Hyperspectral reflectance sensing for quantifying leaf chlorophyll content in wasabi leaves using spectral pre-processing techniques and machine learning algorithms

Hyperspectral reflectance sensing for quantifying leaf chlorophyll content in wasabi leaves using spectral pre–processing techniques and machine learning algorithms

Rei Sonobe

Faculty of Agriculture, Shizuoka University, Shizuoka, Japan

Hiroto Yamashita

Faculty of Agriculture, Shizuoka University, Shizuoka, Japan

United Graduate School of Agricultural Science, Gifu University, Gifu, Japan

Harumi Mihara

Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan

Akio Morita and Takashi Ikka

Faculty of Agriculture, Shizuoka University, Shizuoka, Japan

Corresponding author

Rei sonobe (E–mail: sonobe.rei@shizuoka.ac.jp)

Hyperspectral reflectance sensing for quantifying leaf chlorophyll content in wasabi leaves using spectral pre–processing techniques and machine learning algorithms

Changes in chlorophyll content can be a good indicator of disease as well as nutritional and environmental stresses on plants. Several pre–processing techniques have been proposed for reducing noise from spectral data to identify vegetation properties such as chlorophyll content. Machine learning algorithms have also been applied to assess biochemical properties; however, an approach integrating pre–processing techniques and machine learning algorithms has not been fully evaluated. Therefore, this study evaluates the effectiveness of five pre–processing techniques used in conjunction with five machine learning algorithms for estimating chlorophyll content in two wasabi cultivars. Overall, incorporating pre–processing techniques was effective for obtaining estimated values with high accuracy. Analyses utilizing both pre–processing and machine learning performed best in 88 of 100 repetitions. The kernel–based extreme learning machine (KELM) and Cubist algorithms yielded the highest performance and achieved the highest accuracies in 54 and 26 of 100 repetitions, respectively.

Keywords: *Eutrema japonicum*; hydroponics; Mazuma; Onimidori

1. Introduction

Japanese horseradish (*Eutrema japonicum*), also called 'wasabi', belongs to the Brassicaceae family. Wasabi has been cultivated in Japan for more than a thousand years, and nearly half of the total wasabi rhizome consumed in Japan is produced in the Shizuoka Prefecture. There has been a recent increase in global demand for Japanese cuisine (Hege et al. 2019), leading to an increased demand for wasabi production. Wasabi requires specific growing conditions including north–facing gorges and an abundance of cold and clean flowing water. It takes at least 10 months to cultivate wasabi, and recent environmental climate changes have adversely affected wasabi production. In addition, optimal culture methods are poorly understood and its production depends on the experience of skilled farmers. Enhancing detection of nutritional and environmental stresses as well as diseases that result in lower yields may improve wasabi cultivation and facilitate its production by incipient farmers.

Chlorophyll content can be indicative of plant physiological activity since it is an effective indicator of photosynthesis. Chlorophyll absorbs sunlight and uses the energy to synthesize carbohydrates from CO_2 and H_2O (Gitelson et al. 2006). Chlorophyll within the leaf also has a close relationship with nitrogen, an essential plant nutrient (Bojović and Aca 2009). Thus, changes in chlorophyll content are useful for detecting disease as well as nutritional and environmental stresses on plants (Datt 1999; Sims and Gamon 2002; Sonobe et al. 2018a; Sonobe et al. 2020a).

Chlorophyll content can be precisely quantified using spectroscopic techniques such as ultraviolet and visible–light (UV–Vis) spectroscopy and high–performance liquid chromatography (HPLC). However, these techniques require bulky equipment, which limits their usefulness in the field (i.e. outside the laboratory; Kalaji et al. 2017). More portable equipment has been developed, such as the SPAD–502 Leaf Chlorophyll Meter (Konica Minolta Inc.), which determines the relative amount of chlorophyll by measuring the absorbance of the leaf in two wavelengths. These portable instruments, which can be used by non–experts, provide less expensive and less labour–intensive measurements than UV–Vis or HPLC and have been widely used in previous studies (Jacquemoud and Ustin 2019). Nevertheless, Peng et al. (1993) pointed out that variation in leaf thickness causes a variable relationship between SPAD readings and leaf dry weight, which can be different in cultivars, developmental stages and environmental conditions. On the contrary, leaf chlorophyll content and leaf dry weight are related to leaf reflectance over different spectral regions and some studies proposed methods for simultaneous determination of them (Féret et al. 2008; Féret et al. 2011).

Therefore, this issue may be addressed using hyperspectral remote sensing, an alternative tool for measuring chlorophyll content in the field (Amirruddin et al. 2020; Féret et al. 2008; Golhani et al. 2019; Vahtmae et al. 2018).

To quantify vegetation properties, such as chlorophyll content, from remotely sensed data, regression techniques based on machine learning algorithms are becoming an attractive approach. Random forest (RF) and support vector machine (SVM) algorithms have been successfully applied for both classification and regression (Biau and Scornet 2016; Burges 1998). Powell et al. (2010) compared three statistical techniques including Reduced Major Axis regression, Gradient Nearest Neighbour imputation and RF, for their ability to predict biomass dynamics from Landsat data and showed that the two former techniques generally outperformed RF. Lu et al. (2019) found a high RF performance when comparing RF and Partial least square (PLS) regression for estimating chlorophyll from multispectral and hyperspectral data. On the other hand, Siegmann and Jarmer (2015) compared SVM, RF and partial least squares regression for assessing leaf area index (LAI) from wheat reflectance spectra acquired in 2011 and 2012. They showed that SVM provided the best results in the case of cross validation for the separate years although SVM also showed a clear decline in model performance for independent validation of the data set from both years. Further, a regression model based on SVM has been used for quantifying urban land cover (Okujeni et al. 2017), predicting leaf area index on a tropical grassland (Kiala et al. 2016) and estimating leaf area index and green leaf chlorophyll density of rice using hyperspectral data (Yang et al. 2011). Recently, regression models based on the kernel–based extreme learning machine (KELM) have performed well and KELM has been used for estimating leaf chlorophyll content from tea leaves (Sonobe et al. 2020b; Sonobe et al. 2018b). Cubist and Stochastic Gradient Boosting (SGB) algorithms have also been extended for estimating biomass (Breunig et al. 2020). Houborg and McCabe (2018) conducted LAI estimation via machine–learning and showed that the Cubist algorithm was generally superior to RF in predicting LAI. Wijesingha et al. (2020) showed a Cubist regression model proved best for estimating acid detergent fibre in forage while SVM estimated crude protein with the highest precision and accuracy. Despite the efficacy of these algorithms, insufficient training data can lead to overfitting, which can limit the usefulness of these methods for evaluating biochemical properties based on hyperspectral reflectance. Thus, this study examines the five aforementioned algorithms (RF, SVM, KELM, Cubist and SGM) to determine an optimal approach for analysing reflectance data obtained from wasabi.

Besides regression algorithms, previous studies have applied pre–processing techniques to original data to better determine vegetation properties such as chlorophyll content. In general, pre–processing of original reflectance data is conducted to obtain uncontaminated data for further processing. Common spectral pre–processing techniques include scatter correction methods and spectral derivatives (Xu and Gowen 2020). First derivative reflectance (FDR) analysis, in particular, is an effective technique for removing background effects and enhancing subtle spectral features (Meng et al. 2020) as well as enhancing weak spectral features which are effective for evaluating target parameters (Inoue et al. 2012). Notably, FDR has been used to detect specific points such as the green peak and the red edge inflection point (Cho and Skidmore 2006). As a result, various vegetation indices based on FDR have been proposed to estimate vegetation properties such as chlorophyll content (Penuelas et al. 1994; Zarco–Tejada et al. 2003). Similarly, the continuum–removal (CR) transformation, which is a brightness normalization technique that fits a convex hull over the original reflectance data, has been applied to enhance spectral features and eliminate or reduce effects unrelated to the target's properties of interest as well as improving signal to noise ratio and minimising data redundancy (Miller et al. 2020; Sanches et al. 2014). Mutanga et al. (2005) applied CR to evaluate total digestible nutrients (nitrogen, potassium, phosphorous, calcium and magnesium) of pastures using band depth ratios and stepwise regression and Li et al. (2020) used CR with multiple linear regression, principal component regression, PLS or SVM for leaf nitrogen content estimation. Standard normal variate (SNV) and multiplicative scatter correction (MSC) have also been used to reduce noise in raw reflectance data caused by light scattering and baseline drift (Liang et al. 2020). Feng and Makino (2020) conducted colour analysis in sausages after normalization, SNV, MSC, FDR and second derivative using PLS and then found that MSC made the prediction coefficient of determination up to 0.78. Furthermore, Barnes et al. (1989) reported that SNV and then de–trending (DT) was effective at reducing the effect of additive interference of scattered light from particles. However, Aheto et al. (2020) showed pre–processing with SNV and MSC yielded lower values for predicting values of thiobarbituric acid reactive substances.

