Intercontinental Geoinformation Days

igd.mersin.edu.tr

Prediction chlorophyll content of Zizania latifolia using hyperspectral data and machine learning

Adenan Yandra Nofrizal[*1], **Rei Sonobe** [1], **Hiroto Yamashita** [1], **Akio Morita** [1], **Takashi Ikka** [1]

[1]*Shizuoka University, Faculty of Agriculture, Shizuoka, Japan*

Keywords
Dimensionality Reduction
Zizania latifolia
Hyperspectral
Machine Learning

ABSTRACT

Chlorophyll content can be indicative of plant physiological activity and then changes in chlorophyll content have been used as a good indicator of disease as well as nutritional and environmental stresses on plants. Chlorophyll content estimation is one of the most applications of hyperspectral remote sensing data. Also, Random Forest (RF) has been applied to assess biochemical properties from remote sensing data; however, an approach integrating with dimensionality reduction techniques has not been fully evaluated. A total of 200 leaves were measured for reflectance and chlorophyll content and then the regression models were generated based on RF with three dimensionality reduction methods including principal component analysis, kernel principal component analysis and independent component analysis. This research clarified that PCA is the best method for dimensionality reduction for estimating chlorophyll content in *Zizania Latifolia* with a RMSE value of 5.65 ± 0.58 μg cm^{-2}.

1. INTRODUCTION

Chlorophyll pigments absorb sunlight and then their contents relate closely to primary production (Gitelson et al. 2006). Also, chlorophyll offers the information for assessing leaf nitrogen, an essential plant nutrient, due to the close relationship between them (Ramoelo et al. 2015; Kokali and Skidmore. 2015; Bungard et al. 2000) In addition, changes in the chlorophyll content of leaves are related to the effects of disease and nutritional and environmental stresses (Datt. 1999). Therefore, chlorophyll content is one of the most important indicators of photosynthetic activity among all biochemical variables.

To accurately measure chlorophyll content, spectrophotometric measurements using ultraviolet and visible spectroscopy or high-performance liquid chromatography measurements have been adopted widely, however, these techniques are expensive, labor-intensive and require bulky equipment. Although portable equipment such as the SPAD-502 Leaf Chlorophyll Meter (Konica Minolta Inc.) provides a simpler method of quantifying chlorophyll, leaf structure, water content and leaf pigment distribution make their output obscure (Peng et al. 1993). Thus, they are not suitable for quantifying chlorophyll content in Manchurian wild rice (*Zizania latifolia*), since this plant is one of silicicolous plants and leaf structure would be changed by the silica concentration of irrigated water.

On the other hand, remote sensing is one of the most attractive alternative options for this purpose and it has been revealed that hyperspectral data are useful for evaluating chlorophyll contents. Furthermore, it has been applied and evaluated for monitoring of biochemical properties based on hyperspectral indices and radiative transfer models (Lazaro. 2014; Wang. 2018). Accordingly, appraising hyperspectral reflectance to consider analyze chlorophyll contents with a variety a total slag fertilizer is required for restraint quality.

However, a dimension reduction is required to improve the usability of hyperspectral data, due to the high number of spectral bands (and some of them are highly correlated).

Besides dimension reduction, machine learning has been applied to evaluate vegetation properties (Chen. 2017; Doktor. 2014). Especially, Random Forest (RF), which is a regression technique that combines numerous decision trees to classify or predict the value of variable, has been used and reported its high performances for regression (Biau. 2016).

