable was uncorrelated with one another.

(5) Transform predictor variables using BoxCox (BoxCox)

BoxCox transformation of the predictor variables was executed to approximate the distribution of predictor variables to a normal distribution.

(6) Transform predictor variables by z-score standardization (ZS)

The predictor variables were standardized to convert the mean of the variables to 0 and the variance to 1.
Among the six standard pre-processing methods above, the first two methods, IMP and RMN, were performed as mandatory measures, and the remaining four methods, RMC, PCA, BoxCox and ZS, were performed as additional

measures in 10 different combinations. In this way, a total of 60 different predictive models (6 Machine Learning methods * 10 pre-processing methods) were evaluated.

3. RESULTS AND DISCUSSION

Comparing the mean and 95% confidence interval of kappa statistic between six Machine Learning methods, the highest performance was achieved by RF (0.818), closely followed by XGB (0.814), C5.0 (0.803), SVM (0.799), ANN (0.756), and finally kNN (0.739) (Figure 3). Based on the result, it was found that the RF predictive model was the most effective at creating a regional scale map of bamboo grove distribution using multi-temporal Landsat-8 OLI imagery. Variable importance metrics of the RF predictive model suggested that the NIR and SWIR portions observed during spring and autumn seasons were the key to distinguish bamboo grove from other LULC classes. These findings were consistent with what Koizumi et al (2003) have pointed out. The map created with the RF predictive model would be expected to be used as a source of information for understanding the status of bamboo grove distribution.

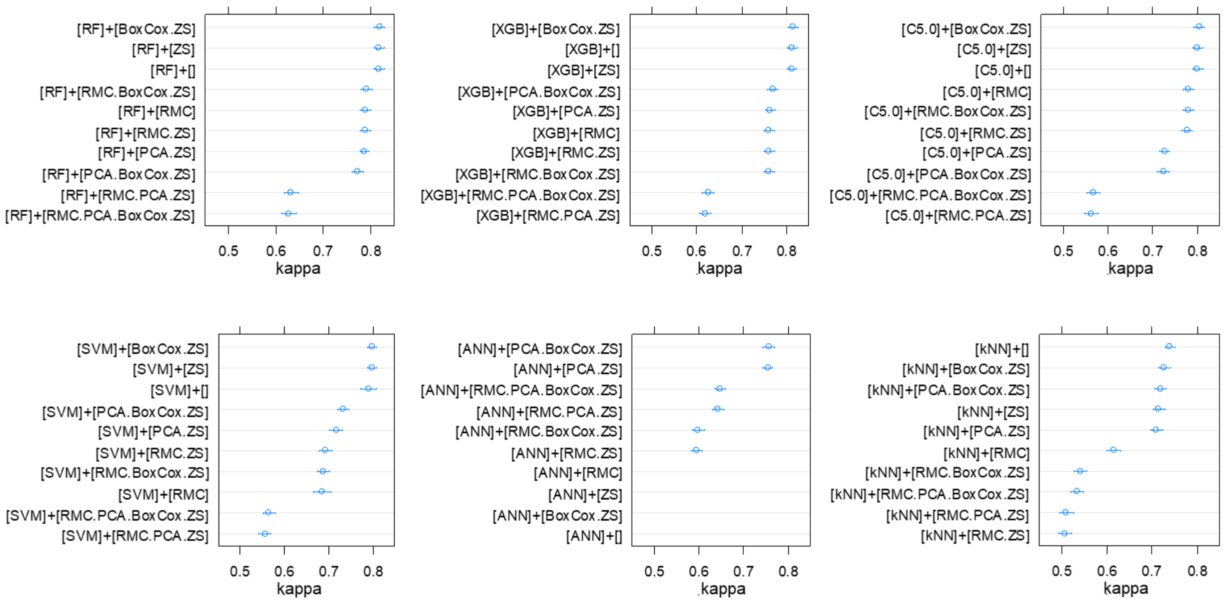


Figure 3. Classification performance of all 60 predictive models. The horizontal axis shows the mean and 95% confidence interval of kappa statistic, while the vertical axis represents model names combining [Machine Learning method] + [pre-processing method].

4. SUMMARY

This study generated 60 predictive models for LULC classification including bamboo grove class using multi-temporal Landsat-8 OLI imagery and six representative Machine Learning methods. When comparing the classification performance among the six Machine Learning methods, the highest performance was achieved by the model generated by Random Forest, of which kappa statistic was significantly higher than 0.8. As for the pre-processing measure for predictor variables, imputation of missing variables and removal of non-informative variables were executed as mandatory measures, and additionally 10 different combinations of four different pre-processing measures were applied. It was found that one of the additional pre-processing measures, removing highly correlated predictor variables could adversely affect the classification performance of all six Machine Learning methods without exception, and effects of the other additional measures differed depending on the type of Machine Learning method. As for the relative importance of predictor variables, it was suggested that the NIR and SWIR portions observed during spring and autumn seasons are important for mapping of bamboo grove distribution. In conclusion, it was found that Random Forest predictive modeling with multi-temporal Landsat-8 OLI imagery being applied proper pre-processing measures would contribute to map the status of bamboo grove distribution. To support a systematic forest management planning, future work includes (1) reconstructing historical distribution of bamboo grove using archived remote sensing data, and (2) spatially explicit estimating of the risk of further expansion of bamboo grove.

ACKNOWLEDGMENTS

The author would like to thank the U.S. Geological Survey for providing Landsat-8 OLI products. Landsat-8 OLI Level-1 precision and terrain corrected product (L1TP) and Level-1 systematic terrain corrected product (L1GT) courtesy of the U.S. Geological Survey.

REFELENCES

Abdi, A. M., 2019. Land cover and land use classification performance of Machine Learning algorithms in a boreal landscape using Sentinel-2 data, GIScience & Remote Sensing, 57(1), pp1-20.

Belward, A. S., Skøien, J. O., 2014. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites, ISPRS Journal of Photogrammetry and Remote Sensing, 103, pp115-128.

Harris, R., Baumann, I., 2015. Open Data Policies and Satellite Earth Observation, Space Policy, 32, pp44-53.

Hashimoto, S., Tadono,T., Onosato, M., Hori, M., Shiomi, K., 2014. A new methods to derive precise Land-use and Land-cover maps using multi-temporal optical data. Journal of The Remote Sensing Society of Japan, 34 (2), pp. 102-112.

Koizumi, K., Tanimoto, C., Piao, C., 2003. Extraction of bamboo stands by observing Landsat 5-TM. Japan Society of Photogrammetry and Remote Sensing, 42(6), pp.42-51.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York, 613pp

Lu, D., Weng Q., 2007. A survey of image classification method and techniques for improving classification performance, International Journal of Remote Sensing, 28(5), pp823-870.

Maxwell, A. E., Warner T. A., Fang F., 2018. Implementation of Machine Learning classification in remote sensing: an applied review, International Journal of Remote Sensing, 39(9), pp2784-2817.

Mochizuki, S., Murakami, T., 2016. Accuracy Comparison of Machine Learning-based Land-cover Classification Using SPOT5/HRG Data, Proceedings of the Institute of Statistical Mathematics, 64(1), pp93-103.

Otukei, J. R., Blaschke T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms, International Journal of Applied Earth Observation and Geoinformation, 12S, ppS27-S31.