Thus, this study evaluates the performance and applicability of five machine learning algorithms in conjunction with five pre–processing techniques for analysing the original reflectance data obtained from wasabi. Data treatment and algorithms were applied to the original data gathered from two wasabi cultivars grown in conditions of varying pH and sulphur ion concentration. Thus, we aim to identify an optimized method for detecting chlorophyll content in response to environmental and nutritional stresses and determine which data treatment strategies may be applicable to other spectral–based detection methods.

2. Materials and methods

2.1.Measurements and datasets

The two popular wasabi cultivars, ‘Onimidori’ and ‘Mazuma’, were cultivated by hydroponics in a greenhouse at Shizuoka University in Shizuoka, Japan between 19 December 2019 and 10 March 2020. A total of 79 leaves (39 and 40 leaves from ‘Onimidori’ and ‘Mazuma’, respectively) were harvested from plant tops among expanding leaves. The wasabi clonal plants were transplanted and single individuals were placed in Wagner pots (0.02 m^2) containing 3 L of tap water adjusted to pH 6.0 using HCl and NaOH, and were continuously aerated. After 1 week, slightly modified 0.1×Hoagland solutions (Hoagland and Arnon, 1950), used as standard nutrient solutions, were supplied stepwise for 1 week each at 1:100 and 1:10 concentrations to allow the plants to adjust to the hydroponic system. After that, the following experiments were performed. Ten samples (5 from each cultivar) were cultivated under standard conditions for wasabi using a nutrient mixture of 0.1×Hoagland solution (pH 6). Different pH values (pH 5, 7, 8 and 9) were applied to samples of 5 leaves for each pH condition and cultivar to assess the influence of pH on mineral absorption changes. In addition to the standard condition of sulphur ion concentration (1×S), which is 0.58 mM SO_4^{2-}, conditions of zero (0×S), half (0.5×S) and 1.5 times (1.5×S) the standard sulphur concentration were applied to 5 leaves for each condition and cultivar, except for 0×S (4 and 5 leaves from ‘Onimidori’ and ‘Mazuma’, respectively).

Reflectance data at 1 nm steps across the entire wavelength domain from 400 to 2500 nm were obtained using the FieldSpec4 (Analytical Spectral Devices Inc., USA) from a leaf clipping. Some spectral drift was observed at two wavelengths (1000 and 1800 nm) due to inherent variations caused by the three detectors including visible and near–infrared (VNIR) portions of the electromagnetic spectrum, short wave infrared

(SWIR_1 and SWIR_2). The splice correction function within the ViewSpec Pro Software (Analytical Spectral Devices Inc., USA) was applied to minimize this inconsistency (Prasad et al. 2015). It is well known that leaf chlorophyll content mainly affects the reflectance in the 400 to 780 nm region (Féret et al. 2017) and this region was used to avoid redundant analyses.

To precisely quantify chlorophyll content (Chl_{a+b}), leaves were freeze–dried, ground and analysed using dual–beam scanning ultra violet–visible spectrophotometers (UV–1900, Shimadzu, Japan) and Porra's method (Porra et al. 1989). Dimethylformamide was used to prepare extracts from which chlorophyll–*a* (Chl–*a*) and *b* (Chl–*b*) contents (in µg ml^{-1}) were calculated according to the following Equations (1 to 3) with the chlorophyll unit converted to µg cm^{-2} using the area of leaf discs:

$$\mathrm{Chl}\text{–}a\ (\mu\mathrm{g\ ml}^{-1}) = 12.00 \times (A_{663.8} - A_{750}) - 3.11 \times (A_{646.8} - A_{750}) \qquad (1)$$

$$\mathrm{Chl}\text{–}b\ (\mu\mathrm{g\ ml}^{-1}) = 20.78 \times (A_{646.8} - A_{750}) - 4.88 \times (A_{663.8} - A_{750}) \qquad (2)$$

$$\mathrm{Chl}_{a+b} = \mathrm{Chl}\text{–}a + \mathrm{Chl}\text{–}b \qquad (3)$$

where *A* is the absorbance, and the subscripts are the wavelengths (in nm).

The measurements were divided into three groups (a training dataset (50%), a validation dataset (25%) and a test data dataset (25%)) using a stratified sampling approach (Hastie et al. 2009). To ensure robust results, this approach was repeated 100 times before pre–processing the original reflectance and generating the regression models based on machine learning algorithms.

2.2. Pre–processing of the raw reflectance data

To generate regression models with high accuracy, the following five pre–processing methods were evaluated based on their success in previous studies. Table 1 includes abbreviations of the pre–processing techniques and machine learning algorithms used in this study.

Please put Table 1 about here

1) First derivative reflectance (FDR) processing was selected to remove background effects from plain spectral features and to reduce systematic errors in spectral data (Tsai and Philpot 1998). As a result, some FDR–based indices (Datt 1999; Zarco–Tejada et al. 2003) and the sum of derivative values (Elvidge and Chen 1995; Filella et al. 1995) have been proposed.

2) The continuum–removed (CR) spectra of plants has been utilized to identify changes in the shapes of absorption features in relation to an abundance of biochemical properties such as chlorophyll content (Sanches et al. 2014).

3) De–trending (DT) is a row–wise transformation that allows correction for wavelength–dependent scattering effects by fitting a second–degree polynomial through each spectrum. It is also effective for accounting for the variation in baseline shift and curvilinearity (Barnes et al. 1989).

4) Multiplicative scatter correction (MSC) has been used to compensate for additive or multiplicative effects or a combination of the two in spectral data (Maleki et al. 2007). MSC requires a reference spectrum and an average spectrum whose wavelength–dependent perturbations are separated from the residual spectral data. This correction is composed of two steps: i) regression of each spectrum against the mean spectrum based on ordinary least squares and ii) the calculation of the corrected spectrum.

5) Standard normal variate (SNV) has been effective for baseline correction and reducing scattering effects by subtracting the averaged value of each spectrum and dividing it by the experimental standard deviation (Genkawa et al. 2015).

To identify which wavelengths were significantly influenced by pH conditions or sulphur ion concentration (S strength) ($p < 0.05$), we employed a stepwise linear discriminant analysis (Burns and Burns 2008; Draper 1998). In this technique, we adopted a combination of forward and backward stepwise regression in a multiple regression model, which was used to select suitable predictor wavelengths among the different treatments. The results were evaluated regarding their overall accuracies (OAs), which is the total classification accuracy. All methods were implemented using R version 3.5.0 (R Core Team 2020).

2.3. Regression model

The regression models based on machine learning algorithms included random forest (RF), support vector machine (SVM), kernel–based extreme learning machine (KELM), Cubist and Stochastic Gradient Boosting (SGB). These were evaluated for estimating chlorophyll content from hyperspectral reflectance data. To remove non–informative variables and generate better and simpler prediction models, we applied variable selection techniques. Previous studies showed that a method based on the genetic algorithm (GA), which is an adaptive heuristic search algorithm centred on the evolutionary ideas of natural selection and genetics, was superior to three variable selection techniques including the Martens uncertainty test and interval partial least square regression (Villar et al. 2017). Thus, we adopted a GA in this study. For tuning the hyperparameters of the machine learning algorithms, Bayesian optimisation has been applied using the Gaussian process (Snoek et al. 2015; Yan 2016).

RF is an ensemble nonparametric technique wherein the samples used for model generation are randomly selected from training data by the bootstrap method, and a decision tree is generalized based on the binomial variance using a Gini index (Breiman 2001). As user–defined hyperparameters, the number of trees and the number of variables used to split the nodes are well known. Additionally, the following three hyperparameters are also optimized: the minimum number of unique cases in a terminal node, the maximum depth to which a tree should be grown, and the number of random splittings. This was done using the “randomForestSRC” package (Ishwaran 2007).

SVM is often applied with the Gaussian radial basis function (RBF) kernel and its efficacy has been demonstrated for resolving problems in high dimensions and with local minima (Ding et al. 2016). In particular, it remains highly functional even with a limited volume of training data (Breunig et al. 2020). The regularisation parameter *C* and the kernel bandwidth σ are considered as user–defined hyperparameters using the “e1071” package (Meyer et al. 2017). The high performance of RF and SVM has been reported in many studies and could thus be considered benchmarks in this study (Hobley et al. 2018; Wang et al. 2013).

The extreme learning machine (ELM), which is based on a single hidden layer feedforward neural network, has been widely applied for prediction, faulty diagnosis, recognition, classification and signal processing (Li et al. 2016). The RBF kernel has been used instead of attempting to fit a non–linear model (Huang et al. 2012), particularly in some earlier studies, which showed that the RBF kernel may be advantageous (Sonobe et al. 2018b). Although some improvements have been documented, KELM already possesses significant robustness due to its use of few hyperparameters (i.e. the regulation coefficient (RC) and the kernel parameter (KP)) and few optimisation constraints, which has been shown to be an advantage in regression applications (Maliha et al. 2018).

Cubist is a rule–based model tree approach where leaves are represented by a multi–variate linear regression model wherein the rules are calibrated based on one or more variables or thresholds of the unique subset of explanatory variables. Although predictions are conducted based on the linear regression model at the terminal node of the tree, the prediction from the linear model in the previous node of the tree is also considered. The number of committee models, i.e. boosting iterations, and the number of neighbours used for correcting the model predictions were thus optimised in this study using the 'Cubist' package (Kuhn et al. 2020).