[*] **Corresponding Author**

[*](adenan.y.nofrizal.20@shizuoka.ac.jp) ORCID ID 0000 – 0003 – 3176 – 0312
(sonobe.rei@shizuoka.ac.jp) ORCID ID 0000 – 0002 – 8330 – 3730
(yamashita.hiroto.16@shizuoka.ac.jp) ORCID ID 0000 – 0003 – 0571 – 8153
(morita.akio@shizuoka.ac.jp) ORCID ID 0000 – 0002 – 8418 – 2109
(ikka.takashi@shizuoka.ac.jp) ORCID ID 0000 – 0002 – 1845 – 5658

Cite this study

Nofrizal A Y, Sonobe R, Yamashita H, Morita A & Ikka T (2021). Prediction of chlorophyll content of Zizania latifolia using hyperspectral and machine learning. 2nd Intercontinental Geoinformation Days (IGD), 139-142, Mersin, Turkey

The objective main study is (1) to evaluate the potential hyperspectral data for estimation the chlorophyll of *Zizania Latifolia* and (2) to investigate the best dimensionality reduction method among Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPC) and Independent Component Analysis (ICA).

2. METHOD

2.1. Study area and measurements

Manchurian wild rice (*Zizania latifolia*) plants were cultivated at within-row distances and inter-row spacing of 100 cm on a paddy field at Shizuoka University (Shizuoka, Japan, Figure 1) and grown in flooded conditions. As a basal fertilization, 18 kg of NH_4Cl, 12 kg of P_2O_5 and 12 kg of K_2O were supplied per 1000 m^2. Two further supplementary fertilizations were administered, consisting of 12 kg of NH_4Cl, 12 kg of P_2O_5 and 12 kg of K_2O, and 6 kg of NH_4Cl, respectively (per 1000 m^2). The soluble silicic acid content of the provided molten slag was 32% and the standard amount of slag fertilizer was 120 kg per 1000 m^2. The experiment included a control without slag and four slag fertilizer treatments: a standard amount of slag (1×Slag), and double (2×Slag), 4 times (4×Slag) and 8 times (8×Slag) the standard concentration. A total of 200 leaves (40 leaves from each treatment) were measured for reflectance and chlorophyll content on 2 and 5 October, 2020.

Hyperspectral reflectance was obtained using the FieldSpec4 (Malvern Panalytical, Almelo, Netherlands) and then a splice correction function was applied to minimize the inconsistency caused by the three detectors using ViewSpec Pro (Analytical Spectral Devices Inc., USA)..

Dimethyl-formamide was used the prepare extracts and their chlorophyll contents were quantified using a dual beam scanning ultraviolet-visible spectrometer (UV-1900, Shimadzu, Japan) and Porra's method (Porra. 1989)

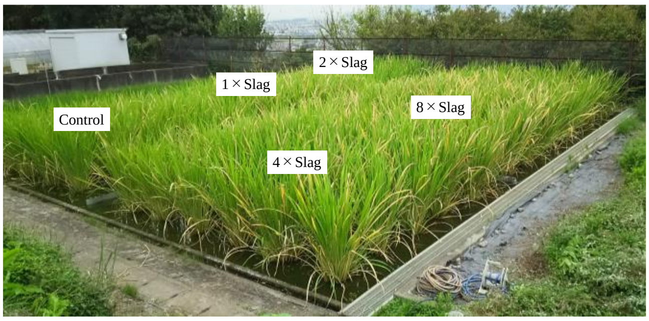

Figure 1. *Zizania latifolia* and location of each treatment

2.2. Data Analysis

Performance evaluation was conducted for RF regression and all processes were implemented using R version 3.5.3 (R Team. 2020). RF regression creates multiple decisions tees called classification and regression trees (CART) based on randomly bootstrapped samples of training data (Breiman 2001) via generalization of the binomial variance (using a Gini index) and by nodes that are using by split variable from a group of randomly selected variable (Liaw. 2002). Since former research has described the effectiveness of RF (Hobbey. 2018; Johannson. 2014), it was also used in this research. RF differs from CART in growing non-deterministically to decorrelate the trees and lessen variance using two-stage randomization scheme related to a bootstrap sample and random variable selection. The number of trees (ntree) and the number of variables used to split the nodes (mtry) are normally established by the user. For tuning these hyperparameters, Bayesian optimization was applied using the Gaussian process.