Regression models based on SGB are sequentially generated from the gradient of the loss function of the previous tree, building a new tree from a random sub–sample of the dataset for each iteration. In other words, a tree is built from a random sub–sample from the training data to improve the model. In this algorithm, a fraction of the training data is used, avoiding over–fitting of the training data as well as improving the computation speed and prediction accuracy (Friedman 2002). SGB was applied using the 'gbm' package (Greenwell et al. 2019) with the following four hyperparameters being optimized: the number of iterations and the number of basis functions in the additive expansion, the maximum depth of each tree, the learning rate and the minimum number of observations in the terminal nodes of the trees.

2.4. Statistical criteria

Evaluations of the estimation accuracy of each method were based on the root mean square error (RMSE) and the ratio of performance to deviation (RPD, Equation (4)) (Williams 1987).

$$\mathrm{RPD} = \frac{(\mathrm{SD})}{(\mathrm{RMSE})} \quad (4)$$

where SD is the standard deviation of the real chlorophyll content as calculated from the test data measurements. Based on the RPD values, the methods are categorised into three groups: category A (RPD > 2.0), category B (1.4 ≤ RPD ≤ 2.0) and category C (RPD < 1.4). For evaluating the efficiencies of the number of variables, a data envelopment analysis (DEA), which is a nonparametric method for the estimation of production frontiers, was conducted using the 'Benchmarking' package (Bogetoft and Otto 2020) and by substituting the number of variables and RPD values for the input and output components. Black box data–based sensitivity analysis (DSA) was also used for the fitted models to clarify which narrow–bands were effective in the supervised learning methods (Cortez and Embrechts 2013).

3. Results

3.1. Chlorophyll content after each treatment

Figure 1 shows box plots of the chlorophyll content at different pH conditions for the two cultivars. Chlorophyll content per leaf area (cm^2) ranged from 2.91 to 35.40 and a significant negative correlation was observed between pH value and chlorophyll content (r = –0.592, p < 0.001 for chlorophyll). The degree of concentration of hydrogen ions (pH) had a stronger effect on Mazuma chlorophyll content (r = –0.740, p < 0.001 for chlorophyll) than on Onimidori (r = –0.416, p < 0.05). As such, significant differences in Mazuma chlorophyll content were observed between pH values of 5 and 8, 5 and 9, 6 and 8 and 6 and 9. By contrast, there were no significant differences in Onimidori chlorophyll content (p < 0.05, Tukey–Kramer test). No significant difference in chlorophyll content was observed between the two cultivars under the same pH conditions (p > 0.05, Tukey–Kramer test).

Please put Figure 1 about here

The relationships between S strength and chlorophyll content for the two cultivars are shown in Figure 2. Without sulphur, the chlorophyll content for Onimidori and Mazuma were 16.69 ± 2.66 and 16.14 ± 1.96 μg cm^{-2} (mean ± SD), respectively. However, 0.5 × S increased the chlorophyll content to 30.88 ± 3.65 and 31.16 ± 3.65 μg cm^{-2} for Onimidori and Mazuma, respectively. Compared to the samples without sulphur, chlorophyll content increased significantly as sulphur concentrations were increased (0 × S; $p < 0.05$, Tukey–Kramer test). However, there were no significant differences between the samples treated with different sulphur concentrations within both cultivars ($p > 0.05$, Tukey–Kramer test). There were also no significant differences between the two cultivars treated with the same S strength ($p > 0.05$, Tukey–Kramer test).

Please put Figure 2 about here

3.2. Reflectance under different conditions

The mean reflectance spectra for each pH and S strength are shown in Figures 3 and 4. The reflectance values of leaves near the green peak (520 to 540 nm) increased with pH for both cultivars until a pH of 8 and decreased at pH 9. Regarding the S strength, reflectance values near the green peak became smaller with increasing S strength except for 0.5×S. Stepwise discriminant analysis ($p < 0.05$) showed that reflectance values at 410, 466, 468 and 683 nm were useful for identifying samples of 5 different pH conditions (OA = 0.660), while reflectance values at 423, 427, 455, 612, 615 and 694 nm were useful for S strengths (OA = 0.897) regardless of the cultivar analysed. When the samples were grouped by cultivar, reflectance values at 402, 414, 466, 473, 480, 759 and 780 nm (OA =1.00) for Onimidori and at 716 nm (OA = 0.460) for Mazuma were selected to discriminate pH conditions. Reflectance values at 502, 658 and 691 nm (OA = 0.842) for Onimidori and at 698 nm (OA = 0.600) for Mazuma were selected to identify S strengths.

Please put Figure 3 about here

Please put Figure 4 about here

3.3. Accuracy validation

Tables 2 and 3 show statistics for the RPD and RMSE values, respectively, calculated using regression models based on machine learning algorithms. The mean RPD values of RF were greater than 1.4 for each regression model, which means the models belong to category B, achieving an RMSE of 3.503 ± 0.525, 2.835 ± 0.560, 2.932 ± 0.513, 2.715 ± 0.488, 2.824 ± 0.607 and 2.619 ± 0.493 μg cm^{-2} for the original reflectance of FDR, CR, DT, MSC and SNV, respectively. As such, RF was the only algorithm capable of estimating chlorophyll content from original reflectance spectra as well as spectra pre–processed by the five techniques. In particular, the combination of RF and DT was consistently greater than 2.063, placing the regression models in category A. However, KELM may be the most promising algorithms when FDRs have been applied for pre–processing since the minimum RPD value was 2.292. For SVM and SGB, their mean RPD values were consistently greater than 1.4; however, they sometimes resulted in poor estimation results.

Please put Table 2 about here

Please put Table 3 about here

3.4. Efficiency and sensitivity analysis

Figure 5 shows the results of DEA and the variables of MSC possessed the lowest efficiencies among the pre–processing techniques. The efficiencies of SVM were not suitable and some outliers were confirmed for all the pre–processing techniques, while those of RF were the most suitable among all the machine learning algorithms.

The importance at 20 nm interval as assessed by DSA is shown in Figure 6. Generally, the importance at wavelengths shorter than 420 nm was smaller, although a peak of importance was confirmed over 420 to 440 nm for SGB when DT or MSC were applied. The highest importance values were observed over the green peak (500 to 560 nm) or REIP (680 to 720 nm) for all algorithms. At the green peak, the importance was larger than that at the REIP for the original reflectance. However, the importance at the REIP was highest when FDR was applied, although this tendency was obscured for SVM and KELM. As with the original reflectance, similar trends in importance were observed for CR, MSC and SNV; however, the peak at 720 to 740 nm was larger for CR than for MSC and SNV.

Please put Figure 5 about here

Please put Figure 6 about here

4. Discussion

4.1. Optimal machine learning algorithms

After 100 repetitions, the best and worst algorithms of each round were determined based on the RPD value (Table 4). Although the RF algorithm consistently gave acceptable results based on RPD values (Table 2) and previous studies have reported on its satisfactory performance for estimating vegetation properties (Liu et al. 2017; Powell et al. 2010), it was determined to be the best algorithm only 14, 7, 8, 19 and 17 times for FDR, CR, DT, MSC and SNV. The KELM and Cubist algorithms consistently performed best owing to their high predication abilities for the test data, which was independent data for generating and optimizing the regression models. The sizes of the training dataset, which was used to generate the regression models, and validation dataset, which was used to optimize the hyperparameters of the machine learning algorithms, were 32 and 16 samples, respectively. KELM performed best for all techniques except for the original reflectance data; however, it was also the worst performer once each for original reflectance, CR and DT, five times for MSC and twice for SNV (however, the RPD values were still more than 1.5 except for MSC). By contrast, SGB and SVM generally performed the worst and SGB was never selected as a best algorithm. Results revealed that SGB showed a low sensitivity to outliers, which is effective for unbalanced training datasets and provides robustness when dealing with interactions (Friedman 2002). Therefore, SGB avoids problems arising from a wrong learning rate, which is an important hyperparameter and determines the size of steps by sampling a fraction of the training data (Friedman 2002). However, this process decreases the sample size for generating regression models, and so, especially when the size of the training data is small, SGB can be inadequate. SVM and KELM are both kernel–based algorithms, although SVM was clearly inferior to KELM. For kernel–

based algorithms, incorrect selection of hyperparameters related to kernel function causes a decrease in the estimation accuracy (Horvath 2003). In this study, the variance of KP appeared smaller than *σ*, which implies that the Bayesian optimization sometimes resulted in local solutions for tuning the hyperparameters of SVM. Thus, there is some potential for improving the SVM estimation accuracies by using other optimization methods which can provide alternative local solutions for tuning hyperparameter combinations.

Please put Table 4 about here

4.2. Best combinations of pre–processing techniques and machine learning algorithms

The optimal combinations of pre–processing techniques and machine learning algorithms are shown in Table 5. After application of the pre–processing techniques, the absolute values of correlation coefficient over the green peak (*r* = –0.923, –0.789, –0.954, –0.961, –0.932 and –0.935 at 550 nm for original reflectance, FDR, CR, DT, MSC and SNV) or REIP (*r* = –0.783, –0.952, –0.957, –0.698, –0.938 and –0.936 at 725 nm for original reflectance, FDR, CR, DT, MSC and SNV) increased. At these wavelengths, the relationships between chlorophyll content and reflectance could be measured. As a result, out of 100 repetitions, using a combination of pre–processing techniques and machine learning algorithms resulted in the best results 88 times whereas applying machine learning algorithms to reflectance data alone (without pre–processing) only resulted in optimal results 12 times (Table 5). The combination of DT and KELM was selected as the best solution for estimating chlorophyll content 15 times and DT was the most selected pre–processing technique: 31 times in total (15 times with KELM, 9 times with Cubist, 4 times with RF and 3 times with SVM). Barnes et al (1989) pointed out that DT was effective for removing the effects of baseline shift and curvilinearity from the original reflectance of powdered or densely packed samples. The potential of DT for estimating chlorophyll content was also confirmed in this study. SNV, which also adjusts for baseline shifts between samples, was effective for removing multiplicative interferences of scatter and particle size (Ng et al. 2020) and SNVs were selected 24 times. Compared to these techniques, the advantages of MSC, which also adjusts for baseline shifts between samples by minimizing additive and multiplicative effects in reflectance (Rinnan et al. 2009), were not corroborated. Some studies have reported successfully using MSC for hyperspectral data over the visible near–infrared and short–wave infrared range with high performance (Liang et al. 2020); however, this technique tends to hide some weaker spectral features, which leads to unsatisfactory results. In our study, leaf–scale measurements were conducted using leaf clippings and additive and multiplicative effects on reflectance might be small. FDR was effective for enhancing resolution and correcting baseline shifts in hyperspectral data (Bruning et al. 2020) and it was selected 13 times due to these advantages. Next, CR was selected as the best 9 times (7 times for KELM and 8 times for Cubist).

Please put Table 5 about here

5. Conclusions

Pre–processing of reflectance is a potentially useful tool for improving estimation accuracies of chlorophyll content. However, the influences of the selected pre–processing techniques and machine learning algorithms have been obscure. The superior usefulness of de–trending (DT) was confirmed and the combination of DT and kernel–

based extreme learning machine (KELM) was most effective for estimating chlorophyll content.

Taken together, these results show that hyperspectral reflectance has great potential as a tool for detecting chlorophyll content within wasabi. Furthermore, treating these data with pre–processing techniques and machine learning algorithms can improve qualifying plant appearance, effects of environmental stresses and effects of ingredients before plant cultivation. These techniques could thus be used to improve the usability of portable devices and subsequent agricultural management, thereby facilitating quality control and plant maintenance for less experienced farmers.

Funding

This study was supported financially by a Kurita Water and Environment Foundation (KWEF, Japan) research grant award (ID: 19B001), Yamazaki Spice Promotion Foundation and JSPS KAKENHI [grant number 19K06313].

Disclosure statement

There are no conflicts of interest.

References

Aheto, J.H., Huang, X.Y., Tian, X.Y., Lv, R.Q., Dai, C.X., Bonah, E. & Chang, X.H. (2020). Evaluation of lipid oxidation and volatile compounds of traditional dry-cured pork belly: The hyperspectral imaging and multi-gas-sensory approaches. Journal of Food Process Engineering, 43, 10

Amirruddin, A.D., Muharam, F.M., Ismail, M.H., Ismail, M.F., Tan, N.P., & Karam, D.S. (2020). Hyperspectral remote sensing for assessment of chlorophyll sufficiency levels in mature oil palm (Elaeis guineensis) based on frond numbers: Analysis of decision tree and random forest. Computers and Electronics in Agriculture, 169, 11

Barnes, R.J., Dhanoa, M.S., & Lister, S.J. (1989). Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Applied Spectroscopy, 43, 772-777

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227

Bojović, B., & Aca, M. (2009). Correlation between nitrogen and chlorophyll content in wheat (Triticum aestivum L.). Kragujevac Journal of Science, 31, 69-74

Bogetoft, P., & Otto, L. (2020). Package 'Benchmarking' Accessed 17 August 2020. https://cran.r-project.org/web/packages/Benchmarking/Benchmarking.pdf

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32

Breunig, F.M., Galvao, L.S., Dalagnol, R., Dauve, C.E., Parraga, A., Santi, A.L., Della Flora, D.P., & Chen, S.S. (2020). Delineation of management zones in agricultural fields using cover crop biomass estimates from PlanetScope data. International Journal of Applied Earth Observation and Geoinformation, 85

Bruning, B., Berger, B., Lewis, M., Liu, H., & Garnett, T. (2020). Approaches, applications, and future directions for hyperspectral vegetation studies: An emphasis on yield ‐ limiting factors in wheat. The Plant Phenome Journal, 3

Burges, C.J.C. (1998). A tutorial on Support Vector Machines for pattern recognition. Data Mining and Knowledge Discovery, 2, 121-167

Burns, R.P., & Burns, R. (2008). Business research methods and statistics using SPSS. SAGE Publications

Cho, M.A. & Skidmore, A.K. (2006). A new technique for extracting the red edge position from hyperspectral data: The linear extrapolation method. Remote Sensing of Environment, 101, 181-193

Cortez, P., & Embrechts, M.J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17

Datt, B. (1999). Visible/near infrared reflectance and chlorophyll content in Eucalyptus leaves. International Journal of Remote Sensing, 20, 2741-2759

Ding, S.F., Shi, Z.Z., Tao, D.C., & An, B. (2016). Recent advances in Support Vector Machines. Neurocomputing, 211, 1-3

Draper, N., H (1998). Applied regression analysis (Wiley Series in Probability and Statistics). Wiley-Interscience

Elvidge, C.D., & Chen, Z.K. (1995). Comparison of Broad-Band and Narrow-Band Red and Near-Infrared Vegetation Indices. Remote Sensing of Environment, 54, 38-48

Feng, C.H. & Makino, Y. (2020). Colour analysis in sausages stuffed in modified casings with different storage days using hyperspectral imaging - A feasibility study. Food Control, 111, 11

Féret, J.-B., Francois, C., Asner, G.P., Gitelson, A.A., Martin, R.E., Bidel, L.P.R., Ustin, S.L., le Maire, G., & Jacquemoud, S. (2008). PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. Remote Sensing of Environment, 112, 3030-3043

Féret, J.-B., Francois, C., Gitelson, A., Asner, G.P., Barry, K.M., Panigada, C., Richardson, A.D. & Jacquemoud, S. (2011). Optimizing spectral indices and chemometric analysis of leaf chemical properties using radiative transfer modeling. Remote Sensing of Environment, 115, 2742-2750

Féret, J.-B., Gitelson, A.A., Noble, S.D. & Jacquemoud, S. (2017). PROSPECT-D: Towards modeling leaf optical properties through a complete lifecycle. Remote Sensing of Environment, 193, 204-215

Filella, I., Serrano, L., Serra, J., & Penuelas, J. (1995). Evaluating wheat nitrogen status with canopy reflectance indices and discriminant analysis. Crop Science, 35, 1400-1405

Friedman, J.H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38, 367-378

Genkawa, T., Shinzawa, H., Kato, H., Ishikawa, D., Murayama, K., Komiyama, M., & Ozaki, Y. (2015). Baseline Correction of Diffuse Reflection Near-Infrared Spectra Using Searching Region Standard Normal Variate (SRSNV). Applied Spectroscopy, 69, 1432-1441

Gitelson, A.A., Keydan, G.P., & Merzlyak, M.N. (2006). Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. Geophysical Research Letters, 33

Golhani, K., Balasundram, S.K., Vadamalai, G., & Pradhan, B. (2019). Estimating chlorophyll content at leaf scale in viroid-inoculated oil palm seedlings (Elaeis guineensis Jacq.) using reflectance spectra (400 nm-1050 nm). International Journal of Remote Sensing, 40, 7647-7662

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. (2019). Package ‘gbm’. Accessed 3 June 2020. https://cran.r-project.org/web/packages/gbm/gbm.pdf /

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. The United States: Springer-Verlag New York

Hege, N., Kobayashi, M., Michiki, N., Takano, T., Baba, F., Kobayashi, K., Ohyanagi, H., Ohgane, J., Yano, K., & Yamane, K. (2019). Complete chloroplast genome sequence and phylogenetic analysis of wasabi (Eutrema japonicum) and its relatives. Scientific Reports, 9

Hoagland, D. R. , & Arnon, D. I. (1950). The water-culture method for growing plants without soil. Circular. California Agricultural Experiment Station, 347

Hobley, E., Steffens, M., Bauke, S.L., & Kogel-Knabner, I. (2018). Hotspots of soil organic carbon storage revealed by laboratory hyperspectral imaging. Scientific Reports, 8, 13

Houborg, R. & Mccabe, M.F. (2018). A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. Isprs Journal of Photogrammetry and Remote Sensing, 135, 173-188

Horvath, G. (2003). CMAC neural network as an SVM with B-spline kernel functions. In, 20th IEEE Instrumentation and Measurement Technology Conference (pp. 1108-1113). Vail, Co

Huang, G.B., Zhou, H.M., Ding, X.J., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 42, 513-529