2.3. Dimension Reduction Techniques

RF-based regression models were generated after dimension reduction techniques including Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA) and Independent Component Analysis (ICA).

2.3.1. Principal Component Analysis (PCA)

PCA is the oldest and best-known technique of multivariate data analysis (Mishra. 2017). It was first coined by (Pearson. 1901), and produced independently by (Hotelling, 1933). PCA is the usual name for a technique which uses sophisticated underlying mathematical principles to transforms several probably correlated variables into smaller number of variables named principle components. The origin PCA lies in multivariate data analysis; however, it has a wide range of other applications. In general terms, PCA uses a vector space transform to reduce the dimensionality of large data sets. Using mathematical forecast, the original dataset, which may involve many variables (i.e the principal component). The central idea of PCA is to reduce the dimensionality of the data set in which there are many interrelated variables.

2.3.2. Kernel Principal Component Analysis (KPCA)

PCA only allows linear dimensionality reduction and then cannot be well represented in a linear subspace if the data has more complicated structures. Kernel PCA is the nonlinear form of PCA, which better exploits the complicated spatial structure of high-dimensional features (Benhart. 1997). The Radial Basis kernel function kernel, which is the typical general-purpose kernel, was applied and the kernel bandwidth was set to 0.1

2.3.3. Independent Component Analysis (ICA)

Independent component analysis (ICA) is closely related to PCA, whereas ICA finds a set of source variable that are mutually independent, PCA finds a set variable that are mutually uncorrelated (Dinesh. 2011). The independent component analysis technique is one of the most well-known algorithms which are used for solving this research. ICA is separating multivariate signal into additive subcomponent. This is done by assuming that the subcomponent is non-Gaussian signal and that they are statistically independent from each order

2.4. Statistical criteria

To evaluate the performance of the regression model, the root-mean-square error (RMSE, equation (1)) was applied.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\widehat{y_i} - y_i)^2}, \quad (1)$$

Where n is number of samples, y_i is measured chlorophyll content and $\widehat{y_i}$ is estimated chlorophyll content.

3. RESULTS AND DISCUSSION

3.1. Chlorophyll content

The measured chlorophyll content per leaf area (cm^2) ranged from 17.53 to 58.02 µg and the maximum value were obtained from the 2 × Slag treatment while the minimum values were from the control. Although there were significant differences in chlorophyll content between 2×Slag and other treatments ($p < 0.05$, Tukey-Kramer test), the other combinations did not differ significantly.

3.2. Spectral Reflectance and Correlation

Spectral reflectance in each fertilizer treatment shown in figure 2. It shown the control samples are highest reflectance values while the lowest values were from the 2 x Slag samples.