Inoue, Y., Sakaiya, E., Zhu, Y. & Takahashi, W. (2012). Diagnostic mapping of canopy nitrogen content in rice based on hyperspectral measurements. Remote Sensing of Environment, 126, 210-221

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. Electronic Journal of Statistics, 1, 519-537

Jacquemoud, S.e.p., & Ustin, S. (2019). Leaf Optical Properties. Cambridge University Press

Kalaji, H.M., Dabrowski, P., Cetner, M.D., Samborska, I.A., Lukasik, I., Brestic, M., Zivcak, M., Tomasz, H., Mojski, J., Kociel, H., & Panchal, B.M. (2017). A comparison between different chlorophyll content meters under nutrient deficiency conditions. Journal of Plant Nutrition, 40, 1024-1034

Kiala, Z., Odindi, J., Mutanga, O. & Peerbhay, K. (2016). Comparison of partial least squares and support vector regressions for predicting leaf area index on a tropical grassland using hyperspectral data. Journal of Applied Remote Sensing, 10, 14

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, R., & Ltd., R.R.P. (2020). Package ‘Cubist’. Accessed 3 June 2020. https://cran.r-project.org/web/packages/Cubist/Cubist.pdf/

Li, L.T., Lin, D., Wang, J., Yang, L. & Wang, Y.L. (2020). Multivariate Analysis Models Based on Full Spectra Range and Effective Wavelengths Using Different Transformation Techniques for Rapid Estimation of Leaf Nitrogen Concentration in Winter Wheat. Frontiers in Plant Science, 11

Li, X.D., Mao, W.J., & Jiang, W. (2016). Multiple-kernel-learning-based extreme learning machine for classification design. Neural Computing & Applications, 27, 175-184

Liang, K., Huang, J.N., He, R.Y., Wang, Q.J., Chai, Y.Y., & Shen, M.X. (2020). Comparison of Vis-NIR and SWIR hyperspectral imaging for the non-destructive detection of DON levels in Fusarium head blight wheat kernels and wheat flour. Infrared Physics & Technology, 106, 9

Liu, K.L., Wang, J.D., Zeng,W., & Song, J. (2017). Comparison of three modeling methods for estimating forest biomass using TM, GLAS and field measurement data. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 5774-5777

Lu, B., He, Y.H. & Dao, P.D. (2019). Comparing the Performance of Multispectral and Hyperspectral Images for Estimating Vegetation Properties. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12, 1784-1797Maleki, M.R., Mouazen, A.M., Ramon, H., & De Baerdemaeker, J. (2007). Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. Biosystems Engineering, 96, 427-433

Maliha, A., Yusof, R., & Shapiai, M.I. (2018). Extreme learning machine for structured output spaces. Neural Computing & Applications, 30, 1251-1264

Meng, X.T., Bao, Y.L., Liu, J.G., Liu, H.J., Zhang, X.L., Zhang, Y., Wang, P., Tang, H.T., & Kong, F.C. (2020). Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. International Journal of Applied Earth Observation and Geoinformation, 89

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2017). Misc Functions of the Department of Statistics, Probability. Accessed 3 June 2020. https://cran.r-project.org/web/packages/e1071/e1071.pdf

Miller, D.L., Alonzo, M., Roberts, D.A., Tague, C.L., & McFadden, J.P. (2020). Drought response of urban trees and turfgrass using airborne imaging spectroscopy.

Remote Sensing of Environment, 240

Mutanga, O., Skidmore, A.K., Kumar, L. & Ferwerda, J. (2005). Estimating tropical pasture quality at canopy level using band depth analysis with continuum removal in the visible domain. International Journal of Remote Sensing, 26, 1093-1108

Ng, W., Minasny, B., & McBratney, A. (2020). Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. Science of the Total Environment, 702

Okujeni, A., Van Der Linden, S., Suess, S. & Hostert, P. (2017). Ensemble Learning From Synthetically Mixed Training Data for Quantifying Urban Land Cover With Support Vector Regression. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 1640-1650

Peng, S., Garcia, F.C., Laza, R.C., Cassman, K.G. (1993). Adjustment for specific leaf weight improves chlorophyll meter's estimation of rice leaf nitrogen concentration. Agronomy Journal, 85, 987-990

Penuelas, J., Gamon, J.A., Fredeen, A.L., Merino, J. & Field, C.B. (1994). Reflectance indices associated with physiological changes in nitrogen- and water-limited sunflower leaves. Remote Sensing of Environment, 48, 135-146

Porra, R.J., Thompson, W.A., & Kriedemann, P.E. (1989). 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, 975, 384-394

Powell, S.L., Cohen, W.B., Healey, S.P., Kennedy, R.E., Moisen, G.G., Pierce, K.B., & Ohmann, J.L. (2010). Quantification of live aboveground forest biomass dynamics with

Landsat time-series and field inventory data: A comparison of empirical modeling approaches. Remote Sensing of Environment, 114, 1053-1068

Prasad, K.A., Gnanappazham, L., Selvam, V., Ramasubramanian, R., & Kar, C.S. (2015). Developing a spectral library of mangrove species of Indian east coast using field spectroscopy. Geocarto International, 30, 580-599

R Core Team (2020): A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Accessed 3 June 2020.https://www.R-project.org/

Rinnan, A., van den Berg, F., & Engelsen, S.B. (2009). Review of the most common pre–processing techniques for near-infrared spectra. Trac-Trends in Analytical Chemistry, 28, 1201-1222

Sanches, I.D.A., Souza Filho, C.R., & Kokaly, R.F. (2014). Spectroscopic remote sensing of plant stress at leaf and canopy levels using the chlorophyll 680 nm absorption feature with continuum removal. ISPRS Journal of Photogrammetry and Remote Sensing, 97, 111-122

Siegmann, B., & Jarmer, T. (2015). Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. International journal of Remote Sensing, 36, 4519-4534

Sims, D.A., & Gamon, J.A. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sensing of Environment, 81, 337-354

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M.M.A., Prabhat, & Adams, R.P. (2015). Scalable Bayesian optimization using deep neural networks. In F. Bach, & D. Blei (Eds.), the 32nd International Conference on

Machine Learning (ICML) (pp. 2171-2180). Paris

Sonobe, R., Hirono, Y. & Oi, A. (2020a). Quantifying chlorophyll-a and b content in tea leaves using hyperspectral reflectance and deep learning. Remote Sensing Letters, 11, 933-942.

Sonobe, R., Hirono, Y., & Oi, A. (2020b). Non-Destructive Detection of Tea Leaf Chlorophyll Content Using Hyperspectral Reflectance and Machine Learning Algorithms. Plants, 9, 368

Sonobe, R., Miura, Y., Sano, T., & Horie, H. (2018a). Monitoring Photosynthetic Pigments of Shade-Grown Tea from Hyperspectral Reflectance. Canadian Journal of Remote Sensing, 44, 104-112

Sonobe, R., Sano, T., & Horie, H. (2018b). Using spectral reflectance to estimate leaf chlorophyll content of tea with shading treatments. Biosystems Engineering, 175, 168-182

Tsai, F., & Philpot, W. (1998). Derivative analysis of hyperspectral data. Remote Sensing of Environment, 66, 41-51

Vahtmae, E., Kotta, J., Orav-Kotta, H., Kotta, I., Parnoja, M., & Kutser, T. (2018). Predicting macroalgal pigments (chlorophyll a, chlorophyll b, chlorophyll a plus b, carotenoids) in various environmental conditions using high-resolution hyperspectral spectroradiometers. International Journal of Remote Sensing, 39, 5716-5738

Villar, A., Vadillo, J., Santos, J.I., Gorritxategi, E., Mabe, J., Arnaiz, A., & Fernandez, L.A. (2017). Cider fermentation process monitoring by Vis-NIR sensor system and chemometrics. Food Chemistry, 221, 100-106

Wang, F.M., Huang, J.F., Wang, Y., Liu, Z.Y., & Zhang, F.Y. (2013). Estimating nitrogen concentration in rape from hyperspectral data at canopy level using support vector machines. Precision Agriculture, 14, 172-183

Wijesingha, J., Astor, T., Schulze-Bruninghoff, D., Wengert, M. & Wachendorf, M. (2020). Predicting Forage Quality of Grasslands Using UAV-Borne Imaging Spectroscopy. Remote Sensing, 12, 23

Williams, P. (1987). Variables affecting near-infraredreflectance spectroscopic analysis. In P. Williams, & K. Norris (Eds.), Near- Infrared Technology in the Agricultural and Food Industries (pp. 143-167): American Association of Cereal Chemists Inc.

Xu, J.L., & Gowen, A.A. (2020). Spatial-spectral analysis method using texture features combined with PCA for information extraction in hyperspectral images. Journal of Chemometrics, 34

Yan, Y. (2016). Bayesian Optimization of Hyperparameters. Accessed 3 June 2020.https://cran.r-project.org/web/packages/rBayesianOptimization/rBayesianOptimization.pdf

Yang, X.H., Huang, J.F., Wu, Y.P., Wang, J.W., Wang, P., Wang, X.M. & Huete, A.R. (2011). Estimating biophysical parameters of rice with remote sensing data using support vector machines. Science China-Life Sciences, 54, 272-281

Zarco-Tejada, P.J., Miller, J.R., Haboudane, D., Tremblay, N., & Apostol, S. (2003). Detection of chlorophyll fluorescence in vegetation from airborne hyperspectral CASI imagery in the red edge spectral region. IGARSS 2003: IEEE International Geoscience and Remote Sensing Symposium, Vols I - Vii, Proceedings: Learning from Earth's Shapes and Sizes, 598-600

Lists of the Figures

Figure 1. Box plots of chlorophyll content as a function of pH for the two cultivars.

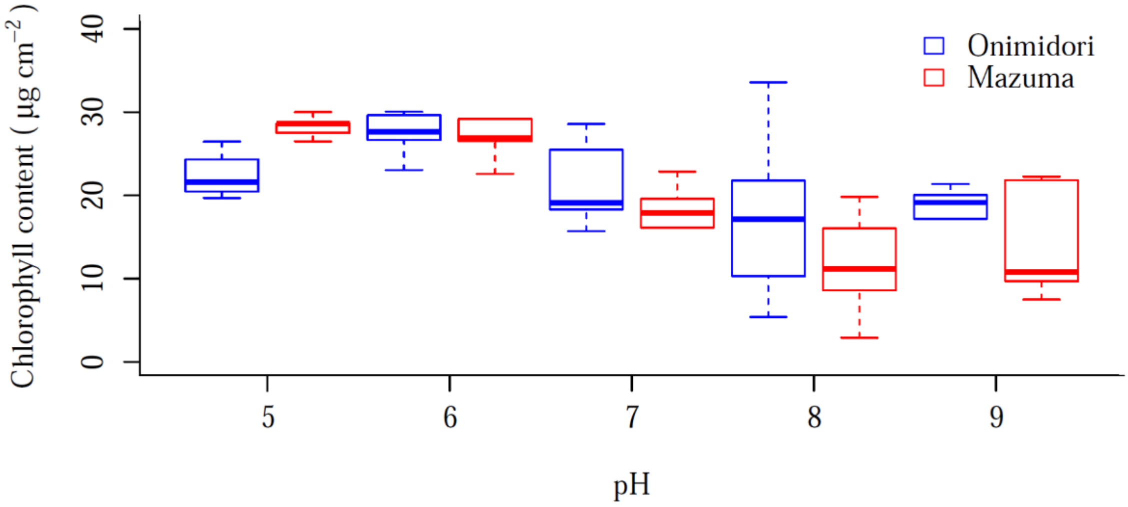


Figure 2. Box plots of chlorophyll content for different sulphur concentrations (S strengths).

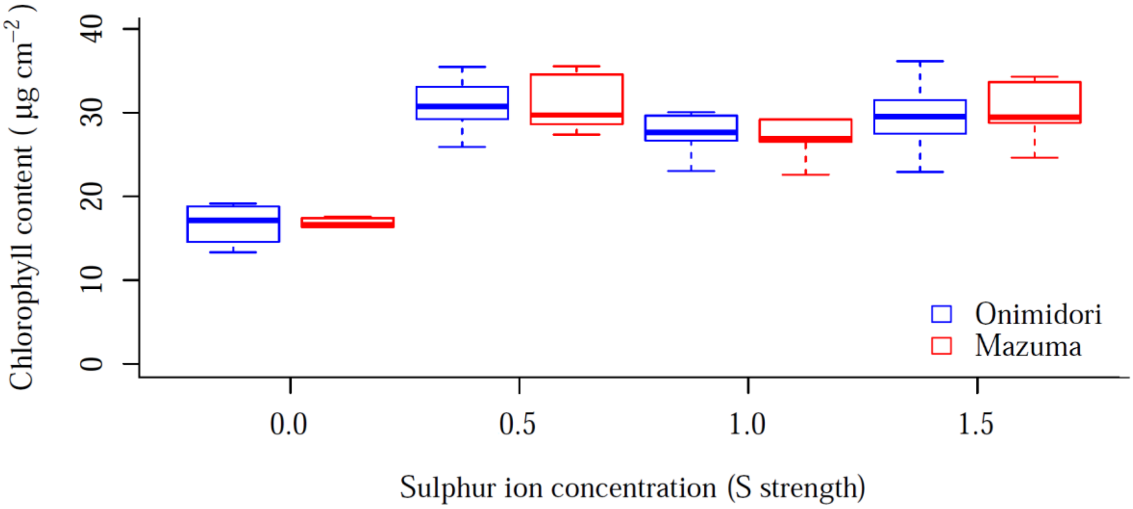


Figure 3. Mean reflectance spectra for each pH condition for (*a*) Onimidori and (*b*) Mazuma.

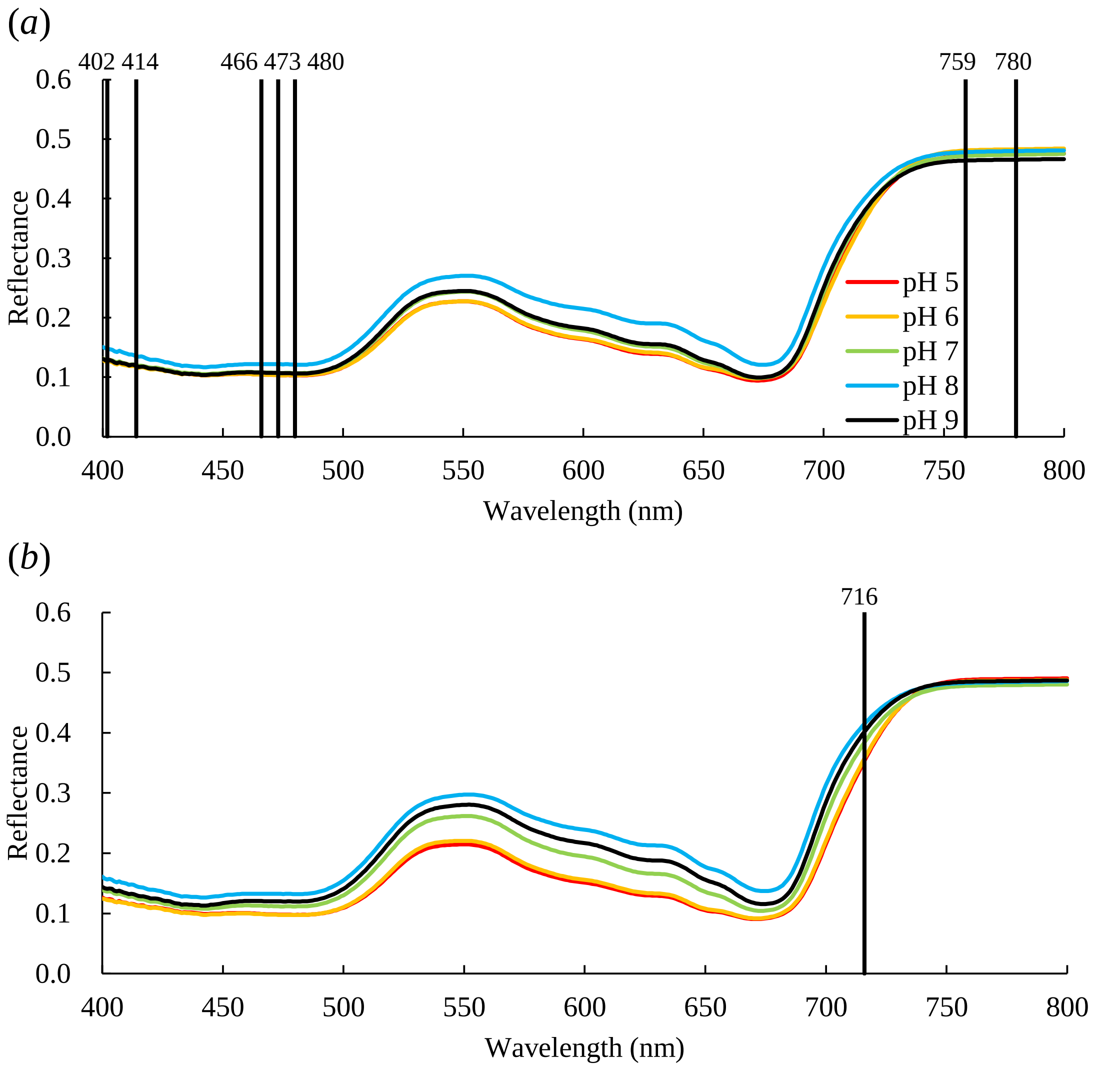


Figure 4. Mean reflectance spectra for each S strength (sulphur concentration) for (*a*) Onimidori and (*b*) Mazuma.