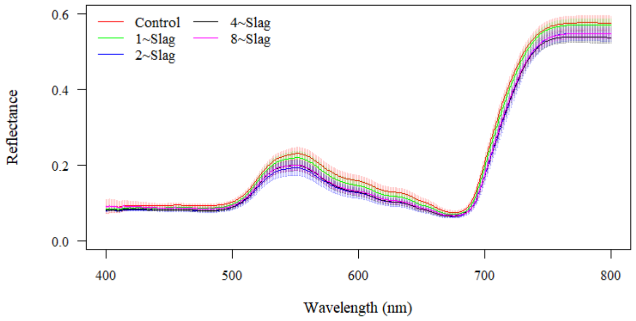

Figure 2. Spectral patterns for each slag fertilizer treatment

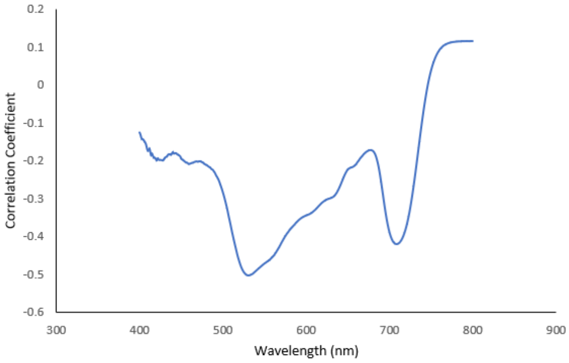

Figure 3. Correlation between spectral reflectance and chlorophyll content

Figure 3 illustrates correlations of each spectral reflectance wavelength with a chlorophyll content. For chlorophyll contents, negative correlations were confirmed near green peak and REIP, and the two bottoms were identified at 531 nm (r = -0.503) and at 709 nm (r = -0.420).

3.3. Accuracy Validation

Tables 1 shows statistics for the RMSE values calculated using regression models. Generally, PCA generally performed the best and PCA was selected as the best solution for estimating chlorophyll content 50 times, while KPCA was selected 12 times. Thus, it is not necessary to use kernel for expressing the relationships between chlorophyll content of *Zizania latifolia* and reflectance data from FielSpec4.

Table 1. Root-mean-square error (RMSE, µg cm^{-2}) for each regression model after 100 repetitions.

	PCA	KPCA	ICA
Minimum	4.49	4.90	4.36
Median	5.65	5.86	5.77
Mean	5.65	5.94	5.76
Maximum	7.33	7.54	7.37
Standard deviation	0.58	0.56	0.59

Figure 4 and 5 show the relationships between measured and estimated chlorophyll contents when the results of 100 repetitions were combined. The coefficient of determinations (R^2) were 0.47, 0.41 and 0.39 for ICA, KPCA, and PCA, respectively and then the advantage of ICA was confirmed. However, the differences were too small to claim that ICA should be applied.

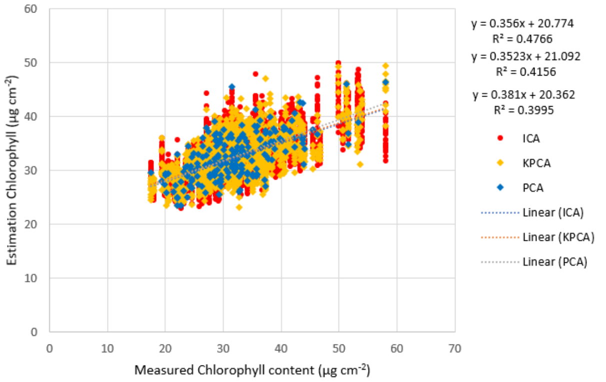

Figure 4. Relationship between estimation and measured chlorophyll contents.

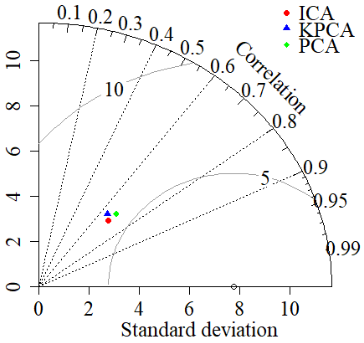

Figure 5. Taylor diagram showing the performance of each dimensionality reduction methods

4. CONCLUSION

This study has evaluated three dimension reduction techniques for estimating chlorophyll contents from reflectance. According of result, spectral reflectance had shown control sample are highest reflectance values more than another slag fertilizer treatment. However, PCA is the best method for dimensionality reduction for the estimation chlorophyll compared another advanced method such as ICA and KPCA.

REFERENCES

Benhart S, (1997). Kernel principal component analysis. Artifical Neural Networks-ICAN'97, 583-588

Biau G (2016). A Random Forest Guided Tour. Test, 25(5), https://doi.org/10.1007/s11749-016-0481-7x

Breiman L (2001). Random Forest. Mach.Math, 45, 4-45, https://doi.org/10.1023/A:1010933404324.