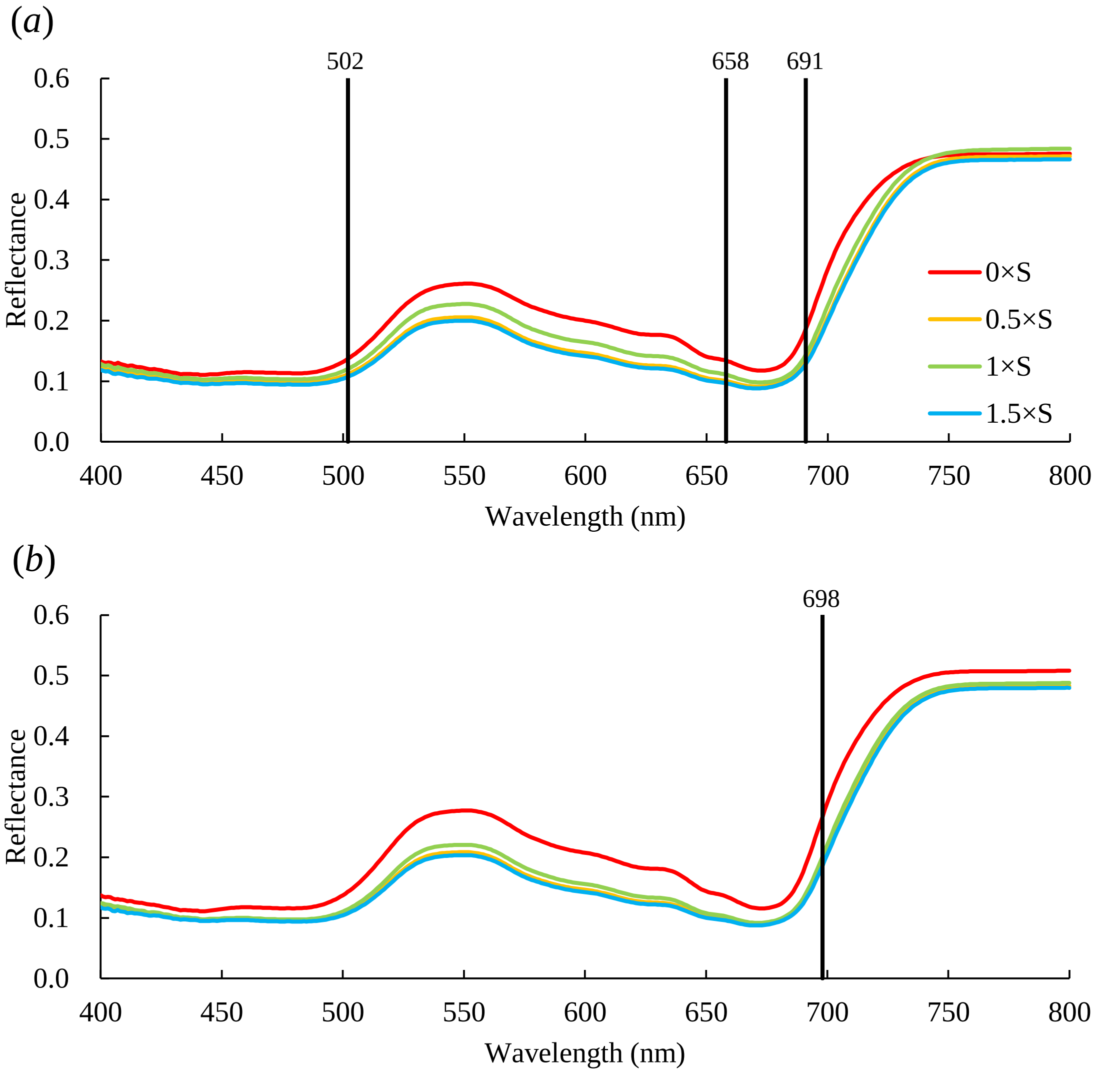


Figure 5. Violin plots of efficiencies of number of variables based on Data envelopment analysis after 100 replicates for random forest (RF), support vector machine (SVM), kernel–based extreme learning machine (KELM), Cubist and Stochastic Gradient Boosting (SGB) from (*a*) original reflectance, (*b*) first derivative reflectance (FDR), (*c*) continuum–removal (CR), (*d*) de–trending (DT), (*e*) multiplicative scatter correction (MSC) and (*f*) standard normal variate (SNV).

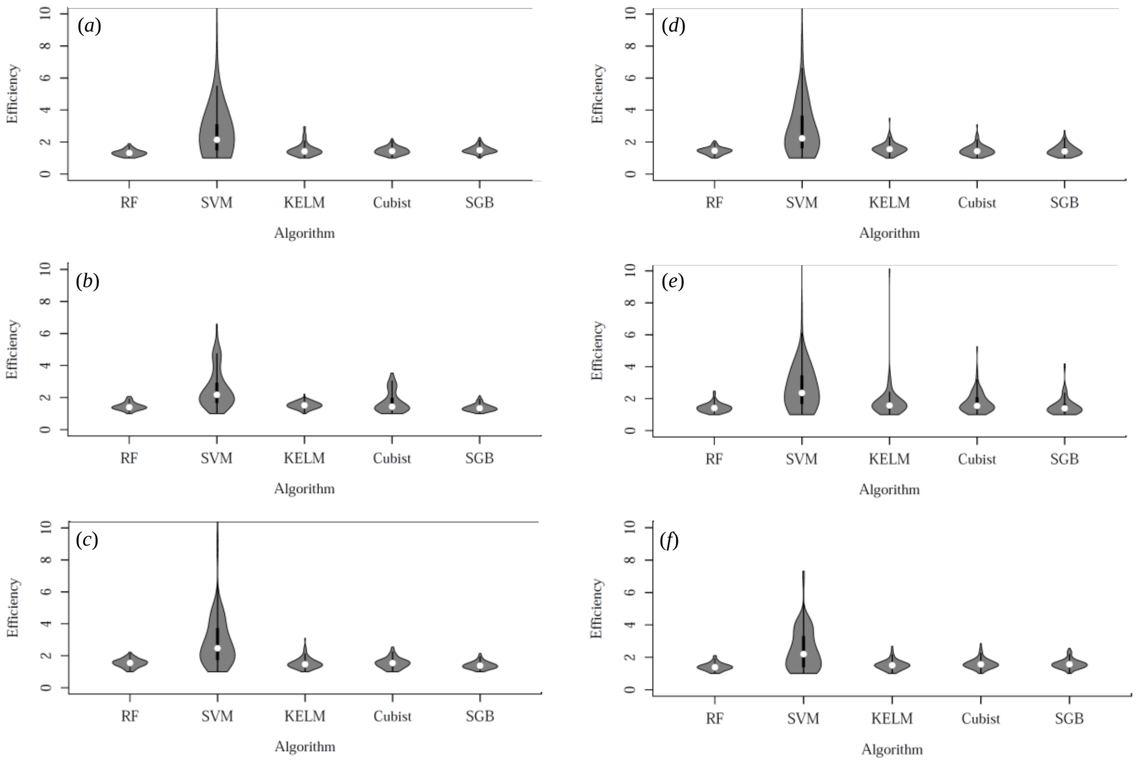
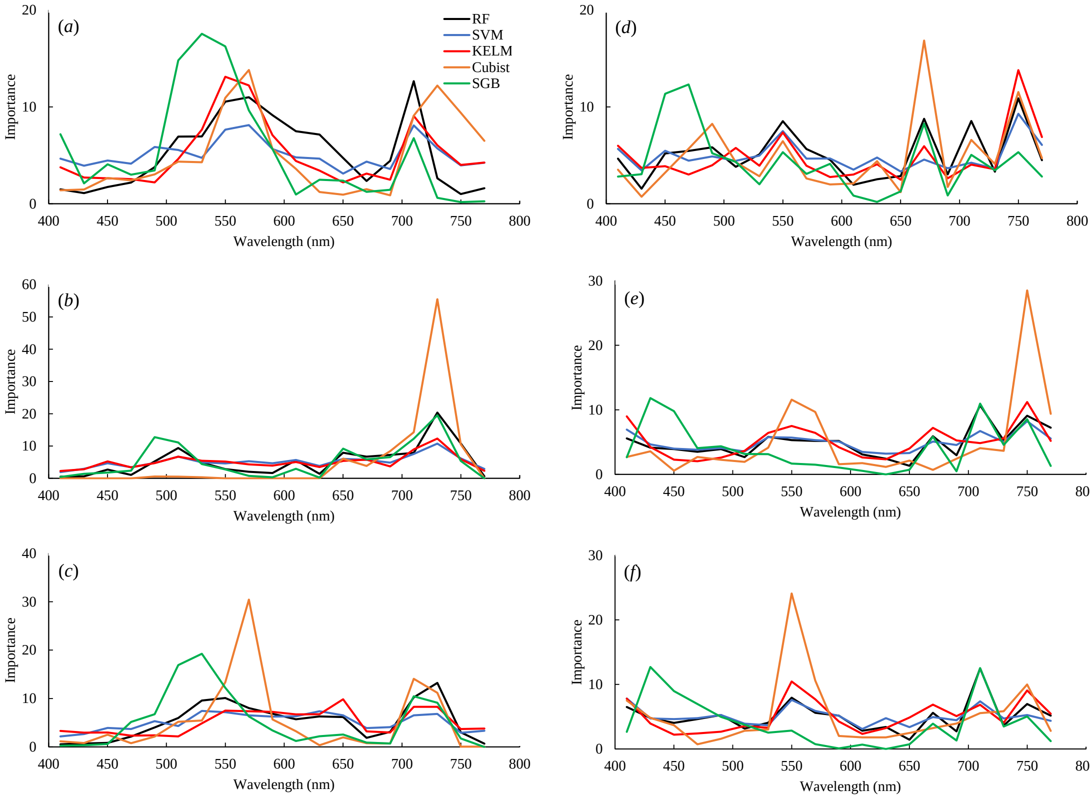


Figure 6. Data-based sensitivity analyses (DSA) results for random forest (RF), support vector machine (SVM), kernel–based extreme learning machine (KELM), Cubist and Stochastic Gradient Boosting (SGB) from (*a*) original reflectance, (*b*) first derivative reflectance (FDR), (*c*) continuum–removal (CR), (*d*) de–trending (DT), (*e*) multiplicative scatter correction (MSC) and (*f*) standard normal variate (SNV). Importance is expressed in percentage and values were averaged over 100 replicates.



Lists of the Tables

Table 1. List of abbreviations of pre–processing techniques and machine learning algorithms.

Abbreviation	Explanation
FDR	first derivative reflectance
CR	continuum-removed spectra
DT	de-trending
MSC	multiplicative scatter correction
SNV	standard normal variate
RF	random forest
SVM	support vector machine
KELM	kernel-based extreme learning machine
Cubist	Cubist
SGB	stochastic gradient boosting

Table 2. Ratio of performance to deviation (RPD) for each regression model (statistical results are based on 100 repetitions).

	Original reflectance	FDR

	RF	SVM	KELM	Cubist	SGB	RF	SVM	KELM	Cubist	SGB
Minimum	1.645	0.087	1.600	1.921	1.187	1.933	0.739	2.292	1.163	1.211
Median	2.303	1.771	3.296	3.348	1.800	2.889	2.240	3.357	2.882	1.918
Mean	2.339	1.980	3.282	3.412	1.800	2.917	2.271	3.415	2.670	1.916
Maximum	3.133	4.405	4.819	5.040	2.706	4.308	4.871	5.092	4.296	2.588
Standard deviation	0.329	0.933	0.663	0.571	0.271	0.444	0.870	0.508	0.795	0.288
Skewness	0.133	0.616	–0.328	0.153	0.395	0.262	0.244	0.706	–0.366	–0.195
Kurtosis	–0.793	–0.370	0.057	0.033	0.820	0.253	–0.486	0.558	–1.049	–0.287
			CR					DT		
	RF	SVM	KELM	Cubist	SGB	RF	SVM	KELM	Cubist	SGB
Minimum	1.924	0.423	1.593	1.813	1.141	2.063	0.244	1.519	1.575	1.075
Median	2.708	1.894	3.308	2.981	1.755	2.931	2.229	3.453	3.399	2.068
Mean	2.819	2.049	3.296	3.079	1.778	3.039	2.291	3.510	3.389	2.036
Maximum	4.371	4.870	5.049	4.623	2.450	4.297	4.974	5.556	4.867	2.934
Standard deviation	0.482	0.908	0.598	0.619	0.280	0.470	1.083	0.670	0.647	0.385
Skewness	0.833	0.482	-0.173	0.405	0.147	0.581	0.309	0.197	–0.187	–0.135
Kurtosis	0.711	–0.441	0.670	–0.314	–0.258	0.131	–0.740	0.772	–0.196	–0.599
			MSC					SNV		
	RF	SVM	KELM	Cubist	SGB	RF	SVM	KELM	Cubist	SGB
Minimum	1.736	0.223	0.512	0.915	0.622	1.782	0.581	1.881	1.884	1.183
Median	2.951	1.820	3.295	3.110	1.834	3.121	1.918	3.354	3.384	1.926
Mean	2.945	1.985	3.245	2.991	1.819	3.151	2.138	3.401	3.361	1.961
Maximum	4.295	4.649	5.251	4.809	2.602	4.516	4.253	5.057	5.419	3.021
Standard deviation	0.499	0.943	0.773	0.829	0.446	0.472	1.004	0.609	0.569	0.339
Skewness	0.021	0.683	–0.413	–0.121	–0.382	0.197	0.531	0.125	0.063	0.285
Kurtosis	–0.173	–0.292	1.203	–0.586	–0.200	0.792	–0.909	0.031	1.220	0.802

Table 3. Root-mean-square error (RMSE, µg cm^{-2}) for each regression model (statistical results are based on 100 repetitions).

			Original reflectance					FDR		
	RF	SVM	KELM	Cubist	SGB	RF	SVM	KELM	Cubist	SGB
Minimum	2.485	1.826	1.747	1.718	2.862	1.689	1.851	1.655	1.878	3.045
Median	3.480	4.471	2.459	2.380	4.506	2.690	3.448	2.375	2.826	4.147
Mean	3.503	5.870	2.566	2.423	4.556	2.835	4.253	2.398	3.378	4.291
Maximum	4.969	91.484	5.267	4.583	6.233	4.494	11.765	3.349	7.208	6.240
Standard deviation	0.525	9.029	0.639	0.448	0.643	0.560	2.133	0.332	1.312	0.689
Skewness	0.649	8.543	1.900	1.557	0.136	0.872	1.315	0.349	1.124	0.831
Kurtosis	0.274	77.882	3.964	4.502	–0.150	0.700	0.902	0.199	0.082	0.229
			CR					DT		
	RF	SVM	KELM	Cubist	SGB	RF	SVM	KELM	Cubist	SGB
Minimum	1.899	1.778	1.670	1.774	2.959	1.739	1.770	1.517	1.602	2.744
Median	2.907	4.018	2.427	2.697	4.489	2.706	3.568	2.286	2.391	3.987
Mean	2.932	4.970	2.524	2.712	4.630	2.715	4.853	2.383	2.476	4.093
Maximum	4.279	20.648	4.957	4.625	6.947	4.449	35.463	6.041	5.280	6.679
Standard deviation	0.513	2.950	0.526	0.548	0.743	0.488	4.144	0.571	0.603	0.814
Skewness	0.210	2.364	2.104	0.781	0.495	0.705	4.478	2.952	1.837	0.974
Kurtosis	–0.353	8.324	6.819	0.718	0.550	1.108	28.162	15.250	4.815	0.936
			MSC					SNV		
	RF	SVM	KELM	Cubist	SGB	RF	SVM	KELM	Cubist	SGB
Minimum	1.933	1.581	1.605	1.721	2.678	1.646	1.718	1.654	1.343	2.431
Median	2.689	4.511	2.467	2.568	4.370	2.597	4.292	2.395	2.393	4.130

Mean	2.824	5.307	2.734	2.957	4.787	2.619	4.758	2.443	2.474	4.214
Maximum	4.988	39.615	16.330	9.465	13.303	4.680	14.806	4.064	4.596	6.299
Standard deviation	0.607	4.172	1.547	1.141	1.659	0.493	2.471	0.484	0.536	0.718
Skewness	1.151	5.639	6.922	2.716	2.248	1.261	1.284	1.157	1.748	0.643
Kurtosis	1.473	43.079	56.931	10.903	7.071	3.081	2.462	1.556	3.938	0.702

Table 4. Best– and worst–performing algorithms after 100 repetitions for original reflectance and five pre–processing techniques. Results presented are number of times per 100 repetitions.

Original reflectance				FDR			
Algorithm	Selected times		Net score (best - worst)	Algorithm	Selected times		Net score (best - worst)
	Best	Worst			Best	Worst	
RF	0	4	–4	RF	14	0	14
SVM	3	49	–46	SVM	7	33	–26
KELM	45	1	44	KELM	65	0	65
Cubist	52	0	52	Cubist	14	17	–3
SGB	0	46	–46	SGB	0	50	–50
CR				DT			
Algorithm	Selected times		Net score (best - worst)	Algorithm	Selected times		Net score (best - worst)
	Best	Worst			Best	Worst	
RF	7	1	6	RF	8	2	6
SVM	8	42	–34	SVM	11	42	–31
KELM	54	1	53	KELM	49	1	48
Cubist	31	0	31	Cubist	32	2	30
SGB	0	56	–56	SGB	0	53	–53
MSC				SNV			
Algorithm	Selected times		Net score (best - worst)	Algorithm	Selected times		Net score (best - worst)
	Best	Worst			Best	Worst	
RF	19	0	19	RF	17	0	17
SVM	7	44	–37	SVM	8	52	–44
KELM	45	5	40	KELM	41	2	39
Cubist	29	2	27	Cubist	34	0	34
SGB	0	49	–49	SGB	0	46	–46

Table 5. Optimal combinations of pre–processing techniques and machine learning algorithms after 100 repetitions.

Pre-processing	Algorithm	Times
DT	KELM	15
SNV	KELM	10
DT	Cubist	9
FDR	KELM	8
SNV	Cubist	8
CR	KELM	7
MSC	KELM	7
Original reflectance	KELM	7
DT	RF	4
SNV	RF	4

DT	SVM	3
MSC	Cubist	3
Original reflectance	Cubist	3
CR	Cubist	2
FDR	RF	2
FDR	SVM	2
Original reflectance	SVM	2
SNV	SVM	2
FDR	Cubist	1
MSC	RF	1