Bungard R A, Press M C & Scholes J D (2000). The influence of nitrogen on rain forest dipterocarp seedlings exposed to a large increase in irradiance. Plant Cell and Environment, 23(1), 1183-1194, https://doi.org/10.1046/j.1365-3040.2000.00642.x.

Chen X J, et all (2017). Spectral Characteristics and Species Identification of Rhododendros Using a Discrimnative Restricted Boltzmann Machine. Spectroscopy Letters, 50(2), 65-72, https://doi.org/10.1080/00387010.2017.1278709

Datt B (1999). Visible/near infrared reflectance and chlorophyll content in Eucalyptus leaves. International Journal of Remote Sensing, 20(14), 2741-2759, https://doi.org/10.1080/014311699211778

Dinesh K K & Naik N R (2011). An Overview of Independent Component Analysis and Its Applications. Informatica, 53, 63-81

Doktor D, et all (2014). Extraction of Plant Physiological Status from Hyperspectral Signatures Using Machine Learning Methods. Remote Sensing 6, 12(2), 12247-12274, https://doi.org/10.3390/rs61212247

Gitelson A, et all (2006). Relationship between gross primary production and chlorophyll content in crops: Implications for the synoptic monitoring of vegetation productivity. Journal of Geophysical Research-Atmospheres, 111(D8), https://doi.org/10.1029/2005jd006017

Hobbey E, et all (2018). Hotspot of soil organic carbon storage revealed by laboratory hyperspectral imaging. Science Report, 8(13), https://doi.org/10.1038/s41598-018-31776-w

Hotelling H (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417-441, https://doi.org/10.1037/h0071325

Johansnon U, et all (2018). Regression conformal prediction with random forest. Machine Learning, 97(1-2), 1-22, https://doi.org/10.1007/s1099-014-5453-0

Kokali R F & Skidmore A K (2015). Plant phenolics and absorption features in vegetation reflectance spectra near 1.66 mu. International Journal of Applied Earth Observation and Geoinformation, 43, 55-83, https://doi.org/10.1016/j.jag.2015.01.010

Lazaro M, et all (2014). Retrieval of Biophysical Parameter with Heteroscedatic Gaussien Processes. IEEE Geosciences and Remote Sensing Letters, 11(4), 838-842, https://doi.org/10.1109/LGRS.2013.2279695.

Liaw A & Wiener M (2002). Classification and regression by random forest. R news, 2, 18-22

Mishra S P, et all (2017). Multivariate statistical data analysis-Principal Component Analysis (PCA). International Journal of Livestock Research, 7(5), 61-78, https://doi.org/10.5455/ijlr.20170415115235

Pearson K, (1901). On lines and planes of closet fit to systems of point in space. Philosophycal Magazine, 6(2), https://doi.org/10.80/14786440109462720

Peng S B, Garcia F V, Laza R C & Cassman K G (1993). Adjustment for Specific Leaf Weight Improves Chlorophyll Meter's Estimate of Rice Leaf Nitrogen Concentration. Agronomy Journal, 85(5), 987-990, https://doi.org/10.2134/agronj1993.00021962008500050005x

Porra R J, Thompson W A & Kriedeman P E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. Biochimica et Biophysica Acta (BBA)-Bioenergitics, 975(3), 384-385, https://doi.org/10.1016/S0005-2728(89)80347-0

Ramoelo A, et all (2015). Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and World View-2 data. International Journal of Applied Earth Observation and Geoinformation, 43, 43-45, https://doi.org/10.1016/j.jag.2014.12.010

R Core Team (2020). R: A Language and environment for Statistic Computing. Vienna: R Foundation for Statistical Computing, Accesed 19 December 2020, http:/www.R-project.org./

Wang Z, et all (2018). Spatio Temporal Variations of Forest Phenology in the Qinling Mountains and Its Response to a Critical Temperature of 10 Degree C. Journal of Applied Remote Sensing, 12(4), https://doi.org/10.1117/1.JRS.12.022202